nature microbiology

Article

https://doi.org/10.1038/s41564-026-02377-5

A deep mutational scanning-informed protein language model predicts SARS-CoV-2 evolution dynamics with spatiotemporal resolution

Received: 17 October 2025

Accepted: 29 April 2026

Published online: 27 May 2026

Check for updates

Sijie Yang[1,2,3,5], Xiaowei Luo[2,5], Jiejian Luo[2], Fanchong Jian[1,4] ✉ & Yunlong Cao[1,2,3,4] ✉

Early identification of emerging dominant variants of pathogens such as SARS-CoV-2 is important for effective public health responses, yet existing approaches are not feasible for real-time surveillance. Here we introduce DeepCoV (DMS-Empowered Evolution Prediction of CoronaVirus), a deep-learning framework for the dynamic identification of emerging variants with high potential to become prevalent at spatiotemporal resolution. It integrates deep mutational scanning (DMS)-derived mutation phenotypes, evolutionary sequence data and epidemiological surveillance data reflecting human immune pressures. Benchmarked against logistic regression-based methods and representative deep-learning approaches in simulated retrospective surveillance scenarios, DeepCoV accurately forecasts the dominance of recently circulating lineages a month in advance, achieving a 90% reduction in false discovery rate while capturing temporal and geographic dynamics of variant spread and reconstructing their regional prevalence trajectories. It also identified mutational hotspots of Omicron-derived backbones in silico, revealing convergent evolution trends. This provides a scalable framework for timely identification of immune-evasive variants and critical mutations, providing actionable insights.

The evolutionary arms race between pathogens and human immunity necessitates proactive surveillance of emerging variants[1–3]. For rapidly evolving viruses such as SARS-CoV-2, early identification of high-growth lineages is essential to pandemic resilience, enabling timely updates to vaccines and informing the development of antibody-based therapeutics.

Although high-throughput experimental approaches such as deep mutational scanning (DMS) can generate valuable data and insights into the functional impact of individual mutations, their substantial resource requirements restrict their application for continuous surveillance[4–9]. Moreover, DMS-based methods are inherently incapable of capturing the evolutionary dynamics of full viral sequences within populations, as they typically probe only a subdomain of the full spike protein or a restricted set of mutations, and face challenges in modelling epistatic interactions, given the prohibitively large mutational combinatorial space[9,10]. These methods proved critical early in the COVID-19 pandemic but have become increasingly impractical (Table 1 and Table 2).

[1]Biomedical Pioneering Innovation Center (BIOPIC), Peking University, Beijing, P. R. China. [2]Changping Laboratory, Beijing, P. R. China. [3]Peking-Tsinghua Center for Life Sciences, Tsinghua University, Beijing, P. R. China. [4]School of Life Sciences, Peking University, Beijing, P. R. China. [5]These authors contributed equally: Sijie Yang, Xiaowei Luo. ✉e-mail: jfc@pku.edu.cn; yunlongcao@pku.edu.cn

Table 1 | Hyperparameters of DeepCoV for JN.1-era prediction

Hyperparameter	Value
Optimizer	AdamW
Warm-up steps	300
Learning rate	0.0001
Weight decay	0.01
Batch size	5
Early stop patience	10
Number of axial attention layers	3
Number of transformer encoder layers	2
Number of LSTM layers	2
Number of total trainable parameters	15.66 million

Table 2 | Hyperparameters of Spike model

Hyperparameter	Value
Optimizer	AdamW
Warm-up steps	300
Learning rate	0.0001
Weight decay	0.01
Batch size	5
Early stop patience	3
Number of axial attention layers	3
Number of transformer encoder layers	2
Number of LSTM layers	2

Statistical models based on sequence frequency dynamics, such as linear growth advantage estimation, offer alternative tools for inferring variant fitness directly based on epidemiological surveillance[11]. However, their predictive reliability deteriorates under data-sparse conditions, especially during the early stage of novel lineage emergence and in the post-pandemic period, when sequencing efforts have markedly decreased. More sophisticated frameworks, including EpiScore and PyR0, incorporate evolutionary constraints by modelling sequence prevalence over time[12,13]. However, they often struggle to pinpoint the most prevalent circulating strains. Without sequence information, such epidemiological analyses remain largely phenomenological and offer limited mechanistic insight into why certain variants rise to dominance while others fade. Meanwhile, reliably capturing sequence-level features remains inherently challenging for statistical approaches. More recently, a dynamic immune landscape framework integrated past lineage dynamics, immune waning and DMS data within an explicit biophysical model to characterize immune-mediated selection[3]. While this approach offers strong mechanistic interpretability, relying solely on limited DMS profiles is often insufficient to capture the full complexity and evolving nature of population-level immune pressure. Furthermore, the lack of full-sequence context or higher-order epistasis modelling limits scalability as viral diversity expands.

Artificial intelligence (AI)-based methods have thus emerged as promising tools for forecasting viral evolution[13–19]. AI-based approaches can overcome the combinatorial explosion arising from multiple mutations within viral sequences and enable the integrated learning of large, diverse strain sets, including capturing amino acid sequence evolutionary patterns. However, existing models remain limited in their ability to prospectively identify emerging dominant variants with sufficient accuracy. Recent work such as CovTransformer has also explored direct forecasting of SARS-CoV-2 lineage dynamics from epidemiological time series using Transformer-based models. CovTransformer demonstrated that such architectures can robustly model noisy lineage frequency trajectories and improve short-term extrapolation of observed Pango lineages[20]. However, because these approaches operate purely on aggregated prevalence signals, they remain inherently retrospective and do not incorporate sequence-level constraints or enable prospective assessment of unobserved variants. Viral protein sequence or structure-based approaches such as EVEscape (variational autoencoder based) and TEMPO (Transformer-based) exhibit strong representational capacity but typically overlook functional data, particularly experimentally derived measurements of antibody escape and other virological phenotypes captured through DMS[14,19]. Methods such as E2VD and CoVFit have advanced by leveraging mutational phenotypes, but they still often neglect the dynamic host immune context, which is critical for capturing the spatiotemporal dimensions of viral transmission[18,21]. Moreover, most current computational models fail to capture the dynamic viral fitness landscape under the co-evolution of host immune pressures, frequently underperforming compared to even simple linear growth advantage estimators in real-world predictive applications.

Despite recent progress, a major challenge remains in jointly capturing the spatiotemporal dynamic fitness landscape for viral evolution under evolving population immune pressures. Even within the same lineage, transmission advantages may diverge substantially across regions and time with distinct immune history. Meanwhile, DMS has been underutilized for predictive purposes, and most applications have remained descriptive, focusing on characterizing escape mutations rather than integrating functional data into dynamic evolutionary modelling[4–7,9].

To bridge this gap, we developed DeepCoV (DMS-Empowered Evolution Prediction of CoronaVirus), a predictive framework that integrates DMS-derived functional phenotypes, evolutionary sequence information and epidemiological data indirectly reflecting immune pressures in human populations. By leveraging Transformer-based architectures, DeepCoV learns the mechanistic relationships between mutation effects and variant fitness while incorporating background epidemiological data and related sequences with pretrained protein language models to model viral evolution at spatiotemporal resolution[3,22,23]. Collectively, DeepCoV provides a scalable and biologically grounded framework for forecasting SARS-CoV-2 evolutionary trajectories, thereby enhancing timely public health interventions.

Results

DeepCoV architecture

Accurately forecasting the evolutionary dynamics of SARS-CoV-2 requires integrating information that reflects both the intrinsic viral fitness for infection and transmission, and the impacts of population immune pressure. We designed DeepCoV, a deep-learning framework that predicts the future prevalence of any SARS-CoV-2 Spike or receptor binding domain (RBD) variant, leveraging three complementary data collected during months before the time of prediction: (1) multiple sequence alignment (MSA) of viral antigen sequences including the variant for prediction and other co-circulating strains competing within the same environment; (2) the proportions of the above strains since 180 days before the day of prediction, capturing recent viral fitness evolution with the population-level immunity and selective pressure considered implicitly, and endowing the model with the ability to learn the spatiotemporal dynamics of variant circulation. We emphasize that population-level immune pressures in this study are indirectly inferred from historical prevalence dynamics, which reflect the combined effects of host immunity, viral fitness and short-term epidemiological trends; and (3) DMS-derived functional mutation phenotypes quantifying the impacts on virological characteristics, thereby grounding the model in experimentally validated datasets and molecular mechanisms (Fig. 1 and Extended Data Fig. 1).

https://doi.org/10.1038/s41564-026-02377-5

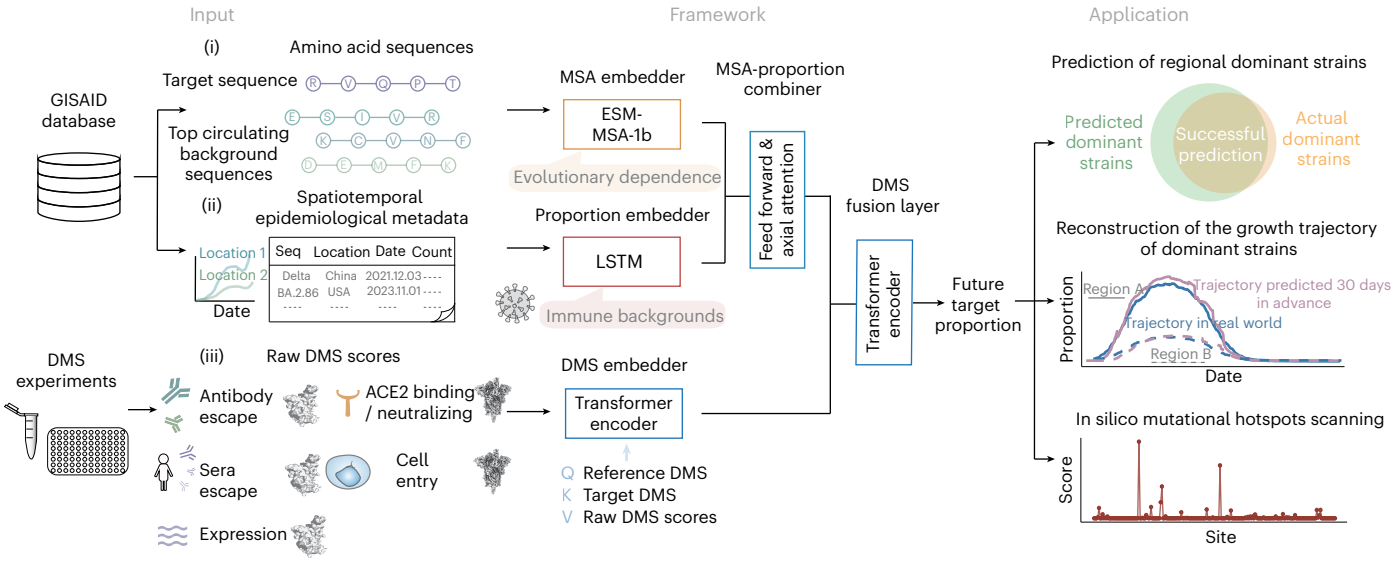


Fig. 1 | Overview of model architecture and predictive applications. The DeepCoV framework integrates (i) MSAs of target variants and co-circulating background strains, (ii) historical lineage prevalence over the preceding 180 days and (iii) DMS-derived mutational phenotypes to predict future variant prevalence. Model outputs are prospective dominance predictions at a future timepoint t_1, which is set to t_0 + 30 days for model training and evaluation in this study, enabling early identification of emerging variants, spatiotemporal trajectory reconstruction and in silico hotspot analysis. The emergence date of JN.1 is defined as the earliest date on which the number of JN.1 RBD sequences exceeds 10.

For sequence modelling, we implemented an evolutionary module using the pretrained ESM-MSA-1b model, which learns evolutionary constraints from MSA[24]. To incorporate temporal dynamics, the historical proportion embedder employs a long short-term memory (LSTM) network to model prevalence trajectories over a sliding window spanning 180 days[25]. These sequence and prevalence representations are concatenated and passed through an axial-attention module to capture residue-prevalence dependencies[22]. This architecture inherently encodes population-level immune histories by integrating background sequence context and prevalence dynamics, reflecting how previous infections and vaccinations shape the viral fitness landscape. Finally, the DMS embedder incorporates quantitative mutational phenotypes, including antibody escape, human antisera evasion, ACE2 binding affinity and protein expression, offering mechanistic insights into variant viability and transmissibility[4–9,26–32]. By integrating these heterogeneous data streams, DeepCoV learns how individual amino acid substitutions and their functional consequences translate into shifts in population-level prevalence, thereby linking molecular evolution with epidemiological outcomes. This unified representation enables the model to infer variant fitness, anticipate lineage competition outcomes, and forecast regional prevalence trends over future time windows. Finally, DeepCoV is expected to predict the future proportion of a certain strain at any timepoint, with the sequences of itself and other co-circulating strains, their historical proportions, and the impacts of mutations carried by the strain from DMS as inputs.

DeepCoV accurately predicts predominant variants

We first evaluated DeepCoV for the early identification of emerging dominant JN.1 variants using a retrospective approach. Specifically, the model was trained exclusively on epidemiological records and RBD sequences collected before October 2023, along with DMS profiles generated before the emergence of JN.1 lineage (Fig. 2a). To prioritize biologically meaningful dynamics and mitigate the severe class imbalance caused by abundant low-frequency variants, training was restricted to lineages exceeding 0.5% frequency on at least 1 day within the t_1 interval. In addition, cluster-based train/validation splitting was applied to prevent temporal leakage.

DeepCoV demonstrated high predictive capacity, evidenced by a strong correlation (Pearson's r = 0.969, Fig. 2b; Pearson's r = 0.893 for Δprevalence, Supplementary Fig. 1a) for historically dominant lineages. We systematically compared DeepCoV with the conventional growth advantage fitting method and state-of-the-art models E2VD and EVEscape, as well as multinomial logistic regression (MLR) to evaluate their performance in meeting real-world pandemic surveillance requirement[14,21,33]. We first assessed how early these methods could correctly prioritize the known dominant variants (JN.1, KP.2 and KP.3) among the top predicted lineages. DeepCoV uniquely identified JN.1, KP.2 and KP.3 as top dominance candidates among all the RBD sequences that appeared since October 2023, well ahead of their observed dominance. In contrast, EVEscape successfully predicted only KP.2, and E2VD failed to detect any of the dominant variants. The MLR model predicted KP.2 and KP.3 well but failed to anticipate the emergence of JN.1. Moreover, growth-advantage-based approaches consistently identified dominant variants later than DeepCoV, indicating a reduced capacity for early and stable detection of emerging dominant lineages (Fig. 2c).

Detailed benchmarking across different numbers of top-ranked predicted variants confirmed DeepCoV's superior ability to identify emerging dominant lineages. To comprehensively assess predictive performance, we evaluated whether the model could identify the dominant lineage within its top-k predictions. For the time-resolved ground truth, we evaluated two metrics across timepoints: the success rate of top-1 versus top-k prediction, and the Jaccard index between the top-k predicted and the top-3/top-5 observed ground-truth variants (Fig. 2d,e and Extended Data Fig. 2). The results indicated that our model consistently outperformed baseline methods, especially at lower k thresholds (Fig. 2e,f and Extended Data Fig. 2). For both top-3 versus top-3 and top-5 versus top-5 comparison, DeepCoV successfully identified most major variants; in contrast, MLR incorrectly prioritized non-dominant strains such as JN.1 + K403R and JN.1 + S408R, and other methods achieved at most two overlapping variants (Fig. 2d and Supplementary Fig. 2).

To investigate the contributions of individual model components and biological data modalities, we performed systematic ablation studies. Four model variants were constructed by selectively excluding key information: (1) evolutionary–epidemiological context, retaining only sequence and prevalence data; (2) DMS phenotypes, preserving sequence and immune trend inputs and replacing DMS module with

linear layers; and (3) evolutionary sequence context, replaced by ESM-2 embeddings to isolate the effect of evolutionary modelling. Removal of any individual module resulted in a marked decrease in predictive performance (Fig. 2g,h and Extended Data Fig. 3). Both the DMS and evolutionary–epidemiological context modules were critical for accurately detecting dominant strains, substantially reducing false discovery rate (FDR) while maintaining recall. Moreover, eliminating the proportion or ESM-MSA-1b modules rendered the model incapable of training. Together, these results highlight the essential and collaborative contributions of background immune landscape dynamics, evolutionary sequence context, historical prevalence trends and functional mutational phenotypes to the overall predictive capacity of DeepCoV.

DeepCoV captures variant spatiotemporal dynamics

Beyond accurate early variant identification, DeepCoV effectively captures the spatiotemporal dynamics of SARS-CoV-2 variant spread, demonstrating substantial improvements over existing methodologies. We further reconstructed the evolutionary trajectories of dominant SARS-CoV-2 variants with high temporal resolution using DeepCoV. The model effectively captured the full expansion and decline cycles of major JN.1 clades, maintaining stable predictive accuracy throughout the JN.1-dominant period (Fig. 3a). DeepCoV maintained consistently high precision over time, with major strain predictions showing slightly larger but acceptable fluctuations (mean absolute error (MAE) < 0.1; root mean square error (RMSE) < 0.15) (Extended Data Fig. 4). Notably, our prospective forecasts anticipated observed dynamics, successfully tracking the turnover of dominant variants. Although the model occasionally yielded conservative estimates of peak magnitude during rapid epidemiological shifts, it outperformed the MLR baseline in signal reliability. Specifically, DeepCoV effectively mitigated spurious growth signals, such as the false expansion of JN.1 + ins483V + K484E reversion predicted by MLR (Supplementary Figs. 3 and 4). Quantitative evaluation confirmed DeepCoV's superior performance, demonstrating lower reconstruction error and greater temporal stability (Supplementary Fig. 5).

Importantly, DeepCoV demonstrated sensitivity to subtle early growth signals, successfully forecasting the rise of dominant variants even from low initial prevalence (<5%). Notably, despite differing from its ancestral BA.2.86 lineage by only a single RBD substitution (L455S), the rapid rise of JN.1 was correctly captured by DeepCoV as the dominant variant. This highlights DeepCoV's capacity to distinguish variants with minimal genetic differences but markedly divergent epidemiological trajectories, underscoring its sensitivity to functionally meaningful mutations. In addition to major lineages, the model faithfully reconstructed the growth trajectories of subdominant variants such as JN.1 + F456L and JN.1 + R346T, as well as high-growth-advantage but ultimately low-prevalence lineages including JN.1 + K403R and JN.1 + N417K (Extended Data Fig. 5). DeepCoV also showed high specificity in handling non-dominant variants, with predicted peak prevalence consistently remaining below 3%. These findings underscore the model's utility for monitoring fine-scale viral evolution and guiding timely public health responses, including vaccine strain selection.

By intrinsically incorporating region-specific prevalence dynamics that reflect inferred immune landscapes, DeepCoV successfully captures geographically distinct transmission patterns. The model accurately reconstructed intercontinental divergence patterns, clearly capturing the sequential emergence of KP.2 from Europe to North America and subsequently Asia (Fig. 3d). It also correctly identified elevated prevalence of variants such as BQ.1.1 and XBB in Western populations relative to Asia while highlighting the regional dominance of HK.3 in Asia, reflecting regional differences in pandemic responses and immune imprinting. Although other models may account for temporal factors, DeepCoV uniquely combines spatiotemporal resolution with proactive forecasting and quantitatively validated accuracy, offering improved interpretability in complex epidemiological settings.

Ablated models were further assessed on growth trajectory reconstruction (Fig. 3e). The 'no-DMS' variant erroneously overestimated the growth advantage of BA.2.86 + K478E, underscoring the essential role of the DMS module in mitigating false positives. Moreover, removing any single module eliminated early prediction of KP.2, indicating that complementary signals from multiple modules are required to support the model's overall performance.

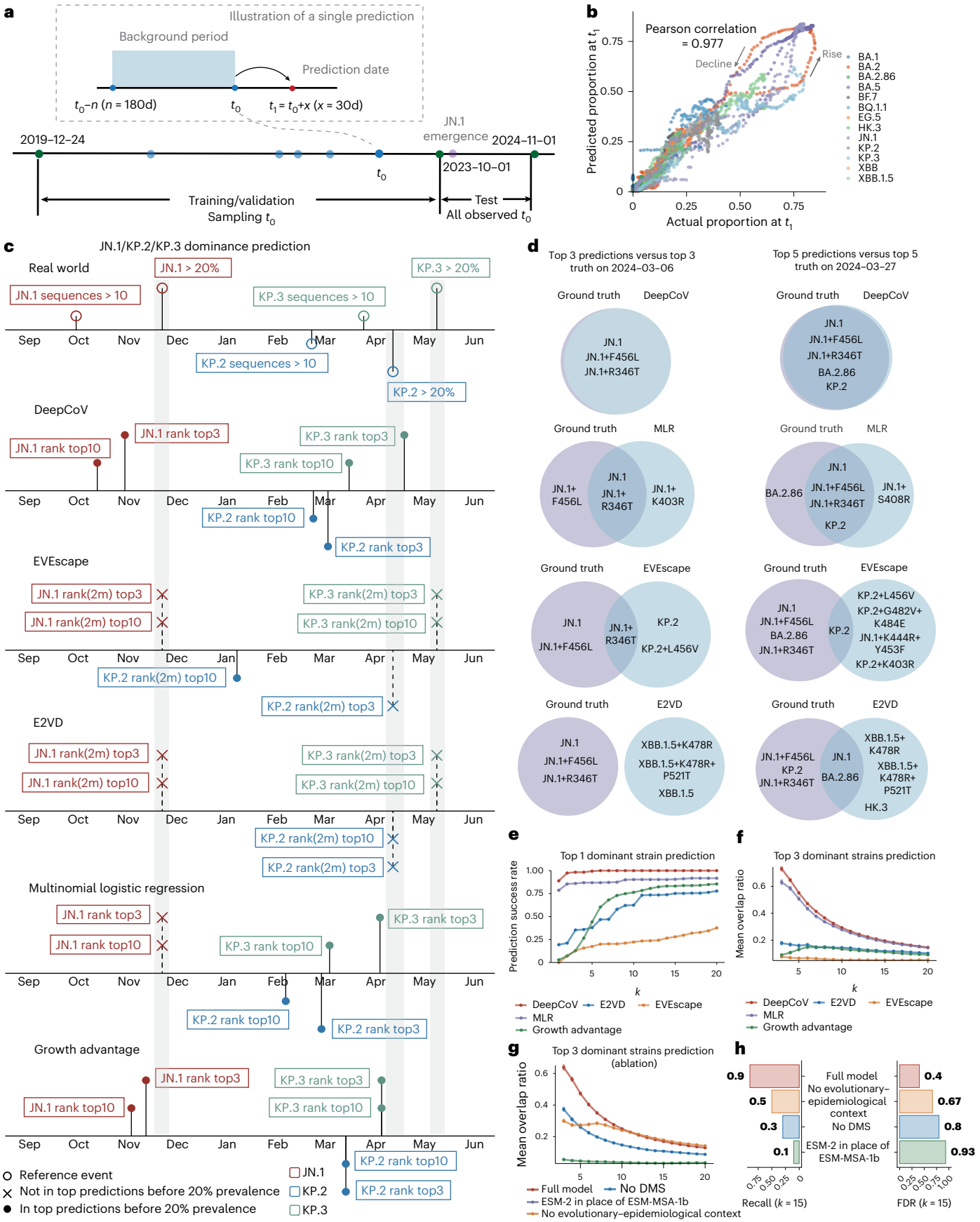
In silico mutational hotspot scanning

Leveraging DeepCoV's ability to capture evolutionary dynamics at single-mutation resolution, we conducted analysis to in silico identify mutational hotspots within the SARS-CoV-2 RBD, to understand driving forces behind immune escape mutations during convergent evolution. We computationally generated all possible single-site RBD mutants for representative convergent lineages, including JN.1 and XBB variants, and applied models trained on temporally matched datasets from the respective JN.1 or XBB eras. By predicting time-resolved evolutionary scores for each mutation, we dynamically mapped site-specific evolutionary pressure and identified candidate hotspots likely to contribute to future adaptation (Fig. 4a).

Our analysis successfully identified mutational hotspots R346T and F456L in JN.1, which later became defining mutations in emerging strains such as KP.2 before their prevalence reached 5% (ref. 34) (Fig. 4c). As expected, sites such as 403, overestimated by DMS-based assays, did not exhibit notable predicted evolutionary potential. Similarly, DeepCoV accurately identified key mutational hotspots associated with subsequent variant dominance, including the S486P substitution, followed by F456L and L455F mutations within the XBB lineage[35–37]. These predictions accurately forecasted the sequential

Fig. 2 | Accurate early detection of predominant strains by DeepCoV. **a**, Dataset construction. Training sequences were collected before 1 October 2023, including variants that exceeded 0.5% prevalence for at least 1 day during the follow-up period. Validation sets were generated using cluster-based sampling at a 1:10 ratio to minimize temporal data leakage. **b**, Scatterplot comparing predicted versus observed variant frequencies at the evaluation timepoint (t_1). Each point represents a unique RBD variant, coloured by lineage. Predictions were generated using only information available before t_0. **c**, Timeline of prediction milestones for DeepCoV and baseline methods (E2VD, EVEscape, growth advantage and MLR), showing the performance (ranking reaching top 3 or top 10) for known dominant variants JN.1, KP.2 and KP.3. Actual emergence events, such as ≥10 sequences, growth advantage >30% and prevalence >20%, are labelled as reference points. For E2VD and EVEscape, rankings include variants with sequences appearing within a 2-month window. Solid circles and lines denote successful predictions; dashed lines with crosses (×) denote failures (inability to rank dominant variants highly before the >20% true prevalence window, shaded grey). **d**, Representative comparisons of top-*k* predicted and observed dominant variants at selected timepoints. Venn diagrams show overlap between the predicted and observed top-*k* RBD variants for *k* = 3 (left; 6 March 2024, predicted 30 days in advance) and *k* = 5 (right; 27 March 2024). Predictions from DeepCoV, MLR, EVEscap and E2VD are compared against the corresponding observed sets. Variant labels indicate RBD lineage with key escape mutations. **e**,**f**, Top-*k* prediction performance over time. Performance of DeepCoV (red), MLR (purple), EVEscape (orange), E2VD (blue) and growth advantage (green). **e**, Success rate for identifying the top predicted dominant variant across varying *k* values. **f**, Jaccard index for the top-3 predicted variants across different *k* values; error bars denote ±1 s.e.m. across independent evaluation timepoints (*n* = 351). Each timepoint represents an independent prediction analysis on distinct variant datasets. **g**,**h**, Ablation analysis. **g**, Mean overlap ratio (Jaccard index) for top-3 predictions across varying *k* values. **h**, FDR and recall when *k* = 15 for the top 10 circulating variants prediction during the JN.1 period. Error bars denote ±1 s.e.m. across evaluation timepoints (*n* = 351). The top 10 true circulating variants include JN.1, KP.2, KP.3, HK.3, JN.1 + R346T, JN.1 + F456L, HK.3 + A475V, KP.2 + G482V + K484E, KP.2 + L456V + K478T and KP.2 + ins483V + K484E. Each timepoint represents an independent prediction analysis on distinct variant datasets.

https://doi.org/10.1038/s41564-026-02377-5



https://doi.org/10.1038/s41564-026-02377-5

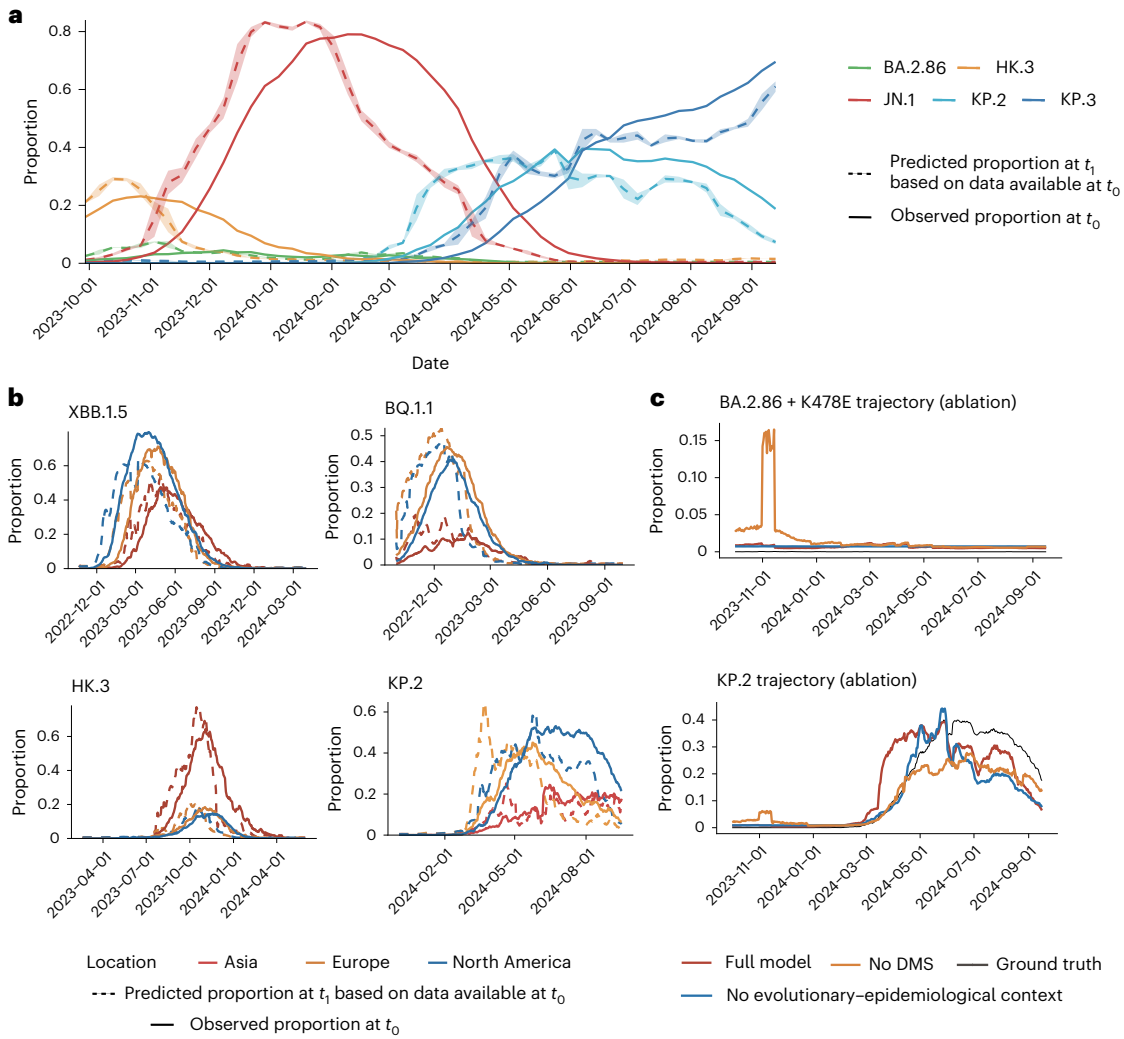


Fig. 3 | DeepCoV captures temporal dynamics and geographic variation in SARS-CoV-2 spread. a, Growth trajectory reconstruction. Weekly aggregated temporal dynamics comparing predicted (dashed lines, proportion at t_1 predicted from t_0) versus actual prevalence (solid lines, observed at t_0). Shaded regions represent mean ± s.d. Predictions are generated using a rolling forecasting strategy, in which the prediction timepoint t_0 is advanced week by week. At each t_0, the model forecasts variant prevalence at $t_1 = t_0 + 30$ days using only information available before t_0. **b**, Regional prevalence patterns across Asia, Europe and North America. Predicted (dashed lines) and observed (solid lines) variant proportions were aggregated weekly. **c**, Ablation study on growth trajectory reconstruction for BA.2.86 + K478E and KP.2. The black line indicates the full model and the coloured lines represent the ablated variants.

'FLip' mutation (F456L + L455F) wave, reflecting real-world evolutionary trajectories. Importantly, DeepCoV identified these mutation patterns ahead of widespread detection; for instance, the S486P hotspot could be predicted before XBB.1.5 (XBB + S486P) became detectable in global sequencing data (Fig. 4b). Subsequently, the emergence of F456L and 'FLip' were identified in early evolutionary stages of variants such as EG.5 and HK.3. Comparison of hotspot-carrying strain proportions between the prediction and peak phases revealed that other hotspots were likewise detectable at early stages of variant emergence, before their subsequent expansion to peak prevalence (Supplementary Fig. 6). Moreover, well before the emergence of HV.1 (EG.5 + L452R), the L452 hotspot could already be identified in the EG.5 background through in silico mutational scanning (Supplementary Fig. 7). Our findings demonstrate that DeepCoV effectively captures intrinsic residue-level drivers of evolutionary convergence. By combining temporal modelling of mutation phenotypes with sequence-based predictions, our approach mirrors the functional resolution provided by DMS experiments but with added temporal insights into mutation dynamics.

We further assessed the contribution of individual modules to JN.1 mutational hotspot detection and forecasting of future evolutionary trajectories (Extended Data Fig. 6). All ablated model variants exhibited a pronounced loss of hotspot discrimination, with only the F456L mutation detected in the 'no-DMS' model, probably reflecting previous selection signals from lineages such as XBB where F456L conferred marked escape potential. Collectively, these complementary modules act synergistically to support reliable prediction of future mutational trends.

DeepCoV generalizes to future SARS-CoV-2 evolution

Recently, increased immune pressure on the spike protein's N-terminal domain (NTD), which facilitates viral entry, has led to elevated mutational activity, establishing it as a secondary hotspot of adaptive evolution[38–40]. Variants such as XEC (T22N/F59S) and KP.3.1.1 (S31del) exemplify this trend, reflecting evolving selective pressures shaping SARS-CoV-2's immune escape mechanisms[34,41–43]. To capture the evolutionary shift, we trained DeepCoV on full spike sequences while preserving the model architecture (Extended Data Fig. 7). Considering that most DMS measurements outside the RBD region are unavailable, the updated model excluded the DMS module. The effect of removing this component could be partially compensated by the expanded number of spike sequences incorporated into the training data.

https://doi.org/10.1038/s41564-026-02377-5

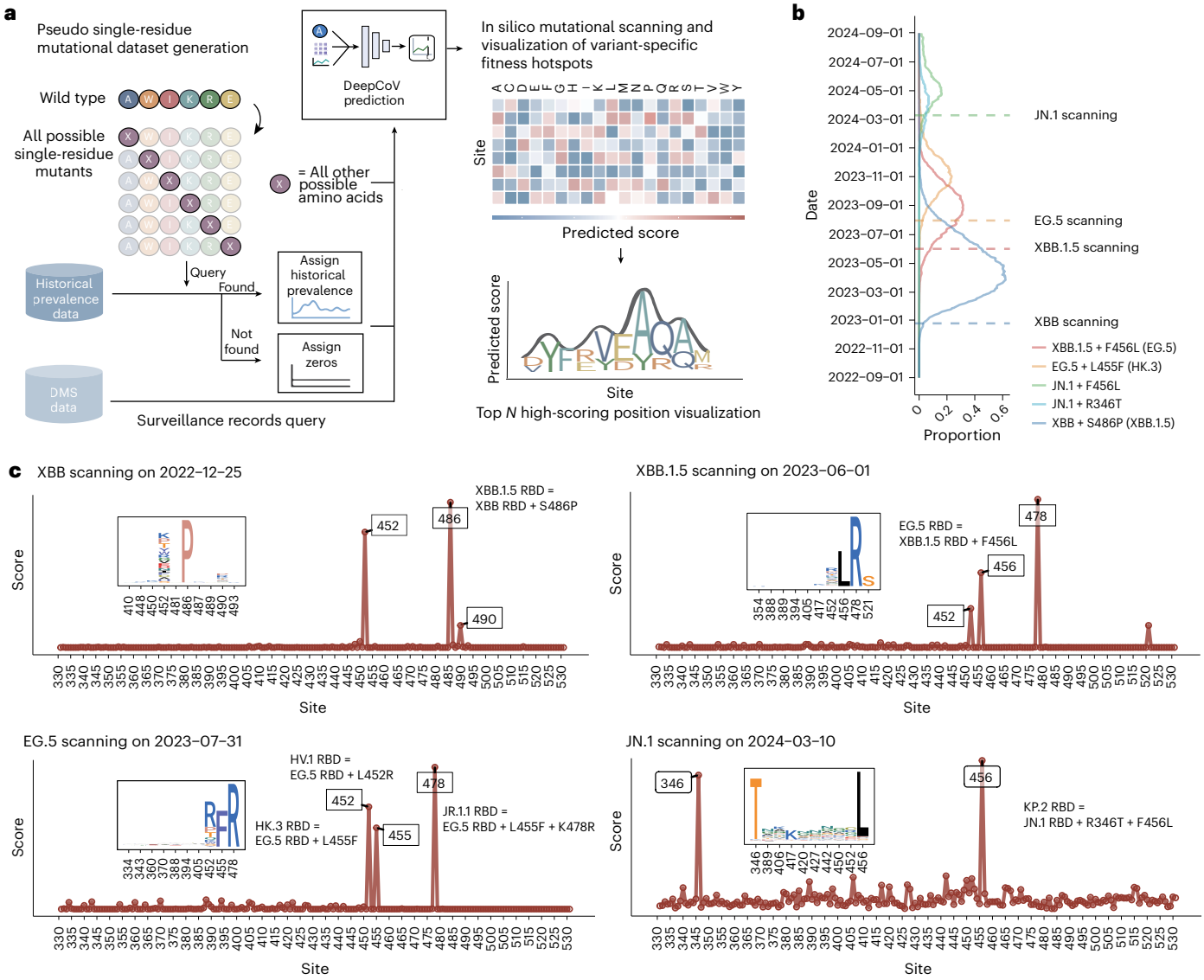


Fig. 4 | Dynamic mutational hotspot scanning in silico. a, Schematic of the mutational scanning workflow. Pseudo single-residue mutational datasets were generated and evaluated using the predictive model to estimate evolutionary potential for each mutation. **b**, Temporal prevalence dynamics of variants containing convergent mutational hotspots. Horizontal dashed lines indicate selected timepoints for in silico scanning. **c**, Predicted mutation preference landscapes for XBB (upper-left, predicted on 2022–12–25), XBB.1.5 (upper-right, predicted on 2023–06–01), EG.5 (lower-left, predicted on 2023–07–31) and JN.1 (lower-right, predicted on 2024–03–10) before the onset of convergent evolution. High-scoring hotspot sites are highlighted. Insets: amino acid substitution profiles are shown as sequence logos, where the *x* axis represents the amino acid position and the *y* axis indicates the predicted mutation score; letter heights are proportional to their respective scores and colours denote different amino acids.

The refined approach maintained excellent prediction accuracy and consistently low FDR for dominant variants. Moreover, the updated model robustly reconstructed the evolutionary trajectories of complex variants such as KP.2.3 (KP.2 + S31del + H146Q) and KP.3.1.1 (KP.3 + S31del), demonstrating strong generalizability. In addition, in silico mutational scanning of NTD successfully identified S31del as a potential immune escape mutation, which had been suggested to enhance immune escape through allosteric modulation of RBD–antibody interactions mediated by additional NTD glycosylation[34]. These results underscore DeepCoV's robust capability to generalize to previously underrepresented structural domains and highlight its utility in modelling future evolutionary adaptations.

We updated the test dataset until May 2025 to evaluate DeepCoV's performance against recently emergent strains, including LP.8 and KP.3 + A435S whose peak prevalence exceeded 10% globally. The model maintained high accuracy (Pearson's $r = 0.968$; Fig. 5b), accurately forecasting LP.8 emergence and continuing to perform well on expanded lineages such as JN.1, KP.2 and KP.3. This predictive accuracy remained robust (Pearson's $r = 0.79$) even after accounting for temporal dependency by analysing prevalence changes rather than absolute values (Supplementary Fig. 1b). In comparison, MLR systematically overestimates the future prevalence of NB.1.8.1 at the global scale (Supplementary Fig. 8). Retrospective in silico mutational scanning of KP.3 and LF.7 revealed early detection of future-dominant mutations (Fig. 5a,d). Notably, A435S was identified as a prominent hotspot ~1 month before the widespread emergence of KP.3 + A435S, while residue 475, harboured by the later-emerging LF.7.2.1 strain, was similarly highlighted during pre-emergence scans. Furthermore, DeepCoV captured the geographic skew of NB.1.8.1, accurately pinpointing its elevated prevalence in Asia[5] (Fig. 5c). Together, these findings demonstrate DeepCoV's continued ability to anticipate the emergence and geographic distribution of newly

https://doi.org/10.1038/s41564-026-02377-5

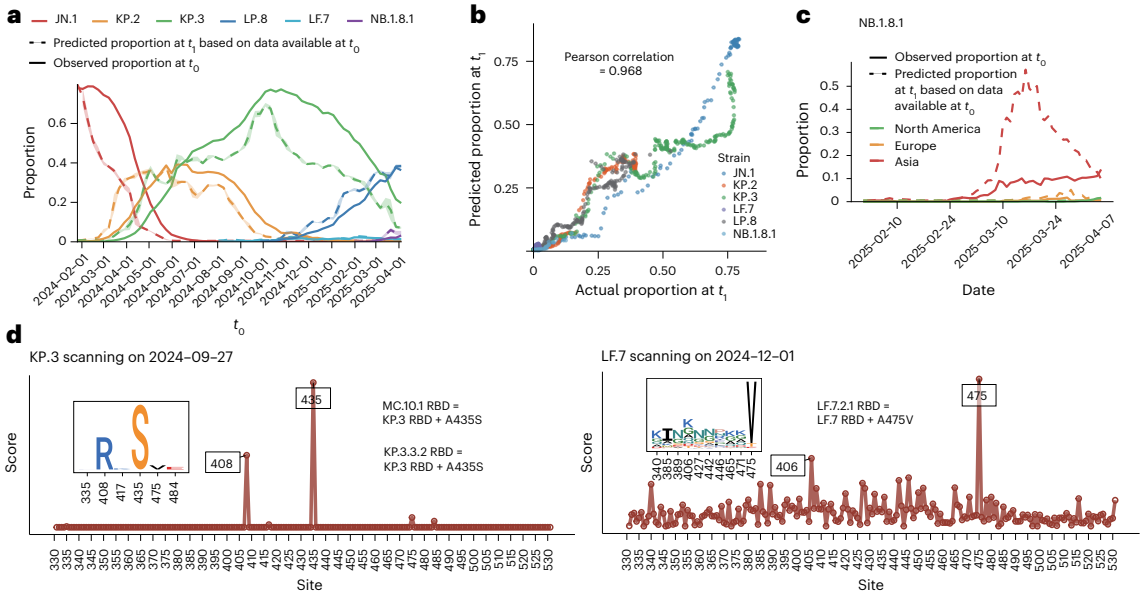


Fig. 5 | Updating SARS-CoV-2 variant data for future predictions. All analyses in this figure are based on an updated test dataset and corresponding model incorporating SARS-CoV-2 data up to May 2025, and are presented as a generalization evaluation distinct from the primary benchmarking framework. **a**, Growth trajectory reconstruction using the renewed test set. Weekly aggregated predictions (dashed lines, t_1 inferred from t_0) are compared with observed prevalence (solid lines, measured at t_0). Shaded regions represent mean ± s.d for days in a week. **b**, Pearson's correlation coefficient of predicted versus observed variant frequencies at t_1 based on data available at t_0, for each strain before reaching its peak prevalence. Colours represent different RBD lineages. Each point represents a variant, coloured by lineage. **c**, Predicted regional differences in NB.1.8.1 prevalence across Asia, Europe and North America. Predicted (dashed) and observed (solid) proportions are aggregated weekly. **d**, Prevalence of site-specific mutations in KP.3 (left, predicted on 2024–09–27) and LF.7 (right, predicted on 2024–12–01) before the emergence of convergent evolution, identifying early mutational hotspots. High-scoring hotspot sites are highlighted. Insets: amino acid substitution profiles are shown as sequence logos, where the *x* axis represents the amino acid position and the *y* axis indicates the predicted mutation score; letter heights are proportional to their respective scores and colours denote different amino acids.

arising variants and mutational hotspots, even beyond its original training horizon.

Discussion

One of the major challenges in SARS-CoV-2 surveillance and vaccine design lies in both the timely identification of emerging dominant variants after their emergence and the anticipation of high-risk mutations before they arise. To address this, we developed DeepCoV, a computational framework that integrates key evolutionary drivers—viral MSA patterns, mutational phenotypes from DMS and epidemiological data reflecting historical immune pressures—to effectively predict SARS-CoV-2 variant prevalence. DeepCoV employs the ESM-MSA-1b to capture evolutionary constraints from sequence alignments, integrates DMS-derived phenotypic data through Transformer-based modules, and models temporal epidemiological trends using LSTM networks. By integrating these evolutionary features and by formulating the training objective to emphasize early prediction of dominant strains, DeepCoV achieves markedly lower FDR and reliably identifies dominant variants ahead of conventional surveillance methods. Moreover, it captures global and regional variant prevalence trajectories and successfully predicts evolutionary mutational hotspots in different dominant strain eras. DeepCoV captures regional prevalence variations, such as for NB.1.8.1, reflecting immune selection at the Class 1 escape site residue 487, which highlights the value of DMS data for local immune profiling[44,45]. Ultimately, DeepCoV offers an early warning system that enhances public health preparedness by supporting timely policy decisions, optimizing vaccine strategies and guiding surveillance efforts.

Despite its strengths, DeepCoV has several limitations. Limited DMS data outside the RBD and inherent sequencing errors in epidemiological datasets may constrain predictive performance. Moreover, our model does not explicitly capture epistatic effects; incorporating combinatorial mutations could improve representation of these interactions[5,7,10,26]. Future refinements might also integrate additional data modalities, such as structural protein information, to further enhance predictive accuracy and generalizability[30,40]. Meanwhile, while DeepCoV indirectly reflects population-level immunity through prevalence, it simplifies the evolving immune landscape. Future in silico virus–immunity co-evolution models that jointly learn viral antigenicity and host immune adaptation may offer a more mechanistic understanding of immune-driven viral evolution. Finally, the reliability of DeepCoV's predictions is influenced by the breadth and representativeness of available sequence prevalence data, highlighting the value of continued global genomic surveillance.

In conclusion, by integrating comprehensive epidemiological and DMS datasets with deep-learning-based protein language models, DeepCoV unifies evolutionary, functional and epidemiological insights to build a reliable platform for the identification and prediction of prevalent SARS-CoV-2 strains. The model could be retrained and utilized in other fast-evolving epidemic viruses, such as influenza and RSV, once corresponding DMS datasets become available. Collectively, these innovations establish DeepCoV as a powerful tool for global health preparedness, enabling proactive responses to emerging infectious threats and informing timely vaccine and surveillance strategies.

Methods

Data preprocessing

SARS-CoV-2 sequence and epidemiological data preprocessing. SARS-CoV-2 viral sequences and related location and date metadata were downloaded from GISAID[46]. For preprocessing, sequences were first filtered on the basis of the metadata to include only original human spike proteins with complete collection dates and submitted dates (YYYY-MM-DD format). After deduplication, sequences were aligned to the reference spike proteins using MAFFT[47], further filtered to >1,230 residues and ≤10 non-standard residues, and quality-controlled RBD

regions were extracted without ambiguous residues. The common insertions ins214:EPE harboured by BA.1 and ins16:MPLF harboured by BA.2.86 substrains were retained in the MSAs. Following quality control, 28,837 unique high-quality RBD sequences and 743,737 unique high-quality spike sequences were used for subsequent procedures. For each unique RBD sequence, cluster index and cluster name were assigned for further usage. Unique RBD were renamed according to their mutations relative to their parental lineage references: WT, Alpha, Beta, Delta, Gamma, Eta, BA.1, BA.2, BA.5, BF.7, BQ.1.1, XBB, XBB.1.5, EG.5, HK.3, BA.2.86, JN.1, KP.2 and KP.3 (Extended Data Fig. 8).

Sequence datasets construction and spatiotemporal stratification. We constructed sequence datasets annotated with spatiotemporal metadata as follows. We computed the count of each unique RBD and the total counts in each region per day. To ensure data stability and reliability, we restricted the analysis to the global total and six representative regions: North America, Europe, Asia, USA, the United Kingdom and Japan, while excluding unique RBD with fewer than 30 total sequences locally. To mitigate high-frequency noise on sequencing day-level variability, we performed 7-day window smoothing on the sequence counts. For efficient indexing and aggregation, we constructed enumerators to map each collection date, country and continent into index spaces, allowing matrix-based count operations.

We constructed spatiotemporal stratified training and validation datasets for variant prevalence modelling as follows. For each spatiotemporal triplet (location, sequence, t_0), background clusters in the past 180 days (from t_0−180d to t_0) were used to indicate immunological pressure. Each candidate (location, sequence) pair was evaluated across all possible t_0 dates using the following inclusion criteria. At time t_0, the number of distinct background clusters circulating within the preceding 180-day window must be ≥16, ensuring sufficient immunological pressure complexity observation. For a given t_0, the model predicts variant growth 30 days later. To ensure the reliability of prediction targets, the cumulative isolate count during the t_1 interval was required to exceed 100. Data before 1 October 2023 were used for the training and validation sets, while data collected afterwards were reserved for the test set. To balance training data quality with comprehensive test set coverage, we applied stricter filtering and sampling criteria to the training data while retaining the test set data as completely as possible. To balance the number of positive and negative samples in the training set and enable the model to better learn the characteristics of major strains, we required that at least 1 day within the t_1 interval exhibit a target cluster ratio above 0.5%. To ensure coverage across the pandemic timeline, t_0 samples were evenly drawn from 5 time bins: 2020.02.01–2021.07.01, 2021.07.01–2022.04.01, 2022.04.01–2022.12.01, 2022.12.01–2023.05.01 and 2023.05.01–2023.10.01. Given the progressive decline in sequencing volumes across regions following 2023, especially after WHO declaration of the end of the pandemic, only global-level sequence counts were included after 1 January 2023 to maintain data quality standards. The data were further split into training and validation sets (90:10 ratio) by stratified random sampling on the (location, sequence) level to prevent data leakage. This yielded 73,081 items for the training and validation sets.

To evaluate the predictive performance of our model on emerging SARS-CoV-2 variants, we constructed two types of test set: a comprehensive full test set and a curated major-strain-specific test set including HK.3, BA.2.86, JN.1, KP.2 and KP.3 RBDs. Both sets were based on observations collected after 1 October 2023. The full test set included all candidate variants that emerged globally after the training cut-off date. There were a total of 741,463 entries in the full test set. Using a curated timeline of variant emergence, we assigned each major strain a reference start date based on its early documented isolation before burst. The full test set emphasizes comprehensive evaluation of overall model performance across a wide range of candidate variants, whereas the major test set specifically assesses the model's ability to forecast the trajectories of high-priority, globally prevalent lineages and allows for detailed, trajectory-level validation of model predictions across variants of notable public health relevance.

Preprocessing of deep mutation scanning features. To integrate functional mutational data into the DeepCoV framework, we curated and standardized multiple DMS datasets from public repositories and internal lab sources. We first collected four major classes of DMS datasets, including: (1) Spike S protein-mediated entry efficiency and ACE2 binding affinity (BA.2 and XBB.1.5) (https://github.com/dms-vep/SARS-CoV-2_Omicron_BA.2_spike_ACE2_binding; https://github.com/dms-vep/SARS-CoV-2_XBB.1.5_spike_DMS)[9]; (2) RBD expression and ACE2 binding data (https://github.com/jbloomlab/SARS2-RBD-escape-calc; https://github.com/jbloomlab/SARS2_RBD_Ab_escape_maps; https://github.com/tstarrlab/SARS-CoV-2-RBD_DMS_Omicron-XBB-BQ; https://github.com/tstarrlab/SARS-CoV-2-RBD_DMS_Omicron-EG5-FLip-BA286)[7]; (3) serum and monoclonal antibody escape profiles including neutralization escape data from XBB.1.5 and Delta (https://github.com/dms-vep/SARS-CoV-2_Delta_spike_DMS_REGN10933; https://github.com/dms-vep/SARS-CoV-2_XBB.1.5_spike_DMS)[9]; and (4) large-scale monoclonal antibody escape data (https://github.com/yunlongcaolab/convergent_RBD_evolution; https://github.com/yunlongcaolab/SARS-CoV-2-reinfection-DMS)[31,32,48–50]. These datasets were retrieved from published GitHub repositories or internal laboratory directories. Raw DMS data were formatted into consistent tabular structures and converted into structured arrays aligned with the reference Spike protein sequence for downstream model input. Each DMS array was indexed along 4–5 axes depending on the dataset: antigen (variant background), feature (phenotype), site (aligned residue index), mutant amino acid, and optionally, antibody identity for antibody escape array. Missing values were filled with NaNs.

To avoid data leakage and enable dynamic, time-aware inference, we indexed DMS features by their antigen source and recorded their earliest availability dates. During training, we applied temporal masking to ensure that only features experimentally available before the target prediction date (t_0) were used. For antibody escape features, we aggregated epitope-level escape scores by antibody cluster and averaged values within each group to mitigate sampling noise. We re-clustered the antibody escape profiles into 56 distinct groups to obtain a more fine-grained representation of the underlying epitope landscape. Only antibodies with previous immunological exposure (that is, sampling time ≤t_0) were retained for analysis. Each input sequence was one-hot encoded and scanned against the aligned DMS matrix to produce a position-specific, feature-aware vector representation.

Model architecture

The overall architecture of DeepCoV is shown in Extended Data Fig. 1, and its aim is to predict the future proportion of certain RBD or Spike. To represent the dynamic immune pressures, MSA of target and top-circulating strains at t_0, and the corresponding sequence counts for certain dates and locations were collected from the GISAID database. Mutation phenotype of evasion (antibody escape, sera escape), fitness (ACE2 binding, spike-mediated entry) as well as expression from deep mutation scanning were generated following refs. 7,31,32,48–50.

First, we adopted pretrained ESM-MSA-1b with frozen parameters to extract amino-acid-level embeddings of RBD or Spike of SARS-CoV-2 target and representative variants. The 3-day-gap-level proportions of variants generated from GISAID, with location information implicitly represented, were fed into LSTM. Subsequently, the MSA embedding and proportion embedding were concatenated together and processed by the self-trained DeepCoV axial-attention module[22]. In this module, a three-layer row-wise and column-wise Transformer was adopted to build local dependencies of the residues and variants. The refined embeddings of the target were then split for further DMS information integration. As for the DMS data, we expressed the sequence-level

score of antibody escape for each epitope by max pooling on the site dimension and reference dimension, and mean across sites and references for other DMS features. To go a step further, the MSA-proportion embedding, concatenated with the DMS features one by one, flows into a Transformer encoder module for feature coupling. The purpose of using a transformer here is to emphasize the impact of protein-level embedding. At the final step, linear layers were applied to generate the proportion value on the target day, or LSTM for the proportion values during a period.

The architecture and algorithms involved in each module are detailed as follows.

Sequence embedding module. To represent protein sequences effectively, we employed state-of-the-art protein language models that capture evolutionary constraints and contextual amino acid relationships. We supported two pretrained models:

$$E_{\text{seq}} = \text{SeqEncoder}(X) \in \mathbb{R}^{B\times(1+N_{\text{bg}})\times L\times D} \tag{1}$$

where E_{seq} is the output sequence MSA embeddings, X is the input sequence tokens (containing the target and background sequences), B is the batch size, N_{bg} is the number of background sequences, L is the length of aligned sequences (one target sequence and N_{bg} background sequences) and D is the embedding dimension. The main model employs ESM-MSA-1b ($D = 768$) as the sequence encoder to exploit MSA information, while ablation variants optionally substitute ESM-2 to assess the impact of evolutionary modelling on downstream performance.

Background sequence frequency encoder. This encoder processes temporal variant frequency data to capture evolutionary dynamics. This component enables the model to learn patterns in how variants rise and fall in prevalence over time. The encoder employs an LSTM architecture that processes time-series data of variant frequencies:

$$h_t, c_t = \text{LSTM}(x_t, h_{t-1}, c_{t-1}) \tag{2}$$

where $x_t \in \mathbb{R}^{B\times(1+N_{\text{bg}})}$ represents the proportion of each target or background sequence at time t, and h_t, c_t are hidden and cell states. To extract a fixed-size representation of the temporal dynamics, the model utilizes the final hidden state as the background frequency embeddings:

$$E_{\text{bg}} = h_{\text{last}} \tag{3}$$

where h_{last} is the output of the LSTM at the final time step. This encoding transforms the input sequence of length (number of time steps) L_{bg} into a high-dimensional feature vector, which serves as the temporal representation of the variant.

Following the initial encoding, the AxialTransformer module from MSA Transformer is employed to fuse the sequence embeddings E_{seq} and background embeddings E_{bg}[24]. The representation of the target sequence is then extracted from the resulting output for downstream computations.

$$E_{\text{target}} = \text{AxialTransformer}(\text{concat}(E_{\text{seq}}, E_{\text{bg}})) \tag{4}$$

Deep mutational scanning encoder. The incorporation of experimental DMS data provides direct measurements of how mutations affect viral properties such as antibody escape, binding affinity and expression levels. The DMS encoder integrates this experimental data with sequence representations. First, we processed sequence embeddings to create a common representation space:

$$E_{\text{combined}} = [E_{\text{target}}; E_{\text{ref}}] \tag{5}$$

$$E_{\text{processed}} = \text{Normalize}(\text{Linear}(E_{\text{combined}})) \tag{6}$$

$$E_{\text{target}^{\text{proc}}}, E_{\text{ref}^{\text{proc}}} = \text{Split}(E_{\text{processed}}, [1, N_{\text{bg}}]) \tag{7}$$

where E_{target} is the target sequence embedding, E_{ref} represents the reference sequence embeddings, E_{combined} is their concatenation, $E_{\text{processed}}$ is the processed embedding after linear projection and normalization, $E_{\text{target}^{\text{proc}}}$ is the processed target embedding and $E_{\text{ref}^{\text{proc}}}$ is the processed reference embedding.

Next, we computed similarity scores between target and reference sequences to create attention weights. This attention mechanism allows the model to focus on reference sequences most similar to the target:

$$r_{i,j,l} = \sum_{d=1}^{D_{\text{seq}}} E_{\text{target}^{\text{proc}}}[i,1,l,d]\, E_{\text{ref}^{\text{proc}}}[i,j,l,d] \tag{8}$$

$$s_{i,j,l} = r_{i,j,l} M_{\text{delay}}[i,j,l] \tag{9}$$

$$w_{i,j,l} = \frac{\exp(s_{i,j,l})}{\sum_{j'=1}^{N_{\text{ref}}} \exp(s_{i,j',l})} \tag{10}$$

where $r_{i,j,l}$ is the dot product similarity between the target sequence (index is 1) and reference sequence j at position l for batch item i, $M_{\text{delay}}[i,j,l]$ is a binary mask that prevents information leakage from future timepoints, $s_{i,j,l}$ is the masked features and $w_{i,j,l}$ is the attention weight derived from softmax normalization across reference sequences. For the standard DMS processing without antibody clusters, we proceed as follows:

$$X_{\text{dms}} \in \mathbb{R}^{B\times N_{\text{ref}}\times L\times 21} \tag{11}$$

$$D_{\text{emb}} = \text{Linear}(X_{\text{dms}}) \in \mathbb{R}^{B\times N_{\text{ref}}\times L\times D_{\text{dms}}} \tag{12}$$

$$D_{\text{trans}} = \text{TransformerEncoder}(D_{\text{emb}}) \tag{13}$$

where X_{dms} is the DMS data with 21 features per position (representing 20 amino acid types and deletion), D_{emb} is the embedded DMS feature and D_{trans} is the transformer-encoded DMS feature.

Finally, we applied the attention weights to aggregate DMS features across reference sequences:

$$D_{\text{weighted}} = \sum_{j=1}^{N_{\text{ref}}} w_{i,j,l} D_{\text{trans}}[i,j,l] \tag{14}$$

$$Y_{\text{dms}} = \text{LayerNorm}(D_{\text{weighted}}) \in \mathbb{R}^{B\times L\times D_{\text{dms}}} \tag{15}$$

where D_{weighted} represents the attention-weighted DMS features and Y_{dms} is the final normalized DMS representation that will be integrated with other features in the model.

This approach enables the model to selectively incorporate DMS information from the most relevant reference sequences, creating a robust representation of mutation effects that informs evolutionary prediction.

Feature integration and output layer. We integrated evolutionary predictions with DMS features through a series of fusion layers. For models with antibody escape data:

$$F_{\text{dms}} = \text{LayerNorm}(\text{Linear}([F_{\text{target}}; Y_{\text{dms}^{\text{ab}}}])) \tag{16}$$

Additional DMS features were integrated sequentially:

$$F_{\text{dms}^{i+1}} = \text{LayerNorm}(\text{Linear}([F_{\text{dms}^{i}}; Y_{\text{dms}^{i}}])) \tag{17}$$

After integrating the DMS features with evolutionary information, the features were further fused through a Transformer Encoder layer. This layer utilizes a self-attention mechanism to capture the long-range dependencies and contextual relationships between amino acid residues. Finally, the representation of the target sequence was derived from the first-position class (CLS) token.

For temporal trajectory prediction, we implemented an LSTM-based output layer that generates predictions for multiple future time points:

$$X_{\text{time}^t} = F_{\text{final}} + \text{Embedding}(t), t \in 0, 1, \ldots, T-1 \quad (18)$$

The time embedding allows the model to distinguish between different prediction horizons. The LSTM processes these time-embedded features:

$$h_t, c_t = \text{LSTM}(X_{\text{time}^t}, h_{t-1}, c_{t-1}) \quad (19)$$

$$p = \sigma(\text{Linear}(h_{T-1})) \quad (20)$$

where F_{target} is the output from the evolutionary prediction module for the target sequence, $Y_{\text{dms}^{\text{ab}}}$ is the antibody escape DMS representation, F_{dms} is the fused representation after integrating evolutionary and antibody escape features, Y_{dms^i} is the representation of the ith DMS feature type, F_{dms^i} is the fused representation after integrating the ith DMS feature, F_{final} is the sequence-level representation extracted from the first token position, X_{time^t} is the time-embedded representation for timepoint t, h_t and c_t are the hidden and cell states of the LSTM at time step t, and p is the predicted prevalence probability. In addition, an alternative optional output module utilizes a two-layer feed-forward neural network to predict prevalence probability p, where the input features are first compressed using a linear layer with Rectified Linear Unit (ReLU) activation, followed by a final linear projection and a sigmoid function to squash the output into a (0,1) probability range.

This architecture enables the model to predict variant prevalence trajectories over time, capturing both immediate and longer-term evolutionary dynamics.

Training and optimization

All models were trained using the AdamW optimizer with a weight decay of 10^{-2} and an initial learning rate of 10^{-4}. A linear warm-up schedule was applied for the first 300 training steps, followed by a linear decay (Table 1). Mixed precision (bfloat16 or float16) was employed under the PyTorch framework on NVIDIA A100 GPUs.

The loss function was a log-transformed, sample-weighted MSE defined as:

$$L_{\text{reg}} = \sum_{i=1}^{N} w_i^{\text{mask}} w_i^{\text{label}} (\log(100\hat{y_i} + 1) - \log(100y_i + 1))^2 \quad (21)$$

where $\hat{y_i}$ is the predicted future prevalence, y_i is the observed ground truth at time t_1 ($t_1 = t_0 + 30$ d), and w_i^{mask} is a weight derived from sequence sampling coverage masking. The mask matrix consists of binary indicators denoting whether the ground truth value y_i is valid (that is, total number of isolates at t_1 is above a defined threshold, for example, 100 in this study). w_i^{label} is determined by the label y_i, to upweight the dominant variants, reducing class imbalance bias during training. To account for the heavy-tailed distribution of the target ratio and the overrepresentation of near-zero values, we applied a logarithmic transformation of the form $\log(100y_i + 1)$ to the target variable. This transformation mitigates the dominance of extremely small values during training and facilitates more balanced gradient propagation throughout optimization.

Extended models

Updated JN.1-era prediction. For the updated JN.1 prediction, the main model—originally trained on data collected before 1 October 2023—was retained, while the test dataset was expanded to include sequences submitted up to 16 May 2025. Unique RBDs were renamed according to their mutations relative to reference lineages: WT, Alpha, Beta, Delta, Gamma, Eta, BA.1, BA.2, BA.5, BF.7, BQ.1.1, XBB, XBB.1.5, EG.5, HK.3, 'FLip' (XBB + S486P + L455F + F456L), BA.2.86, JN.1, KP.2, KP.3, XEC, LF.7, LP.8, NB.1, NB.1.8.1, XFG and XFH. The major-strain-specific test set comprised RBDs from JN.1, KP.2, KP.3, LF.7, LP.8 and NB.1.8.1.

Spike model. For each unique spike sequence, a cluster index and a cluster name were assigned for further usage. Unique spikes were renamed according to their mutations relative to the their parental lineage references: WT, Alpha, Beta, Delta, Gamma, Eta, BA.1, BA.2, BA.5, BF.7, BQ.1.1, XBB, XBB.1.5, EG.5, HK.3, 'FLip' (XBB + S486P + L455F + F456L), BA.2.86, JN.1, KP.2 and KP.3. For the sequencing embedding, considering the input restrictions of the ESM-MSA-1b, only the first 1,023 amino acids of the spike protein were included. Considering the relatively minor contribution of the tail of the C-terminal region in the evolution of SARS-CoV-2, it is an acceptable compromise. The other data processing and model architectures are consistent with those of the main model in all other aspects (see Table 2 for hyperparameter details).

Benchmarking approaches

Growth advantage estimation. The algorithm for calculating growth advantage was adapted from ref. 11, and the daily sequence data were sourced from the GISAID database. Specifically, a logistic regression model was employed to fit the daily frequency of samples for the concerned RBD cluster to estimate the logistic growth rate a and the midpoint t_0 of the sigmoid curve. The growth advantage was defined as $e^{a \times g} - 1$, where g, the generation time, equals 7 days, and a represents the growth rate derived from the logistic model fitting. Confidence intervals were computed with $\alpha = 0.95$.

EVEscape benchmarking. The algorithm for calculating growth advantage was adapted from ref. 14. We first computed the mutations of all RBD sequences in the full test set relative to the wild-type reference. Following the approach described in the original EVEscape publication, we aggregated EVEscape scores based on these relative mutations to obtain a composite score for each RBD sequence.

E2VD benchmarking. The E2VD models were retrained using the publicly available ESM-2 model as the pretrained backbone[21]. Candidate unique RBD sequences served as input to the model, with outputs generated on a 5-fold cross-validated test set. We benchmarked the E2VD framework by aggregating predictions from its three submodules: ACE2 binding affinity, viral expression efficiency and antibody escape potential. Model outputs were generated on a 5-fold cross-validated test set (JN.1-era RBD variants) and averaged across folds for the binding and expression tasks. Escape scores were computed from the last prediction run of the optimization cycle. Following the original study, the model was tasked with predicting escape scores against the BD57-0129 antibody for immune escape evaluation. To prioritize variants with potential fitness and immune evasion advantages, we applied an asymmetric scoring scheme informed by previous biological knowledge: deviations in expression and ACE2 binding were penalized only when falling below functional thresholds, whereas antibody escape was positively weighted when exceeding a permissive cut-off. The final E2VD score was computed as the sum of exponentiated deviations from these empirically defined thresholds:

$$\text{E2VD} = e^{\min(\text{expr}-0.7,0)} + e^{\min(\text{blind}-0.25,0)} + e^{\max(\text{escape}-0.5,0)} \quad (22)$$

Multinomial logistic regression. To provide a comparative baseline for variant frequency forecasting, we implemented an MLR model following

the framework established in ref. 33. MLR models the observed counts of competing SARS-CoV-2 variants over time as draws from a multinomial distribution, with variant frequencies evolving according to variant-specific growth advantage parameters under a fixed-fitness assumption. The vector of linear predictors was exponentiated and normalized to ensure that predicted frequencies sum to unity at each timepoint. The model was fit by maximizing the multinomial likelihood of the observed sequence counts stratified by analysis date, using available genomic surveillance data up to that date. Predicted frequencies at future timepoints (including 30-day forecasts) were obtained by extrapolating the time-dependent linear predictors based on the fitted parameters and normalizing them to the simplex. The 180-day frequency trajectories, as used by DeepCoV, of all RBD clusters were included.

Benchmarking settings

Timeliness of major variant detection by different predictive models. To assess the timeliness of variant prioritization across different computational frameworks, we tracked the earliest timepoints when major RBD variants (JN.1, KP.2 and KP.3) were flagged as high risk by multiple scoring models, including DeepCoV, MLR, growth advantage, EVEscape and E2VD. For each method, we determined the earliest date a given strain entered the top-N ranking (for example, top 10 or top 3) based on model-specific scores. Specifically, rankings of DeepCoV, MLR and growth advantage were derived from the model-predicted target proportion using data available at date t_0. To ensure stability and fairness, growth-advantage predictions were smoothed before ranking using a conservative procedure in which the minimum value within a 7-day window was taken. For EVEscape and E2VD, strain scores were ranked within a ±1-month temporal window centred at each submission date to mimic realistic evaluation scenarios. If a variant was not predicted as a top-ranked candidate before its observed prevalence formally reached 20%, the method was considered to have failed in predicting that variant.

Benchmarking of top-k variant ranking predictions. To enable a fair comparison of different predictive methods for identifying dominant SARS-CoV-2 RBD variants during the JN.1 era, we implemented a dynamic retrospective top-k ranking comparison against several baseline methods. The evaluation was conducted on temporally stratified test data spanning October 2023 to September 2024. We first defined ground-truth dominant variants (top-n_truth) based on their observed proportions at prediction horizon t_1, using a 30-day window of metadata (submit_date $\in [t_0-30, t_0]$) to restrict candidate RBDs to those actively circulating at the time of prediction. For each timepoint t_0, candidate variants were ranked by predicted prevalence using DeepCoV, MLR, growth advantage, and baseline functional scores from EVEscape and E2VD. We evaluated predictions using two criteria:

Prediction success rate. This was used for cases where the number of true dominant variants is equal to 1 and was defined as the proportion of timepoints where the top-ranked predicted variant exactly matched the observed dominant variant at t_1, based on its prevalence.

Mean overlap ratio. This was used for cases where the number of true dominant variants is greater than 1 and was calculated as the temporal average of the Jaccard index across all evaluation points, quantifying the overlap between predicted and true dominant sets. The top-k predictions were computed for a range of values k. Evaluation was conducted independently for each method across all timepoints. XBB.1.5, EG.5, HK.3, BA.2.86, JN.1, KP.2 and KP.3-related variants were included in this evaluation.

Growth trajectory reconstruction and prevalence regional heterogeneity analysis

Growth trajectory reconstruction. To evaluate the temporal dynamics and predictive accuracy of our model across key SARS-CoV-2 lineages, we visualized the predicted and observed relative abundances of major RBD or spike variants over time. For each lineage, model predictions of future proportions (output of target ratio at t_1) were aligned to their corresponding start dates (t_0) and binned biweekly. Observed contemporaneous proportions (target ratio at t_0) were used for reference. Line plots were generated to display predicted trajectories (dashed lines), observed proportions (solid lines) and prediction uncertainty (±1 s.d. as shaded ribbons) across locations. Variants were labelled using standardized nomenclature derived from mutational mappings. We computed temporal prediction error metrics, including RMSE and MAE, to quantify model fidelity across timepoints. These metrics were evaluated both on major variants and the full test set.

Prevalence regional heterogeneity analysis. For several well-known variants exhibiting regional differences in prevalence, we separately modelled their growth trajectories using representative data from individual continents and visualized their predicted dynamics. In addition, to assess the spatiotemporal dynamics of variant emergence, we evaluated the relative ranking of four representative RBD variants across major geographic regions. The test dataset consisted of model outputs spanning from 1 September 2022. For each variant, predictions were grouped by date (t_0) and location, and variants were ranked according to their predicted prevalence at the forecast horizon (t_1).

In silico mutational hotspots scanning

Pseudo single-residue mutational dataset generation. We constructed comprehensive libraries of single amino acid substitutions on representative reference strains. Aligned regions corresponding to the RBD (residues 331–531) or NTD (positions 14–305) were extracted from MSAs of spike protein sequences. For each position in the aligned region and for each reference strain (for example, XBB.1.5, JN.1, KP.2), all 19 possible single amino acid substitutions were introduced. In the region of NTD, extra deletions were also introduced for each position. To simulate synthetic surveillance records, we selected a defined time window (for example, 1 May to 1 June 2023, for XBB.1.5) and globally assigned prevalence value to each mutant based on their empirical prevalence values; unobserved mutants were assigned a prevalence of zero. The combined dataset was used for downstream mutational hotspot scanning and predictive evaluation.

In silico mutational scanning and visualization of variant-specific fitness hotspots. In silico single-mutant scanning of the SARS-CoV-2 RBD or NTD region was conducted on various emerging variant backbones (for example, XBB.1.5, KP.3, LF.7). For reference strains JN.1, KP.3 and LF.7, all single amino acid substitutions were evaluated using a trained JN.1-era predictive model, yielding a fitness score for each synthetic mutant. For reference strains XBB, XBB.1.5 and EG.5, residue substitutions were evaluated using a trained XBB-era predictive model. For each mutation, a normalized contribution score was calculated by subtracting the amino acid-specific global average score over time to control for residue bias. Only mutations with a positive differential score were retained as candidates for fitness advantage. To summarize mutation-level signals into position-level insights, we computed average contribution scores across all substitutions at each site and visualized them using a smoothed line plot with positional annotations. In addition, we generated amino acid logo plots on the top-N (for example, 10) highest-scoring positions, where the height of each letter corresponded to the magnitude of the fitness contribution for that substitution.

Ablation studies

To assess the contribution of different components to the model's predictive performance, we conducted systematic ablation studies by removing specific functional modules:

- Sequence encoder ablation (ESM-2): To evaluate the effect of the MSA-based encoder, we replaced ESM-MSA-1b with the ESM-2 (150 M) model for sequence embedding. All other architectural components were kept unchanged.
- No-DMS model: All DMS-derived features were removed, and the model was trained using only the amino acid sequences and associated epidemiology data. Two feed-forward layers were used in place of the DMS module to achieve the same dimensional transformation.
- No evolutionary–epidemiological context model: Only the target strain and its associated 180-day historical prevalence data were used. No background sequences were included. Sequence features were encoded using the ESM-2 model, and historical prevalence signals were integrated via an LSTM, followed by a Transformer-based feature aggregator.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The SARS-CoV-2 genome sequences and associated metadata analysed in this study were obtained from the GISAID EpiCoV database (www.gisaid.org). Access to GISAID data is subject to their data access agreement, and the collection details for the sequences used in this study are provided in Supplementary Methods and are also available in GitHub at https://github.com/yunlongcaolab/DeepCoV (ref. 51). The deep mutational scanning datasets for single mutations were obtained from published datasets in public repositories (https://github.com/tstarrlab/SARS-CoV-2-RBD_DMS_Omicron-EG5-FLip-BA286; https://github.com/tstarrlab/SARS-CoV-2-RBD_DMS_Omicron-XBB-BQ; https://github.com/jbloomlab/SARS2_RBD_Ab_escape_maps; https://github.com/dms-vep/SARS-CoV-2_Omicron_BA.2_spike_ACE2_binding; https://github.com/dms-vep/SARS-CoV-2_XBB.1.5_spike_DMS; https://github.com/dms-vep/SARS-CoV-2_Delta_spike_DMS_REGN10933; https://github.com/dms-vep/SARS-CoV-2_Omicron_BA.1_spike_DMS_mAbs; https://github.com/jbloomlab/SARS2-RBD-escape-calc; https://github.com/yunlongcaolab/convergent_RBD_evolution; https://github.com/yunlongcaolab/SARS-CoV-2-reinfection-DMS)[7,9,31,32,48–50]. The raw sequence IDs and all derived datasets including training/validation/test sets are available via Zenodo at https://doi.org/10.5281/zenodo.18392647 (ref. 52). Source data are provided with this paper.

Code availability

All scripts used for data preprocessing, model training and evaluation are publicly available in GitHub at https://github.com/yunlongcaolab/DeepCoV (ref. 51).

References

1. Markov, P. V. et al. The evolution of SARS-CoV-2. *Nat. Rev. Microbiol.* **21**, 361–379 (2023).
2. Roemer, C. et al. SARS-CoV-2 evolution in the Omicron era. *Nat. Microbiol.* **8**, 1952–1959 (2023).
3. Raharinirina, N. A. et al. SARS-CoV-2 evolution on a dynamic immune landscape. *Nature* **639**, 196–204 (2025).
4. 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.e14 (2022).
5. Taylor, A. L. & Starr, T. N. Deep mutational scanning of SARS-CoV-2 Omicron BA.2.86 and epistatic emergence of the KP.3 variant. *Virus Evol.* **10**, veae067 (2024).
6. Starr, T. N. et al. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell* **182**, 1295–1310.e20 (2020).
7. Taylor, A. L. & Starr, T. N. Deep mutational scans of XBB.1.5 and BQ.1.1 reveal ongoing epistatic drift during SARS-CoV-2 evolution. *PLoS Pathog.* **19**, e1011901 (2023).
8. Dadonaite, B. et al. A pseudovirus system enables deep mutational scanning of the full SARS-CoV-2 spike. *Cell* **186**, 1263–1278.e20 (2023).
9. Dadonaite, B. et al. Spike deep mutational scanning helps predict success of SARS-CoV-2 clades. *Nature* **631**, 617–626 (2024).
10. Starr, T. N. & Thornton, J. W. Epistasis in protein evolution. *Protein Sci.* **25**, 1204–1218 (2016).
11. Chen, C. et al. Quantification of the spread of SARS-CoV-2 variant B.1.1.7 in Switzerland. *Epidemics* **37**, 100480 (2021).
12. Obermeyer, F. et al. Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness. *Science* **376**, 1327–1332 (2022).
13. Maher, M. C. et al. Predicting the mutational drivers of future SARS-CoV-2 variants of concern. *Sci. Transl. Med.* **14**, eabk3445 (2022).
14. Thadani, N. N. et al. Learning from prepandemic data to forecast viral escape. *Nature* **622**, 818–825 (2023).
15. Hie, B., Zhong, E. D., Berger, B. & Bryson, B. Learning the language of viral evolution and escape. *Science* **371**, 284–288 (2021).
16. Han, W. et al. Predicting the antigenic evolution of SARS-COV-2 with deep learning. *Nat. Commun.* **14**, 3478 (2023).
17. Ma, E. et al. A predictive language model for SARS-CoV-2 evolution. *Signal Transduct. Target. Ther.* **9**, 353 (2024).
18. Ito, J. et al. A protein language model for exploring viral fitness landscapes. *Nat. Commun.* **16**, 4236 (2025).
19. Zhou, B. et al. TEMPO: a transformer-based mutation prediction framework for SARS-CoV-2 evolution. *Comput. Biol. Med.* **152**, 106264 (2023).
20. Feng, Y. et al. CovTransformer: a transformer model for SARS-CoV-2 lineage frequency forecasting. *Virus Evol.* **10**, veae086 (2024).
21. Nie, Z. et al. A unified evolution-driven deep learning framework for virus variation driver prediction. *Nat. Mach. Intell.* **7**, 131–144 (2025).
22. Notin, P., Weitzman, R., Marks, D. S. & Gal, Y. ProteinNPT: improving protein property prediction and design with non-parametric transformers. *Adv. Neural Inf. Process. Syst.* **36**, 33621–33658 (2023).
23. Meijers, M., Ruchnewitz, D., Eberhardt, J., Luksza, M. & Lassig, M. Population immunity predicts evolutionary trajectories of SARS-CoV-2. *Cell* **186**, 5151–5164.e13 (2023).
24. Rao, R. et al. MSA transformer. In *Proc. 38th International Conference on Machine Learning* 8844–8856 (PMLR, 2021).
25. Sherstinsky, A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D* **404**, 132306 (2020).
26. Starr, T. N. et al. Shifting mutational constraints in the SARS-CoV-2 receptor-binding domain during viral evolution. *Science* **377**, 420–424 (2022).
27. Yue, C. et al. ACE2 binding and antibody evasion in enhanced transmissibility of XBB.1.5. *Lancet Infect. Dis.* **23**, 278–280 (2023).
28. Starr, T. N. et al. ACE2 binding is an ancestral and evolvable trait of sarbecoviruses. *Nature* **603**, 913–918 (2022).
29. Ma, W., Fu, H., Jian, F., Cao, Y. & Li, M. Immune evasion and ACE2 binding affinity contribute to SARS-CoV-2 evolution. *Nat. Ecol. Evol.* **7**, 1457–1466 (2023).
30. Jackson, C. B., Farzan, M., Chen, B. & Choe, H. Mechanisms of SARS-CoV-2 entry into cells. *Nat. Rev. Mol. Cell Biol.* **23**, 3–20 (2022).
31. Cao, Y. et al. Omicron escapes the majority of existing SARS-CoV-2 neutralizing antibodies. *Nature* **602**, 657–663 (2022).

32. Yisimayi, A. et al. Repeated Omicron exposures override ancestral SARS-CoV-2 immune imprinting. *Nature* **625**, 148–156 (2024).
33. Abousamra, E., Figgins, M. & Bedford, T. Fitness models provide accurate short-term forecasts of SARS-CoV-2 variant frequency. *PLoS Comput. Biol.* https://doi.org/10.1371/journal.pcbi.1012443 (2024).
34. Liu, J. et al. Enhanced immune evasion of SARS-CoV-2 variants KP.3.1.1 and XEC through N-terminal domain mutations. *Lancet Infect. Dis.* **25**, e6–e7 (2025).
35. Kosugi, Y. et al. Characteristics of the SARS-CoV-2 omicron HK.3 variant harbouring the FLip substitution. *Lancet Microbe* **5**, e313 (2024).
36. Jian, F. et al. Convergent evolution of SARS-CoV-2 XBB lineages on receptor-binding domain 455–456 synergistically enhances antibody evasion and ACE2 binding. *PLoS Pathog.* **19**, e1011868 (2023).
37. Qu, P. et al. Immune evasion, infectivity, and fusogenicity of SARS-CoV-2 BA.2.86 and FLip variants. *Cell* **187**, 585–595.e6 (2024).
38. McCallum, M. et al. N-terminal domain antigenic mapping reveals a site of vulnerability for SARS-CoV-2. *Cell* **184**, 2332–2347.e16 (2021).
39. Meng, B. et al. SARS-CoV-2 spike N-terminal domain modulates TMPRSS2-dependent viral entry and fusogenicity. *Cell Rep.* **40**, 111220 (2022).
40. Zhang, J., Xiao, T., Cai, Y. & Chen, B. Structure of SARS-CoV-2 spike protein. *Curr. Opin. Virol.* **50**, 173–182 (2021).
41. Wang, Q. et al. Antibody evasiveness of SARS-CoV-2 subvariants KP.3.1.1 and XEC. *Cell Rep.* **44**, 115543 (2025).
42. Li, P. et al. Role of glycosylation mutations at the N-terminal domain of SARS-CoV-2 XEC variant in immune evasion, cell–cell fusion, and spike stability. *J. Virol.* **99**, e0024225 (2025).
43. Kaku, Y. et al. Virological characteristics of the SARS-CoV-2 KP.3.1.1 variant. *Lancet Infect. Dis.* **24**, e609 (2024).
44. Niu, X. et al. IGHV3-53 antibody abundance drives divergent SARS-CoV-2 immune imprinting. Preprint at *bioRxiv* https://doi.org/10.64898/2025.12.18.694353 (2025).
45. Guo, C. et al. Antigenic and virological characteristics of SARS-CoV-2 variants BA.3.2, XFG, and NB.1.8.1. *Lancet Infect. Dis.* **25**, e374–e377 (2025).
46. Shu, Y. & McCauley, J. GISAID: global initiative on sharing all influenza data – from vision to reality. *Euro Surveill.* **22**, 30494 (2017).
47. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
48. Jian, F. et al. Evolving antibody response to SARS-CoV-2 antigenic shift from XBB to JN.1. *Nature* **637**, 921–929 (2025).
49. Cao, Y. et al. BA.2.12.1, BA.4 and BA.5 escape antibodies elicited by Omicron infection. *Nature* **608**, 593–602 (2022).
50. Cao, Y. et al. Imprinted SARS-CoV-2 humoral immunity induces convergent Omicron RBD evolution. *Nature* **614**, 521–529 (2023).
51. Yang, S. et al. DeepCoV: DMS-empowered evolution prediction of coronaVirus. *GitHub* https://github.com/yunlongcaolab/DeepCoV (2026).
52. Yang, S., Luo, X., Luo, J. & Jian, F. A deep mutational scanning-informed protein language model predicts SARS-CoV-2 evolution dynamics with spatiotemporal resolution. *Zenodo* https://doi.org/10.5281/zenodo.18392647 (2026).

Acknowledgements

We thank J. D. Bloom and T. N. Starr for making their DMS libraries and data publicly available; J. Wang, Z. Bian and all supporting technicians for their contributions in dataset preparation and preprocessing; the authors and submitting laboratories responsible for generating and sharing the SARS-CoV-2 sequence data via GISAID; and the broader scientific community for their continued efforts in SARS-CoV-2 genomic surveillance and for valuable discussions. F.J. was supported by the Boya Postdoctoral Fellowship Program of Peking University and the Postdoctoral Fellowship Program of China Postdoctoral Science Foundation under Grant Number GZC20250980. Y.C. was financially supported by the National Natural Science Foundation of China (32222030) and Changping Laboratory (2025D0401).

Author contributions

Y.C. designed and supervised the study. S.Y., F.J. and Y.C. wrote the paper. F.J., X.L., S.Y. and J.L. conceived of the computational model architecture. S.Y. and X.L. performed the model training and optimization. S.Y. and J.L. conducted the data preprocessing and curated the experimental datasets. S.Y. executed the comprehensive benchmark evaluations and performed temporal–geospatial analyses.

Competing interests

Y.C. is the inventor of provisional patent applications for anti-SARS-CoV-2 mAbs (China Patent application no. CN202210131235.9). Y.C. is the founder of Singlomics Biopharmaceuticals. The other authors declare no competing interests.

Additional information

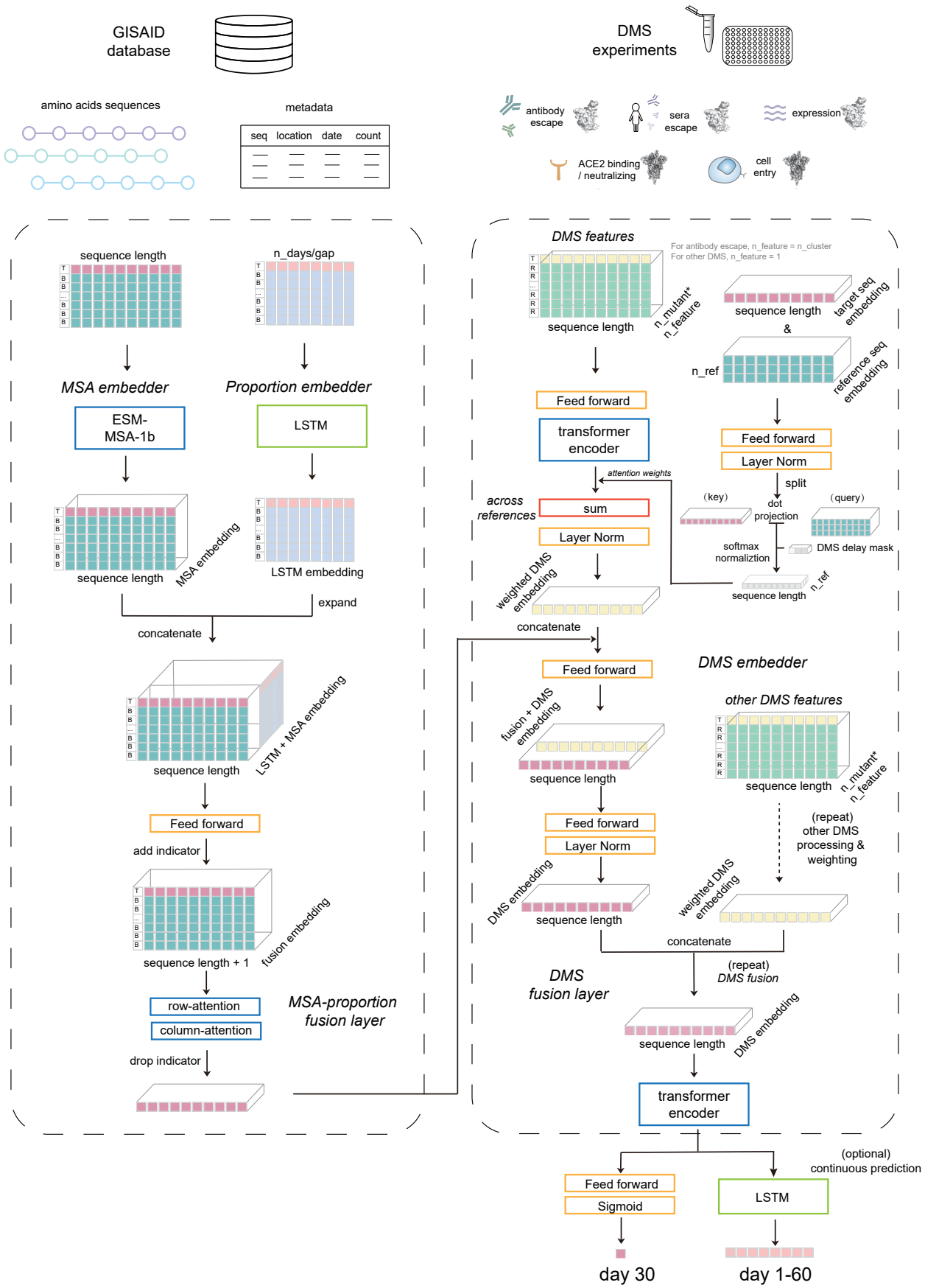
Extended data is available for this paper at https://doi.org/10.1038/s41564-026-02377-5.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41564-026-02377-5.

Correspondence and requests for materials should be addressed to Fanchong Jian or Yunlong Cao.

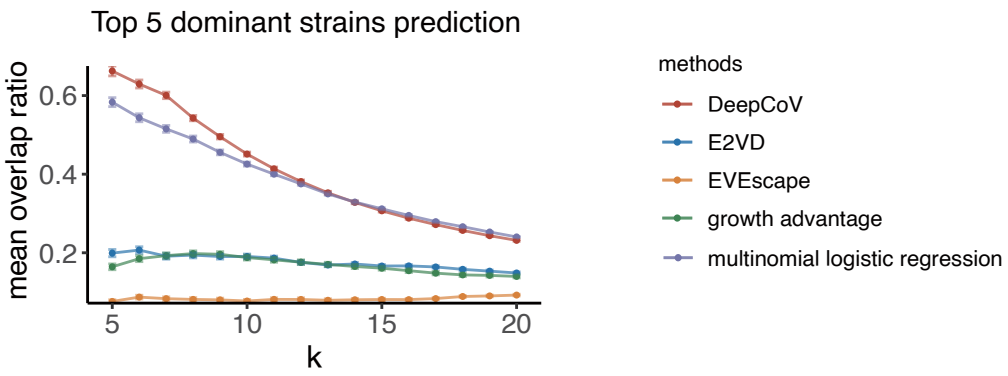
Peer review information *Nature Microbiology* thanks Paul Moss and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s41564-026-02377-5



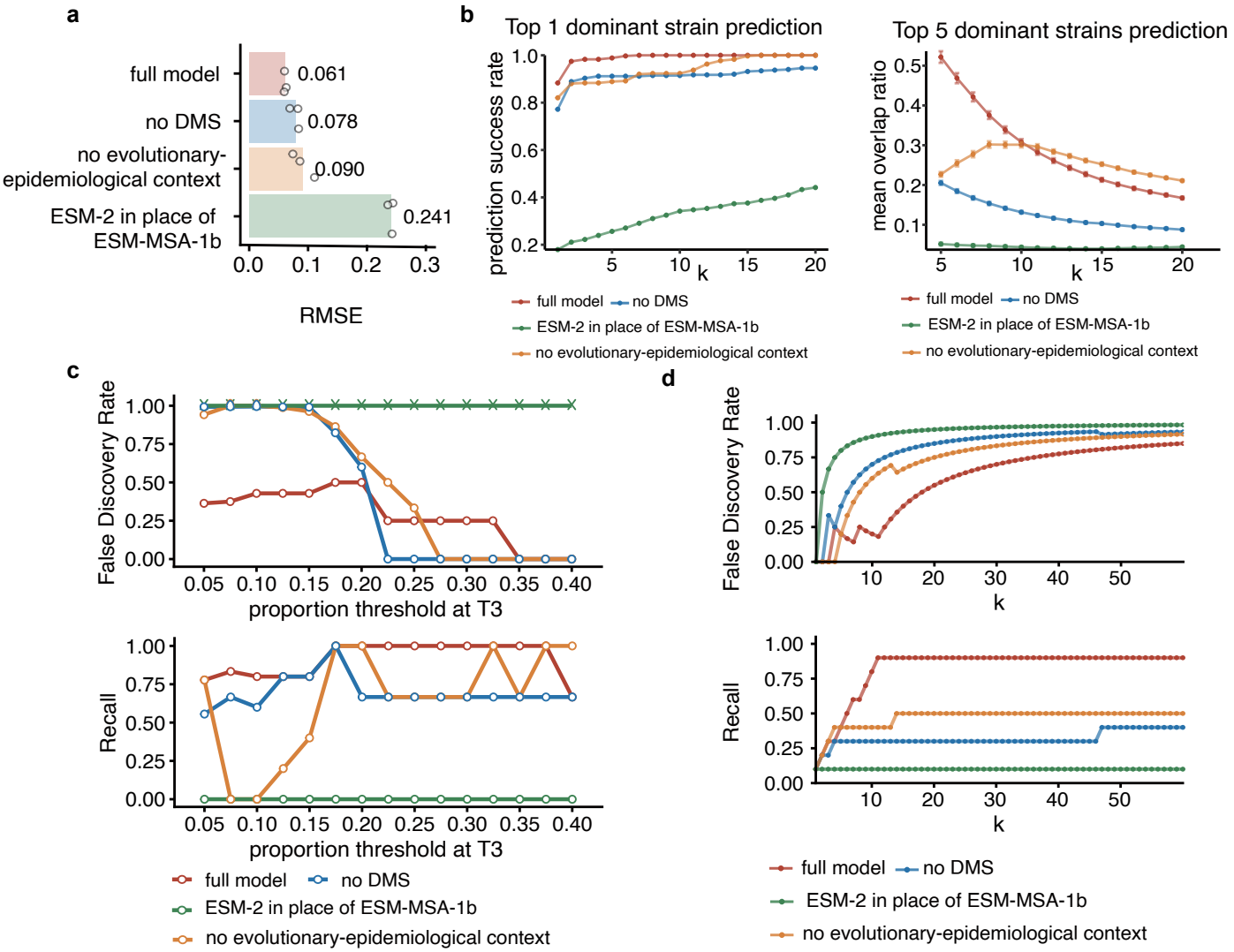
Extended Data Fig. 1 | Detailed schematic of model architecture. Schematic illustrating sequence encoding, background prevalence embedding, axial-attention–based feature fusion, DMS integration, and output layers. Full architectural and implementation details are provided in the Supplementary Methods.

https://doi.org/10.1038/s41564-026-02377-5



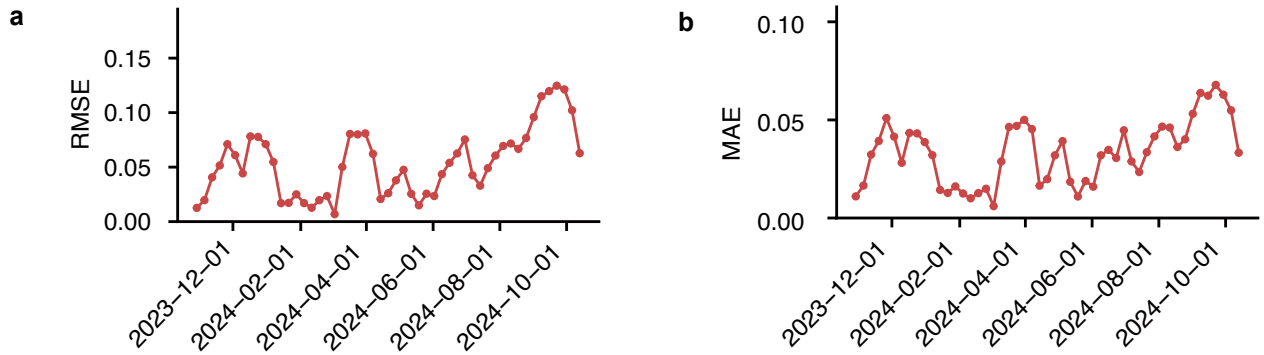
Extended Data Fig. 2 | Comprehensive evaluation of dominant variant prediction accuracy and temporal dynamics. Mean Jaccard overlap ratios between predicted and observed sets of dominant SARS-CoV-2 RBD variants across varying values of k (number of top-ranked predictions considered) for top 5 ground-truth variants. Error bars denote ±1 s.e.m. across evaluation time points (n = 351). Each time point represents an independent prediction analysis on distinct variant datasets.

https://doi.org/10.1038/s41564-026-02377-5



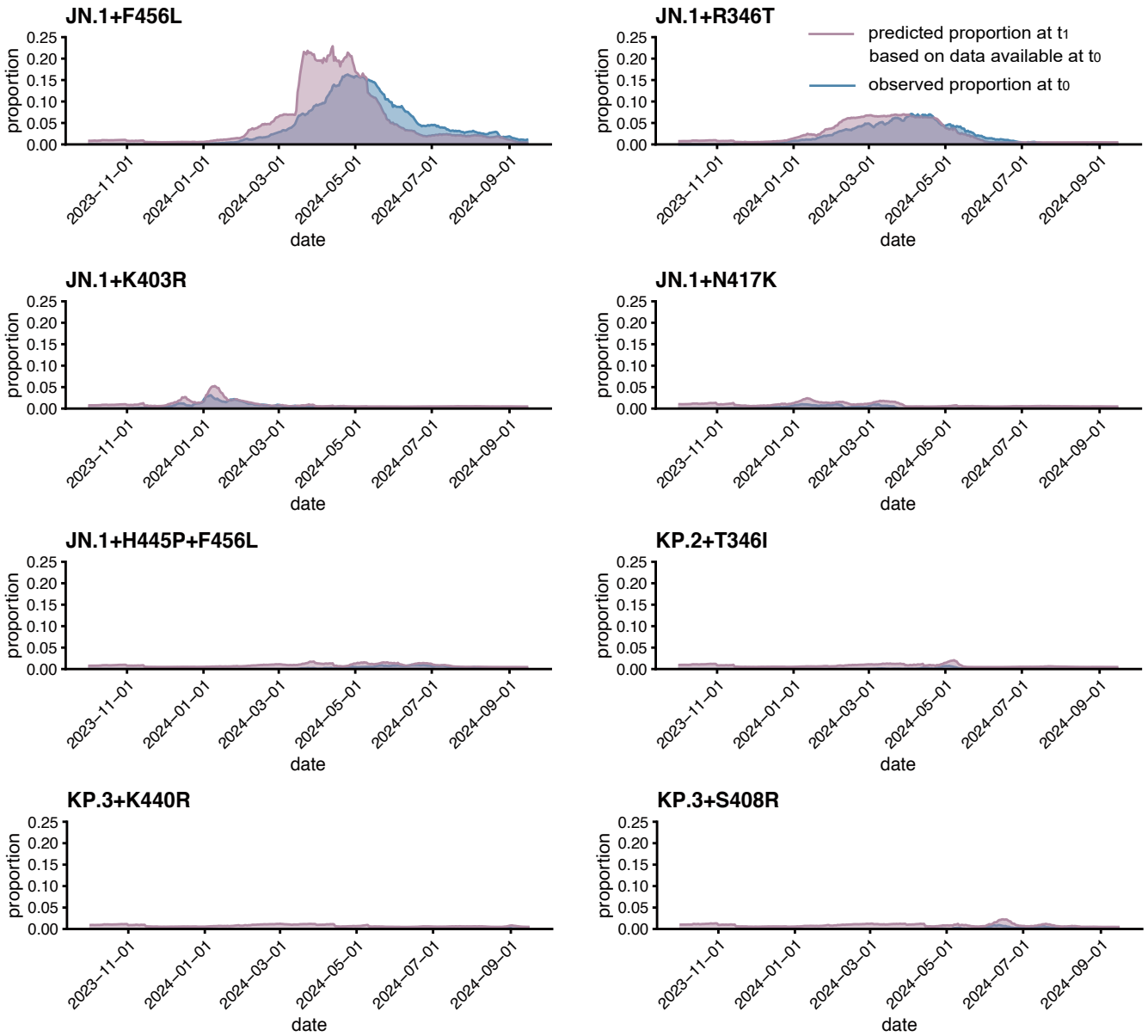
Extended Data Fig. 3 | Ablation study evaluating the contribution of individual components to DeepCoV performance. a, Comparative RMSE across the full model and three ablated models: (i) removal of immune background features, (ii) exclusion of DMS data, and (iii) replacement of evolutionary features with ESM-2 embeddings.**b**, Top-k prediction performance over time for the ablated models. Success rate is reported for the top 1, and jaccard index is reported for the top 5 predicted dominant variants across varying k values (that is, the number of predicted variants considered). Error bars represent ±1 s.e.m. across all evaluation time points. **c**, Predictive metrics (false discovery rate and recall) stratified by variant prevalence thresholds across ablated models. **d**, Performance of ablated model in recovering true dominant variants across increasing top-k prediction thresholds, evaluated using recall, FDR and accuracy computed against top ten major circulating variants (JN.1, KP.2, KP.3, HK.3, JN.1 + R346T, JN.1 + F456L, HK.3 + A475V, KP.2 + G482V + K484E, KP.2 + L456V + K478T, and KP.2+ins483V + K484E).

https://doi.org/10.1038/s41564-026-02377-5



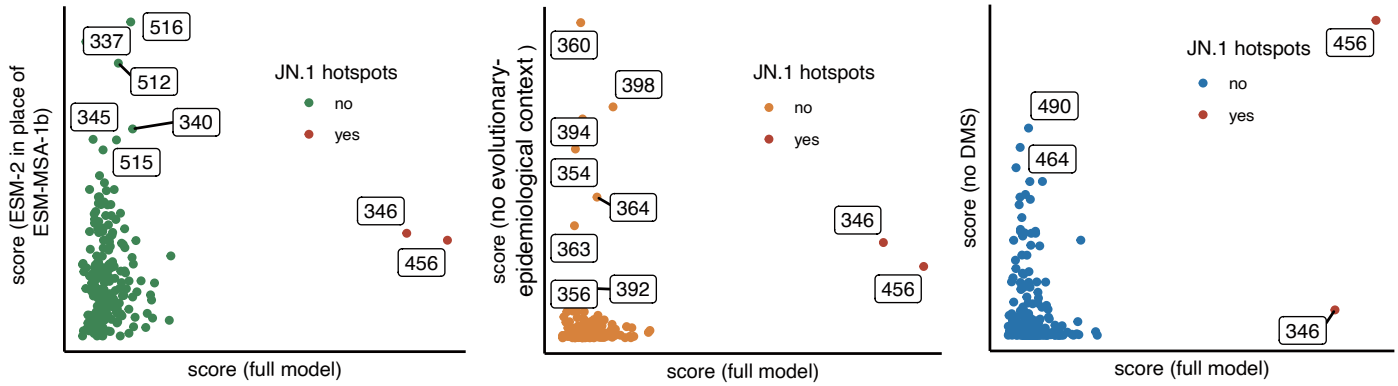
Extended Data Fig. 4 | Comprehensive evaluation of temporal dynamics prediction. a, Root mean squared error (RMSE) and **b**, mean absolute error (MAE) evaluated monthly for dominant variants after 1 October 2023 (HK.3, BA.2.86, JN.1, KP.2, KP.3).

https://doi.org/10.1038/s41564-026-02377-5



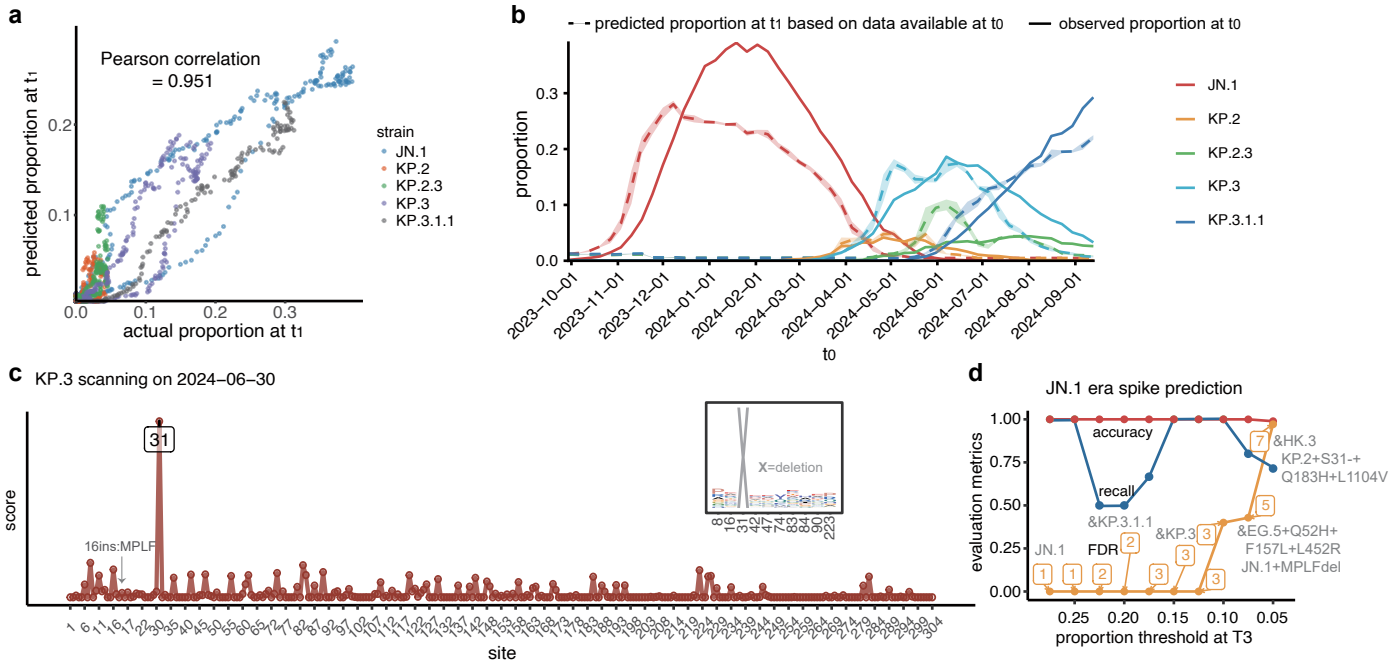
Extended Data Fig. 5 | Model performance in detecting minor SARS-CoV-2 variants. Growth trajectory reconstruction for subdominant and low-prevalence variants. Weekly aggregated predictions (lines colored purple, t_1 inferred from t_0) are compared with observed prevalence (lines colored blue, measured at t_0).

https://doi.org/10.1038/s41564-026-02377-5



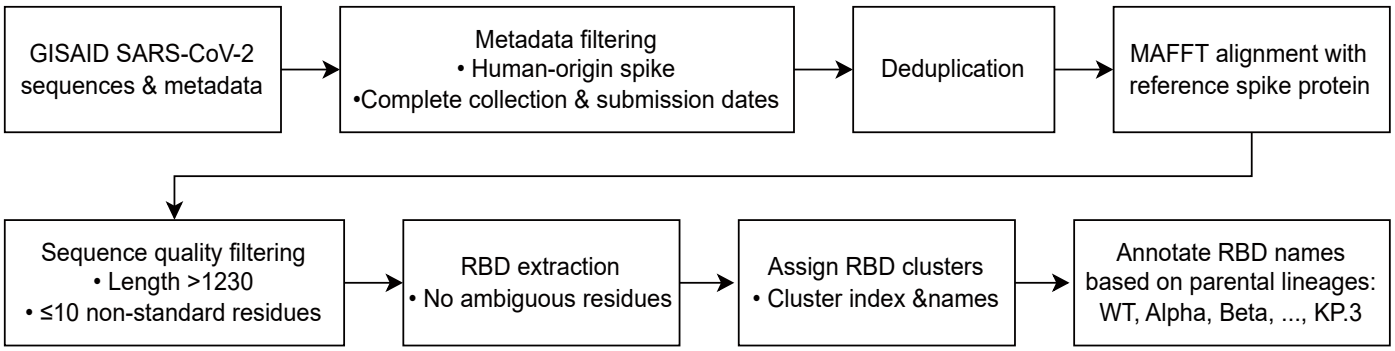
Extended Data Fig. 6 | Ablation on in silico mutational hotspots scanning. Comparison of the full and ablated models in detecting JN.1 mutational hotspots. True hotspots (positions 346 and 456) are highlighted in red. In silico deep mutational scanning of JN.1 single-point mutants at 20 March 2024 illustrates the predicted fitness landscapes under each ablation setting.

https://doi.org/10.1038/s41564-026-02377-5



Extended Data Fig. 7 | DeepCoV performance of model trained on spike protein. a, Pearson's correlation coefficient (r) of predicted versus observed variant frequencies at time t_1 using model trained on SARS-CoV-2 spike. Each point represents a variant, colored by lineage. **b**, Growth trajectory reconstruction using the renewed test set. Weekly aggregated predictions (dashed lines, t_1 inferred from t_0) are compared with observed prevalence (solid lines, measured at t_0). Shaded regions represent mean ± s.d for days in a week. **c**, Prevalence of site-specific mutations and deletions in KP.3 prior to the emergence of convergent evolution, identifying early mutational hotspots. 'X' denotes deletion mutations. **d**, Dynamic assessment of spike-based model performance across varying definitions of dominant variants based on prevalence thresholds. Accuracy, recall, and FDR are reported under each threshold setting. The number of actual dominant variants for each prevalence thresholds are labeled.

https://doi.org/10.1038/s41564-026-02377-5



Extended Data Fig. 8 | Workflow of SARS-CoV-2 sequence data preprocessing. SARS-CoV-2 genomic data and associated metadata retrieved from GISAID were subjected to a systematic, multi-stage curation pipeline to ensure data integrity and high-fidelity sequence quality.

nature portfolio

Corresponding author(s): Yunlong Cao
Fanchong Jian

Last updated by author(s): 01/28/2026

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	SARS-CoV-2 genome sequences and associated metadata were downloaded from GISAID. Full-length Spike coding sequences were extracted and aligned to Wuhan-Hu-1 (NC_045512.2) using MAFFT v7.505. Deep mutational scanning measurements for single-mutation RBD variants were obtained from the published datasets of Starr et al., Bloom et al., and Cao et al., through the public repositories. For sequence embedding, the main model used the pretrained ESM-MSA-1b model (Meta AI) with multiple-sequence alignment inputs. ESM-2 (650M) was used exclusively in ablation experiments as a drop-in replacement for the MSA-based embedding module, without further pretraining.
Data analysis	All computational analyses were performed using Python 3.12 and PyTorch 2.4 with GPU acceleration (CUDA 12.1). Spike sequences were processed using BioPython 1.78, MAFFT v7.505, NumPy 1.26.4, and Pandas 2.2.2. Evaluation metrics were conducted using scikit-learn 1.4.2. Visualization was performed using Matplotlib 3.8.4 and R packages including ggplot2 3.4.4, ggseqlogo 0.1, ggvenn 0.1.10, and ggrepel 0.9.6. All scripts used for data preprocessing, model training, and evaluation are publicly available at: https://github.com/yunlongcaolab/DeepCoV.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The SARS-CoV-2 genome sequences and associated metadata analyzed in this study were obtained from the GISAID EpiCoV database (www.gisaid.org). Access to GISAID data is subject to their data access agreement, and the collection details for the sequences used in this study are provided in Supplementary Methods and are also available at https://github.com/yunlongcaolab/DeepCoV. The raw sequence ids and all derived datasets including training/validation/test set are available can be accessed through 10.5281/zenodo.18392647.

The deep mutational scanning datasets for single mutations were obtained from the published datasets of Starr et al., Bloom et al., and Cao et al., through the public repositories (https://github.com/tstarrlab/SARS-CoV-2-RBD_DMS_Omicron-EG5-FLip-BA286; https://github.com/tstarrlab/SARS-CoV-2-RBD_DMS_Omicron-XBB-BQ; https://github.com/jbloomlab/SARS2_RBD_Ab_escape_maps;https://github.com/dms-vep/SARS-CoV-2_Omicron_BA.2_spike_ACE2_binding; https://github.com/dms-vep/SARS-CoV-2_XBB.1.5_spike_DMS; https://github.com/dms-vep/SARS-CoV-2_Delta_spike_DMS_REGN10933; https://github.com/dms-vep/SARS-CoV-2_Omicron_BA.1_spike_DMS_mAbs; https://github.com/jbloomlab/SARS2-RBD-escape-calc; https://github.com/yunlongcaolab/convergent_RBD_evolution; https://github.com/yunlongcaolab/SARS-CoV-2-reinfection-DMS) .

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	All available SARS-CoV-2 Spike sequences in GISAID were initially retrieved and filtered using metadata to retain human-derived, full-length Spike proteins with complete collection and submission dates (in YYYY-MM-DD format). Sequences with ≤10 non-standard residues and >1230 amino acids were retained to ensure basic sequence quality. To ensure adequate coverage and representativeness for model training and evaluation, we further applied predefined inclusion criteria to each instance, requiring a cumulative isolate count >100 during the 30-day prediction interval, the presence of ≥16 co-circulating clusters in the preceding 180 days, at least one day with the target-cluster proportion >0.5%, and valid geographic and sampling date metadata. Deep mutational scanning datasets included all single-amino-acid RBD mutants quantified in their respective published works. These prospectively defined rules determined the final sample size, and no additional post-hoc subsampling was performed.
Data exclusions	We did not exclude any qualifying data beyond the prospectively defined inclusion criteria above, which address data incompleteness, extremely low viral counts, or uninformative prevalence. No additional post-hoc data removal was performed.
Replication	All primary data used in this analysis was obtained from public repositories, and no experimental replication was performed.
Randomization	No experimental randomization was applicable. For model training and evaluation, we implemented a stratified evaluation design to prevent information leakage and to balance lineage and regional composition. All observations collected after 1 Oct 2023 were held out as an independent prospective test set. Data before this date were split into training and validation sets via stratified random sampling at the RBD-

	cluster level, with the constraint that a given cluster did not appear in more than one split. To ensure temporal balance, prediction start dates for training samples were drawn uniformly from five predefined time bins spanning Feb 2020–Oct 2023.
Blinding	No blinding was necessary because measurements were quantified by automated systems.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	N/A
Novel plant genotypes	N/A
Authentication	N/A