Forecasting dominance of SARS-CoV-2 lineages by anomaly detection using deep AutoEncoders

Simone Rancati[1], Giovanna Nicora[1], Mattia Prosperi[2,3], Riccardo Bellazzi[1]*, Simone Marini[2,3]*, Marco Salemi[3,4]*

1. Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Pavia, Italy
2. Department of Epidemiology, College of Public Health and Health Professions, University of Florida, Gainesville, FL, USA
3. Emerging Pathogens Institute, University of Florida, Gainesville, FL, USA
4. Department of Pathology, Immunology and Laboratory Medicine, College of Medicine, University of Florida, Gainesville, FL, USA

* Corresponding authors:

Dr. Marco Salemi
Emerging Pathogens Institute
University of Florida
2055 Mowry Rd.,
Gainesville, FL 32610, USA
salemi@pathology.ufl.edu

Dr. Simone Marini
Dept. of Epidemiology,
College of Public Health and Health Professions,
University of Florida,
Gainesville, FL, USA
simone.marini@ufl.edu

Dr. Riccardo Bellazzi
Dept. Industrial Engineering & Information,
University of Pavia,
IT-27100
Pavia, Italy
riccardo.bellazzi@unipv.it

Abstract: The COVID-19 pandemic exemplified the need for a rapid, effective genomic-based surveillance system to predict emerging SARS-CoV-2 variants and lineages. Traditional molecular epidemiology methods, which leverage public health surveillance or integrated sequence data repositories, are able to characterize the evolutionary history of infection waves and genetic evolution but fall short in predicting future outlooks in promptly anticipating viral genetic alterations. To bridge this gap, we introduce a novel Deep learning, autoencoder-based method for anomaly detection in SARS-CoV-2 (DeepAutoCov). Trained and updated on the public global SARS-CoV-2 GISAID database. DeepAutoCov identifies Future Dominant Lineages (FDLs), defined as lineages comprising at least 25% of SARS-CoV-2 genomes added on a given week, on a weekly basis, using the Spike (S) protein. Our algorithm is grounded on anomaly detection *via* an unsupervised approach, which is necessary given that FDLs can be known only a posteriori (i.e., after they have become dominant). We developed two concurrent approaches (a linear unsupervised and a posteriori supervised) to evaluate DeepAutoCoV performance. DeepAutoCoV identifies FDL, using the spike (S) protein, with a median lead time of 31 weeks on global data and achieves a positive predictive value ~7x better and 23% higher than the other approaches. Furthermore, it predicts vaccine related FDLs up to 17 months in advance. Finally, DeepAutoCoV is not only predictive but also interpretable, since it can pinpoint specific mutations within FDLs, generating hypotheses on the potential increases in virulence or transmissibility of a lineage. By integrating genomic surveillance with artificial intelligence, our work marks a transformative step that may provide valuable insights for the optimization of public health prevention and intervention strategies.

INTRODUCTION

The COVID-19 pandemic has exerted an unprecedented impact on global health, economy, and daily life causing nearly 7 million deaths between December 2019 and October 2023 [1]. Since its first introduction in our species, the original SARS-CoV-2 Wuhan strain [2, 3] has rapidly diversified in distinct lineages classified by specific genotypic mutations in the Spike (S) glycoprotein [4-6]. The Centers for Disease Control and Prevention (CDC) has also designated several Variants Being Monitored (VBM) and Variants of Concern (VOC) – e.g., Alpha (B.1.1.7), Beta (B.1.351), Gamma (P.1), Delta (B.1.617.2), or the currently circulating Omicron (B.1.1.529) and their lineages – associated with increased transmissibility, more severe disease (e.g., increased hospitalizations or deaths), significant reduction in neutralization by antibodies generated during previous infection or vaccination, or reduced

effectiveness of treatments or vaccines [7-12]. SARS-CoV-2 variants and lineages have been attributed to the occurrence of secondary, tertiary, quaternary, and quinary COVID-19 epidemic waves [13]. As hundreds, if not thousands, of new spike (S) protein variants continue to emerge, it is unfeasible to conduct experimental risk assessments whenever a new variant or lineage is identified. Therefore, one significant public health challenge is the need to develop early warning systems capable of quickly evaluating emerging strains in terms of their epidemic or pandemic threat potential [14, 15].

The World Health Organization has outlined specific criteria to select vaccine reference strains in the context of circulating strains [16]. Unfortunately, the impossibility to predict which virus variant will gain dominance in any area of the world and how long that variant will remain dominant is a practical limit as to how often vaccine composition changes can be implemented either regionally or globally [16].

Artificial Intelligence (AI), an especially Deep Learning methods, has been used to predict SARS-CoV-2 mutations at specific amino acid sites of the S protein [17-20] or characterize drivers of emerging strains evolution by calculating antibody escape and binding affinity of the protein to ACE2, which allows viral entry in host cells [14].

In the present study, we propose DeepAutoCoV: a novel approach based on an AutoEncoder model to predict future dominant lineages (FDLs) of SARS-CoV-2. We define an FDL as any lineage (including its sublineages) that will become predominant in the lineages landscape, i.e., a lineage that at any moment in time will constitute a sizable percentage (e.g., >25%, or any other percentage, depending on user-defined stringency criteria) of all the sequenced genomes in a specific time period. DeepAutoCoV is used to detect anomalies, i.e., data that do not fit learned pattern distributions [21]. Our assumption is that, given a steady pathogenic genomic landscape, represented by all the recently sequenced SARS-CoV-2 genomes, FDLs are detectable as anomalies carrying specific genomic differences compared to the circulating ones, with such differences being the factor making them fitter and, therefore, dominant in the future. The key advantage of DeepAutoCoV is that it does not require any *a priori* assumption on which S protein sites or domains are more likely to mutate to generate FDLs.

RESULTS

Predicting FDLs is different from predicting existing lineages or mutations

The problem of SARS-CoV-2 lineage and mutation prediction has been thoroughly

studied [22-28]. The goal of previous works has been focused, however, on predicting existing lineages, that is, given a new sequence, to label it according to a list of known lineages (a supervised multi-class problem) [23-25]; or predicting which mutations will likely arise [28] or their effects [27]. Our proposed task is a *fundamentally different* binary anomaly detection problem, i.e., given a new sequence, we aim to predict whether it belongs to new (unknown) lineages that in the future will become dominant, reaching the aforementioned 25% threshold.

FDLs: assumptions and definition

Ideally, DeepAutoCoV is designed to serve as part of a global or national genomic surveillance system generating and inputting new sequences into DeepAutoCoV in near-real-time. The model would, then, flag SARS-CoV-2 genomes as FDLs. Such genomes could be considered for further analyses and/or uses, e.g., as potential vaccine reference strains (**Figure S1**). To test its performance in successfully detecting FDLs within a given country/region, we simulated epidemic monitoring programs not only at a global scale, but also at the national level. To this end, we downloaded 5,728,544 S protein sequences annotated with their Pango Lineages [27] available in the Global Initiative on Sharing All Influenza Data (GISAID) [29] database from December 24th, 2019, to February 4th, 2023 (**Table S1**). We assume that collection dates correspond to submission dates. In other words, when a S protein is sequenced, it is assumed that it is immediately uploaded to the GISAID database. Based on the difference between the collection and upload dates, GISAID usually occur a few days to a few weeks from the sampling date [30]. While in a real-time applications such a limitation would be significant, the assumption has no impact for training and validation steps. However, currently available sequence data in GISAID are, mostly, convenience samples with over- or under-representation of different countries that may not reflect actual diffusion of different lineages [31-33]. Therefore, we considered a data set including sequences from all other countries (Global data set), as well as four country-based subsets (**Table S1**): United States of America (USA), United Kingdom (UK), Denmark, and France data sets. The four nations were selected to assess the robustness of our AI model's ability of FDLs early identification, in settings with both high (USA, UK) and low (France, Denmark) sequence volume. We pre-processed each data set to ensure data quality and consistency (see Methods). We define dominant lineages independently for each dataset as the top 15 lineages in terms of fraction of total sequences belonging to the lineage over the whole timeline, which led to different dominant lineages for each dataset. Here we assume a 25% fraction to mark the change of status, from FDL to known dominant lineage, although such threshold is arbitrary and can be changed (see below) to

reflect epidemiology and/or public health considerations.

Surveillance implementation strategy and FDLs prediction

We converted protein sequences into fixed-length numeric features representing small protein subsequences of fixed length *k*, called *k*-mers (see Methods) [34-35]. We focused on the S protein as the target of antibody-mediated immunity and the primary antigen in current vaccines [36]. DeepAutoCoV is an AutoEncoder model, consisting of an encoder that compresses the input, a central layer, and a decoder, which reconstructs the input after compression (**Figure 1**) [21, 37-38]. Anomalies are detected by identifying data points deviating significantly form their learned reconstructions (i.e., over 1.5 standard [39-43] deviations from the median), based on the mean square error (MSE). DeepAutoCoV training process includes checkpointing and early stopping mechanisms to prevent overfitting and saving the model when it has achieved its best performance (see Methods). Each week, it is evaluated whether new dominant lineages reach 25% of frequency in the dataset of the

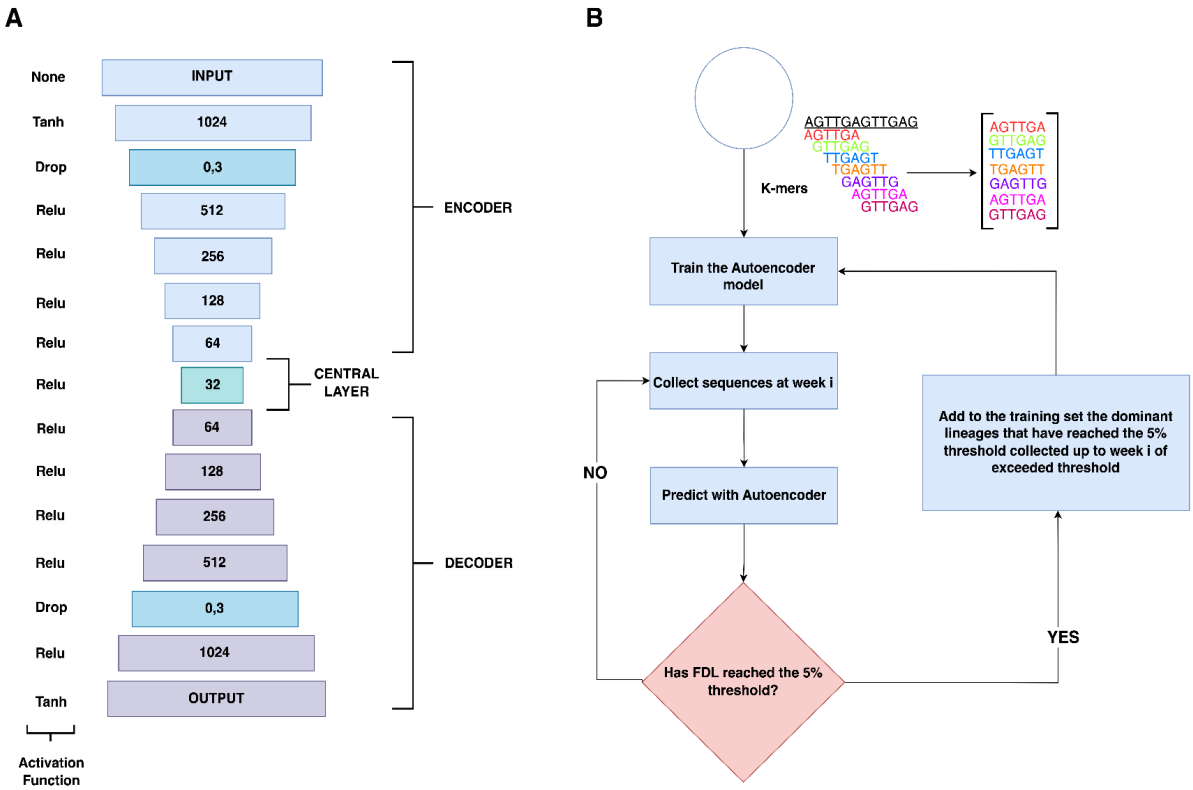

Figure 1. DeepAutoCoV architecture. a) Layered architecture with decreasing neuron counts in the encoding layers (from 1024 to 64) and increasing counts in the decoding layers (from 64 to 1024). A central layer with 32 neurons introduces Gaussian noise to enhance generalization. The decoding layers mirror the encoding structure. The output layer reconstructs the original input data. b) Schematic representation of the automated surveillance system for dominant lineages across various nations. SARS-CoV-2 S protein sequences are gathered, converted into *k*-mers, and filtered. An AutoEncoder outlier detection system is trained on the dataset. Each week, newly uploaded sequences are evaluated as either outliers or not, and these evaluations are recorded as anomalies. During that week, the model is retrained using the initial training dataset in conjunction with all predictive inlier lineages gathered until that time point

corresponding nation. If so, we dynamically update the model by retraining it (see Supplementary Methods for details). As a result, the model continuously updates its understanding of the evolving sequence dynamics, enabling it to capture the latest information about the dominant variants, and becoming more robust against false positives. Since our goal was not to identifying all instances of new dominant variants, but a limited number of sequences that have a high probability of being outliers, we decided to optimize the *Positive Predictive Value (PPV)*, calculated as the ratio between true positives and true positives + false positives. In other words, for the reasons explained above, we aim to optimize the number of genomes that we correctly flag as FDL. We compared DeepAutoCoV with another unsupervised approach, a linear distance-base detection strategy, based on the same features we use, as well as a supervised approach *via* logistic regression (LR), and a dummy method flagging FDLs randomly (See Methods and Supplementary Materials, and **Table S7**).

RESULTS

DeepAutoCoV flags FDLs with a median of 31 weeks in advance

In the global data set, the model achieves a median prediction of 31 weeks (IQR 10; 42) in advance, i.e., 31 weeks before the 25% threshold is reached. For the global data set, USA, UK, and Denmark, the respective (median) values are 47 (IQR 37; 54) weeks, 47.5 (IQR 8; 63), 9 weeks (IQR 8; 23), and 11 weeks (IQR 7; 15) in advance. France, on the other hand, shows performance, with a median value of 16 weeks in advance (IQR 4; 21). Results from the USA dataset (**Figure 2**) show high variability, with difference between 1st and 3d quartile being 54 weeks; while the lowest variability is in the Denmark dataset with a difference of 8 weeks. In contrast, other methods show a lower performance, with the best result on the Global data set being 4 weeks (IQR 0; 8) with the linear approach, and 6 weeks (IQR 0; 8) for the dummy

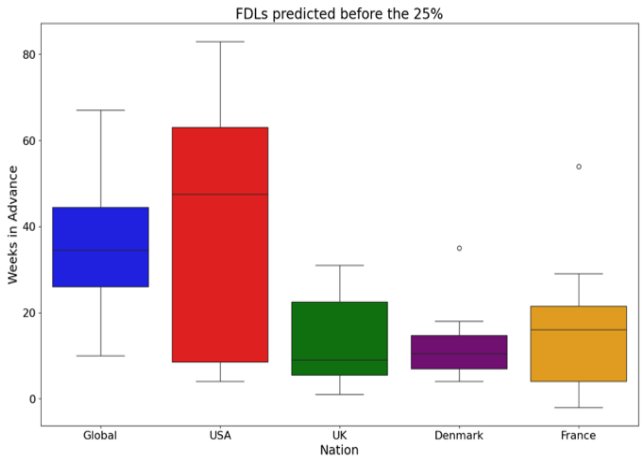

Figure 2: DeepAutoCoV performance in terms of weeks in advance (i.e., before FDLs reach the 25% threshold) on the different datasets.

method. The supervised approach does not correctly flag any FDLs.

DeepAutoCoV flags FDLs during their early emergence, proportional to the data set size (**Table 1**). The median first detection happens when FDLs have a frequency, among all the lineages sampled in given week, from 0.05% (global dataset) to 3% (France). To better understand best- and worst- case scenarios in our simulation,

Table 1: FDL identification: number of weeks in advance for the best and worst fivelineages.

Dataset	General-case FDLs [1]	Worst-case FDLs [1]	Best-case FDLs [1]	FDL frequency at discovery [2]
Global	31 (10, 42)	26 (11, 26)	47 (34, 54)	0.05% (0.03%, 0.3%)
USA	47 (8, 63)	8 (4, 10)	47 (34, 49)	0.9% (0.4%, 3%)
UK	9 (8, 23)	6 (5, 7)	20 (15, 25)	2% (0.6%, 3%)
Denmark	11 (7, 15)	7 (6, 17)	14 (13, 14)	0.9% (0.2%, 3%)
France	16 (4, 21)	2 (3, 5)	20 (20, 23)	3% (1%, 21%)

we focused on the five best and five worst FDL for each data set (**Figure S2** and **Figure S3**), i.e., the FDLs predicted with either the largest (best) or the smallest (worst) advance notice. In the global data set best-case scenario (**Figure S2a**), the median delay between the very first

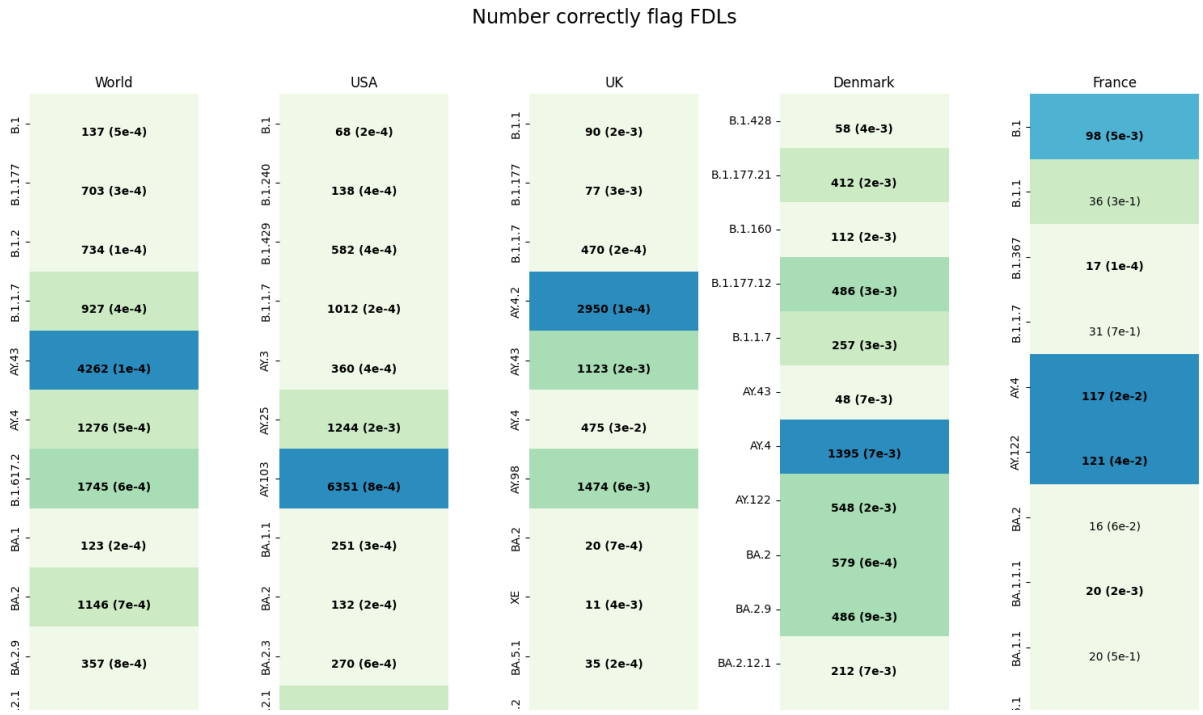

Figure 3: Number of FDLs correctly identified across different datasets.

appearance of an FDL and its discovery is 10 weeks (IQR 10; 15). Conversely, in the worst-case scenarios (**Figure S2b**), the model predicts FDLs with a median of 26 weeks in advance (IQR 11; 26), and again a median delay of 10 weeks (IQR 8; 19). We also calculated the total sequences correctly predicted as FDLs per dataset in their emergence period, i.e., before they reach the 25% frequency threshold (**Table S2-S6**). Lineage AY.43 exhibits the highest count of matches in the World dataset (4,262 identified instances). In the country-specific datasets (**Figure 3**), the top FDL matches are AY.103 for USA (6,351 matches); AY.4.2 for UK (2,940 matches); AY.4 for Denmark (1,395 matches) and AY.4 for France (117 matches). DeepAutoCoV also correctly identifies tens to thousands of FDLs in the global data set. As expected the number of correctly identified FDLs decreases with smaller datasets. Certain FDLs, such as AY.43 and B.1.1.7, become dominant both at the Global level and in country-specific datasets (UK, and Denmark). On the other hand, there are FDLs pertaining only to specific nations that are not listed among the Global dataset FDLs. For instance, XE became an FDL in the UK, but not in the USA, Denmark, or France. The median fraction of correctly identified FDLs, i.e., the PPV in the Global dataset is 0.27 (IQR 0.07; 0.29). In other words, in our genomic surveillance scenario where the top 100 flagged genomes are considered for further analyses, the weekly median of truly identified FDLs is 27 (**Figure S4**). For the other datasets (**Figure S5**), the PPV is 0.28 (IQR 0.09; 0.31) in the USA; 0.26 (IQR 0.07; 0.29) in the UK; 0.26 (IQR 0.05; 0.28) in Denmark and 0.38 (IQR 0; 0.40) in France. DeepAutoCoV outperforms all other approaches: in the Global dataset, other models show a PPV of 0.04 (linear-distance model); 0.02 (dummy approach); and 0 (LR; it does not correctly flag any FDL). Detail performance results are reported in **Table S7**. Other methods to identify FDLs or novel variants being monitored (previously labeled variants of interest or concern) are present in literature, but none of them is based on anomaly detection. In the paper by Begui et al [39] the authors blend Machine Learning with pseudo molecular simulations for novel variant predictions. Unlike our purely data-driven approach, their methodology is based on both injected biological knowledge and transformer-based latent representations (NLP-based classification). We cannot compare their method and ours directly, as there are based on completely different data sets (their dataset is based on a smaller number of sequences and a shorter time (December 2019 - March 2022) and analyzes only global data. we were not able to obtain their dataset). Another relevant approach [28] employs transformers to predict not the emerging lineages, but the mutation probability of each S amino acid.

DeepAutoCoV detects vaccine reference strains months in advance

DeepAutoCoV can help predicting in advance lineages that will become vaccines reference strains. Our simulation shows that, while FDA approved BA.1 vaccine in August 2022 DeepAutoCoV identifies it, as FDL, 17 months earlier. Similarly, BA.4 and BA.5 vaccines were approved in June 2022, but identified by our model 6 and 7 weeks earlier, respectively. In the case of the XBB.1.5 variant, vaccines were approved in June 2023, while the lineage was identified as an anomaly six months earlier.

DeepAutoCoV can identify key FDL mutations

DeepAutoCoV is interpretable and offers insights about key mutations in the S protein contributing to future dominance. In other words, we can identify critical positions in the primary amino acid sequence that are associated to the FDLs. The underlying mechanism revolves around the binary input representation, i.e, a sequence of 0s and 1s (absence or presence of specific *k*-mers). In turn, we can analyze sequences flagged as FDLs and extract *k*-mers that are wrongly reproduced by the model after compression (Figure 1A). By matching these *k*-mers with the S sequence, we can pinpoint critical mutations. For example, while flagging the AY.4 lineage (Delta variant) DeepAutoCoV flags the *k*-mers "NSH", "SHR", and "HRR" which occupy the positions 680, 681, and 682 respectively. This is indeed a critical position characterizing the Delta variant mutation P681R. This is a Furin cleavage site adjuvating Delta variant replication and transmission [40-41]. Another example is the BA.2 (Omicron) lineage, where the model flags a *k*-mer pinpointing the Y505H mutation. This mutation destroys the hydrophobic interactions and has an important role in the diffusion of virus [42]

DISCUSSION AND CONCLUSIONS

This study proposed DeepAutoCoV**,** a novel genomic AutoEncoder model to predict SARS-CoV-2 FDLs, defined as novel lineages that will reach a frequency of 25% or more among the sequenced genomes in any given week. We applied DeepAutoCoV to both global and national (USA, UK, Denmark, and France) SARS-CoV-2 S protein data sets and checked its robustness with decreasing number of available samples (from ~4.5 million to ~130,000).

The USA Centers for Disease Control and Prevention (CDC) uses a four-label system to classify SARS-CoV-2 lineages: Variant of high consequence, Variant of concern, Variant of interest, and Variants being monitored. Typically, a dangerous lineage is flagged as such only after public health data showed its concerning capabilities, such as a spike in infections or hospitalizations. Our main purpose, on the other hand, was robust prediction of FDLs *as*

early as possible. In our envisioned approach, DeepAutoCoV flags FDLs ahead of time, early enough to allow public health interventions, apt to potentially prevent FDL from becoming such and spreading in the population. FDLs detection can, therefore, be the first step in a coordinated genomic surveillance effort (**Figure S1**). Next steps after FDL identification would include (a) molecular simulations, i.e., computational techniques modeling the behavior of the S protein to study if the FDL structures could provide critical advantages to virus ability to spread, such as enhanced ability to bind the receptors; and (b) virulence assays or other phenotypic wetlab validation [37, 49-450]. As resources are limited, it would not be realistic to envision a global or national pandemic surveillance program providing virulence assays or phenotypic wetlab test to include all the sequenced SARS-CoV-2 genomes. For this reason, we designed DeepAutoCoV as the mean to extract a small, manageable number of sequences to be analyzed. In our experiment, DeepAutoCoV is used to retain the top 100 sequences per week in terms of MSE, i.e., the sequences that are most likely FDLs. As a contextual examples, after the end of our considered period (February 2023), new lineages took over the SARS-CoV-2 genomic landscape, with particular XBB.1.5 being the most transmissible Omicron subvariant identified to date. According to CDC data, XBB and XBB.1.5 accounted for 44.1% of COVID-19 cases in the USA on the week of Dec. 31, 2022, compared with 25.9% during the previous week [52]. DeepAutoCoV identified XBB.1.5 as FDL in the World and USA datasets 8 weeks and 3 weeks before they reached the 25% threshold, respectively. Furthermore, DeepAutoCoV provides a list of specific mutations marking the candidate FDLs, providing a hypothesis generating pathway to gain biological insight on these FDLs: in particular our model identifies Q183E mutation that is an unusual mutation as having a global prevalence below 0.01% [54].

DeepAutoCoV has some limitations and makes two main assumptions: (1) the collection date provided with the GISAID sequences corresponds to the submission date (i.e., the sequences are collected and submitted the same day); and (2) when a FDL reaches a frequency of 25% or more among the collected sequences in a given week, it has emerged, i.e., it is no longer an FDL, and its diffusion has been noticed through public health data.

In conclusion, our work marks a transformative step in integrating genomic surveillance with AI for enhanced epidemic surveillance, offering valuable insights for—early—public health responses. Additionally, the DeepAutoCoV framework is suitable not only for SARS-CoV-2, but generalizable to other rapidly evolving viruses.

METHODS

Data sets

We downloaded from GISAID all S protein sequences, along with their Pango Lineages [27], sampled between December 24th, 2019, and February 4th, 2023, and filtered the data by country to assure high quality, by retaining S protein sequences without missing or unrecognized amino acids, and with lengths in the range between the median length of the proteins of their lineage plus or minus 30 amino acids, consistent with S expected lengths (1241 to 1301). After filtering, we generated five distinct datasets: four representing, respectively, each nation of interest, and one including the remaining sequences of all other countries (**Table S1**).

Feature extraction and AutoEncoder Deep Learning model

We focused on the S protein as the target of antibody-mediated immunity and the primary antigen in current vaccines [36]. We converted protein sequences into fixed-length numeric features using *k*-mers as previously described [34, 35]. The AutoEncoder, which is designed to handle *k*-mers in the input layer, consists of: an encoding layer, followed by a series of dense layers, and a dropout layer for reducing overfitting; the central part includes a dense layer with Gaussian noise and 32 neurons; the decoder mirrors the encoder's structure but with an increasing number of neurons, and the output layer reconstructs the input data (see Supplementary Methods for details). The AutoEncoder is trained over ten epochs with a batch size of 256, utilizing the mean square error (MSE) loss function and Adam optimizer, with TensorFlow and Keras Python libraries [37, 38]. Anomalies are identified as data points deviating significantly from learned reconstructions by the MSE plus 1.5 times its standard deviation. The initial training is based on the sequences uploaded on dataset of each nation during the week when sequenced S proteins are available for the first time (**Table S8**). In the first week, each sequence belongs to the lineage that represents the steady-state virus landscape. To reduce the number of features (*k*-mers), we kept only *k*-mers that are present in at least 25% of the sequences in the first training set. In the following weeks, the initially trained AutoEncoder to predict if there are outliers (i.e., new dominant variants that do not belong to the steady state) among the newly introduced sequences. If so, we dynamically update the model by retraining it (see Supplementary Methods for details). For each dataset, the model is evaluated in terms of 1) the ability to detect at least one true new dominant lineage before it reaches 25% of frequency in respective dataset of nation, 2) the positive predictive value

(PPV), i.e., the fraction of new dominant lineages that the model accurately recognizes (see Supplementary Methods for details).

Model comparison

We compared the performance of the AutoEncoder model by developing a supervised approach using Logistic Regression (LR) and a linear distance-based (DB) model i.e., an anomaly detection using linear distance instead of the features extracted by the AutoEncoder. The LR model establishes a relationship between a series of independent variables and a binary dependent variable, providing as output the probability that a particular observation belongs to one of the two categories (in our case Anomaly and non-Anomaly). We used the Logistic Regression function of the Scikit-Learn library [38]. Briefly, each week following the baseline data is considered a test set, with FDL and non-FDL sequences. When an FDL reaches the 25% threshold, its sequences are added to the training set and the class is switched to non-FDL, i.e., the lineage is no longer an anomaly to be detected. The DB model ranks sequences according to the sum of *k*-mers. Sequences whose sum of *k*-mers deviates from the sum of *k*-mers in the training set are defined as anomalies. As with the AutoEncoder model, we considered the top 100 ranked by week to be marked as FDLs. The dummy method was implemented by assigning the FDL class to randomly sampled sequences (See Supplementary Methods for details).

Acknowledgments

NIH NIAID R01 AI170187

Contributions

SM conceived the paper. SM, MS, RB, and MP designed the experiments. SR, GN, and SM, prepared the data and designed the models. SR and GN performed the experiments. All authors analyzed the results and wrote the paper.

References

1. World Health Organization (WHO). WHO coronavirus (COVID-19) dashboard. Available at: https://covid19.who.int. Accessed August 16th, 2023.
2. Zhou, P. et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270-273, doi:10.1038/s41586-020-2012-7 (2020).
3. Wu, F. et al. A new coronavirus associated with human respiratory disease in China. *Nature* **579** (7798), 265-269 (2020).
4. Li, T. et al. Phylogenetic supertree reveals detailed evolution of SARS-CoV-2. *Sci Rep* **10**(1), 22366 (2020).
5. Rambaut , A. et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature Microbiol* **5**, 1403–1407 (2020).
6. O'Toole, A. et al. Pango lineage designation and assignment using SARS-CoV-2 spike gene nucleotide sequences. *BMC Genomics* **23**, 121 (2022).
7. Centers for Disease Control and Prevention (CDC). SARS-CoV-2 Variant Classifications and Definitions, available at: https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-classifications.html. Accessed August 16th, 2023.
8. Goldswain, H. et al. The P323L substitution in the SARS-CoV-2 polymerase (NSP12) confers a selective advantage during infection. *Genome Biol* **24**, 47, doi:10.1186/s13059-023-02881-5 (2023).
9. Plante, J. A. et al. Spike mutation D614G alters SARS-CoV-2 fitness. *Nature* **592**, 116-121, doi:10.1038/s41586-020-2895-3 (2021).
10. Willett, B. J. et al. SARS-CoV-2 Omicron is an immune escape variant with an altered cell entry pathway. *Nat Microbiol* **7**, 1161-1179, doi:10.1038/s41564-022-01143-7 (2022).
11. Prince, T. et al. Analysis of SARS-CoV-2 in Nasopharyngeal Samples from Patients with COVID-19 Illustrates Population Variation and Diverse Phenotypes, Placing the Growth Properties of Variants of Concern in Context with Other Lineages. *mSphere* **7**, e0091321, doi:10.1128/msphere.00913-21 (2022).
12. Otto, S. P. et al. The origins and potential future of SARS-CoV-2 variants of concern in the evolving COVID-19 pandemic. *Curr Biol* **31**(14), R918-R929 (2021).
13. Mabuka, T. et al. The Impact of SARS-CoV-2 Lineages (Variants) and COVID-19 Vaccination on the COVID-19 Epidemic in South Africa: Regression Study. *JMIRx Med* **4**, e34598 (2023).
14. Beguir, K. et al. Early computational detection of potential high-risk SARS-CoV-2 variants. *Comput Biol Med* **155**, 106618 (2023).
15. Subissi, L. et al. An early warning system for emerging SARS-CoV-2 variants, *Nat Med* **28**, 1110-1115 (2022).
16. World Health Organization. TAG-CO-VAC Annex: Statement on the antigen composition of COVID-19 vaccines, May 18th, 2023.
17. Taft, J. M. et al. Deep mutational learning predicts ACE2 binding and antibody escape to combinatorial mutations in the SARS-CoV-2 receptor-binding domain. *Cell* **185**, 4008–4022.e4014 (2022).
18. Maher, M. C. et al. Predicting the mutational drivers of future SARS-CoV-2 variants of concern. *Sci Transl Med* **14**, eabk3445, (2022).

19. Wang, G. et al. Deep-learning-enabled protein–protein interaction analysis for prediction of SARS-CoV-2 infectivity and variant evolution. *Nature Med* **29**, 2007–2018 (2023).

20. Zhou, B. et al. TEMPO: A transformer-based mutation prediction framework for SARS-CoV-2 evolution. *Comput Biol Med.* **152**, 106264. doi: 10.1016/j.compbiomed.2022.106264 (2023).

21. Chen, Z. et al. Autoencoder-based network anomaly detection. Wireless Communications Symposium, Phoenix, AZ (USA), pp. 1-5.

22. Nicora G, Salemi M, Marini S, Bellazzi R. Predicting emerging SARS-CoV-2 variants of concern through a One Class dynamic anomaly detection algorithm. BMJ Health Care Inform. 2022 Dec;29(1): e100643. doi: 10.1136/bmjhci-2022-100643. PMID: 36593658; PMCID: PMC9742845.

23. S. Ali, T. Murad, P. Chourasia and M. Patterson, "Spike2Signal: Classifying Coronavirus Spike Sequences with Deep Learning," *2022 IEEE Eighth International Conference on Big Data Computing Service and Applications (BigDataService)*, Newark, CA, USA, 2022, pp. 81-88, doi: 10.1109/BigDataService55688.2022.00020.

24. Áine O'Toole, Emily Scher, Anthony Underwood, Ben Jackson, Verity Hill, John T McCrone, Rachel Colquhoun, Chris Ruis, Khalil Abu-Dahab, Ben Taylor, Corin Yeats, Louis du Plessis, Daniel Maloney, Nathan Medd, Stephen W Attwood, David M Aanensen, Edward C Holmes, Oliver G Pybus, Andrew Rambaut, Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool, Virus Evolution, Volume 7, Issue 2, December 2021, veab064, https://doi.org/10.1093/ve/veab064S. Ali, "Evaluating COVID-19 Sequence Data Using Nearest-Neighbors Based Network Model," 2022 IEEE International Conference on Big Data (Big Data), Osaka, Japan, 2022, pp. 5182-5188, doi: 10.1109/BigData55660.2022.10020653.

25. S. Ali and M. Patterson, "Spike2Vec: An Efficient and Scalable Embedding Approach for COVID-19 Spike Sequences," *2021 IEEE International Conference on Big Data (Big Data)*, Orlando, FL, USA, 2021, pp. 1533-1540, doi: 10.1109/BigData52589.2021.9671848. S. Ali, T. Murad, P. Chourasia and M. Patterson.

26. Zhao H, Han K, Gao C, et al. VOC-alarm: mutation-based prediction of SARS-CoV-2 variants of concern. Bioinformatics2022; 38:3549–56.

27. O'Toole, Á., et al. Pango lineage designation and assignment using SARS-CoV-2 spike gene nucleotide sequences. *BMC Genomics* **23**, 121 (2022).

28. Binbin Zhou, Hang Zhou, Xue Zhang, Xiaobin Xu, Yi Chai, Zengwei Zheng, Alex Chichung Kot, Zhan Zhou, TEMPO: A transformer-based mutation prediction framework for SARS-CoV-2 evolution, Computers in Biology and Medicine, Volume 152, 2023, 106264, ISSN 0010-4825, https://doi.org/10.1016/j.compbiomed.2022.106264. https://www.sciencedirect.com/science/article/pii/S0010482522009726).

29. Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob Chall* **1**, 33-46 (2017).

30. Gaurav S. The lag in SARS-CoV-2 genome submissions to GISAID *Nature Biotechnology* **39**, 1058-1060 (2021).

31. Simone Marini, Carla Mavian, Alberto Riva, Mattia Prosperi, Marco Salemi, Brittany Rife Magalis, Optimizing viral genome subsampling by genetic diversity and temporal distribution (TARDiS) for phylogenetics, Bioinformatics, Volume 38, Issue 3, February 2022, Pages 856–860, https://doi.org/10.1093/bioinformatics/btab725

32. Mavian C, Marini S, Prosperi M, Salemi M. A Snapshot of SARS-CoV-2 Genome Availability up to April 2020 and its Implications: Data Analysis. JMIR Public Health Surveill. 2020 Jun 1;6(2):e19170. doi: 10.2196/19170. Erratum in: JMIR Public Health Surveill. 2020 Aug 10;6(3):e22853. PMID: 32412415; PMCID: PMC7265655.

33. Carla Mavian, Sergei Kosakovsky Pond, Simone Marini, Brittany Rife Magalis, Anne-Mieke Vandamme, Simon Dellicour, Samuel V. Scarpino, Charlotte Houldcroft, Julian Villabona-Arenas, Taylor K. Paisie, Nídia S. Trovão, Christina Boucher, Yun Zhang, Richard H. Scheuermann, Olivier Gascuel, Tommy Tsan-Yuk Lam, Marc A. Suchard, Ana Abecasis, Eduan Wilkinson, Tulio de Oliveira, Ana I. Bento, Heiko A. Schmidt, Darren Martin, James Hadfield, Nuno Faria, Nathan D. Grubaugh, Richard A. Neher, Guy Baele, Philippe Lemey, Tanja Stadler, Jan Albert, Keith A. Crandall, Thomas Leitner, Alexandros Stamatakis, Mattia Prosperi, and Marco Salemi

34. Marini et al. The K-mer antibiotic resistance gene variant analyzer (KARGVA). *Front Microbiol* **14**, 1060891 1 2023.

35. Ramirez-Mata, A. S. et al. Machine learning prediction and phyloanatomic modeling of viral neuroadaptive signatures in the macaque model of HIV-mediated neuropathology. *Microbiology Spectrum* **11**(2), e03086-22, 2023.

36. McCallum, M. *et al.* N-terminal domain antigenic mapping reveals a site of vulnerability for SARS-CoV-2. *Cell* **184**, 2332–2347, e16 (2021).

37. Wang, H. Y. & Zhao, S. Auto-encoder based dimensionality reduction. *Neurocomputing* **184**, 232-242 (2016).

38. Pedregosa, F., *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**(85), 2825–2830 (2011).

39. Roberto Corizzo, Michael Baron, Nathalie Japkowicz,CPDGA: Change point driven growing auto-encoder for lifelong anomaly detection,Knowledge-Based Systems,Volume 247,2022,108756,ISSN 0950-7051,https://doi.org/10.1016/j.knosys.2022.108756

40. Prifti, E., Buban, J. P., Thind, A. S., Klie, R. F., Variational Convolutional Autoencoders for Anomaly Detection in Scanning Transmission Electron Microscopy. Small 2023, 19, 2205977. https://doi.org/10.1002/smll.202205977.

41. Givnan S, Chalmers C, Fergus P, Ortega-Martorell S, Whalley T. Anomaly Detection Using Autoencoder Reconstruction upon Industrial Motors. Sensors. 2022; 22(9):3166. https://doi.org/10.3390/s22093166.

42. Pegah Sadeghi Vasafi, Olivier Paquet-Durand, Kim Brettschneider, Jörg Hinrichs, Bernd Hitzmann, Anomaly detection during milk processing by autoencoder neural network based on near-infrared spectroscopy,Journal of Food Engineering,Volume 299,2021,110510,ISSN 0260-8774,https://doi.org/10.1016/j.jfoodeng.2021.110510.

43. Y. Wei, J. Jang-Jaccard, W. Xu, F. Sabrina, S. Camtepe and M. Boulic, "LSTM-Autoencoder-Based Anomaly Detection for Indoor Air Quality Time-Series Data," in IEEE Sensors Journal, vol. 23, no. 4, pp. 3787-3800, 15 Feb.15, 2023, doi: 10.1109/JSEN.2022.3230361

44. Beguir K, Skwark MJ, Fu Y, Pierrot T, Carranza NL, Laterre A, Kadri I, Korched A, Lowegard AU, Lui BG, Sänger B, Liu Y, Poran A, Muik A, Şahin U. Early computational detection of potential high-risk SARS-CoV-2 variants. Comput Biol Med. 2023 Mar;155:106618. doi: 10.1016/j.compbiomed.2023.106618. Epub 2023 Feb 2. PMID: 36774893; PMCID: PMC9892295.

45. Zhou Z, Zhu Y, Chu M. Role of COVID-19 Vaccines in SARS-CoV-2 Variants. Front Immunol. 2022 May 20;13:898192. doi: 10.3389/fimmu.2022.898192. PMID: 35669787; PMCID: PMC9165056.

46. David A. Ostrov, Glenn W. Knox, Emerging mutation patterns in SARS-CoV-2 variants, Biochemical and Biophysical Research Communications, Volume 586, 2022, Pages 87-92, ISSN 0006-291X, https://doi.org/10.1016/j.bbrc.2021.11.059. (https://www.sciencedirect.com/science/article/pii/S0006291X21015758)

47. Chakraborty C, Bhattacharya M, Sharma AR, Mallik B. Omicron (B.1.1.529) - A new heavily mutated variant: Mapped location and probable properties of its mutations with an emphasis on S-glycoprotein. Int J Biol Macromol. 2022 Oct 31;219:980-997. doi: 10.1016/j.ijbiomac.2022.07.254. Epub 2022 Aug 8. PMID: 35952818; PMCID: PMC9359758.

48. Martina Zappa, Paolo Verdecchia, Fabio Angeli, Severe acute respiratory syndrome coronavirus 2 evolution: How mutations affect XBB.1.5 variant,European Journal of Internal Medicine, Volume 112,2023, Pages 128-132, ISSN 0953-6205, https://doi.org/10.1016/j.ejim.2023.03.027.(https://www.sciencedirect.com/science/article/pii/S0953620523001012)

49. Abadi, M., *et al*. Tensorflow: A system for large-scale machine learning. In 12th USENIX Symposium on Operating Systems Design and Implementation, pp. 265–283 (2016).

50. Nicora, G., *et al.* Predicting emerging SARS-CoV-2 variants of concern through a One Class dynamic anomaly detection algorithm. *BMJ Health Care Inform* **29**(1), e100643 (2022).

51. Hospital A, Goñi JR, Orozco M, Gelpí JL. Molecular dynamics simulations: advances and applications. Adv Appl Bioinform Chem. 2015 Nov 19;8:37-47. doi: 10.2147/AABC.S70333. PMID: 26604800; PMCID: PMC4655909.

52. F. Xavier López-Labrador, Julianne R. Brown, Nicole Fischer, Heli Harvala, Sander Van Boheemen, Ondrej Cinek, Arzu Sayiner, Tina Vasehus Madsen, Eeva Auvinen, Verena Kufner, Michael Huber, Christophe Rodriguez, Marcel Jonges, Mario Hönemann, Petri Susi, Hugo Sousa, Paul E. Klapper, Alba Pérez-Cataluňa, Marta Hernandez, Richard Molenkamp, Lia van der Hoek, Rob Schuurman, Natacha Couto, Karoline Leuzinger, Peter Simmonds, Martin Beer, Dirk Höper, Sergio Kamminga, Mariet C.W. Feltkamp, Jesús Rodríguez-Díaz, Els Keyaerts, Xiaohui Chen Nielsen, Elisabeth Puchhammer-Stöckl, Aloys C.M. Kroes, Javier Buesa, Judy Breuer, Eric C.J. Claas, Jutte J.C. de Vries, Recommendations for the introduction of metagenomic high-throughput sequencing in clinical virology, part I: Wet lab procedure, Journal of Clinical Virology, Volume 134, 2021, 104691, ISSN 1386-6532, https://doi.org/10.1016/j.jcv.2020.104691.

53. Yue C, Song W, Wang L, Jian F, Chen X, Gao F, Shen Z, Wang Y, Wang X, Cao Y. ACE2 binding and antibody evasion in enhanced transmissibility of XBB.1.5. Lancet Infect Dis. 2023 Mar;23(3):278-280. doi: 10.1016/S1473-3099(23)00010-5. Epub 2023 Feb 3. PMID: 36746173; PMCID: PMC9897732.