Using Text Based Convolutional Neural Networks to Predict and Understand Genetic Population Structure

Sravani Nanduri

September 2020

1 Abstract

Phylogenetics is incredibly important to scientists' understanding of population patterns, dynamics, and mutations, and this knowledge is integral to vaccine research, outbreak prevention, and much more. Phylogenetics currently relies on the creation of a phylognetic tree to understand population structure, a time consuming and inaccurate process that cannot represent recombinant diseases and large populations. In this paper, I study the use of convolutional neural networks to understand population structure within H3H2 Influenza, and analyze both the results of the research and possible next steps. This process, with more research and tuning, has the potential to be a strong tool for a more multifaceted understanding of populations and genetics.

2 Introduction

Phylogenetics, the study of evolution, is important because it underlies how molecular sequences, genes, and characteristics evolve over time. Understanding mutational and hierarchical patterns in a population is a deeper look at the transmission, mutability, and the mechanics behind an outbreak. Beyond this, learning how the virus evolves over time by understanding genetic similarity between strains helps to predict which strains and clades will be most prevalent in the future. This is invaluable information when considering vaccine decisions and accurate prevention measures. In the midst of the COVID-19 pandemic, this data could prove invaluable in understanding the virus' transmission and mutational patterns, which can help create a better vaccine and stop outbreaks.

I present a novel approach of using deep learning to understand population patterns in genomic data through the use of a text based convolutional network that takes a feature map of nucleotides and infers its population structure from it. The neural network was trained using SANTA-SIM, a genetic data simulator for viruses, to understand the generational structure. I begin testing with

genomes that are phylogenetically tractable, and then discuss what must change about the dataset and network in order to test the network on phylogenetically intractable disease.

2.1 Background and Previous Work

2.1.1 Epidemiology

Viral data's evolution is most commonly modeled in a backward-in-time coalescent model. One example of this model is a phylogenetic tree, which makes inferences in genetic mutation and ancestry in disease strains in order to "coalesce" the lineages into a singular common ancestor [11] [6]. The largest issue with coalescent models is their inability to correctly model the evolution of non-tractable diseases, such as MERS and SARS-CoV-2. To combat this, one approach is to split a genome into multiple phylogenies to model the evolution of the nonrecombinant fragments [12] [20] [13] However, modeling the entire evolution of the virus becomes increasingly challenging as the amount of nonrecombinant fragments increases. It would be incredibly helpful to have a network which could "check" the accuracy of the patterns found in the tree. Beyond this, epidemiologists rely on statistics and hundreds of different sources of information to create a full interpretation of how the disease will spread. Because of this, finding a more computationally and mathematically definable way to subset strains and quantify population structure will make epidemiology more mathematically definable than the vague definitions of genetic dissimilarity and population dynamics. This paper will explore different ways of preparing population data into classification labels, and will analyze why they work and do not.

2.1.2 Deep Learning

Deep learning applied to genomics is a fairly new field of study. While machine learning and dimensionality reduction have been used widely in the phenotyping and characteristic identification of plants and animals, it has not been widely used in epidemiology for the characterizing and classification of genomes [21] There has been significant work in text based classification for other purposes [23] [1] [10], and there have been studies concerning hierarchical convolutional neural nets to flesh out patterns in a structured manner from image based sets [24] [22] [14]. The largest problem with using neural networks and deep learning with epidemiology presently is the inherent lack of data due to a public health infrastructure that cannot sample and store many genomic records. Beyond just this, deep learning in epidemiology has been hindered by the process of preparing the data in an epidemiologically useful and accurate way to receive data that can be used to further understand population dynamics. This paper hopes to open the discussion of preparing epidemiological data for deep learning to better understand diseases.

3 Methods

Due to the lack of data stemming from not enough testing and uploading onto NCBI's database, SANTA-SIM seemed like the strongest alternative, a software package that optimizes for a balance between the complexity of the underlying framework and the different evolutionary scenarios that can be modeled. SANTA-SIM is directed towards haploid organisms, and is particularly useful for the study of rapidly evolving pathogens that experience diverse selection pressures and recombination events [7]. SANTA-SIM has a very customizable simulation system, where a start sequence can be added as the starting strain for the population, and the amino acids selection types can also be specified. An epitope,the part of an antigen recognized by antibodies, B cells, or T cells, can undergo different types of selection within a genome that can change population structure. For this project, the signal peptide HA1 and HA2 are mainly under purifying selection. Non-epitope sites in HA1 are under weak purifying selection, but the preferred amino acid at each site can change slowly over time. Some epitope sites were also under exposure-dependent selection, which allows for beneficial new mutations to escape past immunity due to previous exposure. The mutation rate for the population was set to .0001. These attributes create a simulated dataset reminiscent of H3N2 flu data. The full details of the simulation can be found at the github repository linked to this paper.

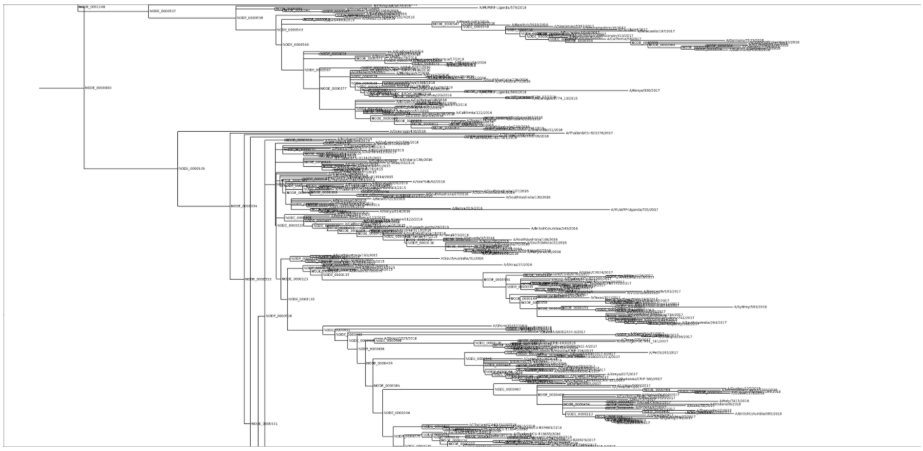
I used a text-based convolutional neural network modeled after the Zhang and Zhaos paper [23]. The hyperparameters of learning rate, dropout rate, batch size, and epochs were tuned using scikit-learn's GridSearchCV. The epoch and batch size that optimized for loss and accuracy were both 10, with a learning rate of .01 and a dropout rate of .001. The embedding size was 1701, as the aligned FASTA files from the SANTA-SIM builds as well as the flu builds were 1701 nucleotides long. The network summary is below.

Layer (type)	Output Shape	Param #
sent_input (InputLayer)	[(None, 1701)]	0
embedding_2 (Embedding)	(None, 1701, 128)	768
conv1d_4 (Conv1D)	(None, 1697, 256)	164096
thresholded_re_lu_8 (Thresho	(None, 1697, 256)	0
conv1d_5 (Conv1D)	(None, 1693, 256)	327936
thresholded_re_lu_9 (Thresho	(None, 1693, 256)	0
flatten_2 (Flatten)	(None, 433408)	0
dense_6 (Dense)	(None, 1024)	443810816
thresholded_re_lu_10 (Thresh	(None, 1024)	0
dropout_4 (Dropout)	(None, 1024)	0
dense_7 (Dense)	(None, 1024)	1049600
thresholded_re_lu_11 (Thresh	(None, 1024)	0
dropout_5 (Dropout)	(None, 1024)	0
dense_8 (Dense)	(None, 49)	50225

Total params: 445,403,441
Trainable params: 445,403,441
Non-trainable params: 0

The network was trained on a dataset simulated using SANTA-SIM. The parameter of "replicator" was tuned for both builds. The build that only simulated point and frameshift mutations used a "clonalreplicator", which simply copies the genome of a single parent with randomly biological mistakes. SANTA-SIM can also replicate recombinant diseases through the use of a "recombinantReplicator", which considers that a child genome can be derived from two parents, where the child genome sequence is created by copying the genome of one or

the other parent, switching between these two genomes at random sites. These random probabilities can be tuned using "dualInfectionProbability" and "recombinationProbability", where higher probabilities simulate an increasingly phylogenetically intractable population. The SANTA-SIM populations for the build used to train the network was around 3 million individual strains, with a 10,000 strain sample rate from each generation of a 10,000 generation build. The metadata from the build returns the generation the strain was sampled from, which the network was trained to classify. To reveal the scientific feasibility of using generation as a metric of population structure, a subset of the SANTA-SIM data from the dataset was reduced via t-SNE [15] and colored by generation to reveal clustering that can be related back to a phylogeny's structure. t-SNE and dimensionality reduction have been used within epidemiology recently to understand geographical introductions of viruses and overall geographical structure, so it can be reasonably inferred that it can also be used to explain population structure [17] [18] [4] A sub-sample of the full phylogeny is below for visual understanding of the tree and branch length.



The dataset was fed into the neural network in different sizes in order to determine the amount of data necessary to optimize for both loss and accuracy. The network was trained on a 10K subset of the build, and later tested on a 120K sub-sample.

The genome data we used for testing the neural networks feasibility on real data was H3N2 HA influenza from the NCBI Influenza database. We analyzed Influenza A/h3n2, and created a FASTA file of multiple sequence alignments with MAFFT v7.407 [9] via augur align [3].

A feature map was used to transform the data. The feature map of values was the aligned FASTA file read in through BioPython [2] split by site into a Pandas dataframe [16].

The classification number to simulate the generation of the strains in these real-world builds were created using the phylogenetic tree created using IQ-TREE v1.6.10 [19] from the aligned FASTA file. The branch lengths were used

as an indicator of generation after being multiplied and rounded to an integer value, and this was used to test the network. This data was split into testing, validation, and training via Scikit Learn [8], and tested on the fitted model.

To analyze the network, the accuracy percentage, confusion matrix, epoch vs loss graph, and t-SNE embedding colored by network classifications were created to understand classifying errors and overall fit.

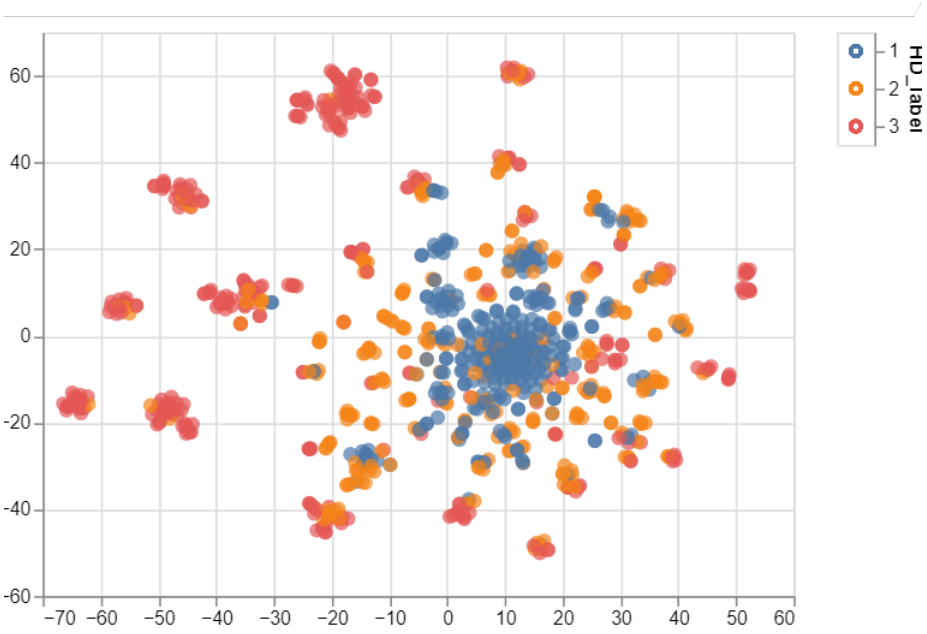
The data given back by the model for each of these instances was reduced via t-SNE and colored by the labels returned by the model. This labeled t-SNE graph was compared with the t-SNE graph colored by the actual labels to reveal issues with the network to be assessed more qualitatively.

4 Results

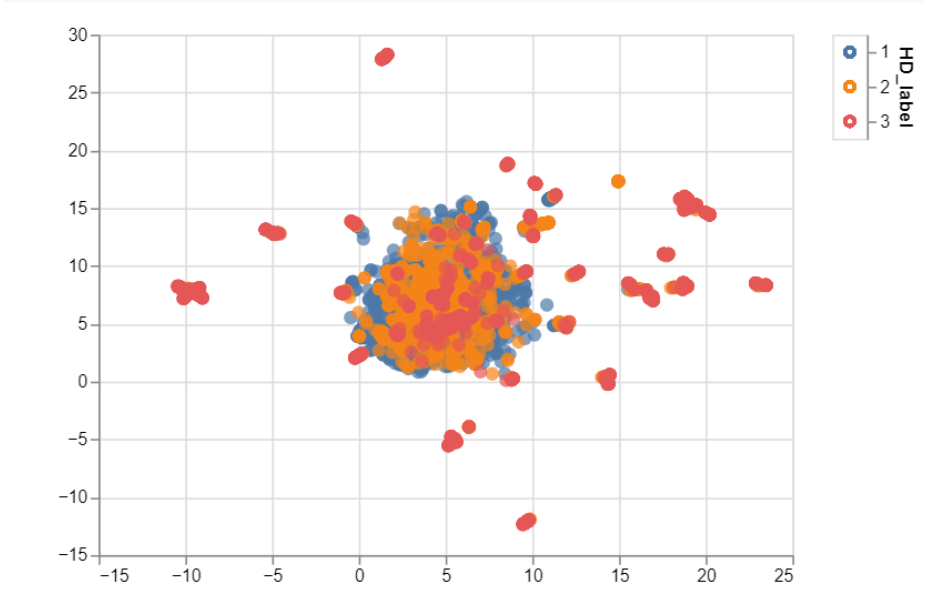
To further understand how generational numbers quantitatively and qualitatively describe important attributes in genomes, t-SNE was used to reduce the dimension of a subset of the SANTA-SIM data for both builds.

As t-SNE reduces distance matrices, a Kimura 2P distance matrix was computed and reduced in order to qualitatively understand the patterns available in the population. By comparing every genome with every other genome and clustering based on their Kimura two parameter distance, the distance-based method groups together genomes with similar differences. This clusters the data by genetic diversity. Each genome was split into separate nucleotides and compared with other nucleotides in the same site on other genomes. The number of transitions vs transversions were counted and plugged into the Kimura two parameter distance equation to create a distance matrix.

Two different samples were created for t-SNE to reduce. Because t-SNE performs best at reducing small amounts of data less than 10K as it emphasizes local patterns over global, the first sample was a 1000 strain subset selected randomly from the 3M strain clonal replicator build, and the other was a 10,000 strain subset selected randomly from the same build. While the 10K reduction would resemble output reminiscent of Multi Dimensional Scaling [5], a reduction technique that preserves global structure, revealing global pattern is incredibly important to assess along with local structure. The global reduction allows for analysis of outliers and their respective branch lengths to further understand the dataset and how well the phylogenetic tree expresses the true population structure.



The reduction with 1K strain sample from the 3M clonal replicator dataset



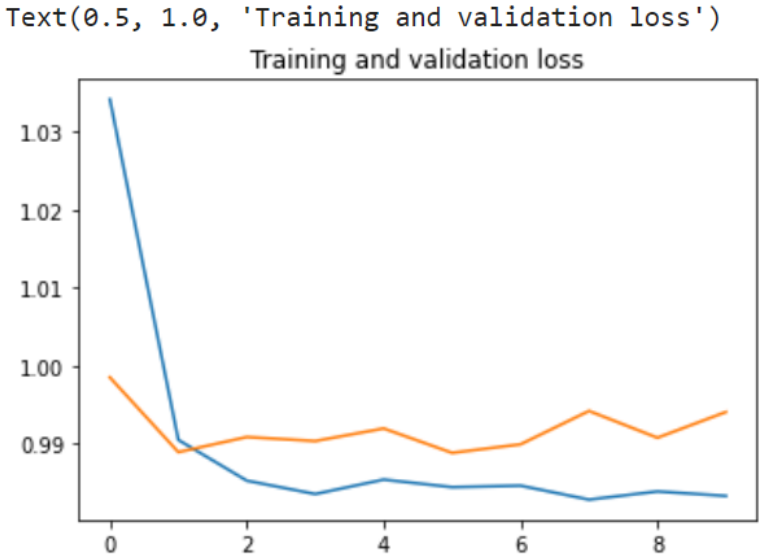
The reduction with 10K strain sample from the 3M clonal replicator dataset

The consistent clustering and rings of different generations is a strong indicator that generation captures genetic dissimilarity between and within the strains in the population. In the 10K strain t-SNE plot, the clustering as indicated is genetically similar strains typically referred to as "clade" structures. They have been defined and exposed via the distance matrix reduction, where clades are related to the distance of different strains from others in a traditional phylogenetic tree. Clades were not used to train networks for the sole reason that there is no mathematically accurate way to define a clade without the help of a phylogenetic tree - as the reason for conducting this research is to understand recombinant disease population, this metric would have become obsolete.

This reduction reveals the genetic structure of a population can be understood via generational structure.

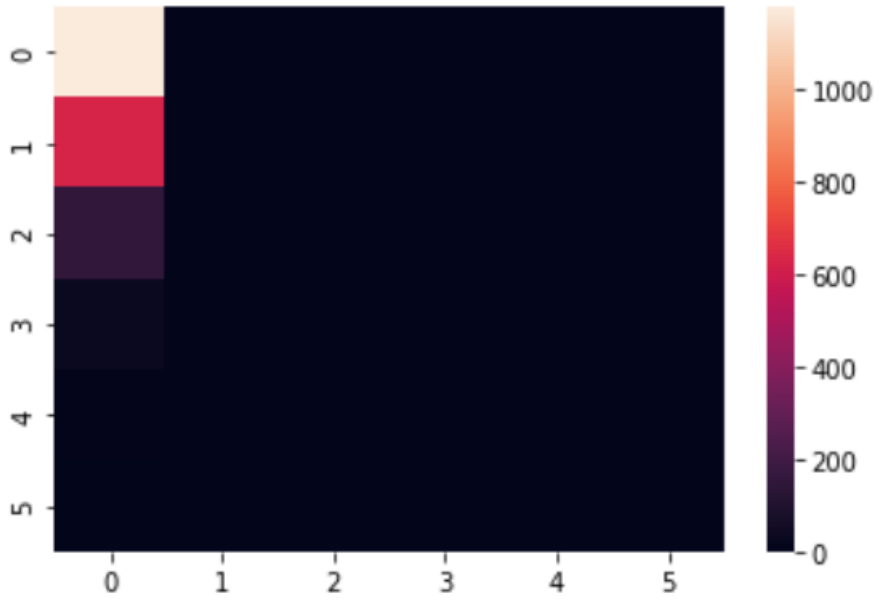
4.1 10K SANTA-SIM clonal replicator build

The neural network trained on 10K strains was simply inaccurate and did not categorize the genomes correctly at all, as shown by the loss vs epoch graph below. The graph clearly displays signs of under fitting, where the training loss is low and the validation loss is much higher and doesn't train over 10 epochs.



A confusion matrix was created to better understand how the neural network was actually classifying the genomes.

```
test loss, test acc: 0.9741735458374023
predictions shape: (2001, 7)
<AxesSubplot:>
```

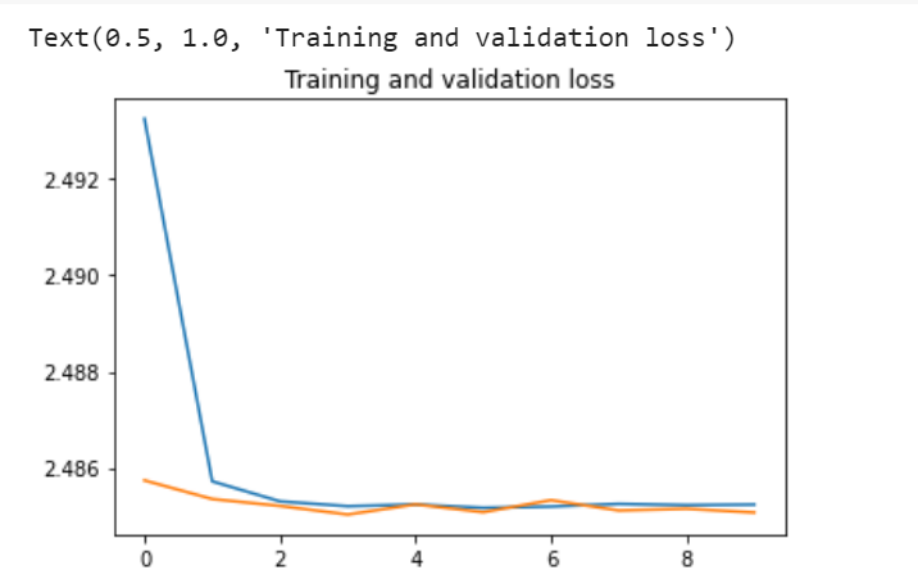


The neural network failed to classify the testing strains from the SANTA-SIM build accurately, and therefore, trying to use this fitted model with real data from the influenza build would also not yield any useful results. This incorrect classification is a direct result of the lack of diversity in the SANTA-SIM dataset, as a sample of 10,000 from a 3M strain build tends to encompass very little diversity, and not enough for a network to learn how 1701 different features on each genomes contributes to overall genetic diversity and mutation rate.

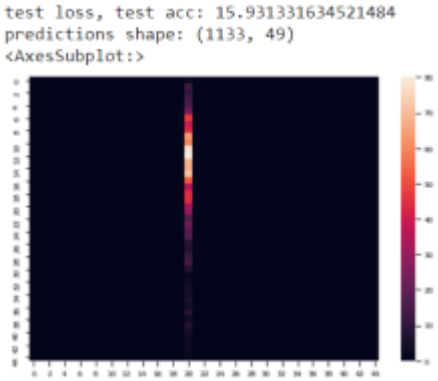
While the test accuracy is very high at 97 percent, this has to do with the amount of data in the smaller classification categories, and lack of data in the higher numbers. This lack of diversity in the dataset leads to the conclusion that more data is required to accurately train this model.

4.2 SANTA-SIM 120K clonal replicator build

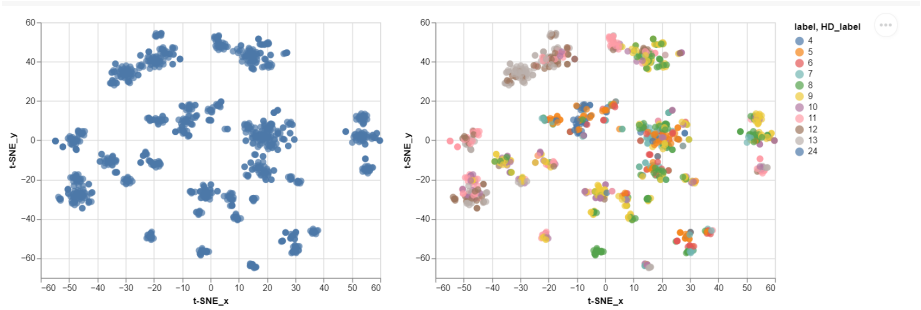
The neural network trained on 120K strains at first did show interesting results. The loss vs epoch graph saw a validation curve that converged with the training loss, a good sign that the data is being fit to the model.



However, the model under performed extremely with the H3n2 dataset, which mirrored a lot of the patterns in the SANTA-SIM build down to the starting strain the build was based off. The accuracy value was around 15 percent, and the confusion matrix showed that the network had simply latched onto a generational value that seemed to create the lowest loss.



Further analysis of the neural network predictions revealed that the network classified every strain in the H3N2 build with a generational value of 24. To further understand this data, a t-SNE reduction of a sub sample of the data was created, coloring the data by both branch length and neural network output.



A few things about this reduction is important to note. This first is that branch length is not a strong indicator of generational variation. This is an issue stemming from epidemiologists incomplete understanding of strain population patterns, as it is impossible to truly know the generation of a strain as is possible in a simulation.

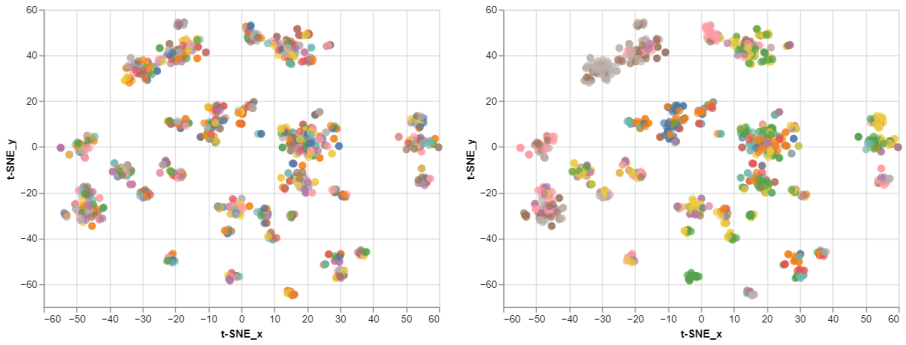
In order to combat this problem, I constructed a UPGMA based output of the H3N2 strain data. A method for constructing a tree from only pairwise distance is the Unweighted Pair-Group Method with Arithmetic mean (UPGMA), which is a popular distance analysis method. UPGMA creates a tree based solely on pairwise distance between strains within a distance matrix, and works backwards to find the oldest ancestor. The steps of UPGMA are as follows: UPGMA clusters the data in multidimensional space, and picks clusters closest together to "group" under the same node. This method continues backwards until there are only two major clusters left. It is the simplest method for constructing a tree, but comes with the disadvantage that it assumes the rate of mutations is constant over time and for all lineages in the tree, which is not true for many viruses. This would mean that all terminal nodes have the same distance from the root. In reality, the individual branches are very unlikely to have the same mutation rate. Therefore, UPGMA frequently generates incorrect tree topologies.

To create the distance matrix necessary to pass into UPGMA to produce a newick tree, we created a Kimura 2 parameter distance based matrix like the one used in the initial t-SNE reduction to understand how generation reveals patterns in a phylogenetic tree. The H3N2 data aligned had gaps (N) where the program did not know what bases went there. Because SANTA-SIM did not have any gaps, differences between the main nucleotide pairs (AGCT) were counted while gaps (N) were not. This is because some sequences were significantly shorter than others, and a shorter strain does not necessarily mean complete genetic dissimilarity, which is what counting gaps implied.

This distance matrix was passed as a lower triangular matrix into a UPGMA tree, and the classification was determined by how nested the strain was in the hierarchy. The number of classes for the network, therefore, did fluctuate between samples.

The hopes with using this approach was to create a better representation of generation than using branch length. To visualize the results, the t-SNE reduction was colored by UPGMA output, and it was compared with branch

length coloring. The UPGMA output is the left graph, and the branch length the right.



Because these results didn't reveal any patterns, I discarded this approach, as branch length is a more pylogenetically accurate way of studying generational structure.

With no new conclusions or solutions coming from changing the classification numbers, I looked at the dataset the neural network was trained on.

The SANTA-SIM data sub sample of 120K strains had the unfortunate consequences of not allowing for a lot of diversity per class. Using a network with over 400 million trainable parameters to train a network with a flexible amount of classes required millions and millions of strains of data, something I did not have the time or resources to create. To go beyond simple computational power, it isn't feasible to need a network with that many parameters when dealing with a science with less than 10,000 strains per disease to train and test a network on. In the field of epidemiology, it becomes clear that a network will need to be easily trained on a small sub sample of data, in order to then use the network on other data.

Another drawback in this project was the SANTA-SIM build. With the sheer size of the build, sub sampling from the dataset became incredibly time consuming. More tuning of population size, sampling size per generation, generation size, and sampling parameters is required to fully flesh out this training method. There should also be greater analysis of the patterns within h3n2 influenza to make sure the parameters in the simulation of mutation rate, epitope and non epitope selection sites, and other epidemiologically important factors are mirroring the population of real data that the network should be tested on.

5 Future Work

Future work with this project can be broken into two sections: computational work, and epidemiological work.

5.1 Computational Work

The network was adapted from a text based network made for NLP purposes, which is a a fields with many datasets containing millions and millions of examples to train the network on. Only a limited set of hyper-parameters were tuned due to lack of computational resources and time. Beyond this, the network has an infeasible amount of parameters to tune for the amount of data feasibly available in this field. There are 2 major areas where computational work is necessary:

1) Creating and tuning the model, using a different model

A Convolutional Neural network, which requires a soft max on the amount of classes it can classify between training and testing, does not seem feasible in a scientific setting. The simulation and the real data the network should predict labels for may have different numbers of classes depending on the phylogenetic structure and branch lengths, and the network should be flexible enough to account for this. An LSTM based network may work better, as it learns off previous examples and creates a memory within itself to recognize patterns.

If a Convolutional Neural Network is used, there should be less than 1 million parameters to tune. Even in an simulation, there is not a strong way to create a dataset with more than 120K strains that will give accurate results. A network with over 3,000 times the amount of parameters than the amount of data given cannot accurately predict or understand any patterns within the dataset it is given.

2) Tuning a larger set of hyper-parameters

There are hundreds more hyper-parameters to be tuned that could drastically improve the performance of the network. If unconstrained by time and computational work, the network should have every single tunable parameter tested in a grid format to test every permutation of possibilities for the best performance. Beyond this, the hyper-parameters should be tuned on both a recombinant and non-recombinant dataset to test if the configuration must change based on the type of genome passed into the network.

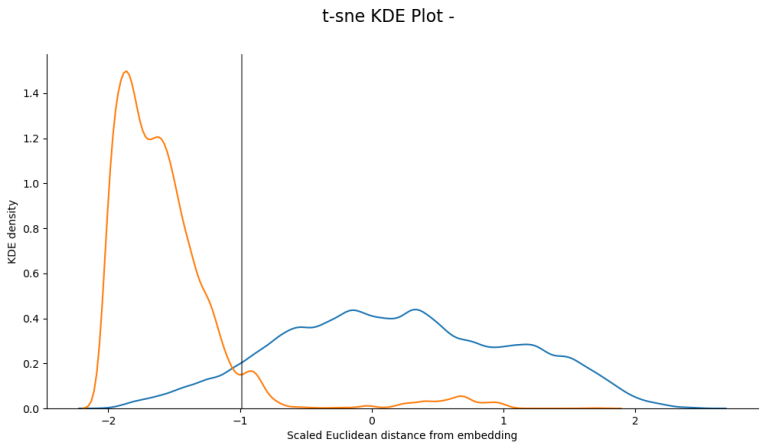
5.2 Epidemiological Work

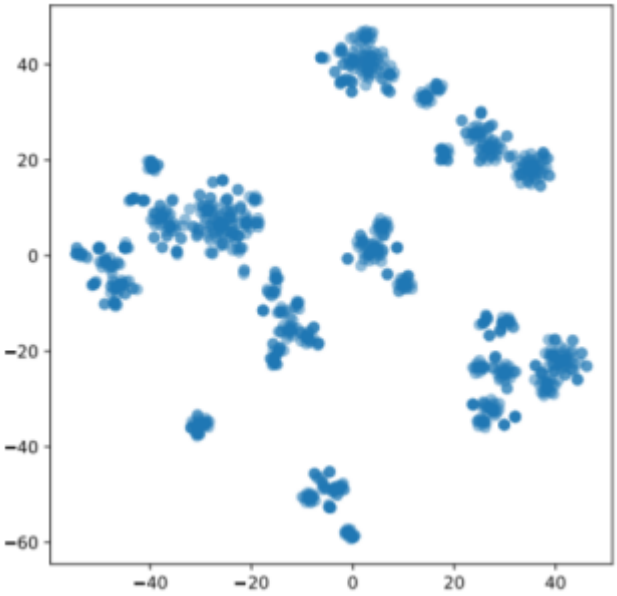
Along with the computation, the science behind the numbers, data, and genomes must be useful and accurate. This can be split into two major sections.

1) Preparing the data, picking biologically meaningful numbers

The data preparation needed in epidemiology is a problem vastly different than the issues most deep learning papers must solve. In epidemiology, the actual data that should be tuned and tested is an incredibly pertinent question. The type of population structure statistics wanted from the network would

completely change how the data is prepared. To study genetic similarity and dissimilarity, some classification or spectrum of values must exist that the network can be trained on. This can come in the form of reducing the data via t-SNE or another dimensionality reduction technique, creating a KDE density plot of some trait, and creating a binary classification problem to define two strains as similar or dissimilar (by euclidean distance, or some other measurable trait within the reduction). An example of this on a much smaller dataset is below.





This type of binary classification holds promise, but it has limitations. t-SNE and other local unsupervised learning techniques do not create strong em-

beddings with more than 10K samples. There are hundreds of other reduction techniques, but t-SNE, a local technique, works best at creating a strong threshold difference. In this KDE density plot, the t-SNE plot's euclidean distances are split by clade, which is an imperfect measure of genetic dissimilarity defined manually by scientists. An interesting endeavor could be creating a better measure of genetic dissimilarity that is more mathematically definable, creating a reduction technique that creates binary classification for each genome relationship, and passing this massive dataset into a neural network to see if it can detect if a genome pair is related or not. Another interesting problem would be to color an unsupervised embedding using HDBSCAN or some other clustering algorithm, and use those numbers as a "clade" classification number to pass into a neural network. These are all possible ways to reveal population structure, but they need to be scaled up in a biologically meaningful way to train the network with.

2) Tuning the SANTA-SIM simulation

More research will need to be done in creating a simulation that accurately describes the real data the network will need to predict. A possible avenue is writing a script pr program that can read a disease population and infer the best simulation to reproduce this output. That would create a more streamlined process for creating a simulation to train the neural network on. The main goal is to remove human bias wherever possible to ensure strong conclusions.

6 Acknowledgments

I would like to thank Dr. Avani Wildani for all of her help, guidance, and suggestions about computations resources, neural network training, and much more. I would like to thank Pioneer Academics for the opportunity to conduct my own research within this field and offering the resources and community to make this process rewarding. I would also like to thank the Bedford Lab, specifically John Huddleston, for the guidance and data needed to complete this project.

References

[1] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

[2] Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

[3] J. Hadfield, C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, and R. A. Neher.

[4] Metsky H.C., Matranga C.B., Wohl S., Schaffner S.F., Freije C.A., Winnicki S.M., West K., Qu J., Baniecki M.L., Gladden-Young A., Lin A.E., Tomkins-Tinch C.H., Ye S.H., Park D.J., Luo C.Y., Barnes K.G., Shah R.R., Chak B., Barbosa-Lima G., Delatorre E., Vieira Y.R., Paul L.M., Tan A.L., Barcellona C.M., Porcelli M.C., Vasquez C., Cannons A.C., Cone M.R., Hogan K.N., Kopp E.W., Anzinger J.J., Garcia K.F., Parham L.A., Gélvez Ramírez R.M., Miranda Montoya M.C., Rojas D.P., Brown C.M., Hennigan S., Sabina B., Scotland S., Gangavarapu K., Grubaugh N.D., Oliveira G., Robles-Sikisaka R., Rambaut A., Gehrke L., Smole S., Halloran M.E., Villar Centeno L.A., Mattar S., Lorenzana I., Cerbino-Neto J., Valim C., Degrave W., Bozza P.T., Gnirke A., Andersen K.G., Isern S., Michael S.F., Bozza F.A., Souza T.M.L., Bosch I., Yozwiak N.L., MacInnis B.L., and Sabeti P.C. Genome sequencing reveals zika virus diversity and spread in the americas. *Nature*, 2017.

[5] Michael C. Hout, Megan H. Papesh, and Stephen D. Goldinger. Multidimensional scaling. *Wiley Online Library*, 2012.

[6] Richard R Hudson. Properties of a neutral allele model with intragenic recombination. *Theoretical population biology*, 23(2):183–201, 1983.

[7] Abbas Jariani, Christopher Warth, Koen Deforche, Pieter Libin, Alexei J Drummond, Andrew Rambaut, Frederick A Matsen Iv, and Kristof Theys. Santa-sim: simulating viral sequence evolution dynamics under selection and recombination. *Virus evolution*, 5(1):vez003, 2019.

[8] Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 2016.

[9] Kazutaka Katoh, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14):3059–3066, 07 2002.

[10] Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. Character-aware neural language models. *arXiv preprint arXiv:1508.06615*, 2015.

[11] John Frank Charles Kingman. The coalescent. *Stochastic processes and their applications*, 13(3):235–248, 1982.

[12] Sergei L Kosakovsky Pond, David Posada, Michael B Gravenor, Christopher H Woelk, and Simon D W Frost. Automated phylogenetic detection of recombination using a genetic algorithm. *Molecular biology and evolution*, 2006.

[13] Sergei L Kosakovsky Pond, David Posada, Michael B Gravenor, Christopher H Woelk, and Simon D W Frost. Gard: a genetic algorithm for recombination detection. *Bioinformatics (Oxford, England)*, 2006.

[14] Riccardo La Grassa, Ignazio Gallo, and Nicola Landro. Learn class hierarchy using convolutional neural networks. *arXiv preprint arXiv:2005.08622*, 2020.

[15] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[16] Wes McKinney et al. pandas: a foundational python library for data analysis and statistics. *Python for High Performance and Scientific Computing*, 14(9), 2011.

[17] I.M. Claro J. Thézé J. G. de Jesus M. Giovanetti M. U. G. Kraemer S. C. Hill A. Black A. C. da Costa L. C. Franco S. P. Silva C.-H. Wu J. Raghwani-S. Cauchemez L. du Plessis M. P. Verotti W. K. de Oliveira E. H. Carmo G. E. Coelho A. C. F. S. Santelli L. C. Vinhal C. M. Henriques J. T. Simpson M. Loose K. G. Andersen N. D. Grubaugh S. Somasekar C. Y. Chiu J. E. Muñoz-Medina C. R. Gonzalez-Bonilla C. F. Arias L. L. Lewis-Ximenez S. A. Baylis A. O. Chieppe S. F. Aguiar C. A. Fernandes P. S. Lemos B. L. S. Nascimento H. A. O. Monteiro I. C. Siqueira M. G. de Queiroz T. R. de Souza J. F. Bezerra M. R. Lemos G. F. Pereira D. Loudal L. C. Moura R. Dhalia R. F. França T. Magalhães E. T. Marques Jr T. Jaenisch G. L. Wallau M. C. de Lima V. Nascimento E. M. de Cerqueira M. M. de Lima D. L. Mascarenhas J. P. Moura Neto A. S. Levin T. R. Tozetto-Mendoza S. N. Fonseca M. C. Mendes-Correa F. P. Milagres A. Segurado E. C. Holmes A. Rambaut T. Bedford M. R. T. Nunes E. C. Sabino L. C. J. Alcantara N. J. Loman N. R. Faria, J. Quick and O. G. Pybus. Establishment and cryptic transmission of zika virus in brazil and the americas. *Nature*, may 2017.

[18] Moritz U. G. Kraemer Gytis Dudas Amanda L. Tan Karthik Gangavarapu Michael R. Wiley Stephen White Julien Thézé Diogo M. Magnani Karla Prieto Daniel Reyes Andrea M. Bingham Lauren M. Paul Refugio Robles-Sikisaka Glenn Oliveira Darryl Pronty Carolyn M. Barcellona Hayden C.

Metsky Mary Lynn Baniecki Kayla G. Barnes Bridget Chak Catherine A. Freije Adrianne Gladden-Young Andreas Gnirke Cynthia Luo Bronwyn MacInnis Christian B. Matranga Daniel J. Park James Qu Stephen F. Schaffner Christopher Tomkins-Tinch Kendra L. West-Sarah M. Winnicki Shirlee Wohl Nathan L. Yozwiak Joshua Quick Joseph R. Fauver Kamran Khan Shannon E. Brent Robert C. Reiner Jr Paola N. Lichtenberger Michael J. Ricciardi Varian K. Bailey David I. Watkins Marshall R. Cone Edgar W. Kopp IV Kelly N. Hogan Andrew C. Cannons Reynald Jean Andrew J. Monaghan Robert F. Garry Nicholas J. Loman Nuno R. Faria Mario C. Porcelli Chalmers Vasquez Elyse R. Nagle Derek A. T. Cummings Danielle Stanek Andrew Rambaut Mariano Sanchez-Lockhart Pardis C. Sabeti Leah D. Gillis Scott F. Michael Trevor Bedford Oliver G. Pybus Sharon Isern Gustavo Palacios & Kristian G. Andersen Nathan D. Grubaugh, Jason T. Ladner. Genomic epidemiology reveals multiple introductions of zika virus into the united states. *Nature*, may 2017.

[19] Lam-Tung Nguyen, Heiko A. Schmidt, Arndt von Haeseler, and Bui Quang Minh. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, 32(1):268–274, 11 2014.

[20] David Posada and Keith A. Crandall. Evaluation of methods for detecting recombination from dna sequences: Computer simulations. *Proceedings of the National Academy of Sciences*, 98(24):13757–13762, 2001.

[21] Asheesh Kumar Singh, Baskar Ganapathysubramanian, Soumik Sarkar, and Arti Singh. Deep learning for plant stress phenotyping: trends and future perspectives. *Trends in plant science*, 23(10):883–898, 2018.

[22] Zhicheng Yan, Hao Zhang, Robinson Piramuthu, Vignesh Jagadeesh, Dennis DeCoste, Wei Di, and Yizhou Yu. Hd-cnn: Hierarchical deep convolutional neural network for large scale visual recognition. *arXiv preprint arXiv:1410.0736*, 2014.

[23] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657, 2015.

[24] Xinqi Zhu and Michael Bain. B-cnn: branch convolutional neural network for hierarchical classification. *arXiv preprint arXiv:1709.09890*, 2017.