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]*, Marco Salemi[3,4]*, Simone Marini[2,3]*

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

The coronavirus disease of 2019 (COVID-19) pandemic is characterized by sequential emergence of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) variants and lineages outcompeting previously circulating ones because of, among other factors, increased transmissibility and immune escape[1-3]. We devised an unsupervised deep learning AutoEncoder for viral genomes anomaly detection to predict future dominant lineages (FDLs), i.e., lineages or sublineages comprising ≥10% of viral sequences added to the GISAID database on a given week[4]. The algorithm was trained and validated by assembling global and country-specific data sets from 16,187,950 Spike protein sequences sampled between December 24th, 2019, and November 8th, 2023. The AutoEncoder flags low frequency FDLs (0.01% – 3%), with median lead times of 4-16 weeks. Over time, positive predictive values oscillate, decreasing linearly with the number of unique sequences per data set, showing average performance up to 30 times better than baseline approaches. The B.1.617.2 vaccine reference strain was flagged as FDL when its frequency was only 0.01%, more than one year earlier of being considered for an updated COVID-19 vaccine. Our AutoEncoder, applicable in principle to any pathogen, also pinpoints specific mutations potentially linked to increased fitness, and may provide significant insights for the optimization of public health *pre-emptive* intervention strategies.

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 January 2024[5]. Since its first introduction in our species, the original SARS-CoV-2 Wuhan strain[6,7] has rapidly diversified in distinct lineages classified by specific genotypic mutations in the Spike glycoprotein[8-10]. 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 sublineages – 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[11,12]. Novel SARS-CoV-2 lineages and sublineages capable of outcompeting the ones in circulation have been attributed to the occurrence of secondary, tertiary, quaternary, and quinary COVID-19 epidemic waves[13]. Thanks to surveillance systems in (near) real-time, like the one successfully implemented in Africa[14], hypermutated strains under heavy immune pressure, e.g., the recently emerged JN.1[15], are nowadays quickly recognized and labelled as VBM or VOC. However, over the past four years, many other emerging sublineages have also reached sizable frequencies in different countries or geographic areas, further complicating the task of assessing their epidemic potential or the need for updated vaccines[16]. The World Health Organization (WHO) has outlined specific criteria to select vaccine reference strains in the context

of circulating variants[16]. Unfortunately, as hundreds, if not thousands SARS-CoV-2 lineages continue to emerge, it is unfeasible to conduct experimental risk assessments or predict which ones will gain dominance in any area of the world; a practical limit as to how often vaccine composition changes need to be implemented, either regionally or globally[17].

Artificial Intelligence (AI), especially deep learning, has been used to predict SARS-CoV-2 mutations at specific amino acid sites of the Spike protein[18-22] 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[23]. In the present study, we propose DeepAutoCoV, a novel approach based on an AutoEncoder model,[23,24] to forecast future dominant lineages/sublineages (FDLs) of SARS-CoV-2. We define an FDL as any lineage or sublineage, including all its polymorphic variants, that will become dominant (or co-dominant) in the virus genomic landscape, i.e., that will constitute a sizable percentage of all genomes sequenced over a specific time interval. DeepAutoCoV is used to detect anomalies, i.e., genomes that do not fit learned pattern distributions. Our assumption is that, given a steady pathogenic genomic landscape, represented by all known SARS-CoV-2 genomes belonging to known lineages, an FDL is detectable as an anomaly carrying specific mutations compared to the circulating ones, with such mutations being the factor that makes them fitter and, therefore, likely dominant in the future. The key advantage of DeepAutoCoV is that it does not require any *a priori* assumption on which protein sites or domains are more likely to mutate to generate FDLs.

DeepAutoCoV implementation and validation strategy

So far, work on SARS-CoV-2 lineages and mutation prediction has been focused on using AI to label new sequences according to a list of known lineages/sublineages (a supervised multi-class problem)[10], to evaluate their probability of becoming VOC[21,25], to predict mutations likely to arise[17-22], or to rank them for immune escape and fitness potential[21]. Herein, our task is fundamentally different, and based on unsupervised anomaly detection, i.e., given a new sequence, we aim to predict whether it belongs to a future lineage or sublineage that will eventually reach or exceed a pre-defined proportion of all circulating ones in given time point. For computational efficiency, we focused on SARS-CoV-2 Spike protein, as the target of antibody-mediated immunity and the primary antigen in current vaccines[3], although our approach can be applied to full genome sequences and, in principle, to any pathogen with sufficient longitudinal sequence data.We downloaded 16,187,950 Spike protein sequences along with their metadata available in the GISAID database[4] from December 24th, 2019, to November 8th, 2023 (**Supplementary Table S1**). We defined as FDL any lineage reaching or exceeding a 10% threshold frequency among all sequences uploaded in a single week. The threshold is arbitrary and can be changed and defined by the user (see below) to reflect epidemiology and/or public health considerations. The data were used to simulate an

epidemic monitoring program, at global and national level, where new sequences are, ideally, uploaded to the GISAID database in real-time and immediately analyzed by DeepAutoCoV. In such an AI/genomic-based surveillance approach, DeepAutoCov would be able to flag sequences as FDLs at regular time intervals (daily, weekly, etc.), depending on the pace at which new sequences are uploaded, substantially reducing the time required for deployment of public health countermeasures based on epidemiology and experimental data (**Supplementary Figure S1**). In fact, differences between sampling and uploading date usually vary from a few days to a few weeks[26], a limitation that would limit the efficacy of a real-time surveillance program but has no impact for the DeepAutoCoV simulation training and validation steps. Available sequences are also, for the most part, the result of convenience sampling, over- or under-representing countries and time periods that may not reflect actual diffusion of different lineages[27]; sampling bias would affect surveillance accuracy but does not affect our simulation. Therefore, we assembled four country-based subsets – United States of America (USA), United Kingdom (UK), Denmark, and France – of varying size and composition, and a Global data set including sequences from all the selected Spike proteins (**Supplementary Table S1**). The four nations were selected to assess the robustness of DeepAutoCov's ability of FDLs early identification in settings with both high (USA, UK) and low (France, Denmark) sequence volume.

We converted protein sequences into fixed-length numeric features using *k*-mers[28,29]. DeepAutoCoV employs both a deep learning model (AutoEncoder), and prior knowledge for its functioning (see Methods). The AutoEncoder model consists of an encoder that compresses the input, a central layer, and a decoder, which reconstructs the input after compression (**Figure 1A**)[24,29-31]. Anomalies are detected by identifying data points that deviate significantly from their learned reconstructions (i.e., over 1.5 standard deviations from the median)[30,31], based on the Mean Squared Error (MSE). Data for training (Spike sequences, sampling times, and sampling locations of Global and country-specific data sets) were pre-processed to ensure quality and consistency (**Supplementary Table S1**). The training process includes checkpointing and early stopping mechanisms to prevent overfitting (see Methods). The model is dynamically updated weekly and unique sequences pertaining to known FDLs (i.e., after the threshold is reached) are stored and used as a prior knowledge for further filtering (**Figures 1B**, **1C**, and Supplementary Methods). When a predicted FDL reaches its status, i.e., once an FDL represents 10% or more of the sequences generated within a week, we assume the FDL to be known and no longer an anomaly (**Figure 1C**). Each true FDL is a discrete cluster comprising many strains, e.g., strains of the B.1.1.7 (alpha) lineage, or AY.3 (delta) sublineage, that are flagged during the interval between first detection and lineage/sublineage frequency reaching 10% (**Supplementary Figures S2, S3**). Our goal is not to identify all genotypic variants of a possible FDL, which given SARS-CoV-2 rapid accumulation of polymorphic sites[2] may be very large and not necessarily useful, but several sequences belonging to a lineage or sublineage with high probability of

being an outlier because of their specific, *a priori* unknown, amino acid signature(s).

DeepAutoCoV flags FDLs with a median of 16 weeks lead time

Each data set is characterized by different lineages or sublineages spreading, over time, in specific areas at different rates. Therefore, our approach resulted in different FDLs for each data set. In the Global data set, FDLs are flagged within an interquartile range (IQR) of 10 to 35 weeks (median 16) before reaching the 10% threshold. In the USA, UK, Denmark, and France data sets, the respective median values are 12 weeks (IQR 7, 43), 4 weeks (IQR 2, 11), 6 weeks (IQR 4, 11), and 10 weeks (IQR 6, 17) in advance. Results from the USA data set (**Figure 2**) show high variability, with difference between 1st and 3rd quartile being 36 weeks, while the lowest variability is in the Denmark data set with a difference of 7 weeks. DeepAutoCoV flagged all FDLs along with their frequency at discovery time, which is very low (**Table 1**), ranging from 0.01% (Global data set) to 3% (UK). In the five FDLs predicted with either the largest (best) or the smallest (worst) lead time, median delay between FDLs first appearance and discovery varies between zero (no delay) and 6 weeks, consistently for best- and worst-case scenarios (**Supplementary Figures S2-S3**).

DeepAutoCoV precision and comparison with other approaches

Since we aim to optimize the number of genomes correctly flagged as FDLs, we calculated for each data set the Positive Predictive Value (PPV) as the ratio between true positives, i.e., FDL correctly flagged as anomalies, and true positives plus false positives, i.e., non-FDL sequences incorrectly flagged as anomalies (see Methods). PPV is equivalent to the true positive rate (precision). DeepAutoCoV median PPV resulted 0.28 (IQR 0.12; 0.37, p = 4e-24) for the Global Data set, 0.48 (IQR 0.27, 0.69; p = 7e-27) for the USA Data set, 0.30 (0.10, 0.38; p = 5e-12) for the UK Data set, 0.38 (0.08, 0.60; p = 4e-08) for the Denmark Data set, and 0.51 (0.20, 0.68; p-value 7e-13) for the France data set. As shown in **Figure 3** and **Supplementary Figures S4** and **S5**, over time PPV moving average follows the same trend in each data set. Starting from week 1, when the first SARS-CoV-2 sequences became available in GISAID, PPVs are unequally distributed between time intervals (in weeks) when most of the FDLs are missed (low PPV), and intervals displaying high (0.65-1.0) PPV. The reason for such a behavior can be found by noting the negative correlation (Pearson coefficient: -0.51) between number of unique sequences and average PPV across the five data sets (**Supplementary Figure S6**). During weeks when most of the sequences uploaded are unique because of few polymorphic sites, usually the result of random genetic drift, flagged outliers are likely to be false positives; the contrary is true in weeks when uploaded sequences belong to multiple tight clusters, especially if the frequency of such clusters is increasing over time. In other words, during weeks when most of the sequences uploaded in GISAID are unique, flagged FDLs tend to have low precision, while weeks when distinct clusters of polymorphic sequences are uploaded, precision significantly increases.

We compared DeepAutoCoV with other prediction algorithms (see **Supplementary Table S2**, and Methods): a linear distance-based (DB) detection strategy (using linear distances instead of features extracted by DeepAutoCoV), a supervised approach *via* logistic regression (LR), and state-of-the-art algorithms Spike2Signal (S2S)[32] and Spike2vec (S2V)[33]. On average, DeepAutoCoV outperforms all other approaches in all the data sets. In the Global data set, other models obtain a PPV of 0.05 (linear model), or 0 (all supervised methods, i.e., LR, S2V, and S2S fail to correctly flag any FDL). Not surprisingly, supervised models seem inadequate to detect FDLs, as by design *emerging* FDLs are unknown to the model. This limitation may cause delays or inability of supervised models to recognize FDLs as anomalies. The use of anomaly detection models[25,34], or models based on biological knowledge[18] is necessary for proper identification of FDLs.

Other methods to identify FDLs or novel variants being monitored (previously labeled variants of interest or concern) are present in literature, but we could not directly compare them to DeepAutoCoV because of different assumptions, test sets, and data granularity, such as discovering variants instead of lineages and sublineages[15,18,34]. For example, a relevant approach[18] employs transformers to predict not the emerging lineages, but the mutation probability of each Spike amino acid.

Advantages and potential applications of DeepAutoCoV

The top ten clusters of flagged sequences in each data set, i.e. the top ten comprising the highest numbers of flagged sequences, correspond to major lineages (alpha, delta, omicron) and several sublineages, all descendants of the highly divergent B.1 parental lineage, itself identified in the Global and French data sets (**Figure 4**), in agreement with the isolation of the earliest known strains in Switzerland and neighboring France[35,36]. Our simulations indicates that in an actual real-time surveillance system, DeepAutoCoV would be capable of early flagging hundreds to thousands of strains belonging to different clusters, which can be easily identified by phylogenetic or dimensionality reduction method[32,33] such as principal component analysis, and provide a solid guess of which ones will become the next dominant lineages/sublineages.

FDLs detected in each filtered data sets may be different for the first time that a new lineage/sublineage appears in GISAID (since earlier low-quality or partial sequences were discarded), as well as they may depend on both SARS-CoV-2 epidemiology patterns in specific locations and potential bias (e.g., sampling bias) in the sequence data (**Supplementary Tables S3-S7**). For example, alpha variant (B.1.1.7) strains were detected in UK for the first time in April 2020, but appear in our filtered data set the 4th week of August 2020, and were flagged as FDL the same week, when their frequency was 0.05%, eleven weeks before the variant reached the 10% threshold (**Supplementary Table S5**). In the US, the first alpha variant sequences appeared in the filtered data set during the 1st week of March 2020, and flagged during the same week, when their frequency was only 0.2%, 52 weeks before they reached the threshold

(**Supplementary Table S4**). The better performance of DeepAutoCoV with the USA data set is likely linked to USA larger number of sequences in GISAID compared to UK, at that time, thus resulting in better training. Denmark and France also performed worse than USA (**Supplementary Tables S6** and **S7**). In the Global data set, alpha was uncovered 2nd week of February 2020 and flagged as FDL the same week, when its frequency was 0.7%, 46 weeks ahead of the threshold (**Supplementary Table S3**). In general, DeepAutoCoV appears to perform consistently well with both high (Global, USA, UK) and low (France, Denmark) sequence volume. For each data set, the median delay between FDL appearance and detection is 0, meaning that most of the FDLs are discovered on the same time point they appear. Reflecting the bias in the data sets, neither beta nor gamma were labeled as FDLs. Despite co-circulating with the alpha variant, beta and gamma infections were limited to regions not well represented in GISAID at that time, such as Brazil[37] and Haiti[38]. The potential usefulness of our approach is also exemplified by its ability to forecast as FDLs several omicron sublineages (**Supplementary Tables S3-S7**).

The heavily mutated omicron variant (B.1.1.529/21K) was first identified by the South African surveillance network at the beginning of the November 2021 4th week[39], and already designated a VOC by the WHO's Technical Advisory Group on SARS-CoV-2 Virus Evolution by November 26th. Hopefully, molecular monitoring programs will continue to detect and assess hypermutated strains, such as omicron, at increasingly fast pace. Yet, several subsequent omicron sublineages became dominant or co-dominant in different parts of the world. Assessing in real-time which ones of all continuously emerging strains might be an FDL, would be unfeasible, if not aided by a prediction algorithm as the one described herein.

DeepAutoCoV might also help predicting in advance lineages that will become vaccines reference strains, such as BA.1. Earliest BA.1 strains were sampled the 1st week of November 2021. In our simulation using the Global data set, BA.1 is detected the same week it appears, when its frequency is only 0.01% and 10 weeks before reaching the threshold. Similarly, BA.5 and XBB.1.5 are flagged 16 (frequency 0.08%) and 23 weeks (frequency 0.1%), respectively, before each variant reaching the threshold. The case of vaccine variant B.617.2 is particularly relevant because it is flagged as FDL April 2020, not only 73 weeks ahead the 10% threshold but also more than one year before Pfizer researchers began to discuss the lineage as a candidate for a vaccine update[40].

DeepAutoCoV can identify key FDL mutations

DeepAutoCoV is interpretable and can provide insights about key mutations contributing to future dominance, i.e., it 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 (a sequence of 0s and 1s indicating absence or presence of specific *k*-mers). Sequences flagged as FDLs can be analyzed by extracting *k*-mers that are wrongly reproduced by the model after compression (**Figure 1A**). By matching

these *k*-mers with the Spike 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, a Furin cleavage site adjuvating Delta variant replication and transmission[3]. Another example is the BA.2 (Omicron) lineage, where the model flags a *k*-mer pinpointing the Y505H mutation, which destroys the hydrophobic interactions and has an important role in viral spread[3].

Discussion and Conclusions

The US CDC typically flags a dangerous lineage after public health data showed its concerning capabilities, such as a Spike in infections or hospitalizations. Our main purpose was the prediction of FDLs *as early as possible*. In an actual implementation as part of a genomic surveillance system, DeepAutoCoV flagging FDLs ahead of time would allow public health interventions, such as early vaccine design, to potentially prevent the lineage from spreading to the population. FDLs detection can, therefore, be the first step in a coordinated genomic surveillance effort (**Supplementary Figure S1**). Next steps after FDL identification would include (a) molecular simulations, i.e., computational techniques modeling the behavior of the Spike 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 wet lab validation. As resources are limited, it would not be realistic to envision a global or national pandemic surveillance program providing virulence assays or phenotypic wet lab test to include *all* the sequenced SARS-CoV-2 genomes. A method such as DeepAutoCov would be therefore critical in selecting likely FDL candidates. Importantly, the DeepAutoCoV framework is suitable not only for SARS-CoV-2, but easily generalizable to other rapidly evolving viruses.

DeepAutoCoV limitations are related to the data used for the simulation: (1) the collection date provided with the GISAID sequences corresponds to the submission date (i.e., new strains are sampled/sequenced and submitted the same day); and (2) when a FDL reaches a frequency of 10% or more among the collected sequences in a given week, we considered it as emerged, i.e., its diffusion has been noticed through public health data. Other potential limitations are due to the approach being solely based on Spike protein sequences, not considering the rest of the genome, which could carry important information for fitness. Similarly, our approach does not consider environmental data, from prevalence per area, to mobility patterns or public health countermeasures (e.g. vaccination campaigns). Integrating these data types would likely increase the model performance. Nevertheless, our work marks a transformative step in integrating genomic surveillance with AI for enhanced epidemic surveillance, offering valuable insights for early public health responses.

METHODS

Data sets

We downloaded from GISAID all Spike protein sequences, along with their Pango Lineages[10], sampled between December 24th, 2019, and November 8th, 2023, and filtered the data by country. To assure high quality, we retained Spike protein sequences without missing or unrecognized amino acids, and with lengths in the range between the median length of their lineage Spike proteins plus or minus 30 amino acids, with a total range of 1241 to 1301 amino acids. After filtering, we generated five data sets: four representing, respectively, each nation of interest, and one including all the filtered sequences (**Supplementary Table S1**). For our simulation we considered the collection date corresponding to the submission date.

Feature extraction and AutoEncoder deep learning model

We focused on the Spike protein as the target of antibody-mediated immunity and the primary antigen in current vaccines[3]. We converted protein sequences into fixed-length numeric features using *k*-mers as previously described[28,29]. 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 and **Figures 1A**, **1B** for details). The AutoEncoder is trained over ten epochs with a batch size of 256, utilizing the MSE loss function and Adam optimizer, with TensorFlow and Keras Python libraries[41,42]. Anomalies are identified as data points deviating from learned reconstructions by the MSE plus 1.5 times its standard deviation[30,31]. The initial training is based on the sequences uploaded on data set of each nation during the week when sequenced Spike proteins are available for the first time (**Supplementary Table S7**). 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 1% 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 and **Figure 1B**). For each data set, the model is evaluated in terms of 1) the ability to detect at least one true new dominant lineage before it reaches 10% of frequency in respective data set 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). The code to reproduce our analysis and use our model is available at https://github.com/simoRancati/DeepAutoCov.

Model comparison

We compared the performance of the AutoEncoder model by developing a supervised approach using Logistic Regression (LR), Spike2Signal (S2S)[32], Spike2Vec (S2V)[33], 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[42]. Briefly, each week following the baseline data is considered a test set, with FDL and non-FDL sequences. When an FDL reaches the 10% 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 (binary) *k*-mers. Sequences whose sum of *k*-mers deviates from the sum of *k*-mers in the training set are defined as anomalies. To benchmark our model, we also utilized embeddings from literature used with machine learning algorithms to represent the Spike protein of SARS-CoV-2. Specifically, we compared DeepAutoCoV with S2V and S2S, which are embeddings for the Spike protein based on random Fourier transform, and the conversion of amino acid sequences into numerical signals, respectively. These embeddings are designed to create a latent representation of the Spike proteins, thereby enabling the distinct lineages of SARS-CoV-2 to be more easily differentiated. Following the generation of these embeddings, we used supervised machine learning algorithms (Naive Bayes and RESNET respectively) to classify the sequences into the correct lineages as described in the respective papers[32,33]. S2S embeddings were initially computed for all sequences in the various data sets, with RESNET serving as the deep learning model for lineage identification. S2V, in contrast, was recalculated each time by retraining the Naive Bayes model, and we also tested the pre-trained embedding provided by the authors (S2VT). The model retraining adheres to the same methodology implemented for DeepAutoCoV.

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. Tegally, H. et al. Dispersal patterns and influence of air travel during the global expansion of SARS-CoV-2 variants of concern. *Cell* **186**, 3277–3290 (2023).
2. Markov, P. V. et al. The evolution of SARS-CoV-2. *Nature Rev Microbiol* **21**, 361–379 (2023).
3. Harvey, W. T. et al. SARS-CoV-2 variants, spike mutations and immune escape. *Nature Rev Microbiol* **19**, 409–424, (2021).
4. Elbe, S. and Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges* **1**, 33–46 (2017).
5. World Health Organization (WHO). WHO coronavirus (COVID-19) dashboard. Available at: https://covid19.who.int. Accessed November 8th, 2023.
6. 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).
7. Wu, F. et al. A new coronavirus associated with human respiratory disease in China. *Nature* **579** (7798), 265-269 (2020).
8. Li, T. et al. Phylogenetic supertree reveals detailed evolution of SARS-CoV-2. *Sci Rep* **10**(1), 22366 (2020).
9. Rambaut , A. et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature Microbiol* **5**, 1403–1407 (2020).
10. O'Toole, A. et al. Pango lineage designation and assignment using SARS-CoV-2 spike gene nucleotide sequences. *BMC Genomics* **23**, 121 (2022).
11. 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.
12. 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).
13. Thakur et al. Waves and variants of SARS-CoV-2: understanding the causes and effect of the COVID-19 catastrophe. *Infection* **50**(2), 309–325 (2022).
14. Msomi, N., Mlisana, K., and de Oliveira, T. A genomics network established to respond rapidly to public health threats in South Africa. *Lancet Microbe* **1**(6), e229–e230 (2020).
15. Yang, S. et al. Fast Msomi, N., Mlisana, K., and de Oliveira, T. A genomics network established to respond rapidly to public health threats in South Africa. *Lancet Infect Dis* S1473-3099(23)00744-2, e229-e230 (2023). Online ahead of print.
16. World Health Organization. TAG-CO-VAC Annex: Statement on the antigen composition of COVID-19 vaccines, May 18th, 2023.
17. Subissi, L. et al. An early warning system for emerging SARS-CoV-2 variants, *Nat Med* **28**, 1110-1115 (2022).

18. Beguir, K. et al. Early computational detection of potential high-risk SARS-CoV-2 variants. *Comput Biol Med* **155**, 106618 (2023).

19. 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).

20. 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).

21. Maher, M. C. et al. Predicting the mutational drivers of future SARS-CoV-2 variants of concern. *Sci Transl Med* **14**, eabk3445, (2022).

22. Zhao, H. et al. VOC-alarm: mutation-based prediction of SARS-CoV-2 variants of concern. *Bioinformatics* **38**, 3549–56 (2022).

23. 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).

24. Chen, Z., C. K. Yeo, B. S. Lee and C. T. Lau. Autoencoder-based network anomaly detection. *Wireless Telecommunications Symposium (WTS),* Phoenix, AZ, USA, pp. 1-5 (2018). doi: 10.1109/WTS.2018.8363930.

25. 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).

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

27. Marini, S. et al. Optimizing viral genome subsampling by genetic diversity and temporal distribution (TARDiS) for phylogenetics. *Bioinformatics* **38**(3), 856–860 (2022).

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

29. 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).

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

31. Corizzo, R., M. Baron, N. Japkowicz. CPDGA: Change point driven growing auto-encoder for lifelong anomaly detection. *Knowledge-Based Systems* **247**,08756 (2022).

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

33. Ali, S. 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, 1533-1540 (2021).

34. Ren, H., Y. Li, T. Huang. Anomaly Detection Models for SARS-CoV-2 Surveillance Based on Genome k-mers. *Microorganisms* **11**, 2773 (2023).

35. Stange, M. et al. SARS-CoV-2 outbreak in a tri-national urban area is dominated by a B.1 lineage variant linked to a mass gathering event. *PLoS Pathog* **17**(3), e1009374 (2021).
36. Coppée, R. et al. Phylodynamics of SARS-CoV-2 in France, Europe, and the world in 2020. *eLife* **12**, e82538 (2023).
37. Naveca, F. G., et al. Amazonas, Brazil, was driven by the persistence of endemic lineages and P.1 emergence. *Nat Med* **27**, 1230–1238 (2021)
38. Tagliamonte, M. et al. Rapid emergence and spread of SARS-CoV-2 gamma (P.1) variant in Haiti. *Clin Infect Dis* **74**(11), 2057-2060 (2022).
39. Tegally, et al. Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature* 592, 438–443 (2021).
40. Pfizer and BioNTech Provide Update on Omicron Variant. https://investors.biontech.de/news-releases/news-release-details/pfizer-and-biontech-provide-update-booster-program-light-delta/ (Jul 8th, 2021).
41. 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).
42. Pedregosa, F., *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**(85), 2825–2830 (2011).

Figures

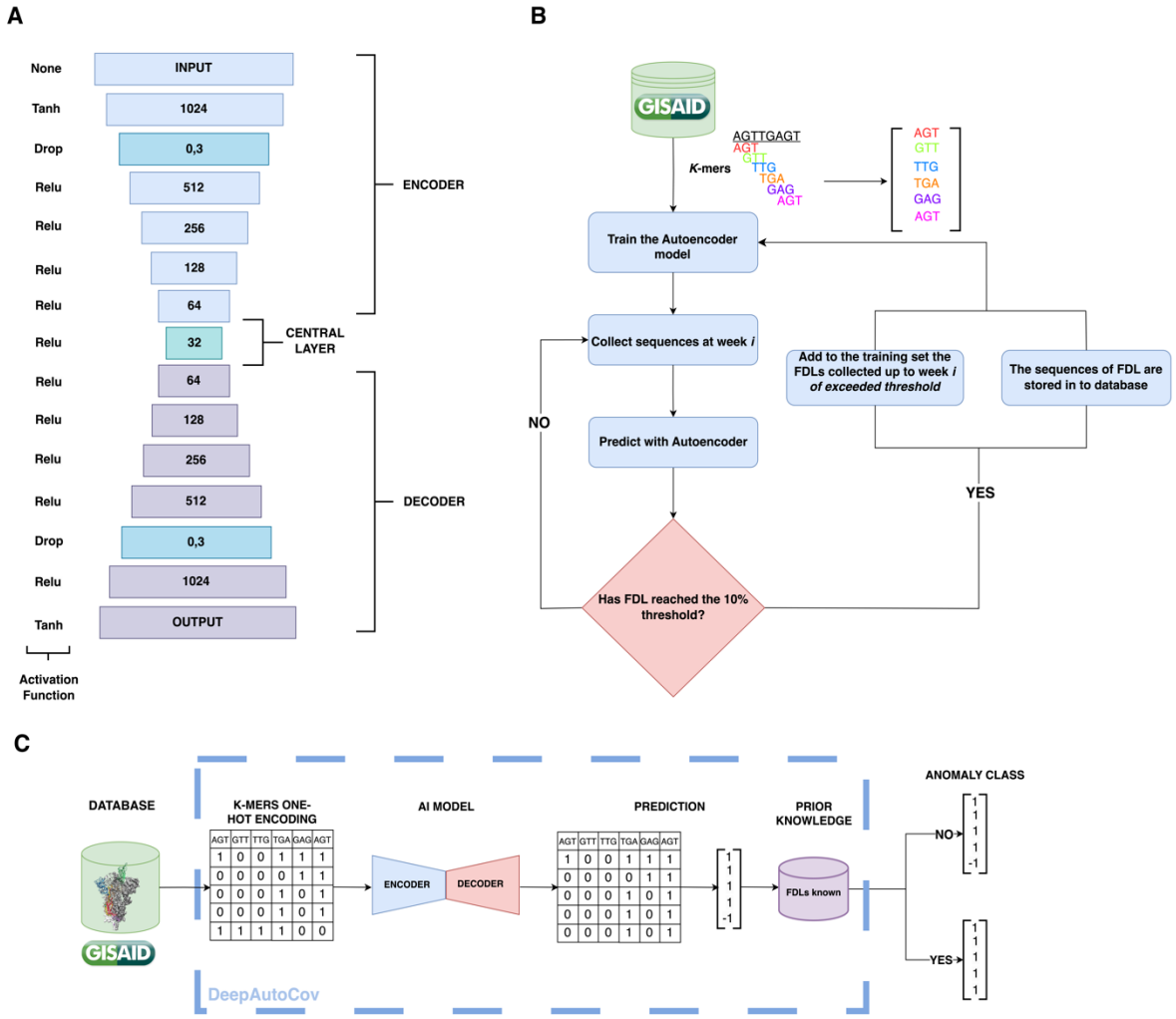


Figure 1. DeepAutoCoV architecture. **A)** Layered architecture with decreasing neuron counts in the encoding layers (from 1024 to 64) and increasing counts in the decoding layers (from 64 to 1024). A central layer with 32 neurons introduces Gaussian noise to enhance generalization. The decoding layers mirror the encoding structure. The output layer reconstructs the original input data. **B)** Schematic representation of our simulation. SARS-CoV-2 Spike protein sequences are gathered, converted into *k*-mers, and filtered. DeepAutoCov is trained on the data set. At each time point (week) newly uploaded sequences are evaluated as either anomalies (FDL) or not. If a FDL reaches the threshold, its sequences up to that time point are added to the training set and DeepAutoCov is retrained. **C)** Filtering scheme. Sequences that are flagged as FDL by DeepAutoCov are compared to sequences pertaining to known past FDLs that have already crossed the 10% threshold and discarded if they match.

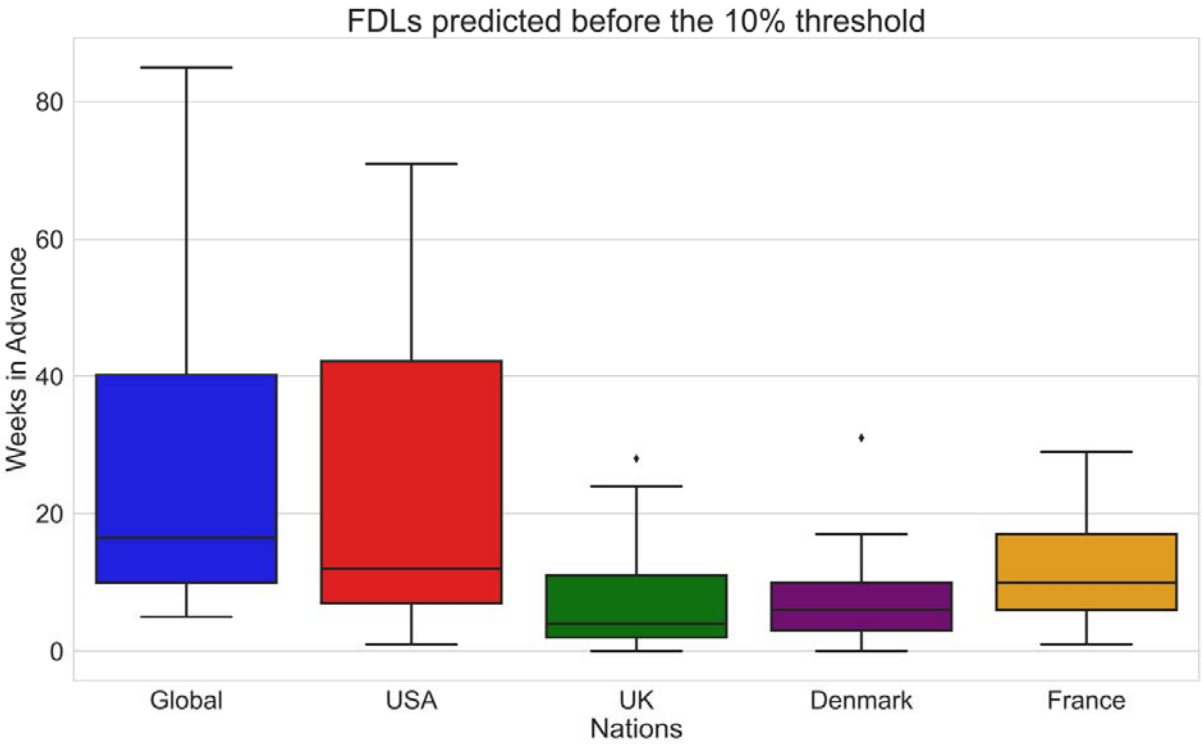
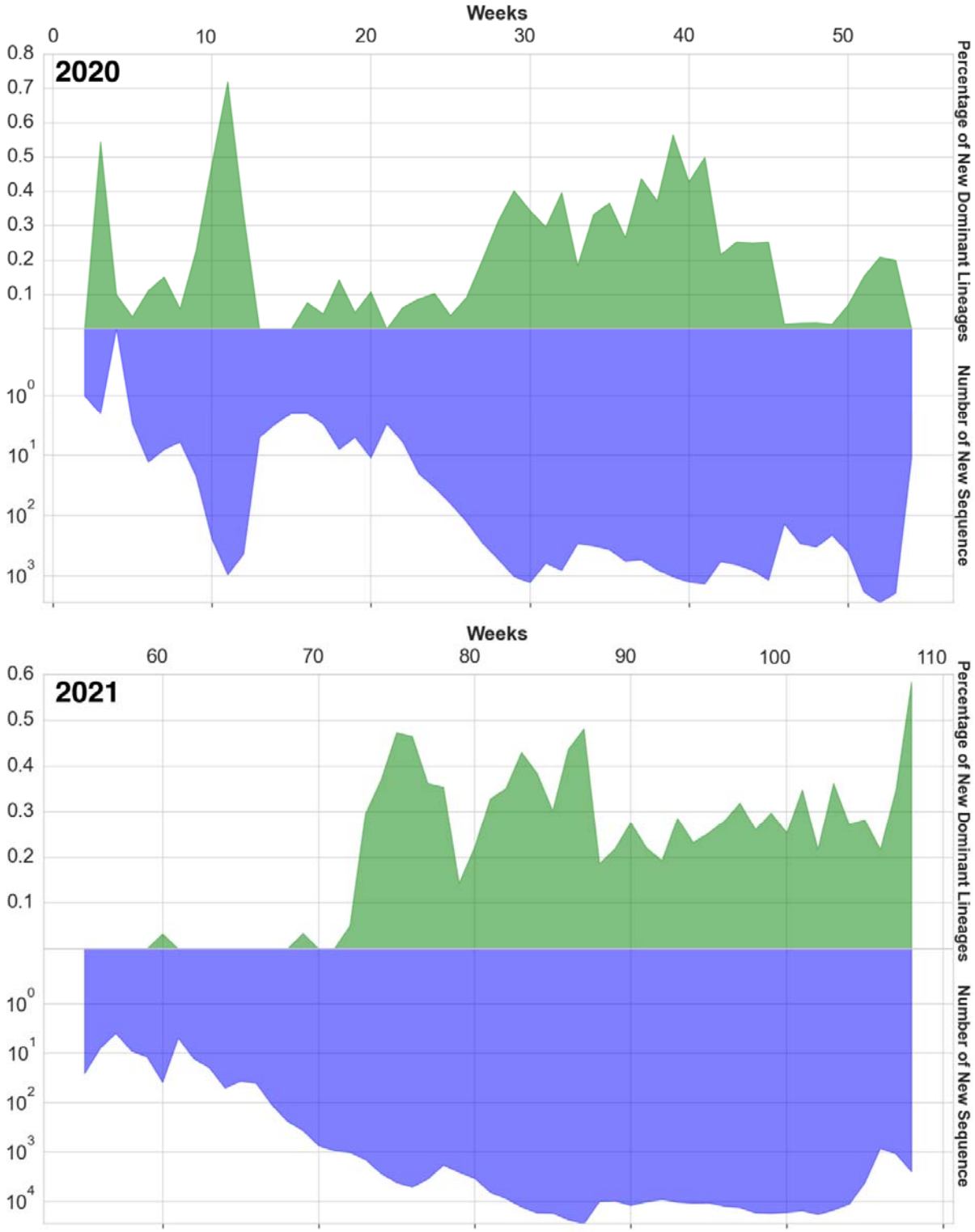


Figure 2: DeepAutoCoV performance in terms of leading time, i.e. weeks in advance detection before FDLs reach the 10% threshold.


Weeks
0
10
20
30
40
50
2020
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Percentage of New Dominant Lineages
10^0
10^1
10^2
10^3
Number of New Sequence
Weeks
60
70
80
90
100
110
2021
0.6
0.5
0.4
0.3
0.2
0.1
Percentage of New Dominant Lineages
10^0
10^1
10^2
10^3
10^4
Number of New Sequence

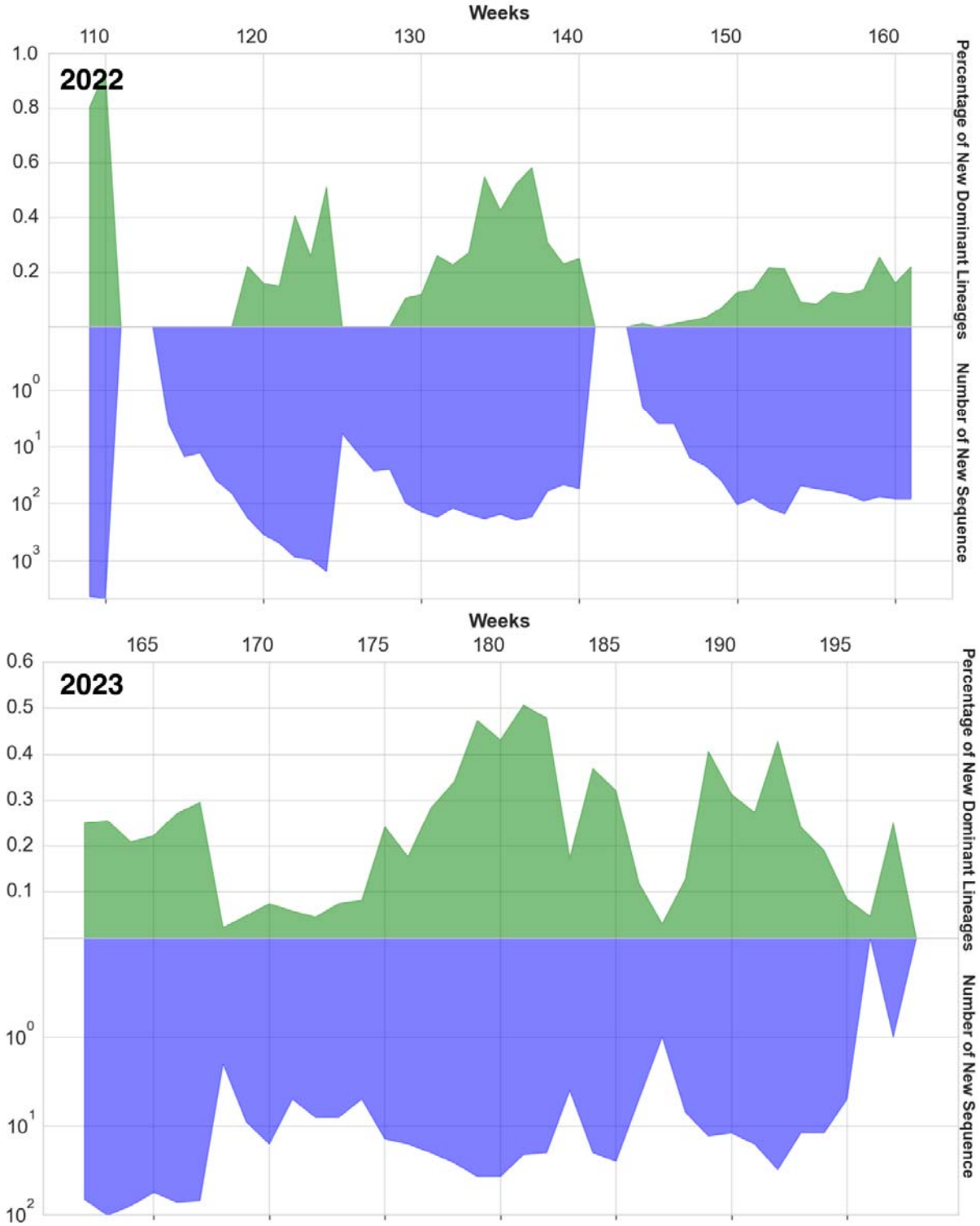


Figure 3: Median of correctly identified FDLs/week for the Global data set. X-axis: simulation weeks; y-axis top side: PPV; y-axis bottom side: number of FDL sequences in semi-logarithmic scale.

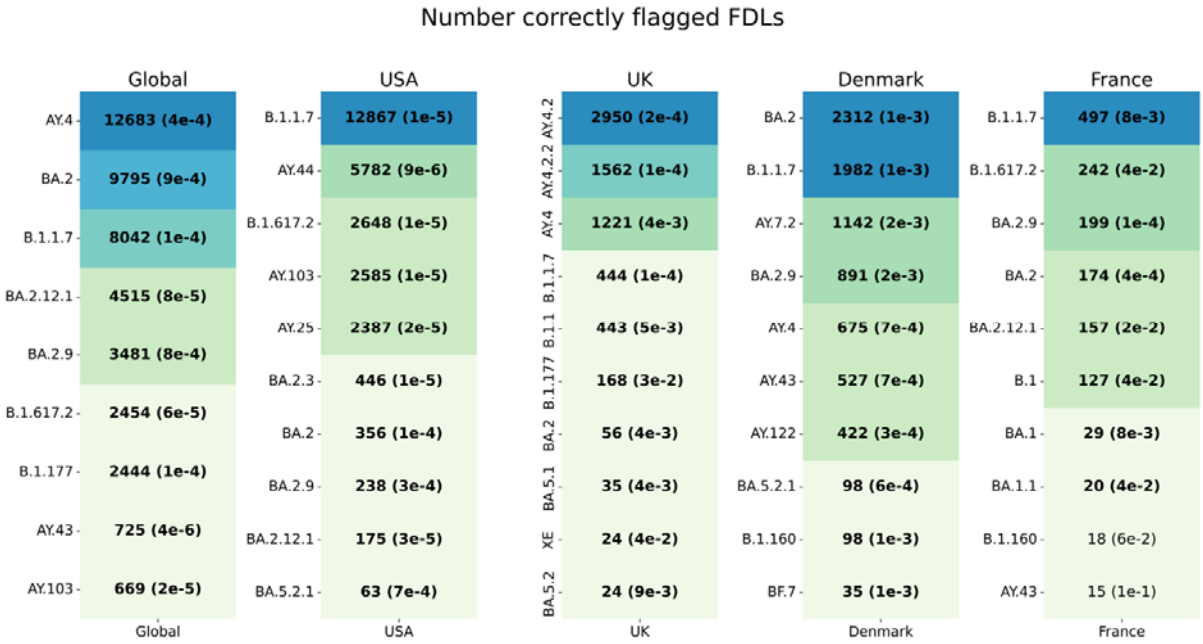


Figure 4: Top 10 FDLs correctly identified across different data sets. We report the total number of detected FDLs (p-value, Wilcoxon paired)

Table 1. DeepAutoCov results over all data sets.

Data set	Leading time (weeks)			FDL frequency at discovery[2]	PPV[3]	p-value[4]
	All FDLs	Best-case FDLs[1]	Worst-case FDLs[1]			
Global	16 (10,35)	73 (57, 78)	8 (7,8)	0.1% (0.02, 0.7)	0.28 (0.12, 0.37)	4e-24
USA	12 (7,43)	64 (52, 66)	3 (1,6)	0.2% (0.05, 3)	0.48 (0.27, 0.69)	7e-27
UK	4 (2,11)	23 (13, 24)	2 (2,2)	3% (0.1, 8)	0.30 (0.10, 0.38)	5e-13
Denmark	6 (4,11)	13 (12, 14)	3 (2, 4)	0.8% (0.2, 5)	0.38 (0.05, 0.60)	4e-08
France	10 (6,17)	28 (24, 29)	3 (2,4)	1.5% (0.4% 7%)	0.51 (0.20, 0.68)	7e-13

1. Best - and worst-case FDLs include the largest and smallest five FDLs in terms of leading time (weeks). 2. The frequency is FDL's proportion of total Spike proteins in the data set in the week the FDL is identified by DeepAutoCov. 3. Positive Predictive Value 4. Wilcoxon paired test (see Methods)