Virus Evolution, 2024, **10(1)**, veae086
DOI: https://doi.org/10.1093/ve/veae086
Advance access publication date: 14 November 2024
Research Article

CovTransformer: A transformer model for SARS-CoV-2 lineage frequency forecasting

Yinan Feng[1,2], Emma E. Goldberg[2], Michael Kupperman[2,3], Xitong Zhang[4], Youzuo Lin[1,5], Ruian Ke[2,*]

[1]Earth and Environmental Sciences Division, Los Alamos National Laboratory, Los Alamos, NM, United States
[2]Theoretical Biology and Biophysics, Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM, United States
[3]Department of Applied Mathematics, University of Washington, Seattle, WA, United States
[4]Department of Computational Mathematics, Science and Engineering, Michigan State University, East Lansing, MI, United States
[5]School of Data Science and Society, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

*Corresponding author. Theoretical Biology and Biophysics, Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM, United States. E-mail: rke@lanl.gov

Abstract

With hundreds of SARS-CoV-2 lineages circulating in the global population, there is an ongoing need for predicting and forecasting lineage frequencies and thus identifying rapidly expanding lineages. Accurate prediction would allow for more focused experimental efforts to understand pathogenicity of future dominating lineages and characterize the extent of their immune escape. Here, we first show that the inherent noise and biases in lineage frequency data make a commonly-used regression-based approach unreliable. To address this weakness, we constructed a machine learning model for SARS-CoV-2 lineage frequency forecasting, called CovTransformer, based on the transformer architecture. We designed our model to navigate challenges such as a limited amount of data with high levels of noise and bias. We first trained and tested the model using data from the UK and the USA, and then tested the generalization ability of the model to many other countries and US states. Remarkably, the trained model makes accurate predictions two months into the future with high levels of accuracy both globally (in 31 countries with high levels of sequencing effort) and at the US-state level. Our model performed substantially better than a widely used forecasting tool, the multinomial regression model implemented in Nextstrain, demonstrating its utility in SARS-CoV-2 monitoring. Assuming a newly emerged lineage is identified and assigned, our test using retrospective data shows that our model is able to identify the dominating lineages 7 weeks in advance on average before they became dominant. Overall, our work demonstrates that transformer models represent a promising approach for SARS-CoV-2 forecasting and pandemic monitoring.

Keywords: SARS-CoV-2, machine learning, viral lineage frequency forecasting, time series

Introduction

SARS-CoV-2 has been continuously evolving new variants that increase transmission fitness and/or evade population immunity (Carabelli et al., 2023; Markov et al., 2023; Meijers et al., 2023; Volz, 2023). As a result, many waves of infection around the world were caused by variants of concern (such as subvariants of the Delta and the Omicron lineages), leading to large numbers of infections and a high death toll (Dong et al., 2020; Ritchie et al., 2020). Vaccination is an important and effective tool to reduce the level of transmission along with morbidity and mortality (Polack et al., 2020; Baden et al., 2021). However, with the frequent origination of immune escape variants (Harvey et al., 2021; Rössler et al., 2023; Wilks et al., 2023), there are repeated waves of infection and hence an ongoing need for predicting the rate and magnitude of spread for new variants that emerge in a population. This has become increasingly important especially with dozens or hundreds of minor variants circulating and competing in the global population (Beesley et al., 2023). Accurate prediction would allow for more focused experimental efforts to understand the pathogenicity and molecular characterization of the extent of immune escape of variants that have high potential to spread (Stockdale et al., 2022).

Existing forecasting tools have focused on predicting future trajectories of the numbers of COVID-19 cases and deaths (Du et al., 2023; Grubaugh et al., 2019; ForecastHub, 2023). However, a predictive tool focusing on forecasting dynamics of variant frequencies is still lacking. The potential for spread of a variant when it first emerges has been predicted based on the extent of structural change on the spike protein leading to immune escape (Harvey et al., 2021; Chen et al., 2022), but the potential for an immune escape mutant to efficiently transmit in the population is unknown. Regression or mechanistic models predict the fitness of a lineage, and thus the potential for spread, from time series data of variant frequencies (Volz, 2023; Beesley et al., 2023; Nextstrain, 2023; Abousamra et al., 2023; Figgins & Bedford, 2022). However, due to the noise arising from data collection and deposition (as we show below), this general approach suffers from large uncertainty in prediction. We previously found that robust and accurate predictions require several weeks of data from multiple countries (van Dorp et al., 2021).

In this work, we constructed a transformer model (Vaswani et al., 2017; Dosovitskiy et al., 2021), called CovTransformer, to forecast the future frequency of existing SARS-CoV-2 lineages from noisy lineage frequency time series data. Previously, machine learning approaches have been successfully applied to time series analysis (Che et al., 2018; Song et al., 2018) and more recently, they were applied to various problems in COVID-19 pandemic response (Syrowatka et al., 2021). In particular, state-of-the-art transformers (Vaswani et al., 2017), distinct for their self-attention mechanism, are renowned for their ability to capture complex sequential patterns and long-term relationships. This enables the nuanced detection of intricate patterns and long-term dependencies within data sequences, achieving success in many problems such as natural language processing (Vaswani et al., 2017; Brown et al., 2020; Devlin et al., 2018), computer vision (Dosovitskiy et al., 2021; Khan et al., 2022; Kirillov et al., 2023; Girdhar et al., 2019), multimodal learning (Xu et al., 2023; Radford et al., 2021), and time series analysis (Wen et al., 2023; Zhang & Yan, 2022; Gao et al., 2022; Song et al., 2018). Our approach integrates transformers within a broader strategy to address the complex task of forecasting pandemic lineage frequencies. This methodology systematically tackles inherent challenges such as large noise, reporting delays, input biases, variable label quality, and the scarcity of comprehensive data across a wide array of lineages. It sets a precedent for utilizing machine learning tools in lineage-level frequency forecasting. Here, we first show that simple forecasting of emerging variant frequencies based on regression leads to erroneous predictions due to the inherently noisy nature of the data. Then we demonstrate that our transformer model accurately forecasts lineage frequencies and, importantly, identifies lineages that eventually reach high frequencies from a collection of newly emerged lineages.

Materials and methods

SARS-CoV-2 data

All available SARS-CoV-2 sequence metadata was downloaded from GISAID (Elbe & Buckland-Merrett, 2017) on February. 26, 2024. We used the Pango lineage designations provided by GISAID in this metadata, discarding sequences that lacked Pango or country information. We then summarized the data as the number of records for each combination of collection date, submission date, and country. These counts were used to compute the variant frequencies used in our analysis. Our model was trained and evaluated using datasets from the UK and the USA, regions with the highest collection of SARS-CoV-2 genomic sequences. We divided the dataset of all lineages into training and validation sets comprising data up to Dec. 31, 2022, and a testing set with data following this date. This partition resulted in 107,712 entries for training and validation, and 37,707 entries for testing.

Regression model for lineage frequency forecasting for individual lineages

The frequency of a variant typically follows a linear increase in the logit transformed space (van Dorp et al., 2021; Bedford, 2023). Therefore we first transformed the frequency of each lineage with a logit transformation, $p = \log[p/(1-p)]$. We then performed a linear regression on the time series of the transformed lineage frequencies for the past 42 days. To forecast the lineage frequency into the future, we extrapolated lineage frequency based on the parameters from the linear regression.

Data processing for the transformer models

To improve the data quality, we first removed isolated data points during pre-processing. Specifically, a data point is identified as potentially anomalous and removed if there are five or more days without any records within a seven-day period centered on that point. The input of the machine learning model is constructed as a $5n$ dimensional vector, where n is the number of input days. For each day, we incorporate five features: the frequency of the specific lineage, the number of sequences of the lineage, the total number of sequences on that day, the number of days elapsed since January 1, 2020, and the elapsed days since the lineage's first recorded collection day. For the missing records, we use a fixed negative token as a placeholder. By using special tokens, transformers can effectively manage missing or masked data, making them robust tools for a wide range of sequence-related tasks. In particular, we treat each one-day feature as a discrete element (referred to as a token or patch) for input processing. Thus, the patch size is 1×5. The model's output is the predicted frequency of the lineage for a future time point, specified as T days ahead (Fig. 1a). Given the presence of noise and missing data in our dataset, we employ a 1-D smoothing spline following interpolation to smooth the ground truth, which we then use as our labels.

Machine learning framework

We employed a single-layer transformer with a linear input layer preceding it and a linear output layer following it. The transformer contains a self-attention layer, followed by a multi-layer perceptron (MLP). The overall model architecture is illustrated in Fig. 1b. In particular, the model uses embedding dimensions of 8 and 2 attention heads without dropouts. There is a layer normalization between the transformer and the input/output linear layer, separately. For the position embedding, we employed the fixed sin-cos embedding described in the original paper (Vaswani et al., 2017). Here, the primary challenge in our problem lies in the presence of significant data noise. Our primary objective is not to extract deeply hierarchical or intricate semantic information, but rather to uncover valuable signals within this noisy data. Given the relatively low amount of data for model training, our foremost considerations are robustness, simplicity, and resistance to overfitting. Consequently, we have deliberately opted for a shallower network architecture.

Because it is easier for the model to learn the short-term prediction, we used an incremental method to enhance long-term predictions by including short-term predictions (Fig. 1b). We first trained a model that can predict 14 days into the future. Then, the 14-day predictions (ensemble of five models' results from five-fold cross-validation) are concatenated with the input feature. Additionally, we introduce a shortcut connection that adds the short-term prediction to the model's output to calculate the final long-term predictions. This approach combines short-term insights with long-term forecasting, resulting in improved accuracy.

For the purpose of training, validating, and testing our model, we partition the data as of January 1, 2023. Data recorded prior to this date are allocated for model training and validation with five-fold cross-validation. Then, the best-performing models are

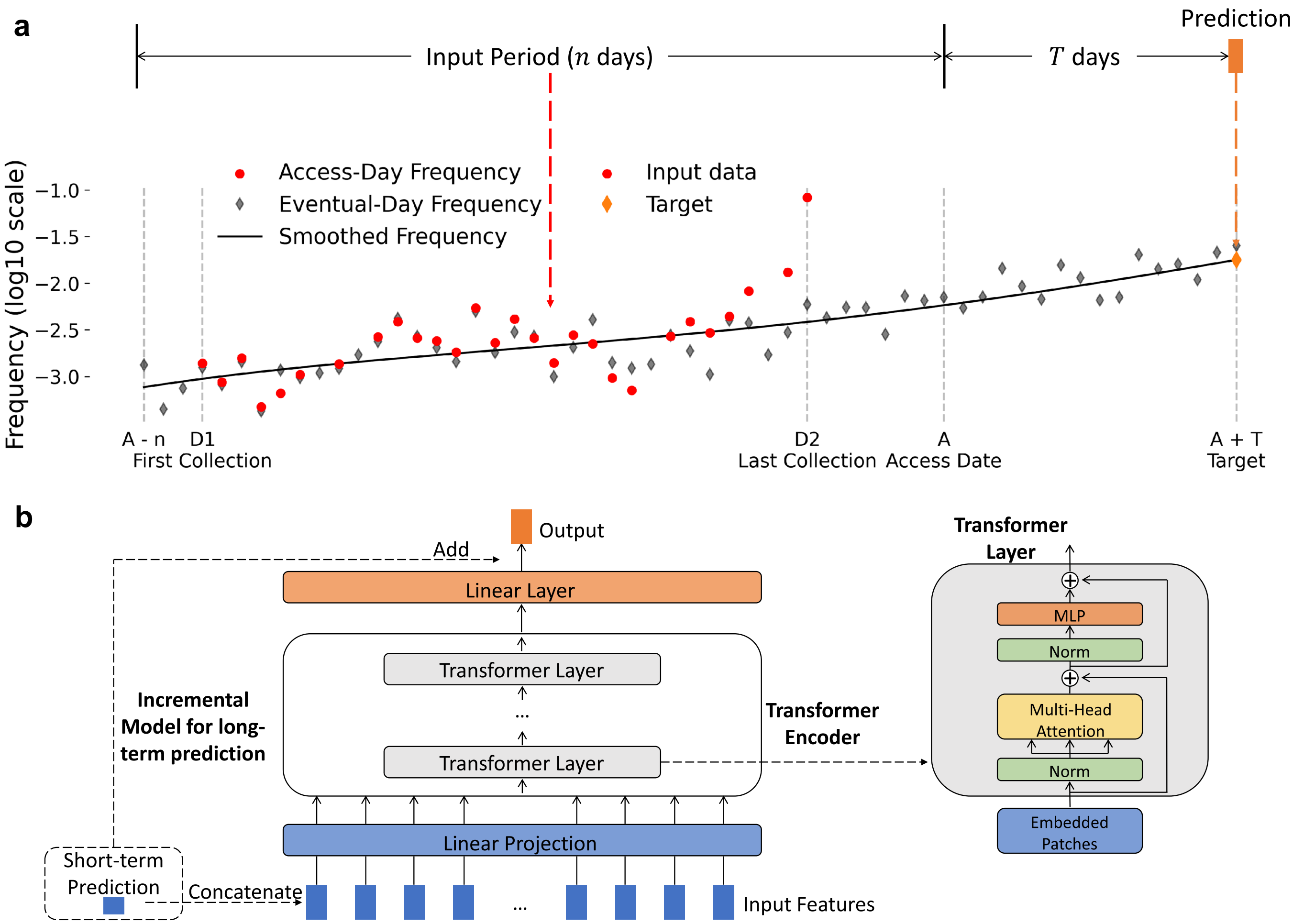

Figure 1. Schematics of model input, output, and architecture. (a) An example of input and target data for the transformer model, using lineage XBB.1.9.1 in the USA. The access date A is when the data were assumed to be accessed in this example (February 23, 2023). Red dots show the access-day frequency time series available on that day. The data input to our model is these access-day frequencies, for the input period of $n = 42$ days before A. In this example, note that the first three days of the input period do not contain data on the access day, so the day of first collection, $D1$, is later than the start of the input period, $A - n$. Similarly, the last seven days of the input period do not contain data on the access day, so the day of last collection, $D2$, is earlier than A. Frequencies on these days only become available after more data are collected later; gray dots show the frequency time series that is eventually measured, using all data collected by our single data download on February 26, 2024. The model makes predictions of the lineage frequency T days after the access date. To compute this target frequency for training, we fit a smoothing spline (black line) to the eventual frequency time series. The error of the model is calculated as the difference (in log10) between the model-predicted frequency and the target frequency. (b) Model architecture. We employed a transformer with a linear input layer preceding it and a linear output layer following it. The transformer layer contains a multi-head self-attention layer, followed by a multi-layer perceptron (MLP). We used an incremental method to enhance long-term predictions by including short-term (i.e. 14-day) predictions.

ensembled and tested on data subsequent to this date as the testing data set. Our model is trained on combined datasets from the USA and UK. Moreover, we utilize a noise injection technique (Orvieto et al., 2022) to improve the model generalization ability, which perturbs the model parameters by random Gaussian noise (zero means and $1e-4$ std) in each training iteration. This prevents the model from converging to sharp local minima.

To address the challenge of data imbalance inherent in our dataset, we have employed an approach wherein we apply an exponential function to weight the loss function during model training. This weighting strategy takes into account the varying frequencies of the target day, assigning higher weights to datapoints corresponding to high frequency labels. By doing so, the model places greater emphasis on accurately predicting the high frequency variants, which are often of paramount importance in epidemiological and public health contexts. This adaptive weighting mechanism ensures that the model focuses on effectively capturing the dynamics of the most prevalent variants, ultimately leading to improved overall forecasting performance and a more balanced predictive outcome. The overall loss function combines l_1 and l_2 loss, with an exponential weighting scheme. This loss function is designed to give more emphasis to lineages with higher frequency. The loss function can be mathematically represented as:

$$\mathcal{L}(label, pred) = \frac{1}{N}\sum_{i=1}^{N}\left(w_i \cdot |label_i - pred_i| + w_i \cdot (label_i - pred_i)^2\right),$$

where N represents the number of samples, $label_i$ is the true label of the i-th sample, $pred_i$ is the model's prediction for the ith sample, and w_i is the weight associated with the ith sample, which is computed as $w_i = 2e^{label_i}$.

To train the model, we employed the AdamW (Loshchilov & Hutter, 2018) optimizer with momentum parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a weight decay of 0.05. The initial learning rate was set to be 1×10^{-3}, and we modified the learning rate with a cosine annealing (Loshchilov & Hutter, 2016). We set the batch size to 256 and trained the model for 1000 epochs. We implemented our models in Pytorch and trained them on 1 NVIDIA Tesla V100 GPU.

Results

Data collection

To predict variant frequency trajectories, we used the metadata for SARS-CoV-2 sequences from GISAID [(Elbe & Buckland-Merrett, 2017) downloaded on February 26, 2024, see Methods section for details]. For each viral sequence, we noted both the collection date when a viral sample was collected from an infected individual, and the submission date when the sequence was submitted to the database. In general, there is a wide distribution of the delay in reporting, defined as the difference in days between the collection date and the submission date (Fig. S1). The mean and median delays in reporting are 30.1 and 17 days, respectively, whereas the standard deviation is 51.6 days. As we show below, the long reporting delays leads to large uncertainties and biases in lineage frequency time series especially for days immediately before the date of access to the database (see Fig. S2 for example).

To form a dataset for training our models below, we first constructed an input dataset retrospectively where we assumed the database was accessed on each day between January 1, 2020 and February 22, 2024. For each day of access, we calculated the frequency time series of each Pango lineage (Rambaut et al., 2020) for the past two months by only considering the viral genetic sequences submitted to the database before or on the access date. Overall, there are a total of 145,419 time series for two countries' data (USA and UK), which serve as an input dataset for our models. Note that we used the lineage assignment for each sequence at the time of our one metadata download (February 26, 2024). It is possible that some lineages were not assigned yet at the assumed day of access, for example, if the size of a to-be new lineage was not large enough to be assigned as a new lineage. This means that the training of our model benefits from the knowledge of the contemporary lineage assignment, which improves the training datasets especially at low lineage frequencies, and thus enhances model performance especially for newly emerged lineages once they are assigned. It also means however, when we deploy this method, its prediction of newly emerged lineages relies on Pango lineage designation and assignment (O'Toole et al., 2022).

To construct a target dataset for model evaluation and testing, we calculated the frequency of each Pango lineage on each day, using all the sequences available when we downloaded the GISAID metadata. We termed these the 'final' frequencies. For the ground truth frequencies used as targets to train and test the model, we fit a smoothing spline to the final frequencies to remove the inherent noise from the data due to day-to-day variations in sampling. Our strategy for constructing time series datasets is shown in Fig. 1a.

Limitation of regression-based estimations due to noise and biases in data

One common forecasting approach is to estimate the growth rate of a lineage by applying regression-based methods on lineage frequency time series, and then project future lineage frequency based on the estimated growth rate. We found that because of the reporting delay, the Pango frequencies within two weeks' time of the access date were often substantially different than the final frequencies, and typically there was no viral genome reported within 1–3 days prior to the access date (Fig. S2). This high level of uncertainty in the lineage frequency time series due to reporting delay makes the regression-based approach highly unreliable (Fig. S2) especially when a lineage was at low frequencies. For example, we used a linear regression on the logit transform of the lineage frequency data (since it has been shown that the SARS-CoV-2 lineage frequency follows a linear growth or decline on the logit transform; van Dorp et al. 2021; Bedford 2023). We found that for many lineages, the projection suffers from either large noise or from systematic biases in the lineage frequency data (Fig. S2). We used this regression-based prediction as a baseline model below for our machine learning model to compete against.

Development of a transformer model

We developed a transformer model to forecast future frequencies of each Pango lineage (see Methods section for details). The transformer is characterized by its distinctive attention mechanism, known as multi-head self-attention. This mechanism enables the model to weigh the significance of each element in a time series against all others, enabling it to capture intricate relationships and long-term dependencies within data sequences, whether in text, time series, or other temporal data (Vaswani et al., 2017; Xu et al., 2023; Wen et al., 2023).

In brief, our model employs a shallow transformer network with a linear input layer preceding it for patch embedding and a linear output layer. The model architecture is illustrated in Fig. 1b. It uses the frequency time series of a Pango lineage in a country as input to predict the frequency of the lineage at a future date (14, 21, 28, 35, 42 and 60 days into the future). We employed an incremental approach to enhance long-term model prediction, developing six model variants to make predictions on the six future dates. The first model variant is trained to forecast on day 14, and then the 14-day predictions are concatenated with the input feature as inputs for the other five model variants. This approach combines short-term insights with long-term forecasting, resulting in improved accuracy. Note that the model at this stage concerns predictions for a single lineage in one country only. It does not consider predictions from other extant lineages at the same time period, although we address this consideration further below.

The error of the model is calculated as the mean absolute error (MAE) between the log10 of the model predicted frequency and the log10 of the target frequency. We calculate the error on the log10 transform of the frequencies because in general, the size of each of the lineages increases or decreases exponentially. One characteristic of the dataset is that there are only a few lineages that rose to a high frequency to become the dominant lineage in the population, while most lineages stayed at relatively low frequencies (Fig. 2b). To address this issue of data imbalance, we applied an exponential function to weight the loss function during model training, such that the errors between target and model prediction from lineages that rose to a high frequency were weighted heavily in the training. In this way, our model is trained to better identify lineages that will become dominant when their frequencies are still low (Fig. 2b).

We also developed and tested a two-layer long short-term memory (LSTM) model (Sak et al., 2014) to check how this earlier and simpler model architecture performs. This model is designed to make 14-day predictions. It has approximately 1.4k parameters (compared to 1.6k parameters in the transformer model), and it is trained and tested using the same dataset as we used above.

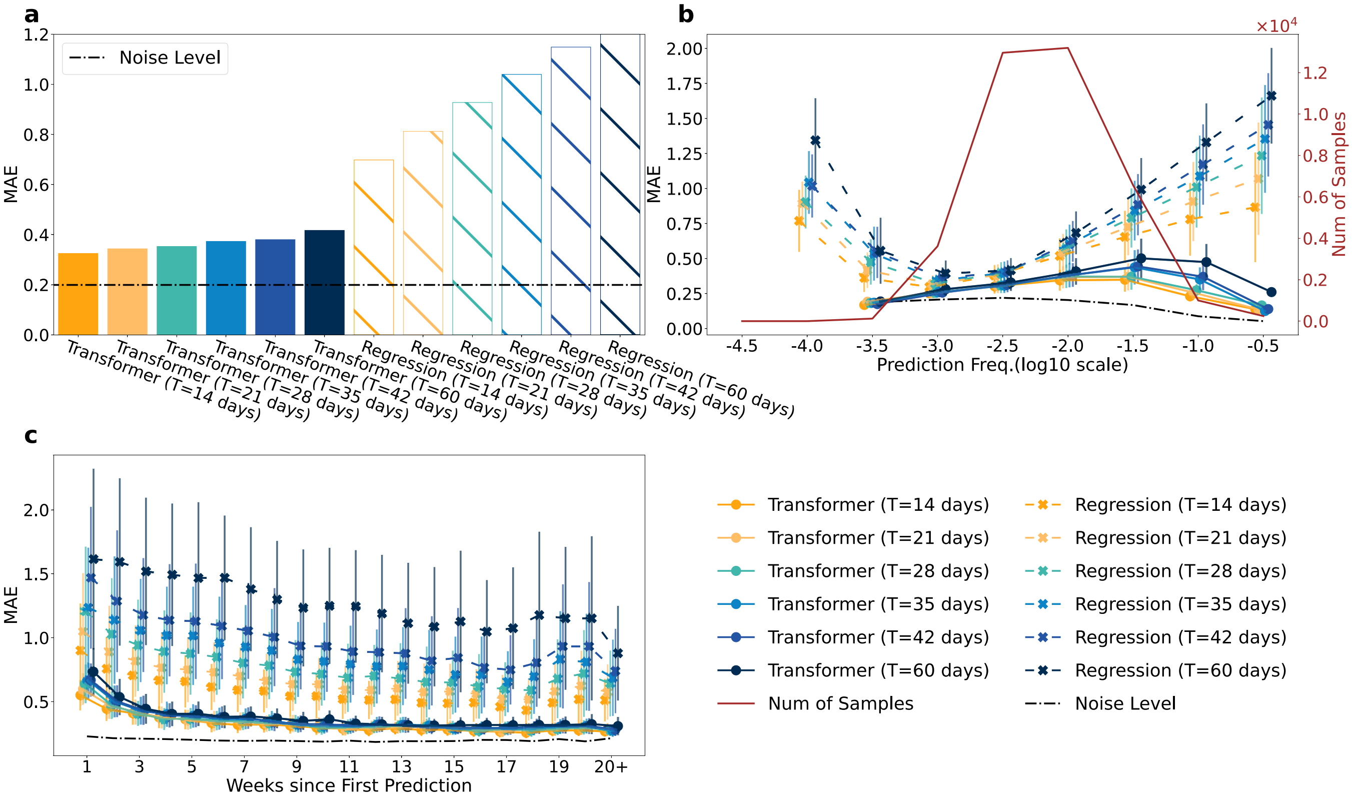

Figure 2. The transformer model outperforms linear regression models on data from UK and USA across evaluation categories. In all panels, the noise level (dashed line) is calculated as the MAE between the smoothed final frequency and the final frequency (on a log10 scale). (a) Overall model performance of the transformer model versus the regression model for six different future days. (b) MAE categorized by the model-predicted frequencies. Results were binned by each 0.5 interval on the log10 of the frequencies. The transformer model performed substantially better than the regression model especially for high frequency predictions: it identifies lineages that would rise to a high frequency much better than the regression model. The number of samples (red line) indicates the number of total valid input time series in each bin. (c) MAE categorized by the week since the first prediction day for each lineage (the first day when there are 14 non-zero frequencies in the model input). The transformer has superior performance compared to the regression model, especially in the initial weeks following the emergence of a new variant, when data is scarce.

As expected, our transformer model makes more accurate predictions than the LSTM model. The MAE of the Transformer was 0.4239, compared to 0.6098 for the LSTM model. This confirms that transformer architecture is suited to handle this type of datasets and make accurate forecasts.

The transformer-based model performs well

We first trained and tested our model using data derived from the UK and the USA, where most SARS-CoV-2 genomic sequences have been collected. We split the dataset containing all lineages such that the data before January 1, 2023 were used as the training and validation dataset and the data between January 1, 2023 and January 1, 2024 were used as the testing dataset. This leads to 107,712 training data and 37,707 testing data.

We tested all the model variants and compared their performance with model predictions from linear regressions on the logit transform of lineage frequencies. Our model substantially outperforms the regression model across all prediction time points Fig. 2. This performance advantage is especially pronounced in long-term forecasts (e.g. 60 days into the future) where the regression models exhibit escalating error rates. In contrast, the transformer models maintain a lower and more stable loss growth, indicating robustness in handling extended prediction horizons. Remarkably, the MAE for the transformer model predictions ranges between 0.32 and 0.42 for the six future dates. These MAEs are only slightly above the mean noise level of the data, 0.2 calculated as the mean difference between the actual final frequencies derived from the sequence data and the smoothed target frequencies. This means we expect the average difference between our predictions and the ground truth frequencies to be approximately 0.42 on a log10 scale (i.e. 2.6-fold difference on a linear scale) even for predictions 2 months into the future.

We next examined the MAEs using two different categorizations. First, we derived statistics of MAEs according to the predicted frequency of the lineage (Fig. 2b). Our models show remarkable consistency across various prediction frequencies, even though the number of training datasets is very low when the predicted frequency is high. This is critical for lineage surveillance because of the importance of accurate and early identification of lineages that eventually rise to a high frequency in the population. The regression models, in contrast, exhibit pronounced inaccuracies and false positive rates, particularly in these categories. Second, we derived statistics of MAEs according to the time since our model was able to make the first prediction, defined as the day when a lineage has at least 14 days of non-zero data points in the past 42 days. Figure 2c shows that even when the number of input data is low, such as two weeks' data points, our model makes accurate predictions, whereas the regression model performed poorly.

We further evaluated how early our model is able to identify future-dominating lineages. For each variant that became the dominant variant at some time (the lineage with the highest frequency in the population) in our testing dataset, we calculated the difference between the date when it actually became dominant and the date when our 60-day prediction model first predicted it to be the dominant variant. The mean and median of the difference in days are 48.7 and 54 days, respectively (Fig. S3). This suggests our model is able to identify the future dominating lineages 7 weeks in advance on average. Note that here we assumed that at the time of forecasting, the newly emerged lineages were already identified and assigned a unique name, although in reality, this may not be the case. Therefore, the ability of our model to make early forecasts is potentially constrained by the time of Pango lineage assignment.

Transformer models accurately forecast lineage trends in the USA and the UK, and more generally

The results above showed that the transformer models perform well in making predictions for individual lineages. However, in practice, dozens of lineages may coexist in a population at the same time. Therefore, when implementing our models to deal with real-time data, we normalized the predictions of individual lineages such that the sum of all extant lineage frequencies is 1. Figure 3 shows comparisons between the raw data and our model predictions for the USA and the UK during the period between November 30, 2022 and February 18, 2024. Model predictions shown on each day were made using lineage frequency time series collected for an access date 28 or 60 days prior to the day of prediction. The model predictions agree well with the raw data. In particular, our models correctly predicted the rise of XBB lineages in the first half of 2023 and the JN lineages in late 2023, emphasizing their utility in making accurate real-time forecasts and identifying highly transmissible SARS-CoV-2 lineages. Again, these forecasts rely upon the identification and assignment of new lineages from their parent lineages.

Similarly, we normalized the predictions from the logit regression model, and Fig. S4 shows comparisons between the raw data and the projected frequencies from the model. As expected, the predicted frequencies aligned poorly with raw data.

Given the highly accurate performance of our models on the data sets from the USA and the UK, we further tested our models using data collected from over 100 other countries as well as the USA state level data, to assess their generalization ability (4). Our model performance is surprisingly robust, at least for countries with relatively high numbers of available genomic sequences (Fig. 4a). For example, for 31 countries (each with >20,000 sequences reported), the MAEs of our model predictions are below 0.5. Moreover, 24 of 27 countries with >70,000 sequences reported have MAEs below 0.5, and the MAEs of the other three countries are close to 0.5. There is a clear correlation between the intensity of a country's sequencing efforts and the performance of our model, with a Pearson correlation coefficient of −0.82 (Fig. 4a). We found similar patterns with the USA state-level data. Our model performs very well for states with >20,000 sequences reported (Fig. 4b). The performance drops when fewer genomic sequences are available. Overall, the robust performance of our model across data from other countries and from the USA states suggests the wide applicability of our model. The clear correlation between model performance and the number of genomic sequences underscores the importance of comprehensive sequencing in enhancing the precision of predictive models in viral genomics.

Model design ablation test

During model development, we tailored our model specifically to characteristics of the input data and the need to identify lineages that may rise to a high frequency. This includes estimating the target frequency by smoothing the time series to remove noise, using a single-layer transformer to reduce model complexity, and using an exponential weighting function to focus on learning high-frequency lineages (see Methods section). To test the effectiveness of these model settings, we performed ablation tests using alternative models where either raw frequencies were used as target frequencies, a two-layer transformer was implemented, or no weighting function was applied. We found that none of the alternative models performed better than our original model (Fig. S5), emphasizing that our model design choices indeed improve overall performance.

Transformer models outperform Nextstrain predictions

Currently, a widely accepted tool for lineage frequency forecasting is Nextstrain's multinomial logistic regression (MLR) model (Hadfield et al., 2018; Abousamra et al., 2023; Nextstrain, 2023). The Nextstrain MLR model has stringent data criteria, requiring for example that a lineage have at least 5000 sequences in the past 150 days for a given day of access. In contrast, our model makes predictions on lineages that have at least 14 days of records in the past, and has no formal requirement for the minimum number of sequences. Therefore, our model is able to make predictions for many more lineages than the Nextstrain model. Furthermore, for those lineages that were included in Nextstrain model prediction, our model can make predictions much earlier than Nextstrain. The caveat for early predictions remains that both of these models rely on newly emerging lineages being assigned a Pango designation at the time of forecasting.

To test relative prediction accuracy, we compared our model predictions on the lineages and days where predictions from the Nextstrain MLR model were available. In general, our model had much better performance (Fig. 5). For example, after we normalized the frequencies of all existing lineages predicted from our models, the MAEs of our model are between 0.28 and 0.36 for 14, 21, and 28-day predictions, only slightly above the noise level. In contrast, the Nextstrain MLR model had MAE >0.5 across all predictions (Fig. 5a). Our transformer model performed particularly well when the predicted frequency was high. That is, it can identify lineages that reached high frequencies, whereas the Nextstrain MLR model performed poorly (Fig. 5b). We also provide visualization examples of 14-day predictions, including 12 lineages from four countries (GBR, USA, CHN, and AUS), in Fig. S6. Overall, these results highlight the strength of our approach in accuracy and stability.

Discussion

Here, we developed a transformer-based model to make accurate forecasts on SARS-CoV-2 lineage frequencies up to 60 days into the future. The model excels in robustness, accuracy, and generalization ability across datasets. Its ability to outperform existing models, including the widely-used Nextstrain multinomial regression model, and adapt to diverse datasets makes it a powerful tool for SAR-CoV-2 lineage frequency forecasting and pandemic monitoring. Assuming the newly emerged lineage is identified and named, our test using retrospective data shows that our model is able to identify the future dominating lineage 7 weeks in advance on average.

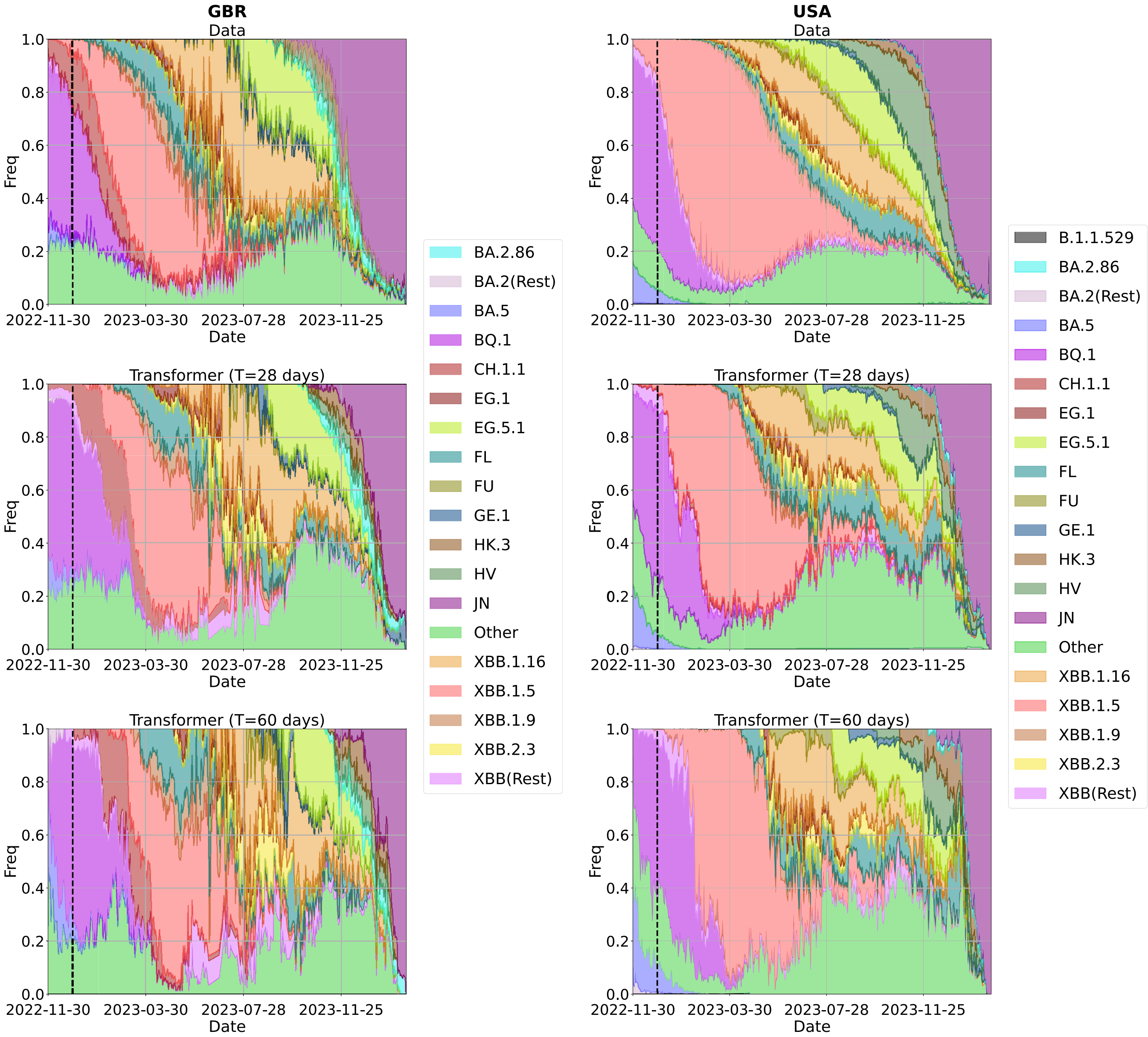

Figure 3. Retrospective predictions from the transformer model agree well with raw data from the UK and the USA. Upper panels show the lineage frequencies over time derived from the raw data. Middle and lower panels show the frequencies from the 28 and 60 day predictions from the transformer model. The model-predicted frequencies on a day were calculated by first applying the transformer model to time series of each extant lineage assuming an access date of 28 or 60 days prior to the prediction date, and then normalizing the predictions for all lineages such that the sum of frequencies is 1 on each day. The dashed line denote the date (December 31, 2022) after which the raw data was not used for model training.

Although trained only on partial data from the USA and the UK, our model demonstrated remarkable generalization ability across datasets from 31 countries in distinct geographic regions across the globe and datasets from the USA states. This is especially important for pandemic monitoring purposes, because it enables early identification of a highly transmissible variant that can appear in any geographic region. Thus, we expect our model could be implemented and used to make continuous forecasts using data collected globally and at the state or provincial level (for regions where sufficient genomic data is available to make reliable predictions), such that highly transmissible variants can be identified as early as possible. This identification would provide researchers and public health officials with a list of candidate highly transmissible variants to be further monitored and investigated, for example, through experimental evaluation of its pathogenic and their potential of immune escape. In addition, when our model forecasts that a lineage may become dominant in multiple countries or multiple states/provinces, it serves as a strong indication that the lineage may become a global variant of concern. In actual implementation for monitoring purposes, we could define a risk score for each variant calculated from, for example, the weighted sum of predicted future frequencies of the variant across regions/countries considered in the model. This would provide public health officials a direct measure of risk of spread for each circulating minor variant.

Existing tools for lineage frequency forecasting, such as the multinomial regression approach implemented in the Nextstrain model (Nextstrain, 2023; Abousamra et al., 2023), mostly adopt

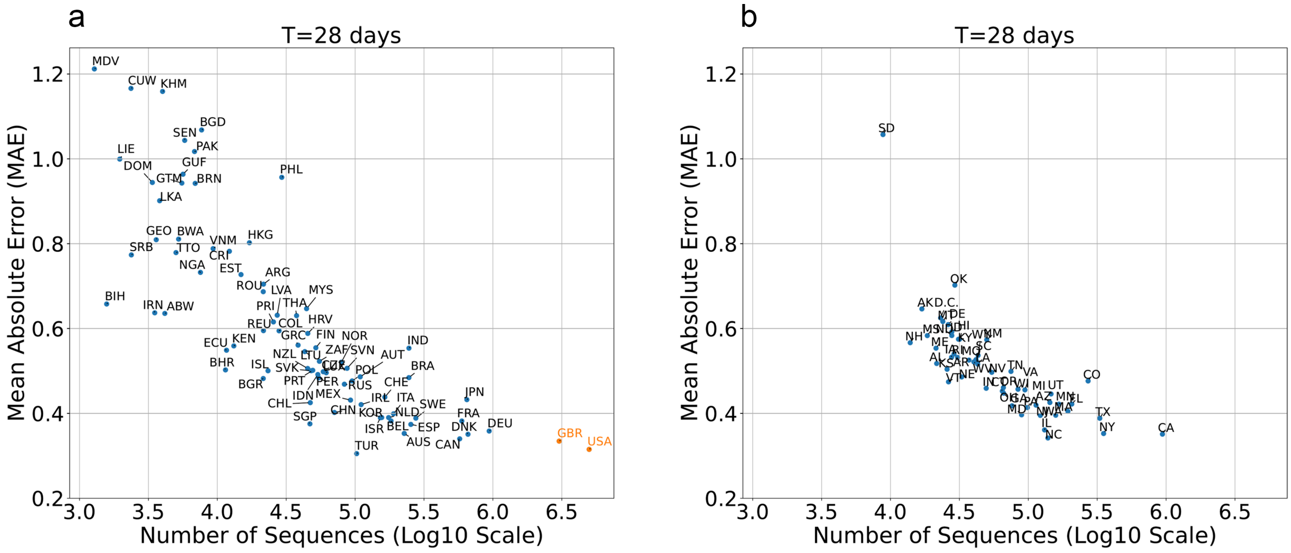

Figure 4. The transformer model exhibits remarkable generalization ability when tested on data from countries across the globe (a) and data from the US states (b). (a) The MAE for the 28-day prediction against the total number of sequences collected before February 26, 2024 in each country. Country codes are indicated around the data points. There is a significant linear correlation between the total number of sequences (on the log10 scale) and the MAE of our model predictions (Pearson correlation coefficient −0.82). (b) The MAE for the 28-day prediction against the total number of sequences collected before February 26, 2024 in each US state. State names were indicated around the data points. There is a significant linear correlation between the total number of sequences (on the log10 scale) and the MAE of our model predictions (Pearson correlation coefficient is −0.77).

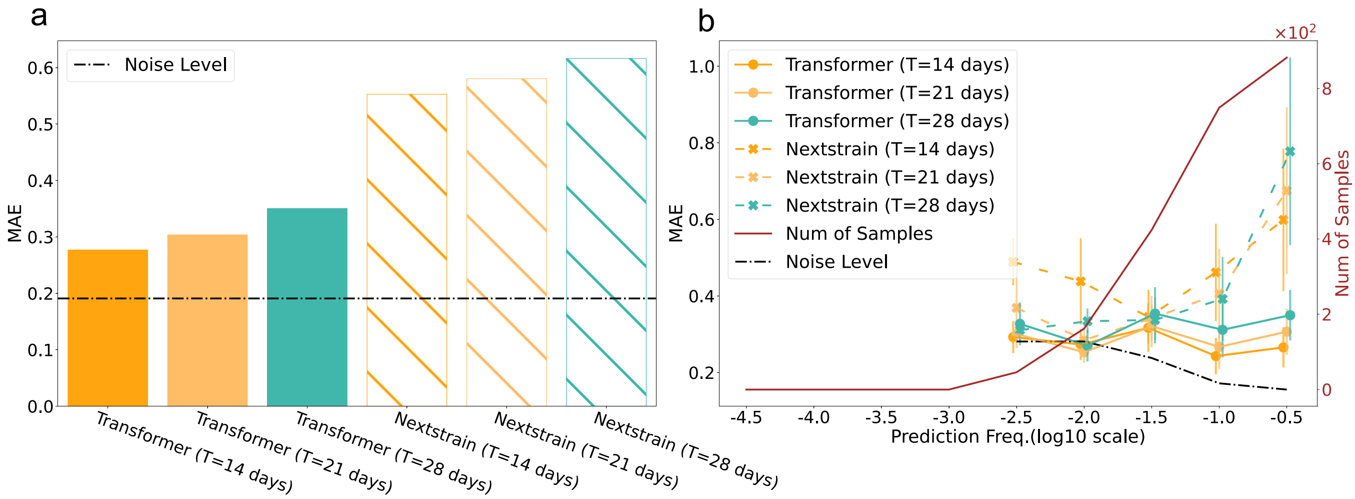

Figure 5. The transformer model outperforms Nextstrain's Multinomial Logistic Regression (MLR) model. The MAEs for the transformer model were calculated for countries and dates where the Nextstrain MLR model predictions exist. Only results for 14, 21, and 28 days' forecasting are shown, because the Nextstrain MLR model forecast a maximum of 30 days into the future. Our models outperform Nextstrain's MLR model both in terms of overall results (a) and stability across different prediction frequencies (b).

a regression-based approach. This type of approach implements regression on past lineage time series and projects future trajectories based on estimated growth rates. As we demonstrated in this work, these approaches suffer from large noise and biases in the recently collected data points due to reporting delay. In addition, in our previous work, we found that long time series are needed for regression-based approaches to make reliable predictions of relative variant fitness due to variation during the prediction interval (van Dorp et al., 2021). Indeed, the multinomial regression model implemented in Nextstrain has stringent criteria for prediction (a lineage must have at least 5000 sequences in the past 150 days) to ensure prediction accuracy. In contrast, the transformer-based models we developed here overcome these challenges and make reliable predictions from a minimum of only 14 days of non-zero data in the past 42 days. Furthermore, from a computational cost perspective, our model is relatively small for a deep learning model, containing only 1600 parameters. Training our model takes approximately one hour on an Nvidia V100 GPU with a batch size of 256. Once the model is trained, it can be used to make predictions much more quickly. For making forecasts, it requires just 0.0007 seconds to predict a single lineage on a CPU. Thus, the trained model can be easily implemented on a laptop computer to make rapid forecasts. Retraining will be desirable if model performance can be improved by new data coming in. However, we observed similar performance even after adding an additional six months of data to train the model (between January 1, 2023 and

June 30, 2023; results not shown). Given also the robust model performance we observed against datasets that were not used for model training, we expect that frequent retraining of the model will not be needed.

Our model's exceptional performance and ability to generalize stem from several strategies specifically designed according to the unique features of the dataset. First, early identification of emerging lineages that eventually become dominant in the population is critical for any lineage forecasting tool. One inherent issue with the data is that lineages that eventually dominated the population are few compared to those that did not, leading to a data imbalance for machine learning model training. By introducing an adaptive loss weighting mechanism to address the data imbalance, we ensured our model prioritizes accurate predictions for rapidly growing lineages that will rise to a high frequency. Second, to increase the accuracy of long-term prediction, we employed an incremental learning strategy that uses the trained short-term predictions to facilitate long-term predictions. Third, we implemented cross-validation and noise injection techniques (Orvieto et al., 2022) to increase the stability of the model, which works well for countries with a relatively high sequencing effort. The model performance becomes poorer for countries with less sequencing data available (for example, countries with < 30,000 sequences collected before 2024), suggesting the noise and bias in the datasets from these countries are too large for our current model.

Despite the good performance, there are limitations and areas of improvement to our model. First, in the lineage frequency data used to train our model, we assumed that the metadata for each sequence is available when it appears in the database; however, in reality, a newly emerged lineage is assigned as a new Pango lineage only after enough sequences on that lineage have been obtained, and Pango lineage designations may continue to change over time. Therefore, although we showed that our model provides good predictions at low frequencies as a lineage is newly emerging, predictions only are possible after lineage assignment in a real-time setting. We were unable to recover the full history of Pango lineage assignments in order to evaluate how early our model would be able to forecast for each newly emerged lineage. Nonetheless, we demonstrated here that our model is able to make accurate predictions once a newly emerged lineage is defined and there are at least 14 days of non-zero data available. Second, our model only makes predictions on lineages already existing in the database. As with other forecasting models, it does not predict the origination of new lineages. The origin of a new lineage that transmits efficiently in the population may completely change the dynamics of existing lineages. For example, for the Omicron BA.1 outbreak in late 2021, our model does not make predictions about it before it is assigned as a new lineage; however, the model makes accurate predictions of this and other Omicron lineage frequencies once it is assigned and there is enough time series data available. Third, our transformer model was trained on the time series of lineages individually. Although we normalized the predicted frequencies of all existing lineages to 1 for real-time predictions, our model fundamentally only focuses on the dynamics of individual lineages without considering the interactions among co-existing ones. One potential future direction is to develop models to make predictions based on all existing lineages. However, data collected on a single day only lead to one training data point for the model, in this case. Currently, the amount of available data (on the order of a thousand data points) is not sufficient to train such a transformer-based lineage interaction model. Fourth, our model was trained on data collected from the UK and the USA. One potential improvement to the model is to add a country token in the input data stream, train the model using data from more countries, and make country-specific predictions. However, currently, the amount of data from most countries is insufficient to train a model with a country token to make substantial improvements compared to our existing model. As more and more data becomes available, the transformer model can be improved by implementing this strategy. This also points towards the importance of continuing and broadening the genomic surveillance of SARS-CoV-2 lineage evolution.

Overall, our transformer-based model is adept at navigating the challenges posed by substantial noise, bias, and missing data inherent in lineage viral frequency datasets. It not only demonstrates substantially improved forecasting accuracy relative to other methods, it also exhibits remarkable generalization ability across numerous countries worldwide. It thus demonstrates that modern machine learning-based approaches represent a promising framework going forward to advance the field of outbreak analysis and epidemiological forecasting.

Acknowledgements

We thank C. Römer for pointing out the issue of lineage name definitions changing over time, and T. Bedford for discussions and the suggestion of performing logit transform on lineage frequency. We gratefully acknowledge all data contributors, i.e. the Authors and their Originating laboratories responsible for obtaining the specimens, and their Submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based.

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Conflict of interest: None declared.

Funding

Research presented in this article was supported by the Laboratory Directed Research and Development program of Los Alamos National Laboratory under project number 20230830ER, and an NIH grant to R.K. (Grant No. R01-AI152703). This research used resources provided by the Darwin testbed at Los Alamos National Laboratory (LANL) which is funded by the Computational Systems and Software Environments subprogram of LANL's Advanced Simulation and Computing program (NNSA/DOE).

Data availability

The findings of this study are based on metadata associated with 15,450,997 sequences available on GISAID up to February 26, 2024, via gisaid.org/EPI_SET_240904ek or doi.org/10.55876/gis8.240904ek.

Code availability

Code underlying this work is available at https://github.com/ruianke/CovTransformer.

References

Nextstrain. Nextstrain SARS-CoV-2 forecasts. https://nextstrain.org/sars-cov-2/forecasts (2023).

Abousamra E, Figgins M, and Bedford T. Fitness models provide accurate short-term forecasts of SARS-CoV-2 variant frequency. *Medrxiv* 2023;2023–11.

Baden LR, El Sahly HM, Essink B. et al. Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine. *N Engl J Med* 2021;384:403–416.

Beaufays F Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

Bedford T. Evolutionary forecasting for seasonal influenza and SARS-CoV-2. https://bedford.io/talks/evolutionary-forecasting-fda-dvp/#/36 (2023).

Beesley LJ, Moran K R, Wagh K. et al. SARS-CoV-2 variant transition dynamics are associated with vaccination rates, number of co-circulating variants, and convalescent immunity. *EBioMedicine* 2023;91.

Brown T, Mann B, Ryder N. et al. Language models are few-shot learners. *Adv Neural Inform Process Syst* 2020;33:1877–1901.

Elbe S, and Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob Challeng* 2017;1: 33–46.

Carabelli AM, Peacock TP, Thorne LG. et al. SARS-CoV-2 variant biology: immune escape, transmission and fitness. *Nat Rev Microbiol* 2023;21:162–177.

Che Z, Purushotham S, Cho K. et al. Recurrent neural networks for multivariate time series with missing values. *Sci Rep* 2018; 8:6085.

Chen J, Wang R, Gilby NB. et al. Omicron variant (B.1.1.529): infectivity, vaccine breakthrough, and antibody resistance. *J Chem Inform Model* 2022;62: 412–422.

Colijn C, Liu P & Colijn C The potential of genomics for infectious disease forecasting. *Nat Microbiol* 2022;7:1736–1743.

Devlin J., Chang M.-W., Lee K. et al. *Pre-Training of Deep Bidirectional Transformers for Language understanding*. 2018, *arXiv:1810.04805: arXiv preprint*.

Dong E, Du H, and Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect Dis* 2020;20:533–534.

Dosovitskiy A Lucas B, Alexander K. et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR* 2021.

Du H, Dong E, Badr HS. et al. Incorporating variant frequencies data into short-term forecasting for COVID-19 cases and deaths in the USA: a deep learning approach. *Ebiomedicine* 2023;89: 104482.

Edouard M, Ritchie H, Lucas R. et al. Coronavirus pandemic (COVID-19). *Our World in Data* 2020.

Figgins MD, and Bedford T. Sars-cov-2 variant dynamics across us states show consistent differences in effective reproduction numbers. *medRxiv* 2022.

ForecastHub C. COVID-19 forecasthub. https://covid19forecasthub.org (2023).

Gao Z Shi X, Wang H. et al. Earthformer: exploring space-time transformers for earth system forecasting. *Adv Neural Inform Process Syst* 2022;35:25390–25403.

Girdhar R, Carreira J, Doersch C. et al. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Grubaugh ND, Ladner JT, Lemey P, et al. Tracking virus outbreaks in the twenty-first century. *Nat Microbiol* 2019;4:10–19.

Hadfield J &, Megill C, Bell S *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* 2018;34: 4121–4123.

Harvey WT, Carabelli AM, Jackson B. et al. SARS-CoV-2 variants, spike mutations and immune escape. *Nat Rev Microbiol* 2021;19:409–424.

Khan S, Naseer M, Hayat M. et al. Transformers in vision: A survey. *ACM computing surveys (CSUR)* 2022;54:1–41.

Kirillov A, Mintun E, Ravi N. et al. *Segment anything*. 2023, *arXiv:2304.02643. arXiv preprint*.

Lässig M, Ruchnewitz D, Eberhardt J. *et al* Population immunity predicts evolutionary trajectories of SARS-CoV-2. *Cell* 2023;186:5151–5164.

Loshchilov I & Hutter F. *Sgdr: Stochastic Gradient Descent With Warm restarts*. 2016, *arXiv:1608.03983: arXiv preprint*.

Loshchilov I & Hutter F Decoupled weight decay regularization. In *Sixth International Conference on Learning Representations (ICLR)*, 2018.

Lucchi A, Kersting H, Proske F. Anticorrelated noise injection for improved generalization. In *International Conference on Machine Learning*, PMLR, 2022, pp.17094–17116.

Markov PV, Ghafari M, Beer M. *et al* The evolution of SARS-CoV-2. *Nat Rev Microbiol* 2023;21:361–379 .

O'Toole A, Pybus OG, Abram ME. et al. Pango lineage designation and assignment using SARS-CoV-2 spike gene nucleotide sequences. *BMC Genomics* 2022;23:121.

Polack FP, Thomas SJ, Kitchin N. et al. Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine. *N Engl J Med* 2020;383:2603–2615.

Radford A, Kim JW, Hallacy C. et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, PMLR, 2021, pp.8748–8763.

Rambaut A, Holmes EC, O'Toole Á. et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol* 2020;5:1403–1407.

Rössler A, Netzl A, Knabl L, Bante D. et al. Characterizing SARS-CoV-2 neutralization profiles after bivalent boosting using antigenic cartography. *Nat Commun* 2023;14:5224.

Song H, Rajan D, Thiagarajan J. et al. Attend and diagnose: Clinical time series analysis using attention models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018;32.

Syrowatka A, Kuznetsova M, Alsubai A. et al. Leveraging artificial intelligence for pandemic preparedness and response: a scoping review to identify key use cases. *NPJ Digit Med* 2021;4.

Van Dorp C. H., Goldberg EE, Hengartner N et al. Estimating the strength of selection for new SARS-CoV-2 variants. *Nat Commun* 2021;12:7239.

Vaswani A, Noam S, Niki P. *et al.* Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, California, USA, Vol. 30, pp.6000–6010. 2017.

Volz E Fitness, growth and transmissibility of SARS-CoV-2 genetic variants. *Nat Rev Genet* 2023;24:724–734.

Wen Q Zhou T, Zhang C. et al. Transformers in time series: a survey. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2023.

Wilks SH, Mühlemann B, Shen X. et al. Mapping SARS-CoV-2 antigenic relationships and serological responses. *Science* 2023;382:eadj0070.

Yan J & Yan J. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The Eleventh International Conference on Learning Representations*, 2022.

Zisserman A Video action transformer network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp.244–253.

Virus Evolution, 2024, 10(1), veae086 , DOI: https://doi.org/10.1093/ve/veae086, Advance access publication date: 14 November 2024, Research Article