Title: Learning genotype-fitness landscape of SARS-CoV-2 in embedding space

Authors: Y. X. Liu[1]†, Y. Luo[4]†, X. Lu[1]†, H. Gao[7]†, R. K. He[7], X. Zhang[7]*, X. G. Zhang[5]*, Y. X. Li[1,2,3,6,7,8,9]*

Affiliations:

[1]Guangzhou Laboratory, Guangzhou 510005, Guangdong Province, Guangzhou, 510005, China

[2]GZMU-GIBH Joint School of Life Sciences, The Guangdong-Hong Kong-Macau Joint Laboratory for Cell Fate Regulation and Diseases, Guangzhou Medical University, Guangzhou, 511436, China

[3]Key Laboratory of Systems Health Science of Zhejiang Province, School of Life Science, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou, 310024, China

[4]National University of Singapore, 21 Lower Kent Ridge Road, Singapore 119077

[5]Mengniu Institute of Nutrition Science, Shanghai, 200126, China

[6]School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, 200240, China

[7]Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, Shanghai, 200030, China

[8]Collaborative Innovation Center for Genetics and Development, Fudan University, Shanghai, 200433, China

[9]Shanghai Institute for Biomedical and Pharmaceutical Technologies, Shanghai,

200032, China

†These authors contributed equally to this work.

*Corresponding author. Email: xzhang@picb.ac.cn (X.Z.); fxgzhang@mengniu.cn (X.G.Z.); li_yixue@gzlab.ac.cn (Y.X.L.)

Abstract: We present a Phenotype-Embedding theorem and prove that the virus fitness can be calculated under the Encoder-Decoder framework. Building on the Phenotype-Embedding theorem and a modified Transformer model, we obtain a calculable quantitative relationship between "immune escape" mutations and the virus fitness and plot a genotype-fitness landscape in the embedded space. Using only the sequence data of the SARS-CoV-2 spike protein, we accurately calculated the virus fitness and the basic reproduction number (R_0). In addition, our model can simulate viral neutral evolution and spatio-temporal selection, decipher the effects of epistasis and recombination, and more accurately predict viral mutations associated with immune escape. Our work provides a theoretical framework for constructing genotype-phenotype landscapes and a paradigm for the interpretability of deep learning in studying virus evolution.

Keywords: Genotype-phenotype landscape; genotype-fitness landscape; interpretable deep learning; immune escape; SARS-CoV-2

One-Sentence Summary: Computing virus immune escape mutations and the basic reproduction number (R_0) in embedding space to construct genotype-fitness landscapes.

Introduction

The evolutionary mutations of viruses often lead to vaccine failure and the escape of neutralizing antibody recognition, resulting in weakened or ineffective drug effects and a substantial increase in infection rates. Phenotypic changes caused by these mutations enhance the fitness of the virus and become fixed in the population. To detect the evolutionary mutations that drive virus "immune escape", researchers have employed various high-throughput experimental techniques (*1*, *2*). However, these techniques have faced challenges in considering the possibility of epistasis and recombination across virus strains and in directly linking "immune escape" mutations to virus fitness. Since the global COVID-19 outbreak, the GISAID database [https://gisaid.org] has collected more than 10 million SARS-CoV-2 genome sequences with spatio-temporal information. This massive data set provides an opportunity to uncover the biological mechanisms behind virus evolution, test the role of epistatic and recombination constraints (*3*), identify and predict virus "immune escape" mutations, and obtain virus epidemic-related fitness information. Such findings can then be used to provide evidence for further experimental verification.

B. Hie et al. have made groundbreaking strides in the field of virus sequence analysis (*4*). By utilizing a natural language model, they have developed an "embedding" of a given virus sequence, which quantitatively links mutated sequence structures to language regulations and correlates them to the virus' ability to evade the immune system. This model can accurately predict whether a mutation will occur in a given sequence context.

F. Obermeyer et al. (*5*) developed a hierarchical Bayesian multinomial logistic regression model, PyR0, to estimate the per-lineage fitness of SARS-CoV-2. The model can infer the relative prevalence of all virus lineages within a geographic area, detect fitness-related mutations and lineages with increased prevalence rates, and identify amino acid substitutions most significantly associated with increased fitness. Significance is defined as the posterior mean or posterior standard deviation.

M. C. Maher et al. (*6*) developed statistical models incorporating epidemiological covariates to account for the effects of drivers and the associated fitness of different viruses.

Our work focuses on quantifying virus fitness by constructing a computable representation of the genotype-fitness landscape. To do this, we developed the P-E theorem, a theoretical framework that allows us to accurately calculate the fitness of a virus population by selecting appropriate embedding expressions for virus genome mutation under the Encoder-Decoder Seq2Seq framework. We then used our developed deep learning natural language model CoT2G-F to accurately calculate virus population fitness without introducing and artificially coupling macro-epidemiological information, resulting in a precise mathematical representation of the virus genotype-fitness landscape in embedding space.

Using the CoT2G-F model and the P-E theorem, we have derived a quantitative relationship between the fitness of SARS-CoV-2 lineages and "immune escape" mutations. This fitness, a macroscopic biological variable, can be accurately described mathematically as the expectation of a latent variable function according to a hidden

state distribution. This result can be seen as a manifestation in the biology of the core concept in statistical mechanics that "macroscopically observable physical quantities are the ensemble averages of a corresponding microscopic state variables function". We call this the Genotype-Phenotype (G-P) principle，it is an important corollary of the P-E theorem.

The G-P principle and its mathematical formulation offer a novel, valuable perspective on the application of deep learning methods to the study of virus evolution. Our research paradigm of directly calculating R_0 through "immune escape" mutations provides a promising approach to constructing genotype-phenotype landscapes.

Our model can more accurately predict virus "immune escape" mutations due to its ability to simulate virus-neutral evolution and spatiotemporal selection, as well as decipher the effects of epistasis and recombination on immune escape. We applied the model to publicly available SARS-CoV-2 spike proteins and obtained better, earlier, and more comprehensive predictions for "immune escape" mutations. Our model can also be used as a generative model to predict the likely occurrence of "immune escape" mutations within the next three to six months and to calculate the fitness of these virus strains carrying "immune escape" mutations after forming lineages. Our prospective and retrospective calculations (Fig. 3B) have confirmed these results. To ensure more accurate predictions when using our model to predict possible "immune escape" mutations in the future, it is necessary to fine-tune the model using current data.

Results

Phenotype-Embedding (P-E) theorem for constructing genotype-fitness landscape

115 Unravelling the evolutionary causes and consequences of the genotype-fitness landscape is a fundamental yet challenging task. The features of the genotype-fitness landscape have a profound influence on the evolutionary process and the predictability of evolution. However, the impact of the genotype-fitness landscape concept on evolutionary biology has been limited due to the lack of empirical
120 information about the topography of real genotype-fitness landscapes and the precise and quantified relationship between genomic variation and fitness (*7*). To address this issue, we propose a novel approach combining basic statistical theory and deep learning models to computationally quantify the genotype-fitness landscape of the SARS-CoV-2 virus based on its genome sequence.

125 Starting from the definitions of evolutionary biology (*7*, *8*), we derive a formal expression for the fitness of a virus population, which is consistent with the mathematical expression given in (5). Under the Encoder-Decoder Seq2Seq framework, we introduce the Bayes theorem, and combine it with Gaussian Mixed Model (GMM) expansion and Expectation-Maximum (EM) algorithm to deduce the
130 P-E theorem: "An observable macrobiological phenotype can be calculated under the Encoder-Decoder Seq2Seq framework if we can find a reasonable embedded representation of the related microscopic genotype". The P-E theorem provides a way to establish a deep learning model to calculate virus fitness (see supplemental note 1,

formulas (1-11)). This means that, under the Encoder-Decoder Seq2Seq framework with a reasonable embedding representation, we can accurately calculate the fitness of the virus population as an observable macrobiological phenotype by using only virus genome sequence data and constructing the genotype-fitness landscape of the virus population (Box 1 and Fig. 1A).

Modelling of virus evolution mechanisms based on virus sequences

Based on the P-E theorem, we have developed a natural language model within the Encoder-Decoder Seq2Seq framework to measure the fitness of a virus population. This model seeks to understand the "grammar" of the biological world that governs the structure of a virus genome sequence, allowing us to uncover the evolutionary laws and composition rules of virus genomic sequences. By doing so, we can identify driver mutations that enable the virus to evade the immune system, providing us with valuable insights into virus fitness. In recent years, natural language models have made great strides in learning the composition rules of DNA and amino acid sequences and exploring the evolution of biological species (*4*, *9–11*).

Referring to B. Hie's work (4), we construct a natural language model CoT2G-F, which is a Co-attention-based Transformer model to bridge Genotype and Fitness: (i) Introducing co-attention and self-attention mechanisms (*12*) to extract the Spatio-temporal correlation and long-range site interaction information within and across virus sequences. These mechanisms enable the model to extract epistatic and recombination signals that promote the occurrence of immune escape mutations. (ii) Dividing training process into two stages: pre-training, simulating the neutral

evolution of the virus by randomly masking one or more continuous bases at any position of the input virus sequence, learning the "grammar" rules of sequence composition; and fine-tuning, in chronological order construct the evolutionary map of the virus sequences dynamically, perform supervised fine-tuning according to the Spatial-temporal correlation constraints and target sequences, reproduce the role of environmental selection, bringing the model closer to the virus evolution process (fig. S4). (iii) Using Transformer to replace BiLSTM (4) as the kernel of the natural language pre-training model. When constructing the "semantic" representation or "embedding" of the input virus sequences, the Transformer model with self-attention mechanism can further extract the upstream and downstream long-range related mutation information of the sequence itself.

Drawing on B. Hie's work (4), we developed CoT2G-F, a co-attention-based Transformer model that bridges genotype and fitness. This model introduces co-attention and self-attention mechanisms to extract spatio-temporal correlation and long-range site interaction information within and across virus sequences, enabling it to capture epistatic and recombination signals that promote the occurrence of immune escape mutations. The training process is divided into two stages: pre-training, which simulates the neutral evolution of the virus by randomly masking one or more continuous bases at any position of the input virus sequence, and fine-tuning, which constructs the evolutionary map of the virus sequences dynamically and performs supervised fine-tuning according to the spatio-temporal correlation constraints and target sequences, thereby reproducing the role of environmental selection (Fig. S4).

Transformer is used to replace BiLSTM (4) as the kernel of the natural language pre-training model, allowing the model to further extract upstream and downstream long-range related mutation information of the sequence itself when constructing the "semantic" representation or "embedding" of the input virus sequences.

As of November 2022, 10.2 million SARS-CoV-2 spike proteins from GISAID (www.gisaid.org) were selected and subjected to strict quality control to train the CoT2G-F model (table S1). This data was used for pre-training, with further stages of fine-tuning to verify the model's predictive ability. The CoT2G-F model can obtain the hidden state distribution of DNA bases or amino acids, and a sequence "embedding" that satisfies this hidden state distribution is guaranteed to conform to the "grammar" rules even in the presence of "semantic" changes. This allows for the identification and inference of "immune escape" mutations from the degree of "semantic" change (Fig. 1B, 1C and 1B and see supplemental notes 2-4). Future work will focus on using the CoT2G-F model to determine the micro-state function of a latent variable corresponding to the virus fitness in the hidden space, starting from the P-E theorem (Formula (11)).

Identifying and predicting "immune escape" mutations

Following the CSCS decision rule for semantics and syntax changes proposed by B. Hie (*4*), we tested the performance of our model CoT2G-F against two related mainstream models from the framework design in three technical dimensions. Our model demonstrated superior performance (see supplementary materials). As shown in Fig. 2A, our model outperformed the BiLSTM and Vanilla Transformer models,

200 and the ability to identify and predict immune escape mutations was significantly improved. By considering the Spatio-temporal dynamics of virus evolution during training processes, the model was able to identify existing and predict future "immune escape" mutations. We tested this using more than 2 million SARS-CoV-2 genomic sequence data from the United Kingdom as an example and obtained better and more
205 comprehensive prediction results (Fig. 2B, 2C and 2D). This provides a foundation for subsequent fitness calculation.

The usual immune escape mutations

Our model, based on SARS-CoV-2 genome data collected and submitted by the UK to the GISAID database from August 2021 to October 2021 (prior to the emergence of
210 the Omicron lineage as a pandemic), shows good reproducibility for immune escape mutations, including most of the mutations carried by virus lineages that emerged before October 2021, such as B.1.1.7(H69del, V70del, Y144del, N501YA570D, etc.), B.1.351(K417N, E484K, etc.), B.1.617.2(G142D, E156del, F157del, L452R, T478KP681R, D950N, etc.), P.1(T20N, P26S, D138Y, R190S, K417T, H655Y,
215 T1027I, V1176F, etc.), as well as those that were present early on in other virus lineages but then mutated in the Omicron lineage, such as BA.1 (T95I, G142del, V143del, N211del, G339D, S371L, S373P, S375F, N440K, G446S, E484A, Q493R, Q496S, Q498R, Y505H, T547K, N679K, etc.), B.A.2(P25del, P26del, G142D, P384L, etc.), BA.2.75(P25del, P26del, G142D, G446S,etc.) , BQ.1(P25del, P26del,
220 H69del, V70del, L452R, etc.), XBB(P25del, P26del, E156del, etc.) (Fig. S4).

To verify our model's predictive ability, we used data from the UK submitted to GISAID after March and before June 2022 as input and step-by-step predicted the future occurrence of immune escape mutations. Our model correctly forecasted almost all of the immune escape mutations carried by the Omicron lineage, which emerged after March 2022. Typically, we observe mutations in the Omicron lineage that emerge and spread in late 2022, such as B.1.621 (N501Y, P681H, D950N), BF.7 (P25del, P26del, A27del, G142d, G339D, N440K, D796Y, etc.), BJ.1 (Q498R), BA.2.12.1 (L452Q), BQ.1.1 (G339D), and XBB (Q954H, N969K) （Fig. S5C and S5D）. Notably, our model was trained using data from GISAID as of March 31, 2022, demonstrating its ability to give accurate predictions up to six months in advance without dynamic fine-tuning.

It is worth noting that our model can also be used as a generative model to generate possible immune escape mutations for screening and early warning study (Fig. S6).

The convergent immune escape mutations

Convergent evolution is a common phenomenon in the evolutionary process of viruses. This evolution typically occurs under strong environmental selection constraints (*13*). From March to June 2023, virus lineages with convergent evolutionary mutations were selected as supervised input data for fine-tuning our model, and specific environmental selection constraints were simulated. After training was completed, data from the United Kingdom from June to November were used for prediction and replication. The results show that most convergent evolutionary

mutations can be predicted and reproduced (Fig. S6).

The "catastrophes" evolutionary events

Our pre-training model, which randomly masks a sequence span with an average amino acid length of 3, is capable of simulating various patterns of continuous evolution-related genetic drift, as well as reproducing and predicting the occurrence of discrete evolutionary events caused by so-called "catastrophes" events (see Supplementary Notes 2-4). Our prediction results contain a large number of predicted indels, including those that occur in the Omicron virus lineage (see Supplementary Materials, Table S2). Although we have not yet conducted a detailed analysis due to the limited space and scope of this paper, it is certainly worthy of further study.

Deciphering the intrinsic correlation between "Semantic" change related "immune escape" mutations and virus fitness

The immune escape ability of a virus lineage determines its fitness or the basic reproduction number (R_0). The problem is how to establish the quantitative relationship between the "immune escape" mutation and the fitness of the virus conforming to the P-E theorem under the framework of the natural language model. B. Hie et al. propose in their natural language model to quantify the relationship between the "immune escape" mutation of a virus lineage and its fitness. The model suggests that the degree of "semantic" change is related to the virus's immune evasion ability, while the "grammatical" fitness is related to its fitness (*4*). Meanwhile, F. Obermeyer et al. used a hierarchical Bayesian regression model to fit the relative growth of virus

lineages through multinomial likelihood (*5*). They found that the regression coefficient of their model was equivalent to the per-lineage fitness of the virus.

Very interestingly, the per-lineage fitness defined by regression coefficients is the product of two parameters. The first relates to virus mutations, corresponding to "semantic" changes in the natural language model. The second is a macroscopic observable measure related to virus sequence classification and occurrence frequency. This has no corresponding concept in the existing natural language model but appears in the P-E theorem as the transformation coefficient λ associated with the posterior distribution of a latent variable connecting real space and embedded space (see Box 1, Formula (9) and Supplemental Note 1, Formula (11)). According to our model, λ should relate to the "synthetic" fitness corresponding to the hidden state distribution of the "embedded" virus sequence.

If evolution by natural selection is to occur, there must be a genotypic change in the virus population across generations. This change in fitness is dependent on the genotype; if the genotype does not change, there is no change in fitness. The syntactic nature of the stable genome composition does not directly contribute to the virus's fitness. However, the "semantic" changes that affect the gain of genomic function and contribute to the virus's fitness change. The "grammar" rules of the natural language model determine the global composition of the sequence to ensure the virus's survival, and the "semantic" changes correspond to the changes in the "immune escape" ability. Therefore, "semantics" and "grammar" together determine the virus's fitness. Under the theoretical framework of deep learning, the above property can be expressed

mathematically as the convolution of "semantic" and "grammatical" terms of the CSCS criterion (*4*). When multiplied by the coefficient λ, the final mathematical representation is consistent with the P-E theorem (see Box 1, Formula (9) and Supplemental Note 1 and Note 4, Formulas (11) and (31)). These results suggest that the term related to the "semantics" change of the model CoT2G-F and the CSCS criterion (*4*) may be the micro-state function of the latent variable corresponding to the virus's fitness that we hope to obtain.

Viewing the correlation between the trends of virus epidemics and the absolute mean distribution of conditional semantic change score CSC

The P-E theorem proves that we can construct the function of a latent variable through the Encoder-Decoder Seq2Seq model or its extension model. The mathematical expectation of this function in terms of the hidden state distribution is the corresponding macroscopic observable variable that we hope to obtain. We can extend and define a Conditional Semantic Change Score (CSC Score) from the CSCS Score presented by B. Hie et al. (*4*). The Conditional Semantic Change Score (CSC), as a function of latent variables describing "semantic" changes, is used to measure the "immune escape" ability related to virus sequence mutation. Thinking that the absolute mean of the CSC Score is an index of the immune escape ability of a virus lineage, we want to see how this index changes across different virus lineages (see Supplemental Notes 2-4, Formula (27)). We selected 2.7 million SARS-CoV-2 spike protein sequence data from the UK to calculate the absolute mean distribution of the CSC Score according to the hidden state probability distribution (Fig. 3).

Fig. 3A and 3B clearly show that, from the Wuhan lineage to the Delta lineage and then to the Omicron lineage, the immune evasion ability of the emerging virus lineages continues to increase and even presents an accelerated trend. The absolute mean of the "immune escape" ability of the virus strain is a calculable and quantifiable parameter directly related to the virus genome sequence mutation, which can be used to assess the cumulative effects of the immune escape ability of an emerging virus lineage. The absolute mean of the CSC score accurately describes the changing trend of the "immune escape" ability of virus lineages and thus reveals the epidemic potency of the virus. This result further suggests that the CSC score is the microstate function corresponding to the virus fitness indicated by the P-E theorem.

Expressing the fitness of a virus lineage as a convolution of "semantics" and "grammar" and the mathematical expectation of a semantic change function according to a hidden state probability distribution

It has long been a goal in biology to create genotype-fitness landscapes by mapping DNA sequences to mutation combinations observed in phylogeny or evolutionary experiments. However, due to the consideration of the epistatic effects of mutations, when the mutation number is large, the model becomes complicated and difficult to calculate and verify experimentally (*7*, *8*, *14*) . E. D. Vaishnav et al. (*15*) argue that the complete fitness landscape defined by a fitness function maps every sequence (including mutations) in the sequence space to its associated fitness. However, no theoretical model has yet been able to provide a concise computational model and framework for the genotype-fitness landscape while fully considering

epistasis and recombination potential. Under the framework of the P-E theorem, we attempt to solve this problem through our deep learning natural language model CoT2G-F combined with CSC scores.

Based on the functional relationship between the CSC score and the latent variable of the CoT2G-F model (see Supplemental Note 4, Formula (22)), we know that the CSC score is a relative measure of semantic change corresponding to the relative immune escape ability of the virus and the "semantic" mutation corresponding to genomic mutation. Therefore, the CSC score measures the degree of the genomic mutation of emerging virus strains relative to wild-type virus strains. Furthermore, the CSC score and the "grammatical" coincidence degree determined by the mutation state distribution – the hidden state distribution – together form a measure of the virus's immune escape ability in the form of convolution (see Supplemental Note 4, Formula (31)). Referring to the P-E theorem and by the properties of the CSC score and the formal equivalence of formulas (11) and (31), the CSC score should be the corresponding hidden space microstate function of the macroscopic observable variable R_0. Formula (30) is the concrete realization of the P-E theorem (see formula (11)) under the CoT2G-F model. Finally, we acquire the basic formula for calculating R_0 (see Supplemental Note 4 and Formula (33)). The correctness of the derived calculation formula for the per-lineage fitness is evident and verified (Fig. 3A and 3B, Fig. 4). This result and the P-E theorem provide a mathematical basis for redefining macro-phenotypes such as R_0 and and the resulting computability.

The genotype-fitness landscape

The calculation formula of the R_0 has a novel meaning (see Supplemental Note 4, Formulae (31), (32) and (33)): it is the convolution of the semantic change function and the syntactic state distribution function in the embedding space. Here, convolution means giving the average of the cumulative effects of the contribution of all sampled virus variants to the R_0 according to the reference sequence (we selected Wuhan-Hu-1 as a reference sequence; see Supplementary Materials). More importantly, the R_0 can be precisely rewritten as a mathematical expectation of a semantic change function according to a hidden state probability distribution, giving the quantitative relationship between the "immune escape" ability of the virus and the fitness of the virus lineage. Finally, we present a comprehensive and precise formulation of the virus genotype-fitness landscape. The consistency between these two different representations provides a solid mathematical basis and research paradigm for applying deep learning models to biological problems and gaining interpretability (see Supplemental Notes 1-4).

Studying the intrinsic relationship between statistical mechanics and deep learning has always been an intriguing subject (*16–18*). We know that the core concept of statistical mechanics is: "The macroscopic physical observables can be characterized as the ensemble average of the corresponding microstate variable functions". Inspired by this, the inference is naturally drawn: "Under the Encoder-Decoder framework, a macro-biological observable can be expressed as the mathematical expectation of a function of a latent variable according to a hidden state

distribution in the decoder space". We call this framework bridging genotype-phenotype landscapes the G-P principle, which is an important corollary of the P-E theorem and gives a novel interpretable application of deep learning theory to life science (Fig. 5, Supplemental Note 1 and 4, equations (11), (34) and (35)). Finally, we can plot the genotype-fitness landscape in the embedded space based on the variation state of the viral genome sequence according to equation (30). It is a two-dimensional hypersurface, and we can obtain the fitness of the virus lineage by integrating the surface density of the specific region corresponding to the virus lineage on this hypersurface, thereby defining the immune escape force of the virus (see Supplemental Note 4, Fig. 5B and Fig. S9).

Inferring the R_0 of the virus lineage

According to our theoretical framework, computing the fitness of each virus lineage and the fold increase in relative fitness to obtain R_0 requires adequate sampling for each virus lineage in a specific region and time interval. In this region and time interval, the virus begins to emerge, sustain its spread, increase in number, and eventually reach a plateau. Based on the biological definition of fitness, the selected sampling time interval needs to ensure the virus is transmitted for enough generations to reduce the violent fluctuation of the signal (5). It is also crucial for determining the occurrence frequency of the largest cluster of the different lineages and simultaneously calculating the absolute mean of the CSC score (see Supplemental Note 4, Formula (30), Figures S3A and S3B). Currently, the GISAID (www.gisaid.org) contains over 2 million genome sequences of SARS-CoV-2 from the United Kingdom

[http://gisaid.org], and the spatio-temporal distribution of the virus lineages reflects the major trend of the world epidemic of SARS-CoV-2, which can serve as a model system for a reliable and feasible G-P principle.

We computed the virus per-lineage fitness using 2.7 million SARS-CoV-2 spike proteins from the GISAID database (www.gisaid.org) and adopted the Pango lineage designation and assignment. To calculate the factor λ_k in the per-lineage fitness (see Box 1, Formula (9) and Supplemental Note 4, Formula (30)), we first performed lineage assignment with Pangolin. We then selected the k_{th} lineage and its nearest sublineages to form a population, sampling for a given time interval t_{bin}, and λ_k was the occurrence frequency of the k_{th} lineage in this small population. The calculation result is shown in Fig. 4, where the time interval we chose was 10 days (see Fig. S6). Our model correctly inferred that the WHO classification variant Omicron lineages (Pango lineage BA.2.X) had very high relative fitness in a pandemic, significantly higher than the result from F. Obermeyer *et al.* (*5*), but more consistent with the actual and official monitoring results. It was approximately: ~ 16-17 times higher than the original Wuhan and lambda lineages (Fig. 4, [95% confidence interval (CI) 16.78 to 16.84], Fig. S7), accurately predicting its rise in the spread regions.

The general trends of the R_0 are consistent with the results of (5), but the R_0 value is higher overall. This is because our model considers all possible interactions between all sequence sites of the virus and mutations, taking into account temporal-spatial correlations and recombination effects when calculating the

contribution of immune escape mutation sites to immune escape abilities and R0. This result validates our proposed P-E theorem and G-P principle, allowing for the precise computation of the relative fitness of the virus lineage and R_0.

Discussion and conclusion

Applying AI and deep learning methods to biology has two key points. The first is how to make the model more in line with biological logic, and the second is to delve into the connotation of the model and provide interpretability. This study explores these two aspects and presents an initial research paradigm. We first proposed a general theorem, the P-E theorem, and in the second step, we constructed a model, CoT2G-F, that more closely matches the evolutionary biology scenario. The calculation of R_0 was realized by the model under the guidance of the P-E theorem.

In our model the R_0 is a mathematical expectation of the "semantic" change according to a hidden state distribution. It is also a convolution of the "semantic" change function and a hidden state distribution function. The hidden state distribution as a prior probability distribution, determines the "grammatical" term in the CSCS criterion (Fig. 5, supplementary notes 3 and 4). This directly gives quantitative relationships among "semantics" change and "grammatical" fitness, "immune escape" mutation and the fitness of virus lineages (see Supplementary Materials, Supplementary Note 4, formulas (30) and (33)). This result can be seen as the correspondence in the biology of the core concept in statistical mechanics that "macroscopic observable physical quantities are the ensemble average of the corresponding microscopic state variable functions". We call this the G-P

(Genotype-Phenotype) principle and it can be regarded as a corollary of the P-E theorem. The G-P principle provides a new valuable perspective and interpretability for applying deep learning to virus evolution. It reveals the intrinsic correlation between deep learning theory and statistical mechanics studied by many researchers (*16–18*).

Currently, most deep learning algorithms for biological studies attempt to establish the relationship between microscopic molecular biology characteristics and macroscopic biology variables to gain the interpretability of the models (*19–25*). However, it is always challenging to establish a direct, quantitative and computable correlation between them. The practice and results of the R_0 calculation in the present work address this question and provide a feasible paradigm for the computational modelling of genotype-phenotype landscapes, which is a concrete and successful example of the G-P principle or the P-E theorem. We believe that, following the research path proposed in the present work, many research examples that adhere to the G-P principle or the P-E theorem will emerge in the future. Our model, the derived mathematical representation of R_0, and computational results provide a compelling commentary on the first principle of "selection imprinting recorded in viral genomic mutation".

In addition, the model uses the Transformer as a kernel and further introduces the co-attention and continuous span masking mechanism to consider the spatial-temporal related virus evolution. The model has advantages in extracting the upstream and downstream long-range correlation mutation information from the sequence itself and

the ability to obtain the spatio-temporal features related to virus evolution. With the introduction of a continuous mask mechanism, the model can mimic genetic drift and determine what kind of mutations will be fixed by selection. All together, by inputting the virus genome sequence alone, the model can more accurately identify and predict possible future "immune escape" mutations by calculating the conditional "semantic" change related to the latent variables. The model's power is demonstrated in Figures 2A, 2B, 2C, and 2D.

Due to the introduction of a continuous masking mechanism, our model is able to reproduce and predict the occurrence of discontinuous evolutionary events caused by so-called "catastrophes" to some extent. In our prediction results, there is a large number of predicted indels, including quite a few amino acid insertions of the spike protein presented by the Omicron lineage. These insertions conform to the semantic and grammatical constraints of the sequence composition (see Supplementary Materials, Table S2). Further research and experimental verification are warranted.

Importantly, we plotted a virus genotype-fitness landscape in a formal three-dimensional embedding space, starting from the quantitative relationship between "immune escape" mutations and viral lineage fitness. This is the first precise mathematical representation of a virus' genotype-fitness landscape, and its topological properties can describe the virus evolution and the driving force of immune escape. The realization of this description method provides a groundbreaking theoretical and technical framework for future research, and the research paradigm has certain universality.

Finally, it is worth noting that the proposed P-E theorem is universal in biological research. Many fundamental states of biology, such as brain homeostasis, metabolic homeostasis, cell fate state, tumour microenvironment, advanced ageing state, etc., can be effectively studied based on the P-E theorem. The P-E theorem provides a solid mathematical foundation for the deterministic description of biological homeostasis and for deciphering its stability mechanism.

490 **References and Notes**

1. D. Mathew, J. R. Giles, A. E. Baxter, D. A. Oldridge, A. R. Greenplate, J. E. Wu, C. Alanio, L. Kuri-Cervantes, M. B. Pampena, K. D'Andrea, S. Manne, Z. Chen, Y. J. Huang, J. P. Reilly, A. R. Weisman, C. A. G. Ittner, O. Kuthuru, J. Dougherty, K. Nzingha, N. Han, J. Kim, A. Pattekar, E. C. Goodwin, E. M. Anderson, M. E. Weirick, S. Gouma, C. P. Arevalo, M. J. Bolton, F. Chen, S. F. Lacey, H. Ramage, S. Cherry, S. E. Hensley, S. A. Apostolidis, A. C. Huang, L. A. Vella, The UPenn COVID Processing Unit, M. R. Betts, N. J. Meyer, E. J. Wherry, Z. Alam, M. M. Addison, K. T. Byrne, A. Chandra, H. C. Descamps, Y. Kaminskiy, J. T. Hamilton, J. H. Noll, D. K. Omran, E. Perkey, E. M. Prager, D. Pueschl, J. B. Shah, J. S. Shilan, A. N. Vanderbeck, Deep immune profiling of COVID-19 patients reveals distinct immunotypes with therapeutic implications. *Science*. **369**, eabc8511 (2020).
2. T. N. Starr, A. J. Greaney, S. K. Hilton, D. Ellis, K. H. D. Crawford, A. S. Dingens, M. J. Navarro, J. E. Bowen, M. A. Tortorici, A. C. Walls, N. P. King, D. Veesler, J. D. Bloom, Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell*. **182**, 1295-1310.e20 (2020).
3. J. A. G. M. de Visser, S. F. Elena, I. Fragata, S. Matuszewski, The utility of fitness landscapes and big data for predicting evolution. *Heredity*. **121**, 401–405 (2018).
4. B. Hie, E. D. Zhong, B. Berger, B. Bryson, Learning the language of viral

evolution and escape. *Science*. **371**, 284–288 (2021).

5. F. Obermeyer, M. Jankowiak, N. Barkas, S. F. Schaffner, J. D. Pyle, L. Yurkovetskiy, M. Bosso, D. J. Park, M. Babadi, B. L. MacInnis, J. Luban, P. C. Sabeti, J. E. Lemieux, Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness. *Science*. **376**, 1327–1332 (2022).
6. M. C. Maher, I. Bartha, S. Weaver, J. di Iulio, E. Ferri, L. Soriaga, F. A. Lempp, B. L. Hie, B. Bryson, B. Berger, D. L. Robertson, G. Snell, D. Corti, H. W. Virgin, S. L. Kosakovsky Pond, A. Telenti, Predicting the mutational drivers of future SARS-CoV-2 variants of concern. *Sci. Transl. Med.* **14**, eabk3445 (2022).
7. J. A. G. M. de Visser, J. Krug, Empirical fitness landscapes and the predictability of evolution. *Nature Reviews Genetics*. **15**, 480–490 (2014).
8. I. Fragata, A. Blanckaert, M. A. Dias Louro, D. A. Liberles, C. Bank, Evolution in the light of fitness landscape theory. *Trends in Ecology & Evolution*. **34**, 69–82 (2019).
9. E. C. Alley, G. Khimulya, S. Biswas, M. AlQuraishi, G. M. Church, Unified rational protein engineering with sequence-based deep representation learning. *Nat Methods*. **16**, 1315–1322 (2019).
10. T. Bepler, B. Berger, Learning protein sequence embeddings using information from structure (2019), , doi:10.48550/arXiv.1902.08661.
11. R. Rao, N. Bhattacharya, N. Thomas, Y. Duan, P. Chen, J. Canny, P. Abbeel, Y. Song, "Evaluating Protein Transfer Learning with TAPE" in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2019;

https://papers.nips.cc/paper/2019/hash/37f65c068b7723cd7809ee2d31d7861c-Abstract.html), vol. 32.

12. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, "Attention is All you Need" in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett, Eds. (Curran Associates, Inc., 2017; https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf), vol. 30.
13. Y. Cao, F. Jian, J. Wang, Y. Yu, W. Song, A. Yisimayi, J. Wang, R. An, X. Chen, N. Zhang, Y. Wang, P. Wang, L. Zhao, H. Sun, L. Yu, S. Yang, X. Niu, T. Xiao, Q. Gu, F. Shao, X. Hao, Y. Xu, R. Jin, Z. Shen, Y. Wang, X. S. Xie, Imprinted SARS-CoV-2 humoral immunity induces convergent Omicron RBD evolution. *Nature* (2022), doi:10.1038/s41586-022-05644-7.
14. D. A. Kondrashov, F. A. Kondrashov, Topological features of rugged fitness landscapes in sequence space. *Trends in Genetics*. **31**, 24–33 (2015).
15. E. D. Vaishnav, C. G. de Boer, J. Molinet, M. Yassour, L. Fan, X. Adiconis, D. A. Thompson, J. Z. Levin, F. A. Cubillos, A. Regev, The evolution, evolvability and engineering of gene regulatory DNA. *Nature*. **603**, 455–463 (2022).
16. Y. Bahri, J. Kadmon, J. Pennington, S. S. Schoenholz, J. Sohl-Dickstein, S. Ganguli, Statistical Mechanics of Deep Learning. *Annual Review of Condensed Matter Physics*. **11**, 501–528 (2020).

17. G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, L. Zdeborová, Machine learning and the physical sciences. *Rev. Mod. Phys.* **91**, 045002 (2019).

18. Y. Liu, X. Yao, Ensemble learning via negative correlation. *Neural Networks*. **12**, 1399–1404 (1999).

19. L. Qiu, L. Lin, V. M. Chinchilli, Variational Interpretable Learning from Multi-view Data (2022), , doi:10.48550/arXiv.2202.13503.

20. K. M. Boehm, P. Khosravi, R. Vanguri, J. Gao, S. P. Shah, Harnessing multimodal data integration to advance precision oncology. *Nat Rev Cancer*. **22**, 114–126 (2022).

21. I. Cetin, O. Camara, M. A. G. Ballester, Attri-VAE: attribute-based, disentangled and interpretable representations of medical images with variational autoencoders (2022), , doi:10.48550/arXiv.2203.10417.

22. L. Seninge, I. Anastopoulos, H. Ding, J. Stuart, VEGA is an interpretable generative model for inferring biological network activity in single-cell transcriptomics. *Nat Commun*. **12**, 5684 (2021).

23. H. A. Elmarakeby, J. Hwang, R. Arafeh, J. Crowdis, S. Gang, D. Liu, S. H. AlDubayan, K. Salari, S. Kregel, C. Richter, T. E. Arnoff, J. Park, W. C. Hahn, E. M. Van Allen, Biologically informed deep neural network for prostate cancer discovery. *Nature*. **598**, 348–352 (2021).

24. A. Cheerla, O. Gevaert, Deep learning with multimodal representation for pancancer prognosis prediction. *Bioinformatics*. **35**, i446–i454 (2019).

25. S. Ainsworth, N. Foti, A. K. Lee, E. Fox, Interpretable VAEs for nonlinear group factor analysis (2018), , doi:10.48550/arXiv.1802.06765.

26. A. Rambaut, E. C. Holmes, Á. O'Toole, V. Hill, J. T. McCrone, C. Ruis, L. du Plessis, O. G. Pybus, A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature Microbiology*. **5**, 1403–1407 (2020).

27. F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei, M.-L. Yuan, Y.-L. Zhang, F.-H. Dai, Y. Liu, Q.-M. Wang, J.-J. Zheng, L. Xu, E. C. Holmes, Y.-Z. Zhang, A new coronavirus associated with human respiratory disease in China. *Nature*. **579**, 265–269 (2020).

28. K. Katoh, D. M. Standley, MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*. **30**, 772–780 (2013).

29. D. J. Futuyma, M. Kirkpatrick, "The Evolution of a Population's Mean Fitness" in *Evolution* (Sinauer Associates, Inc, Sunderland, Massachusetts, Fourth., 2017), pp. 103–134.

30. K. Hornik, M. Stinchcombe, H. White, Multilayer feedforward networks are universal approximators. *Neural Networks*. **2**, 359–366 (1989).

31. A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, R. Fergus, Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*. **118**, e2016239118 (2021).

32. J. Meier, R. Rao, R. Verkuil, J. Liu, T. Sercu, A. Rives, "Language models enable zero-shot prediction of the effects of mutations on protein function" in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, J. W. Vaughan, Eds. (Curran Associates, Inc., 2021; https://proceedings.neurips.cc/paper/2021/file/f51338d736f95dd42427296047067694-Paper.pdf), vol. 34, pp. 29287–29303.

33. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2018), doi:10.48550/arXiv.1810.04805.

34. S. R. Bowman, L. Vilnis, "Generating Sentences from a Continuous Space" in *the 20th SIGNLL Conference* (Berlin, Germany, 2016), pp. 10–21.

35. A. Tsoularis, J. Wallace, Analysis of logistic growth models. *Mathematical Biosciences*. **179**, 21–55 (2002).

Acknowledgements:

We thank G. P. Zhao, L. Jin, J. R. Wu, P. J. Zeng, R. B. Liu and Z. J. Lang for their deep and helpful discussions. We thank T. Zeng and J. W. Liu for their assistance with the manuscript. We are also particularly grateful to B. Hie for his constructive responses to our questions about the details of his related excellent models.

Funding:

CAS Research Fund, Grant No. XDB38050200

Self-supporting Program of Guangzhou Laboratory, Grant No. SRPG22-001

Self-supporting Program of Guangzhou Laboratory, Grant No. SRPG22-007

Author contributions:

Conceptualization: Y. X. Li

Methodology: Y. Luo, Y. X. Liu, X. Lu, H. Gao, R. K. He, X. Zhang, X. G. Zhang and Y.X. Li

Investigation: Y. Luo, Y. X. Liu, X. Lu, H. Gao and R. K. He

Visualization: Y. X. Liu, Y. Luo, X. Lu and H. Gao

Funding acquisition: Y. X. Li

Project administration: X. Zhang, X. G. Zhang and Y.X. Li

Supervision: Y. X. Li

Writing – original draft: Y. X. Liu, Y. Luo, X. Lu, H. Gao, R. K. He, X. Zhang, X. G. Zhang and Y.X. Li

Writing – review & editing: Y. X. Li

Competing interests: All the authors have no conflicts of interest to declare.

Data and materials availability: Code, scripts for plotting and visualizing, and pre-trained models are available at https://github.com/martyLY/CoT2G-F. We used the following publicly available datasets for model training: SARS-CoV-2 spike protein sequences from GISAID (www.gisaid.org); training and validation datasets are deposited to https://zenodo.org/deposit/7388491, and links to this data are also available at https://github.com/brianhie/viral-mutation.

Supplementary Materials for

Learning genotype-fitness landscape of SARS-CoV-2 in embedding space

Y. X. Liu†, Y. Luo†, X. Lu†, H. Gao†, R. K. He , X. Zhang*, X. G. Zhang*, Y. X. Li*

Correspondence to: xzhang@picb.ac.cn (X.Z.); fxgzhang@mengniu.cn (X.G.Z); li_yixue@gzlab.ac.cn (Y.X.L)

This PDF file includes:

Materials and Methods

Supplementary Text

Figs. S1 to S5

Tables S1 to S2

References

Content

660 Materials and Methods 4

Data and Code Availability 4

Data access and preprocessing 4

Preparation of the dataset for pre-training and fine-tuning 5

Co-attention based Transformer model bridging Genotype and Fitness 5

665 Comparison with other major natural language models 5

Model training based on the fundamental principles of evolutionary biology 7

"Immune escape" mutations prediction and R_0 inference 9

Supplementary Text 10

Supplemental Note 1: The formal expression of fitness in Encoder-Decoder

670 Seq2Seq framework and the P-E theorem 10

1. Mathematical description of the average and per-lineage fitness of the virus population 10

2. Fitness calculation based on the Bayes' theorem and deep learning model of the Encoder-Decoder Seq2Seq architecture 13

675 Supplemental Note 2: The fulcrum of building model system and the introduction of CSC scores 17

Supplemental Note 3: Detailed description of the CoT2G-F model 21

Supplemental Note 4: The formal logic of the Genotype-Fitness landscapes and R_0 calculation 26

680 1. The "immune escape" ability score CSC 26

2. The mathematical expectation of the CSC score according to the hidden state probability distribution **Pz** is consistent with the formal expression of the average and the per-lineage fitness of the virus population. 27

3. The prediction of immune escape mutation sites and calculation of R_0 based on the CoT2G-F model 32

4. G-P principle: Genotype-phenotype landscapes in embedding space based on the deep learning framework 34

Fig. S1. Preprocessing of SARS-CoV-2 sequences. 36

Fig. S2. Summary of the preprocessed SARS-CoV-2 sequences. 37

Fig. S3. Growth rate of representative SARS-CoV-2 lineages in the UK 38

Fig. S4. Reproducing known immune escape mutations 40

Fig. S5. The schematic diagram for the target sequences search of the dynamic fine-tuning step 43

Fig. S6. Model-Generated Spectrum of Immune Escape Mutants of SARS-CoV-2 44

Fig. S7. Relative fitness derived from the natural language model CoT2G-F versus date of lineage emergence 45

Fig. S8. The confidence interval (CI) distribution for the R_0 calculation 46

Fig. S9. Plot virus genotype fitness landscape 47

Materials and Methods

Data and Code Availability

The source code for data preprocessing and modelling and available at https://github.com/martyLY/CoT2G-F. GISAID sequence data is publicly available at https://gisaid.org. The data set sampled from the UK region according to the Pango lineage designation and assignment, used for the R_0 calculation is available at https://zenodo.org/deposit/7388491.

Data access and preprocessing

Virial sequence and metadata were obtained from ViGTK (https://www.biosino.org/ViGTK), which is a mirror site for SARS-CoV-2 genome data of the GISAID (https://gisaid.org) and NCBI (https://www.ncbi.nlm.nih.gov) database. We download 16,395,521 genome sequences from ViGTK as of November 26, 2022. Records with missing time or location were discarded. After removing low quality sequences of length less than 29000 bp or with ambiguous bases more than 400, we obtained 14,959,290 genome sequences altogether (Table S1). Lineage assignment was performed with Pangolin (version 4.1.2) (*26*) subsequently. To get the spike protein sequences, genome sequences were firstly aligned to SARS-CoV-2 reference genome (Wuhan-Hu-1, NC_045512.2) (*27*) using the MAFFT7 (version 7.505) (*28*). Then, coding sequences were extracted from the alignment and translated into protein sequences. We only considered spike protein sequences whose sequence length were between 1265 and 1275 amino acid residues. Additionally, spike proteins with ambiguous amino acids were excluded, which yielded high quality dataset of

10,184, 273 sequences in total (Fig. S1). A total of 2,713,327 SARS-CoV-2 spike protein sequences that belong to UK was subset from processed dataset.

Preparation of the dataset for pre-training and fine-tuning

After preprocessing, we bin time intervals into one-month segments, resulting in 35 time-bins (Fig. S4). In each month, sequences were arranged by their evolution distance relative to Wuhan-Hu-1 strain (NC_045512.2) of SARS-CoV-2 (Fig. S5). In order to find nearest neighborhood for the sequence in t time, all sequences in t-1 month were aligned to this sequence. The spike protein sequence in t-1 month with minimum edit distance relative to Wuhan-Hu-1 strain were selected as the nearest neighborhood.

Co-attention based Transformer model bridging Genotype and Fitness

We built an Encoder-Decoder Seq2Seq architecture natural language model CoT2G-F with a Transformer as the technical framework and introduced the co-attention and contiguous spans masking mechanisms. Pre-trained with 10.2 million SARS-CoV-2 Spike protein sequence data from GISAID, the data set with 2.7 million sequences from the UK region was selected to validate our model. Details have been described in the supplemental notes below.

Comparison with other major natural language models

Fig.2A shows a comparative experiment result with our method (CoT2G-F) and Vanilla Transformer without the introduction of the cross-attention mechanism, and Bi-LSTM, which is used in B. Hi (*4*). to capture the semantic and syntactic variations of protein sequences. It is worth noting that B. Hi (4) and our method (CoT2G-F)

solve different problems: B. Hi (*4*). uses Bi-LSTM as an encoder to extract semantic variation and syntactic information and predict Immune escape, and our objective is to generate the protein sequence of immune escape in the next time slice, so we connected the decoder module in Bi-LSTM to generate protein sequence. In this figure, the comparison experiment is based on Bi-LSTM, Vanilla Transformer, and our method (CoT2G-F). We compare the effect of the model from tow dimensions, the first is the identification of sequence mutation sites, we hope that the model can accurately predict the mutated sites (i.e., high recall) without randomly predicting the mutated sites (i.e., high precise), so the f1 score combine with recall and precise is used to compare the identification of mutated sites by different models, denoted as 'Site' in this figure, compared with Bi-LSTM and Vanilla transformer, our method (CoT2G-F) improved f1 score of mutation sites by 1.53% and 0.29% respectively. The second dimensions is to identify the amino acid mutated into the site based on the identification of the mutation site in the first dimension, denoted as 'Amino acid' in this figure, we find the effect of Bi-LSTM is not ideal when dealing with the task of generating long protein sequences, and cannot accurately identify the mutated amino acids. However, the attention mechanism in Vanilla Transformer and our method (CoT2G-F) can capture the long-range dependencies of long protein sequences and can learn the mutation logic of amino acids relatively better. Compared with Vanilla Transformer, our method (CoT2G-F) improves the identification rate of mutated amino acids by nearly 1.64%, thanks to the information gain brought by the introduction of spatiotemporal information in our method (Fig. 2A) .

Model training based on the fundamental principles of evolutionary biology

CoT2G-F is an Encoder-Decoder Seq2Seq architecture natural language pre-training model (see supplemental note 1) that introduces a co-attention and contiguous spans masking mechanisms and uses a Transformer as a basic framework. To make the model better reflect the real virus evolution scenario, we adopt a two-step training mode. The first step is the pre-training process. First, from the outbreak of the virus to the present, the time cross-section is divided into months, and the global SARS-CoV-2 genome sequences in one month are arranged according to the evolutionary distance to form a two-dimensional genome sequence matrix $\mathbf{X_t} = (\mathbf{x}_{(t,1)}, \dots, \mathbf{x}_{(t,p)}, \dots, \mathbf{x}_{(t,N)})$, the row of the submatrix $\mathbf{x_{t,p}}$ represents the genome sequence of a virus around spatial position p, and t is the corresponding time of a certain time cross-section. During pre-training, the co-attention module of CoT2G-F starts from the first row of the matrix $\mathbf{X_t}$, and sequentially takes three spatially connected Spike protein sequences $(\mathbf{x}_{(t,p-1)}, \mathbf{x}_{(t,p)}, \mathbf{x}_{(t,p+1)})$, get $(\mathbf{x}_{(t,p-1)}, \mathbf{x}_{(t,p)})$ and $(\mathbf{x}_{(t,p)}, \mathbf{x}_{(t,p+1)})$ pairwise co-attention signals, and then the Transformer module randomly masks any amino acid sites of the Spike protein with a proportion under 15%, which is completely different from (*5*). We adopt such a pre-training process that can not only replicate the "neutral evolution" source of virus variation but also consider the long-range associations between sites in the viral genome sequence composition rules, as well as possible recombination effects. Therefore, our pre-training model that reflects the "semantics" and "grammar" of viral sequence composition is more reasonable. After completing the first step of training, we can

already calculate the "semantic" changes of the Spike protein of each input virus genome sequence as defined in (see supplemental note 4) and select the target sequences according to the CSCS defined in (*5*) for the second step training.

The second step of training is the dynamic fine-tuning process. The occurrence of virus "immune escape" mutations are the result of selection. Pre-training can also indirectly sense the evolution of a virus over time. It is certain that if the evolution direction of the virus over time is highlighted in the training, it will be able to get a stronger "selection" signal and can more accurately identify and predict "immune escape" mutations. According to the time arrow, we take the three adjacent virus genome sequence matrices $\mathbf{X_{t-1}}$, $\mathbf{X_t}$, $\mathbf{X_{t+1}}$, the five selected sequences in $\mathbf{X_{t+1}}$ with the largest CSCS scores used as the training targets, and then the three spatially correlated sequences （$\mathbf{x}_{(\mathrm{t,p-1})}$, $\mathbf{x}_{(\mathrm{t,p})}$, $\mathbf{x}_{(\mathrm{t,p+1})}$） are selected according to the same pre-training mode in $\mathbf{X_t}$, and finally in $\mathbf{X_{t-1}}$, select a sequence $\mathbf{x}_{(\mathrm{t-1,p})}$ with the closest evolution distance to the sequence $\mathbf{x}_{(\mathrm{t,p})}$ in $\mathbf{X_t}$, when calculating the pairwise co-attention mechanism between （$\mathbf{x}_{(\mathrm{t,p-1})}$, $\mathbf{x}_{(t,p)}$, $\mathbf{x}_{(t,p+1)}$）, while computing the co-attention mechanism between **x**(t-1,p) and **x**(t,p). The sequence input of the dynamic fine-tuning process is sequentially expanded in Spatio-temporal order, from the outbreak to the present. Finally, we obtained a trained model CoT2G-F that can accurately reflect the "mutation + selection" mechanism during virus evolution, can predict and identify "immune escape" mutations, and then compute R_0 (see supplemental note 4).

"Immune escape" mutations prediction and R_0 inference

As in (*4*), our model CoT2G-F can construct a semantic and grammatical "embedded" representation for given virus genome sequences and even has significant advantages to extract the upstream and downstream long-range associated variation information of the sequence itself and obtain the epistatic effects of "immune escape" mutations and recombination signals of virus strains. The model identifies and predicts virus "immune escape" mutational patterns by the learned semantic and grammatical rules. We further prove that the basic reproduction number (R_0) or a fold increase in relative fitness of a virus lineage over the Wuhan lineage as a macro-observable variable is a mathematical expectation of a latent variable function according to a micro-hidden state probability distribution (see supplemental note 4).

When computing R_0, we find this microscopic latent variable function is what we define as the conditional semantic change score (CSC). In our model, the CSC score counts the semantic change of the embedded virus sequence, reflecting the immune escape ability of the virus strain (see supplemental note 4). We applied CoT2G-F to all publicly available SARS-CoV-2 spike protein sequences following the Pango lineage designation and assignment to identify and predict "immune escape" mutations and calculate the per-lineage fitness of the virus without introducing additional epidemiological parameters and then obtain the R_0 of the virus lineage. The introduced conditional semantic change score CSC=$\|\hat{\mathbf{z}}-\hat{\mathbf{z}}[\tilde{x}_i]\|_{\ell_x}$, $\|\cdot\|_{\ell_x}$ is the ℓ_x norm, x = 1, 2 (see supplemental note 1), in which $\hat{\mathbf{z}}=\frac{1}{N}\sum_{i=1}^{N}\hat{\mathbf{z}}_i$. Considering the

history of virus evolution and the reference of the relative fitness, we selected the virus sequence Wuhan-Hu-1 [GenBank: MN908947.3] as a reference corresponding to the latent variable $\hat{\mathbf{z}}$ to align and normalize the CSC score (see supplemental note 4).

Supplementary Text

Supplemental Note 1: The formal expression of fitness in Encoder-Decoder Seq2Seq framework and the P-E theorem

1. Mathematical description of the average and per-lineage fitness of the virus population

From the basic principles and definitions of population genetics (*29*), it can be known that the mean fitness $\bar{\omega}$ of the virus population (the total number of viruses is N) containing K virus lineages can be expressed by the following formula:

$$\bar{\omega} = \lambda_1\omega_1 + \lambda_2\omega_2 + \ldots\ldots + \lambda_K\omega_K = \sum_{k=1}^{K}\lambda_k\omega_k = \boldsymbol{\lambda}^T\boldsymbol{\omega} \quad (1)$$

Here, $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots\ldots, \lambda_K)$, λ_k is the occurrence frequency of the k_{th} virus lineage in the whole population, which is an observable measure related to amino acid substitution characteristics, and $\boldsymbol{\omega}=(\omega_1, \omega_2, \ldots\ldots, \omega_K)$, ω_k is the fitness of the k_{th} virus lineage. Apparently, $\boldsymbol{\omega}=\boldsymbol{\omega}(\mathbf{x})$ is a very complex high-order function of a random variable $\boldsymbol{x}$ that describes the variation state of a virus genome sequence sites, and $\boldsymbol{x} = (x_1, x_2, \ldots, x_i, \ldots, x_L)$ represent L variation sites. Thus, each ω_k can involve the epistatic interaction of multiple mutation sites in the input sequence (*7*) as following:

$$\omega_k(\boldsymbol{x}) =$$

$$=a^{(0)}+\sum a_i^{(1)} x^k{}_i +\sum a_{ij}^{(2)} x^k{}_i x^k{}_j +\sum a_{ijk}^{(3)} x^k{}_i x^k{}_j x^k{}_k +\cdots+\sum a_{1,2,\ldots L}^{(L)} x^k{}_1 x^k{}_2 \cdots x^k{}_L \quad (2)$$

According to the above definitions, $\lambda_k\omega_k$ is a component of the fitness of the k_{th} virus lineage. Given we take the Wuhan lineage as a reference [Wuhan-Hu-1, GenBank: MN908947.3], ω_k can also regard as the relative fitness of the k_{th} virus lineage. Usually we also call ω_k as the basic reproduction number (R_0) of the k_{th} virus lineage. The basic definition of the average fitness given by formula (1) is still valid for virus lineages or sub-lineages with a sufficient number of viruses.

On the other hand, we can define a function $P(\boldsymbol{x})$ as a state probability distribution of the sequence's variation states in real space. This variation causes changes in viral fitness. The $P(\boldsymbol{x})$ can have an expanded form using a Gaussian Mixture Model (GMM).

$$P(\mathbf{x}) = \sum_{k=1}^{K} \pi_k N(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (3)$$

$\pi_k N(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is the k_{th} component in the mixture model, π_k is the mixture coefficient, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the mean and variance values of the relative fitness. These coefficients π_k, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$ can obtain indirectly through the EM algorithm or deep neural network.

Combining formulas (1) and (3), the formal representation of the average fitness of the virus population can obtain as follows:

$$\bar{\omega} = E(\omega) = \int \boldsymbol{\omega}(\boldsymbol{x}) P(\boldsymbol{x}) d\boldsymbol{x} = \sum_{k=1}^{K} \pi_k \bar{\omega}_k \quad (4)$$

According to formula (3) and by the EM algorithm: $\pi_k = N_k/N$, $\sum_{k=1}^{K} \pi_k = 1$, N_k is the number of viruses in the k_{th} virus lineage, N is the total number of the virus

population. Here, π_k is equal to λ_k, and $\bar{\omega}_k$ is the fitness of the k_{th} virus lineage. And we get formula (1) again.

$$\sum_{k=1}^{K} \pi_k \bar{\omega}_k = \sum_{k=1}^{K} \lambda_k \bar{\omega}_k$$

The per-lineage fitness $\bar{\omega}_k$ in real space can be derived when repeating the processes (3) and (4) for the whole individual strain of the k_{th} virus lineage

$$\sum_{k=1}^{K} \pi_k \bar{\omega}_k = \sum_{k=1}^{K} \pi_k \left(\frac{1}{n_k} \sum_{i=1}^{n_k} \omega_i^k\right)$$

and

$$\bar{\omega}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \omega_i^k \qquad (5)$$

Formula (5) is consistent with the definition by F. Obermeyer et al in their work (*5*). According to the basic principles of population genetics (*29*) that the fitness of a virus lineage can be considered to be determined by the largest virus cluster in the lineage. Let λ_k be the occurrence frequency of the dominant cluster in the k_{th} virus lineage when sampling virus strains, we can redefine the per-lineage fitness through the contribution of the dominant virus cluster of a lineage as follows:

$$\bar{\omega}_k \stackrel{\text{def}}{=} \frac{\lambda_k}{n_k} \sum_{i=1}^{n_k} \omega_i^k \qquad (6)$$

Because of $\omega_i^k \geq 0$, formula (6) can be rewritten as follows for the logical needs of the following proof:

$$\bar{\omega}_k = \frac{\lambda_k}{n_k} \sum_{i=1}^{n_k} \omega_i^k = \frac{\lambda_k}{n_k} \sum_{i=1}^{n_k} |\omega_i^k| = |\omega_k| \qquad (7)$$

$|\omega_i^k|$ represents the contribution of the sampled virus strain to the fitness of the k_{th} virus lineage. Now the problem of finding the per-lineage fitness is transformed into calculating the variable $|\omega_i^k|$.

2. Fitness calculation based on the Bayes' theorem and deep learning model of the Encoder-Decoder Seq2Seq architecture

According to the definition of the virus fitness (1) , formula (6) and (7),

$$\bar{\omega}_k = \frac{\lambda_k}{n_k}\sum_{i=1}^{n_k} |\omega_i^k|$$

is the per-lineage fitness of the virus population and a mathematical expectation of the contribution of a single virus strain to the fitness. Referring to formula (3), let $\boldsymbol{\omega}(\mathbf{x}) = \boldsymbol{\omega}_{\boldsymbol{k}}(\mathbf{x})$, $P(\mathbf{x}) = N(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, the per-lineage fitness of the virus can express formally again as:

$$\bar{\omega}_k = \bar{\omega}_{|P(x)} = \int \boldsymbol{\omega}(\mathbf{x})\, P(\mathbf{x})\, d\mathbf{x} \tag{8}$$

The subscript P (·) represents different spaces corresponding to k. As long as we find a suitable functional form of $\boldsymbol{\omega}(\mathbf{x})$ and a corresponding probability distribution $P(\mathbf{x})$, we can obtain the per-lineage fitness of viruses by calculating their mathematical expectations by the formula (8), then get the average fitness of the virus population by formula (4). We can now try to do this within the context of deep learning, despite the fact that it is typically difficult to obtain and calculate $\boldsymbol{\omega}(\mathbf{x})$ and $P(\mathbf{x})$ directly from real-world data with noise.

Let $Q_\theta(\mathbf{z}|\mathbf{x})$ and $P(\mathbf{z})$ be the hidden state probability distribution in encoder-decoder spaces, and $P(\mathbf{x})$ and $P_\varphi(\mathbf{x}|\mathbf{z})$ be the prior and posterior probability related to the input and output of the model. The relationship among $P(\mathbf{x})$, $Q_\theta(\mathbf{z}|\mathbf{x})$, $P_\varphi(\mathbf{x}|\mathbf{z})$ and $P(\mathbf{z})$ is given by Bayes' theorem:

$$P(\mathbf{x})Q_\theta(\mathbf{z}|\mathbf{x}) = P_\varphi(\mathbf{x}|\mathbf{z})P(\mathbf{z}) \tag{9}$$

The function relationship defined by formula (1) still holds when $P(\mathbf{x})Q_\theta(\mathbf{z}|\mathbf{x})$ maps to $P_\varphi(\mathbf{x}|\mathbf{z})P(\mathbf{z})$. When the prior probability $P(\mathbf{x})$ is unknown, obtaining the fitness of virus lineages in the encoder or decoder space is transformed into finding the hidden state probability distributions $Q_\theta(\mathbf{z}|\mathbf{x})$ and $P(\mathbf{z})$, the posterior probability distribution $P_\varphi(\mathbf{x}|\mathbf{z})$ and the counterpart $\boldsymbol{\omega}(\mathbf{z})$ of $\boldsymbol{\omega}(\mathbf{x})$.

Under the Encoder-Decoder Seq2Seq framework together with Bayes' theorem and Universality Theorem (*30*), because

$$\boldsymbol{x} \quad \rightarrow \quad \boldsymbol{x}' = W\boldsymbol{z} + \boldsymbol{B}$$

we have

$$\int \boldsymbol{\omega}(\mathbf{x})\, P(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int \boldsymbol{\omega}(W\boldsymbol{z})\, P(W\boldsymbol{z})W\mathrm{d}\boldsymbol{z}$$

$$= \int \boldsymbol{\omega}(\boldsymbol{z})\, P(\boldsymbol{z})(P_\varphi(\mathbf{x} \mid \mathbf{z})/Q_\theta(\mathbf{z} \mid \mathbf{x}))\mathrm{d}\boldsymbol{z}$$

Based on a standard Encoder-Decoder Seq2Seq architecture,

$$P_\varphi(\mathbf{x} \mid \mathbf{z})/Q_\theta(\mathbf{z} \mid \mathbf{x}) = W$$

Here, the ratio of the two likelihoods appears as the transformation coefficient. W is a learnable embedding representation constant matrix of **x**, therefore, for the per-lineage fitness we have

$$\overline{\omega}_{|P(\boldsymbol{x})} = \int \boldsymbol{\omega}(\mathbf{x})\, P(\mathbf{x})\, \mathrm{d}\mathbf{x} = \int W\boldsymbol{\omega}(\boldsymbol{z})\, P(\boldsymbol{z})\, \mathrm{d}\boldsymbol{z}$$

Considering that the average of states is by virus lineage, the fitness of a virus lineage is determined by the dominant virus cluster in the lineage，and then by doing a simple matrix operation we can get

$$W\boldsymbol{\omega}(\boldsymbol{z}) \rightarrow \lambda\boldsymbol{\omega}(\boldsymbol{z})$$

and λ is a diagonal matrix. Then, we have

$$\bar{\omega}_{|\mathrm{P}(\boldsymbol{x})} = \int \boldsymbol{\omega}(\mathbf{x})\,\mathrm{P}(\mathbf{x})\,\mathrm{d}\mathbf{x} = \int W\boldsymbol{\omega}(\boldsymbol{z})\,\mathrm{P}(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z} = \int \lambda\boldsymbol{\omega}(\boldsymbol{z})\,\mathrm{P}(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z} \qquad (10)$$

Obviously, formula (10) is the corresponding formal representation of formula (1) under the Encoder-Decoder Seq2Seq framework of deep learning, then, let $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots\ldots, \lambda_{\mathrm{K}})$ is a vector, which components are made up of diagonal elements of the diagonal matrix λ. Finally we get

$$\bar{\omega}_{|\mathrm{P}(\boldsymbol{x})} = \boldsymbol{\lambda}\bar{\boldsymbol{\omega}}_{|\mathrm{P}(\boldsymbol{z})} \qquad (11)$$

Because of the hierarchical structure of the population, lineage and cluster of the virus, and referring to the formula(1), (3), (4), (6) and (10) and based on the basic principles of population genetics (*29*), the elements of the vector $\boldsymbol{\lambda}$ are the corresponding elements of mixture coefficient in formula (6). The element λ is an occurrence frequency of a dominant cluster in virus lineage when sampling virus strains, and is a macroscopic parameter that links the hidden space with the actual space.

Under Encoder-Decoder Seq2Seq framework, $\mathrm{P}(\boldsymbol{z})$ is a prior probability distribution of hidden space. Now our task is to find a $\boldsymbol{\omega}(\boldsymbol{z})$ corresponding to the $\boldsymbol{\omega}(\boldsymbol{x})$ in hidden space.

Based on formula (10), we know that $\boldsymbol{\omega}(\boldsymbol{x})$ and $\boldsymbol{\omega}(\boldsymbol{z})$ belong to two different spaces. One is the real-world space, where the input is the mutation information of the virus sequence $\boldsymbol{x}$. The $\boldsymbol{\omega}(\boldsymbol{x})$ is a very complex high-order function involving the epistatic interaction of multiple mutation sites in the input sequence (see formula (2)) However, based on the embedded mode of deep learning, the learning mechanism of fully connected network linear transformation and dimension reduction characteristics, the $\boldsymbol{\omega}(\boldsymbol{z})$ in hidden space can be a low-order simple quasi-linear function of a latent variable $\boldsymbol{z}$ to achieve accurate fitness calculation. Therefore, the formula (10) and (11) show that the fitness of the virus lineage can be computed as a mathematical expectation according to a hidden state probability distribution under the Encoder-Decoder Seq2Seq framework if we can find a reasonable embedded representation of the related microscopic genome sequence variation.

In this way, to calculate the fitness of the virus population, according to formulas (10) and (11), we need to establish a Seq2Seq deep learning model under the Encoder-Decoder Seq2Seq framework to realize our calculation of the fitness of the virus population.

Considering that the fitness of the virus population is an observable macrobiological phenotype, and the genome sequence variation is the relevant microscopic genotype, the formulas (10) and (11) can further lead to the following Phenotype-Embedding (P-E) theorem:

The P-E theorem: "An observable macro-biological phenotype can be computed under the Encoder-Decoder Seq2Seq framework if we can find a reasonable embedded representation of the related microscopic genotype."

Amazingly, the P-E theorem may tell us how the human brain works. When we receive the signals corresponding to $\overline{\omega}_{|P(\boldsymbol{x})}$, these signals are input into the brain through our senses. The learning knowledge stored in neurons of the brain will embed these signals and convert them into latent variables, which will trigger some mechanism to transform the state of these latent variables through a specific function $\boldsymbol{\omega}(\boldsymbol{z})$ to form a corresponding structure. A stable representation of the mathematical expectation of this function, then output as $\boldsymbol{\lambda}\overline{\boldsymbol{\omega}}_{|P(\boldsymbol{z})}$, and then we image $\overline{\omega}_{|P(\boldsymbol{x})}$.

<u>Supplemental Note 2: The fulcrum of building model system and the introduction of CSC scores</u>

Our goal is to find virus mutations that cause a high degree of semantic change but are grammatically acceptable, from which to identify and predict "immune escape" mutations, and then calculate the fitness of virus lineages. Because of B. Hie's pioneering work [4], and the need for a clear and similar pivot to construct our theoretical system, in addition to some modifications required by our model, we adopted the basic formulation of DNA base or amino acid embedding given in (*4*), and it is paraphrased as follows:

For a virus sequence defined as $\mathbf{x} \stackrel{\text{def}}{=} (x_1, \dots, x_N)$ such that $x_i \in \mathcal{X}$, $i \in [N]$, where $\mathcal{X}$ is a finite set of characters, $\mathbf{x} \in V$, V is virus population set and $\mathbf{x}_k \in V_k$, V_k is virus lineage set. Let $\tilde{x}_{\{i\}} = (\tilde{x}_{i-\lambda+1}, \dots, \tilde{x}_i, \dots, \tilde{x}_{i+\lambda-1})$ denote mutations around

position i and the mutated sequence is $\mathbf{x}[\tilde{x}_{i-\lambda+1}, \ldots, \tilde{x}_{i+\lambda-1}] \stackrel{\text{def}}{=} (x_1, \ldots, x_{i-\lambda}, \tilde{x}_{i-\lambda+1}, \ldots, \tilde{x}_i, \ldots, \tilde{x}_{i+\lambda-1}, x_{i+\lambda}, \ldots, x_N)$, here, 1<=λ<=Length(protein sequence) $*15\%$ and for SARS-CoV-2 spike protein $\bar{\lambda}$=3. The syntax for a mutated sequence $\mathbf{x}$ is described as a probability distribution function:

$$p(\tilde{x}_{\{i\}}|\mathbf{x}) \tag{12}$$

For brevity, we will use $\mathbf{x}[\tilde{x}_{\{i\}}]$ instead of $\mathbf{x}[x_1, \ldots, x_{i-\lambda}, \tilde{x}_{i-\lambda+1}, \ldots, \tilde{x}_i, \ldots, \tilde{x}_{i+\lambda-1}, x_{i+\lambda}, \ldots, x_N]$ in the following. Then, $p(\tilde{x}_{\{i\}}|\mathbf{x})$ is close to zero if $\mathbf{x}[\tilde{x}_{\{i\}}]$ does not conform to the grammar rules, and $p(\tilde{x}_{\{i\}}|\mathbf{x})$ is close to one if $\mathbf{x}[\tilde{x}_{\{i\}}]$ conforms to the grammar rules. In a language composed of a given sequence, violation of these grammatical rules results in a loss of grammaticality of the sequence. To establish the correlation between the above definition and the natural language model, a latent variable distribution function $\hat{p}$ and the mapping $\hat{f}_s$ can be introduced, and for all $i \in [N]$, let the latent variable $\hat{\mathbf{z}}_{\{i\}}$ is encoded from the context $\mathbf{x}_{[N]\setminus\{i\}} \stackrel{\text{def}}{=} (\ldots, x_{i-\lambda}, x_{i+\lambda}, \ldots)$, in which the viral genome mutates at position $(x_{i-\lambda+1}, \ldots, x_{i+\lambda-1})$. In our model CoT2G-F, $\tilde{x}_{\{i\}}$ in $\mathbf{x}_{[N]\setminus\{i\}}$ is randomly masked, and then in the original position $\tilde{x}_{\{i\}}$ re-gives the character according to the "grammar rules" learned by the model, besides, the multiple consecutive sites ($\lambda > 1$) have been masked and they are treated as a "span" and a single unique mask token is used to replace the entire span. Thus, we have

$$\hat{p}(x_{\{i\}}|\mathbf{x}_{[N]\setminus\{i\}}, \hat{\mathbf{z}}_{\{i\}}) = \hat{p}(x_{\{i\}}|\hat{\mathbf{z}}_{\{i\}}) \quad \text{and} \quad \hat{\mathbf{z}}_{\{i\}} = \hat{f}_s(\mathbf{x}_{[N]\setminus\{i\}})$$

We define:

$$\Delta\mathbf{z}[\tilde{x}_{\{i\}}] \stackrel{\text{def}}{=} \|\hat{\mathbf{z}} - \hat{\mathbf{z}}[\tilde{x}_{\{i\}}]\|_{\ell_x} \tag{13}$$

and $\hat{p}(x_i|\hat{\mathbf{z}}_{\{i\}})$ as the mapping representation of $p(\tilde{x}_{\{i\}}|\mathbf{x})$ into the encoding hidden state space under our model, and $p(\tilde{x}_{\{i\}}|\mathbf{x}')$ is the mapping representation of $\hat{p}(\tilde{x}_{\{i\}}|\hat{\mathbf{z}}_{\{i\}})$ in the decoding recovery state space and $\mathbf{x}'$ is the output sequence of our Se2Seq model,

$$\text{CoT2G} - \text{F encoder: } p(\tilde{x}_{\{i\}}|\mathbf{x}) \rightarrow \hat{p}(\tilde{x}_{\{i\}}|\hat{\mathbf{z}}_{\{i\}})$$

$$\text{CoT2G} - \text{F decoder: } \hat{p}(\tilde{x}_{\{i\}}|\hat{\mathbf{z}}_{\{i\}}) \rightarrow p(\tilde{x}_{\{i\}}|\mathbf{x}') \tag{14}$$

Here (*4*) defines

$$\hat{\mathbf{z}} \stackrel{\text{def}}{=} \frac{1}{N}\sum_1^N \hat{\mathbf{z}}_{\{i\}},$$

which is the average of semantic embeddings at all positions. Formula (2) shows that $\hat{\mathbf{z}}$ is a reference value for embedded "semantic" changes, and what virus lineage is chosen as a reference has biological significance. The Wuhan lineage was selected as a reference in this work. $\hat{\mathbf{z}}[\tilde{x}_{\{i\}}]$ is the sequence embedding that does not contain the variant site $x_{\{i\}}$. $\|\cdot\|_{\ell_x}$ is the ℓ_x norm, x = 1, 2, and $\hat{p}(\tilde{x}_{\{i\}}|\hat{\mathbf{z}}_{\{i\}})$ can be obtained through model training.

In our model CoT2G-F (see supplemental note 2)

$$\hat{\mathbf{z}}_{\{i\}} \stackrel{\text{def}}{=} \text{CoT2G} - \text{F}_{\hat{f}_s}(\mathbf{x}[\tilde{x}_{\{i\}}])$$

When there are multiple mutations in the sequence, as defined

$$\mathcal{M}(\mathbf{x}, \tilde{\mathbf{x}}) \stackrel{\text{def}}{=} \{\tilde{x}_{\{i\}}|\tilde{x}_{\{i\}} \neq x_{\{i\}}\}$$

According to formula (3), the hidden state probability distribution function of the decoding space is rewritten as

$$P_\varphi(\mathbf{x}|\hat{\mathbf{z}}) = P_\varphi(\mathbf{x}[\tilde{x}_{\{i\}}]|\hat{\mathbf{z}}) = \prod_{\tilde{x}_{\{i\}} \in \mathcal{M}(x,\tilde{x})} \hat{p}(\tilde{x}_{\{i\}}|\hat{\mathbf{z}}) = \prod_{\tilde{x}_{\{i\}} \in \mathcal{M}(x,\tilde{x})} p(\tilde{x}_{\{i\}}|\mathbf{x}')$$

In our CoT2G-F model, we can further get

$$P_{\varphi}(\mathbf{x}|\hat{\mathbf{z}}) \stackrel{\text{def}}{=} \prod_{\tilde{x}_{\{i\}} \in \mathcal{M}(x,\tilde{x})} \text{softmax}(\mathbf{W}(\varphi)\hat{\mathbf{z}}[\tilde{x}_{\{i\}}] + \mathbf{b}(\varphi)).$$

Where $\mathbf{W}$ and $\mathbf{b}$ are both the parameters the model needs to learn and $\hat{\mathbf{z}}[\tilde{x}_{\{i\}}]$ obey a hidden state probability distribution. In this way, when judging semantic and grammatical changes and identifying and predicting "immune escape" mutational patterns, the "semantic" change $\Delta\mathbf{z}\ [\tilde{x}_{\{i\}}]$ remains unchanged, and the rules from (*4*) still hold:

$$\alpha'(\tilde{x}_{\{i\}}; \mathbf{x}) \stackrel{\text{def}}{=} \text{rank}(\mathbf{\Delta z}[\tilde{x}_{\{i\}}]) + \beta \text{rank}(P_{\varphi}(\mathbf{x}|\hat{\mathbf{z}})) \tag{15}$$

The distance $\mathbf{\Delta z}[\tilde{x}_{\{i\}}]$ in the embedding space describes the semantic change, and the hidden state probability distribution $P(\hat{\mathbf{z}})$ and the related posterior probability $P_{\varphi}(\mathbf{x}|\hat{\mathbf{z}})$ gives the grammatical rules that constrain the sequence composition to be grammatical.

Based on the principles of evolutionary biology, fitness is a concept of change. If the genome does not change, there is no fitness change. Therefore, the "grammaticality" of the genome composition will not directly affect the fitness of the virus. If the genome composition is already following "grammaticality", only the "semantic" changes corresponding to the sequence variations affecting the genome function contribute to the fitness of the virus. Referring to the study of (*4*), we think that the first term $\mathbf{\Delta z}[\tilde{x}_{\{i\}}]$ on the right side of equation (4) is a variable related to the fitness of the virus under the condition that its sequence composition is grammatical. $\mathbf{\Delta}[\tilde{x}_{\{i\}}]$ can also regard as a conditional "semantic" change（CSC =$\mathbf{\Delta z}[\tilde{x}_{\{i\}}]$）score. Late we will show that the per-lineage fitness can derive from the measure CSC with a hidden state distribution function $P(\mathbf{z})$ and $P_{\varphi}(\mathbf{x}|\hat{\mathbf{z}})$ in the decoder space (see

supplemental notes 3 and 4).

Supplemental Note 3: Detailed description of the CoT2G-F model

For our CoT2G-F model, we use a vanilla encoder-decoder Transformer (*12*), which has emerged as a powerful general-purpose language model architecture for representation learning and generative modelling, outperforming recurrent and convolutional architectures in natural language settings.

We take as input amino acid character sequences. Protein sequences are composed of amino acid characters and natural language sequences are composed of words that can both be processed in language models based on different vocabulary. Researchers (*4*, *31*, *32*) have demonstrated that protein language models trained in diverse sequences can learn the structure and function of proteins, the metric structure in the protein representation space that accords with organizing principles at scales from physicochemical to remote homology.

Language mode is commonly trained with two steps: pre-train and fine-tune, which corresponds to neutral evolution and environmental selection during virus evolution. The first stage is pre-train, where a high-capacity language model is trained on a large amount of unlabeled data, and such a language model can capture general semantic and syntactic information from sequence data alone. Then for the second fine-tuning stage, models apply the first stage pre-trained language model to a particular task, which means a task-specific input and objective are introduced. In other words, the model is fine-tuned on the task data from pre-trained parameters. In our model's pre-training step, we not only leverage the 300 million GISAID database

of virus sequences but also introduce a co-attention mechanism to capture the spatial correlations of virus evolution. The self-attention mechanism in the vanilla Transformer only focuses on the internal context information of the sequence itself. The query, key, and value in the attention function are all from the sequence itself. By contrast, in our CoT2G-F model, the co-attention mechanism allows our model to introduce Spatio-temporal neighbour sequences that participate in the key and value of the attention function. Then the learned representation from pre-training has

- a multi-scale organization reflecting structure from the level of biochemical properties of amino acids to the remote homology of proteins.
- a spatial correlation reflecting the sequence structure composition rules of virus genomes.

Both properties contribute to our downstream tasks: identify the immune evasion sites of viruses and calculate their relative fitness only from sequence information.

Pre-training is done using a denoising objective, where some parts of the amino acid characters are randomly replaced with a special token, and the model needs to predict the identity of those amino acid characters, which is also called the "masked language modelling" objective. Inspired by BERT's (*33*) "masked language modelling" objective and the "word dropout" regularization technique (*34*), we design an objective that randomly samples and drops out 15% of consecutive amino acid characters in the input protein sequence, during the pre-train step, the multiple consecutive amino acid characters have been masked and they are treated as a "span" and a single unique mask token is used to replace the entire span. All spans of

discarded amino acid characters are replaced by a single marker, each assigned a mask token ID that is unique to the sequence of amino acid characters. These mask token IDs are special tokens which are added to our new vocabulary, we choose to mask contiguous spans of amino acid characters and only predict discarded spans that can reduce the computational cost of pre-training. The loss function to predict the missing amino acid characters from the corrupted sequence is followed:

$$\mathcal{L}_{\mathcal{M}ask} = \sum_{\mathrm{M}\in\mathrm{N}} -\log p\left(\mathbf{x}_{\mathrm{M}/\mathrm{N}} \mid \mathbf{x}_{\mathrm{N}/\mathrm{M}}\right) \tag{16}$$

For each sequence x, sample a set of indices N contiguous spans of amino acid characters to mask, replacing the true amino acid characters with the mask token IDs. We minimize the negative log-likelihood of the unmasked amino acid characters $\mathrm{x}_{\mathrm{M}/\mathrm{N}}$ given the masked sequence $\mathrm{x}_{\mathrm{N}/\mathrm{M}}$ as context for every masked token ID.

To discover the implicit alignments between two spatial contiguous sequences of protein sequences, we introduce a co-attention mechanism. For each sequence $\mathbf{x}_{(\mathrm{t,p})}$ at t month slice, $\mathbf{x}_{(\mathrm{t,p-1})}$ and $\mathbf{x}_{(\mathrm{t,p+1})}$ is the spatial pre-order and follow-up contiguous sequences of $\mathbf{x}_{(\mathbf{t,p})}$, [:] means concatenate two sequences, so the

$$\mathcal{L}_{\mathcal{P}re-\mathcal{M}ask} = \sum_{\mathrm{M}\in\mathrm{N}} -\log p\ \left(\left[\mathrm{x}_{(\mathrm{t,p})}:\mathrm{x}_{(\mathrm{t,p-1})}\right]_{\mathrm{M}/\mathrm{N}} \middle| \left[\mathrm{x}_{(\mathrm{t,p})}:\mathrm{x}_{(\mathrm{t,p-1})}\right]_{\mathrm{N}/\mathrm{M}}\right)$$

$$\mathcal{L}_{\mathcal{F}ollow-\mathcal{M}ask} = \sum_{\mathrm{M}\in\mathrm{N}} -\log p\ \left(\left[\mathrm{x}_{(\mathrm{t,p})}:\mathrm{x}_{(\mathrm{t,p+1})}\right]_{\mathrm{M}/\mathrm{N}} \middle| \left[\mathrm{x}_{(\mathrm{t,p})}:\mathrm{x}_{(\mathrm{t,p+1})}\right]_{\mathrm{N}/\mathrm{M}}\right) \tag{17}$$

The pre-training is done using a combination of pre-order denoising objective and follow-up denoising objective, to speed up the convergence of the model, we adopted a staged training strategy, i.e. our method first set $\alpha_{\mathcal{P}re-\mathcal{M}ask} = 1$, $\alpha_{\mathcal{F}ollow-\mathcal{M}ask} = 0$, then set $\alpha_{\mathcal{P}re-\mathcal{M}ask} = 0$, $\alpha_{\mathcal{F}ollow-\mathcal{M}ask} = 1$ and continue training on the checkpoint of the previous model.

$$\mathcal{L}_{Pre-train} = \alpha_{Pre-Mask} * \mathcal{L}_{Pre-Mask} + \alpha_{Follow-Mask} * \mathcal{L}_{Follow-Mask} \quad (18)$$

In the fine-tuning step, we follow the Constrained Semantic Change Search (CSCS) criterion proposed by B. Hi (*4*), CSCS is a sorting algorithm that sorts the semantic and grammatical changes of a base sequence to multiple sequences that may be mutated and find the mutation sequences that are most prone to 'immune escape' from highest to lowest. We denote such a base sequence as $\mathbf{x}$ and the mutated sequences as $\tilde{\mathbf{x}} = (\tilde{x}_1, \ldots, \tilde{x}_N)$, we assume there are N mutant sequences, for one mutant sequence $\tilde{\mathbf{x}}_j$, which consists of amino acid characters in $\tilde{\mathbf{x}}_{ji}$ that disagree with those at the same position in the base sequence $\mathbf{x}$, which we denote as

$$\mathcal{M}(\tilde{\mathbf{x}}_j, \mathbf{x}) \stackrel{\text{def}}{=} (\tilde{\mathbf{x}}_{ji} \mid \tilde{\mathbf{x}}_{ji} \neq \mathbf{x}_i)$$

All possible mutations $\tilde{Y}$ are given priority based on the corresponding values of CSCS, from highest to lowest.

$$\boldsymbol{CSCS} = \mathcal{S}_{rank}(\tilde{\mathbf{x}}_j; \mathbf{x}) + \beta * \mathcal{G}_{rank}(\tilde{\mathbf{x}}_j; \mathbf{x}) \quad (19)$$

$$= \boldsymbol{CSC} + \beta * \mathcal{G}_{rank}(\tilde{\mathbf{x}}_j; \mathbf{x})$$

where $\mathcal{S}_{rank}(\tilde{\mathbf{x}}_j; \mathbf{x})$ and $\mathcal{G}_{rank}(\tilde{\mathbf{x}}_j; \mathbf{x})$ denote semantic and grammatical rank respectively, the consistently well-calibrated around β =1 equally weighting both semantic and grammatical change), which we used in our experiments.

The semantic change is measured by the following:

$$\mathcal{S}_{rank}(\tilde{\mathbf{x}}_j; \mathbf{x}) = \left|\left|\hat{\mathbf{z}} - \hat{\mathbf{z}}[\tilde{\mathbf{x}}_j]\right|\right|_1$$

Where $\hat{\mathbf{z}}$ is the hidden states of the Transformer's last decoder block and $||\cdot||_1$ is the ℓ_1 norm.

The distance in the embedding space encodes semantic changes, while

probabilistic emanations encode grammatical changes:

$$\mathcal{G}_{rank}(\tilde{\mathbf{x}}_j;\mathbf{x}) = \prod_{\mathcal{M}(\tilde{\mathbf{x}}_j;\mathbf{x})} \mathrm{p}(\tilde{\mathbf{x}}_{ji} \mid \mathbf{x}_{(j,[0:i-1])},)$$

where p is the mutation probability that the i site corresponds to the $\tilde{\mathbf{x}}_{ji}$ is mutated to the amino acid corresponding to the i site in $\mathbf{x}$, and the grammatical change is equal to the product of the probabilities of the individual mutations.

Following the above CSCS criterion, we perform protein sequences set slicing monthly, select the virus strain with the largest semantic change but conform to the grammar in the next month, these selected protein sequences can "immune escape" most likely, from which the labelled sequence-to-sequence dataset during fine-tune step is constructed. The objective of fine-tuning step is as follows:

$$\mathcal{L}_{Seq} = \sum_{i=1}^{|\tilde{x}|} -\log q(\tilde{\mathbf{x}}_i \mid \mathbf{x}, \tilde{\mathbf{x}}_{<i})$$

For each sequence $\mathbf{x}$ at one month slice, the model needs to generate a list of target sequences, donated as $\tilde{\mathbf{x}}$, which is the virus strain that may emerge "immune escape" mutations in the next month, the $|\tilde{\mathbf{x}}|$ means all sites of $\tilde{\mathbf{x}}$, and due to the autoregressive property of the decoder, when predicting $\tilde{\mathbf{x}}_i$, the sites before the i site can also be introduced, i.e. $\tilde{\mathbf{x}}_{<i}$. We need to independently minimize the negative log-likelihood of the amino acid $\tilde{\mathbf{x}}_i$ given the $\mathbf{x}$ and $\tilde{\mathbf{x}}_{<i}$ as context. Then based on pre-trained language model parameters, iterative training and dynamic inference of immune evasion strains are performed simultaneously, thus, a temporal correlation reflecting the selection mechanism is introduced at this fine-tuning step.

$$\mathcal{L}_{SP} = \sum_{i=1}^{|\tilde{x}|} -\log q\left(\tilde{\mathbf{x}}_i \mid \left[\mathbf{x}_{(tp)}:\mathbf{x}_{(tp-1)}\right], \tilde{\mathbf{x}}_{<i}\right)$$

$$\mathcal{L}_{\mathcal{SF}} = \sum_{i=1}^{|\tilde{x}|} -\log q\left(\tilde{\mathbf{x}}_i \,\middle|\, \left[\mathbf{x}_{(t,p)} : \mathbf{x}_{(t,p+1)}\right], \tilde{\mathbf{x}}_{<i}\right)$$

$$\mathcal{L}_{\mathcal{T}} = \sum_{i=1}^{|\tilde{x}|} -\log q\left(\tilde{\mathbf{x}}_i \,\middle|\, \left[\mathbf{x}_{(t,p)} : \mathbf{x}_{(t-1,p)}\right], \tilde{\mathbf{x}}_{<i}\right)$$

We endow the language model with the ability to predict new immune escape sites arising from environmental selection and time migration following a neutral evolution event. Fine-tune is done using a combination of three objectives, refer to the staged training strategy mentioned above, we set $\alpha_{\mathcal{SP}} = 1$, $\alpha_{\mathcal{SF}} = 1$ and $\alpha_{\mathcal{T}} = 1$ successively.

$$\mathcal{L}_{Fine-tune} = \alpha_{\mathcal{SP}} * \mathcal{L}_{\mathcal{SP}} + \alpha_{\mathcal{SF}} * \mathcal{L}_{\mathcal{SF}} + \alpha_{\mathcal{T}} * \mathcal{L}_{\mathcal{T}} \tag{20}$$

<u>Supplemental Note 4: The formal logic of the Genotype-Fitness landscapes and R_0 calculation</u>

1. The "immune escape" ability score CSC

B. Hie et al. proposed a semantic-grammatical criterion CSCS to determine the ability of the virus "immune escape", which is expressed by the following formula (*4*):

$$\text{CSCS} = \alpha\,(\tilde{x}_{\{i\}}\,;\mathbf{x}) \stackrel{\text{def}}{=} \|\hat{\mathbf{z}} - \hat{\mathbf{z}}_{\mathbf{N}\setminus\{\mathcal{M}(\vec{\mathbf{X}},\tilde{\mathbf{X}})\}}\|_{\ell_x} + \beta\hat{p}(\mathbf{x}|\hat{\mathbf{z}}_{\mathbf{N}\setminus\{\mathcal{M}(\vec{\mathbf{X}},\tilde{\mathbf{X}})\}}) \tag{21}$$

From the logic of evolutionary biology, combined with our model,

$$\hat{\mathbf{z}} = \frac{1}{N}\sum_{i=1}^{N} \hat{\mathbf{z}}_{\{i\}}$$

is the mean of all possible "semantic" changes in a sequence of N bases. Formula (1), gives the fundamental definition of the mean fitness of the virus population and the fitness of the virus lineage. If we select the Wuhan lineage as the reference [Wuhan-Hu-1, GenBank: MN908947.3], $\hat{\mathbf{z}}_{\mathbf{N}\setminus\{\mathcal{M}(\vec{\mathbf{X}},\tilde{\mathbf{X}})\}}$ is an absolute semantic change

of a virus strain, and $\|\hat{\mathbf{z}} - \hat{\mathbf{z}}_{\mathbf{N}\setminus\{\mathcal{M}(\vec{\mathbf{X}},\tilde{\mathbf{X}})\}}\|_{\ell_{\mathbf{x}}}$ is a relative semantic change of a virus strain according to the Wuhan virus lineage. In this way, the conditional semantic change score (CSC) under the condition that the sequence composition conforms to the grammar can be defined as:

$$\mathrm{CSC} = \Delta\hat{\mathbf{z}}_{\mathbf{N}\setminus\{\mathcal{M}(\vec{\mathbf{X}},\tilde{\mathbf{X}})\}} = \|\hat{\mathbf{z}} - \hat{\mathbf{z}}_{\mathbf{N}\setminus\{\mathcal{M}(\vec{\mathbf{X}},\tilde{\mathbf{X}})\}}\|_{\ell_{\mathbf{x}}} \tag{22}$$

and we can rewrite (21) as follows

$$\mathrm{CSCS} = \mathrm{CSC} + \beta\hat{\mathrm{p}}(\mathbf{x}|\hat{\mathbf{z}}_{\mathbf{N}\setminus\{\mathcal{M}(\vec{\mathbf{X}},\tilde{\mathbf{X}})\}}) \tag{23}$$

CSC score is the measurement of the "immune escape" ability of the virus strain referring to the Wuhan lineage when the virus mutates. It should note that this is the fundamental starting point of the definition of fitness in evolutionary biology. Here, starting from the basic definition of fitness and homology in population genetics and considering the computational complexity, we adopt the $\boldsymbol{\ell_1}$ norm to define the CSC score. Because our model has better avoiding overfitting characteristics (Fig. 3).

2. The mathematical expectation of the CSC score according to the hidden state probability distribution $\mathbf{P}(\mathbf{z})$ is consistent with the formal expression of the average and the per-lineage fitness of the virus population.

Based onur proposed Encoder-Decoder Seq2Seq architecture model, the encoder module created an embedded state distribution $\mathrm{Q}_\theta(\mathbf{z}|\mathbf{x})$ related to the virus fitness from the input genome sequence $\mathbf{x}$. The decoder module can characterize the "immune escape" of the virus and decode a hidden state probability distribution $\mathrm{P}(\mathbf{z})$ of a latent variable $\mathbf{z}$. $\mathrm{P}(\mathbf{z})$ can also be formally expanded by the Gaussian Mixture Model (GMM) kernel function:

$$P(\mathbf{z}) = \sum_{k=1}^{K} \hat{\pi}_k N\left(\mathbf{z} \mid \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k\right) \tag{24}$$

Where $\hat{\pi}_k N(\mathbf{z} \mid \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k)$ is called the k_{th} component in the mixture model, corresponding to the probability distribution of the k_{th} virus lineage, and $\hat{\pi}_k$ is the mixture coefficient satisfying $\hat{\pi}_k \geq 0$, and $\sum_{k=1}^{K} \hat{\pi}_k = 1$.

Definition

$$\boldsymbol{\Delta} \stackrel{\text{def}}{=} |\hat{\mathbf{z}} - \hat{\mathbf{z}}_{\mathbf{N}\setminus\{\mathcal{M}(\vec{\mathbf{X}},\tilde{\mathbf{X}})\}}|_{\ell_1}, \quad \vec{\mathbf{X}} \in V \tag{25}$$

$$\boldsymbol{\Delta}_{\mathbf{k}} \stackrel{\text{def}}{=} |\hat{\mathbf{z}}^k - \hat{\mathbf{z}}^{\mathbf{k}}_{\mathbf{N}\setminus\{\mathcal{M}(\vec{\mathbf{X}},\tilde{\mathbf{X}})\}}|_{\ell_1}, \quad \vec{\mathbf{X}} \in V_k \tag{26}$$

$\bar{\Delta}$ and $\bar{\Delta}_k$ are the expected values of $\boldsymbol{\Delta}$ and $\boldsymbol{\Delta}_{\mathbf{k}}$ according to the whole distribution of viruses and the sub-distribution of the k_{th} virus lineage in the decoder space, respectively. In our Encoder-Decoder Seq2Seq architecture model, combining formulas (22), (25) and (26), the average "immune escape" ability of the virus population can express as:

$$\overline{CSC} = \bar{\Delta} = \sum_{k=1}^{K} \hat{\pi}_k \int \Delta_{\mathbf{k}} P^k(\mathbf{z}) d\mathbf{z} = \sum_{k=1}^{K} \hat{\pi}_k \int \Delta_{\mathbf{k}} N\left(\mathbf{z} \mid \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k\right) d\mathbf{z}$$

$$= \sum_{k=1}^{K} \hat{\pi}_k \bar{\Delta}_k = \sum_{k=1}^{K} \hat{\pi}_k \hat{\omega}_k \tag{27}$$

Here, the component $P^k(\boldsymbol{z})$ is the prior probability distribution of the k_{th} virus lineage, and together with the formula (22), the average "immune escape" ability of the k_{th} virus lineage is:

$$\bar{\Delta}_k = \hat{\omega}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} |\hat{\omega}_i^k| = \frac{1}{n_k} \sum_{i=1}^{n_k} CSC_i^k = \overline{CSC}^k \tag{28}$$

The $\hat{\omega}_k$ has a mathematical representation consistent with (7). Here,

$$\hat{\omega}_i^k = CSC_i^k = |\hat{\mathbf{z}}^k - \hat{\mathbf{z}}^{\mathbf{k}}_{\mathbf{N}\setminus\{\mathcal{M}(\vec{\mathbf{X}},\tilde{\mathbf{X}})\}}|_{\ell_1}, \quad \mathbf{x} \in V_k.$$

$\hat{\omega}_i^k$ is the "immune escape" ability determined by "semantic changes" of a mutated virus strain in the k_{th} virus lineage compared with the Wuhan virus lineage, and

$$\overline{\mathrm{CSC}}^{\mathrm{k}} = \frac{1}{\mathrm{n_k}} \sum_{\mathrm{i=1}}^{\mathrm{n_k}} |\widehat{\omega}_{\mathrm{i}}^{\mathrm{k}}| \tag{29}$$

is an absolute mean of the CSC_i^k, a mean of the cumulative effects of the "immune escape" ability of a virus strain for the $\mathrm{k_{th}}$ virus lineage. We define $\overline{\mathrm{CSC}}^{\mathrm{k}}$ as the immune escape ability of the $\mathrm{k_{th}}$ virus lineage.

Let us look at the distribution of the absolute mean "semantic" changes computed by sampling different virus lineages (Fig. 3B). The trend of this distribution is highly consistent with the fold increase in relative fitness over the Wuhan (A) lineage given by F. Obermeyer, et al. (*5*). Let us now consider the following four facts: (i) The model determined that latent variable **z** reflected the semantic and grammatical changes of virus sequence site variation and the correlation between variation sites given by formula (2) and (*4*); (ii) After the selection of a reference virus lineage (in this case, the Wuhan lineage as the reference lineage), from formula (22) the CSC score reflects the degree of semantic change of a virus strain after a new sequence variation, indicating its possible contribution to the fitness related to the immune escape ability of the lineage; (iii) The CSC score correctly reflected the changes in immune escape ability caused by virus sequence mutation under syntactic fitness constraints (CSCS constraints (*4*)) (Fig. 2B); (iv) The mean value of the CSC score as the convolution of "semantic" and "grammatical" terms of CSCS criterion correctly reflected the whole immune escape ability of the virus lineage (Fig. 3), and the variation trend of the virus across lineages given by mean CSC score was consistent with the R_0 calculation by the work (*5*) and general epidemiological observation.

Based on the above four facts, we believe that $\boldsymbol{CSC} = \boldsymbol{\omega}(\boldsymbol{z})$ is the corresponding function of $\boldsymbol{\omega}(\boldsymbol{x})$ in hidden space. It's a lower-order function of the latent variable $\boldsymbol{z}$. Based on Bayes' theorem (9) and together with formulas (1), (11) , (22) and (29), we have

$$\bar{\omega}_k = \hat{\lambda}_{\mathrm{k}}\hat{\omega}_{\mathrm{k}} = \frac{\hat{\lambda}_{\mathrm{k}}}{\mathrm{n_k}}\sum_{\mathrm{i}=1}^{\mathrm{n_k}}|\hat{\omega}_{\mathrm{i}}^{\mathrm{k}}| = \frac{\hat{\lambda}_{\mathrm{k}}}{\mathrm{n_k}}\sum_{\mathrm{i}=1}^{\mathrm{n_k}} CSC_i^k \tag{30}$$

Here, $\bar{\omega}$ is the per-lineage fitness of the virus and it can be calculated directly under the framework of deep learning. In this way, we prove that the item of the degree of "semantic" change in the CSCS criterion proposed by B. Hie, et. al., (*4*) is related to the per-cluster fitness of the virus given by F. Obermeyer, et al. (*5*). Although all the parameters involved in the calculation of the per-lineage fitness $\hat{\lambda}_{\mathrm{k}}\hat{\omega}_{\mathrm{k}}$ can be computed directly by the model, it is clear that the changes of the parameter $\hat{\lambda}_{\mathrm{k}}$ and latent variable $\mathbf{z}$ indirectly reflect the effects of selection. Formula (30) is the concrete realization of the P-E theorem (see formula (11)) under the CoT2G-F model.

We can carefully compare formulas (15), (22) and (30) to find there exists an intrinsic correlation between the semantic change term and the grammatical term. This association can express as a convolution of the semantic variation term with the hidden state probability distribution because the grammar term in our model is the posterior distribution of the hidden state probability distribution. This convolution can also express as a mathematical expectation of the CSC according to the hidden state probability distribution $\mathrm{P}(\mathbf{z})$. Interpreted in the sense of convolution, it is precisely the mean value of the accumulation of "semantic" changes of the virus lineage under the constraints of "grammar". And this mean is the same in form as the fitness

calculation formula (15) we got from evolutionary biology, and the calculation formula of the immune escape ability of the k_{th} virus lineage is deduced.

$$\overline{CSC}^{k} = (f * \boldsymbol{g})(\mathbf{z}) = (\boldsymbol{g} * f)(\mathbf{z}) = \int \boldsymbol{g}(\hat{\mathbf{z}} - \mathbf{z}) f(\mathbf{z}) d\mathbf{z}$$

$$= \mathbf{CSC^k} * \hat{p}^{k}(\mathbf{z}) = \int \mathbf{CSC^k}\, P^{k}(\mathbf{z}) d\mathbf{z}$$

$$= \hat{\omega}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} |\hat{\omega}_i^k| \tag{31}$$

Here,

$$f(\mathbf{z}) = P^{k}(\mathbf{z})$$

$$\boldsymbol{g}(\hat{\mathbf{z}} - \mathbf{z}) = |\hat{\mathbf{z}}^{\mathbf{k}} - \hat{\mathbf{z}}^{\mathbf{k}}_{\mathbf{N}\setminus\{\mathcal{M}(\vec{\mathbf{X}},\tilde{\mathbf{X}})\}}|_{\ell_1}$$

At this point, all we need to do is start from actual data and calculate the correctness of our mathematical framework. Therefore, we need to calculate the per-lineage fitness of the k_{th} virus in the unit area R and unit time t_{bin} according to the following formula by sampling:

$$\overline{\omega}_k = \hat{\lambda}_k \hat{\omega}_k \stackrel{\text{def}}{=} \frac{\hat{\lambda}_k}{n_k} \sum_{i=1}^{n_k} |\hat{\omega}_i^k| / (R \times t_{bin}) = \frac{\hat{\lambda}_k}{n_k} \sum_{i=1}^{n_k} CSC_i^k / (R \times t_{bin}) \tag{32}$$

R is an area, such as the United Kingdom, and t_{bin} is a passage time from occurrence to an epidemic of a virus lineage. n_k is the sampling number of the k_{th} virus lineage within the range of $(R \times t_{bin})$, CSC_i^k is the conditional semantic change score of the i_{th} virus strain of the k_{th} virus lineage, and $\hat{\lambda}_k$ is the occurrence frequency of the largest cluster in the k_{th} virus lineage. A reasonable sampling time interval t_{bin} can be selected for calculating the $\hat{\lambda}_k$ (see supplementary materials, fig. S3A and 3B). At this time, the problem of finding R_0 of the k_{th} virus lineage transforms into computing the mathematical expectation $\overline{CSC}^{k}$ of the CSC_i^k of the virus sequence using the deep learning method according to the following formula:

$$R_0 = \frac{\hat{\lambda}_k}{\hat{\lambda}_A}(\hat{\omega}_k/\hat{\omega}_A) = \frac{\hat{\lambda}_k}{\hat{\lambda}_A}\int(\boldsymbol{CSC^k}/\hat{\omega}_A)P^k(\boldsymbol{z})d\boldsymbol{z} = (\frac{\hat{\lambda}_k}{n_k}\sum_{i=1}^{n_k} CSC_i^k)/(\hat{\lambda}_A\hat{\omega}_A)$$

$$= \overline{\omega}_k/\overline{\omega}_A \quad (33)$$

Here, $\overline{\omega}_A = \hat{\lambda}_A\hat{\omega}_A$ is the fitness of the Wuhan lineage, $\overline{\omega}_k = \hat{\lambda}_k\hat{\omega}_k$ is the fitness of the k_{th} virus lineage, and the basic reproduction number R_0 is the fold increase of the per-lineage fitness $\overline{\omega}_k$ over $\overline{\omega}_A$. At this point, we finally used the deep learning model to realize the calculation of formula (1) for SARS-CoV-2 virus population. Now the relative fitness or the basic reproduction number (R_0) of the virus lineage are precisely computable, which is proportional to the occurrence frequency of the largest cluster in the virus lineage times the absolute mean of the CSC score according to the hidden state probability distribution.

3. The prediction of immune escape mutation sites and calculation of R_0 based on the CoT2G-F model

The methods are significantly different when using the CoT2G-F model to predict immune escape mutation sites and calculate R_0. The detailed steps are as follows:

Prediction of immune escape mutations: When predicting immune escape mutation sites, we first need to continuously and randomly mask the amino acids in the input sequence, then based on the CSCS criterion proposed by B. Hie (4), we have designated the model output as the protein sequences most prone to immune escape in the training phase. So in the prediction phase, the model would naturally compound this paradigm and output the most likely protein sequences for immune escape in the next time slice. Afterwards, based on the comparison results with the Wuhan sequence, we can determine the possible immune escape mutation sites in the generated protein

sequence i.e., we regard the CoT2G-F model as a generation model. After generating various sequence mutations, we use the probability of mutations learned by the model to determine immune escape mutation sites through CSCS scoring.

R_0 calculation: When calculating R_0, we no longer regard the CoT2G-F model as a generation model but as a scoring model. For the input SARS-CoV-2 spike protein sequence, the model does not generate any mutation sites, and only uses the trained model as a feature extractor to get the corresponding sequence embedding. Through the model, according to the Wuhan lineage, each input sequence is given a CSC score calculated by the model. Then we calculate the probability average of the CSC to obtain the fitness of each lineage and compare it with the fitness of the Wuhan lineage to obtain the relative fitness, getting R_0.

Genotype-Fitness landscape derived by the P-E theorem: Based on the P-E theorem and from formula (30), we get

$$\overline{\omega}_k = \frac{\widehat{\lambda}_k}{n_k} \sum_{i=1}^{n_k} CSC_i^k$$

Here, CSC_i^k is the conditional semantic change score of the i_{th} virus strain of the k_{th} virus lineage, and is the contribution of a single virus strain to the fitness.

Finally, we can plot the genotype-fitness landscape in the embedded space based on the variation state of the viral genome sequence. Starting from CSCS criteria (*4*), we can regard the CSC score as the "immune escape" potential of the virus, which is a function of latent variables and time. As a two-dimensional hypersurface, the genotype-fitness landscape can plot in a three-dimensional mapping space with "semantics", "syntax", and time as axes. The peaks on the hypersurface correspond to

the genomic variation states that contribute the most to the "immune escape" ability of the virus lineages (see Fig. S9) . We can obtain the fitness of the virus lineage by integrating the surface density of the specific region corresponding to the virus lineage on this hypersurface (Fig. 4A). Now, if we take the gradient of the CSC score along the time and the syntax axis on a two-dimensional hypersurface and take a negative value, and we got and defined the immune escape force $\boldsymbol{F}$ of the virus

$$\boldsymbol{F} = - \nabla CSC(z,t)$$

4. G-P principle: Genotype-phenotype landscapes in embedding space based on the deep learning framework

In equation (33), the R_0 on the left side is a macroscopically observable biological variable, and

$$\frac{\hat{\lambda}_k}{\bar{\lambda}_A}\int \left(\frac{\boldsymbol{CSC^k}}{\hat{\omega}_A}\right)P^k(\boldsymbol{z})d\boldsymbol{z} = \left(\frac{\hat{\lambda}_k}{n_k}\sum_{i=1}^{n_k} CSC_i^k\right)/\left(\hat{\lambda}_A\hat{\omega}_A\right)$$

on the right side is a mathematical expectation of a function $\hat{\lambda}_k \mathrm{CSC}^{\mathrm{k}}$ of the latent variable $\mathbf{z}$, which is related to genome sequence variation, according to the hidden state probability distribution $\mathrm{P}(\mathbf{z})$. From equation (33), we can see that deep learning can help us build a bridge between Genotype-Phenotype, and we can draw a natural inference: "For a microscopic molecular biological variable $\mathbf{x}$ and its embedded latent variable $\mathbf{z}$, a macroscopically observable biology variable $\overline{\mathrm{B}}$ can express as a mathematical expectation of a function $\mathbf{b}(\mathbf{z})$ according to a hidden state probability distribution $\mathrm{P}(\mathbf{z})$". Both $\mathbf{b}(\mathbf{z})$ and $\mathrm{P}(\mathbf{z})$ can obtain by the deep learning model. This statement we call the G-P (Genotype-Phenotype landscapes) principle, it can represent by the following equation:

$$\overline{B} = \int \mathbf{b}(\mathbf{z})\, P(\mathbf{z}) d\mathbf{z} \tag{34}$$

In recent years, many research works have been devoted to discovering the intrinsic theoretical relationship and mathematic basis of deep learning theory and statistical mechanics (*31*, *32*). We know that the core concept of statistical mechanics is: "macroscopic physical observables can characterize as ensemble averages of functions of corresponding microscopic variables", which can be commonly expressed by the following formula:

$$\bar{f} = \int f(\Gamma)\rho(\Gamma)d\Gamma \tag{35}$$

Here: $\bar{f}$ is the macro-observable physical quantity, such as temperature, air pressure, etc., $f(\Gamma)$ is the function of the micro-state space variables, and $\rho(\Gamma)$ is the ensemble distribution function of the micro-state space. We see that equations (34) and (35) have the same form. Referring to formulas (10) and (11)), it is clear the G-P principle is an important corollary of the P-E theorem.

Starting from the calculation of virus fitness, based on the basic mathematical framework of deep learning theory, we obtained the genotype-fitness landscape in embedding space. Current work is a practical example that conforms to the G-P principle. It opens up a new field for applying deep learning theory in biological research, giving a new route for the interpretability of models, and presenting theoretical guidance on establishing a genotype-phenotype landscape.

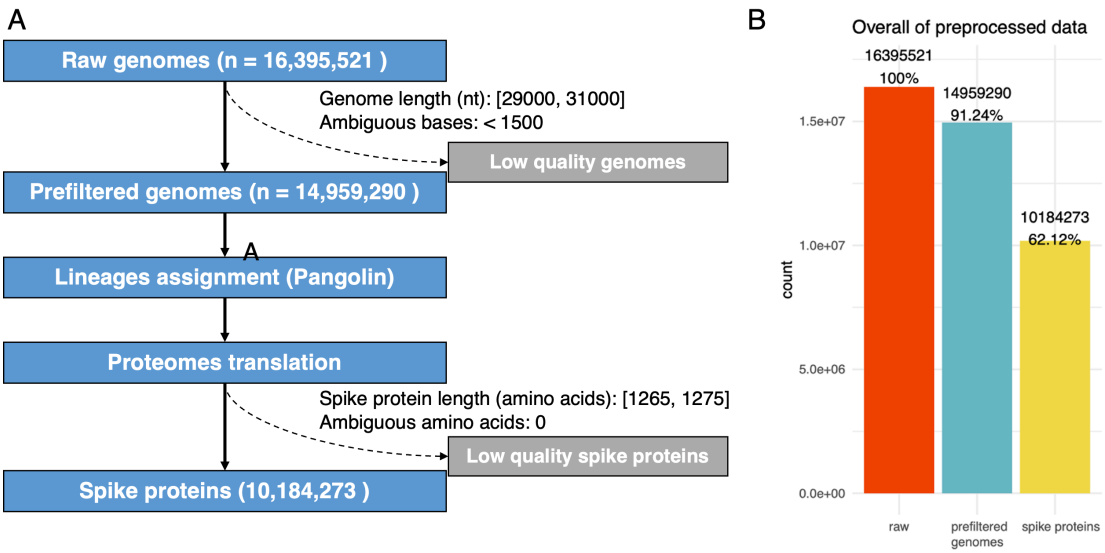


Fig. S1. Preprocessing of SARS-CoV-2 sequences.

(A) Workflow of the SARS-CoV-2 sequences preprocessing. **(B)** Summary of the preprocessed SARS-CoV-2 sequences.

1390

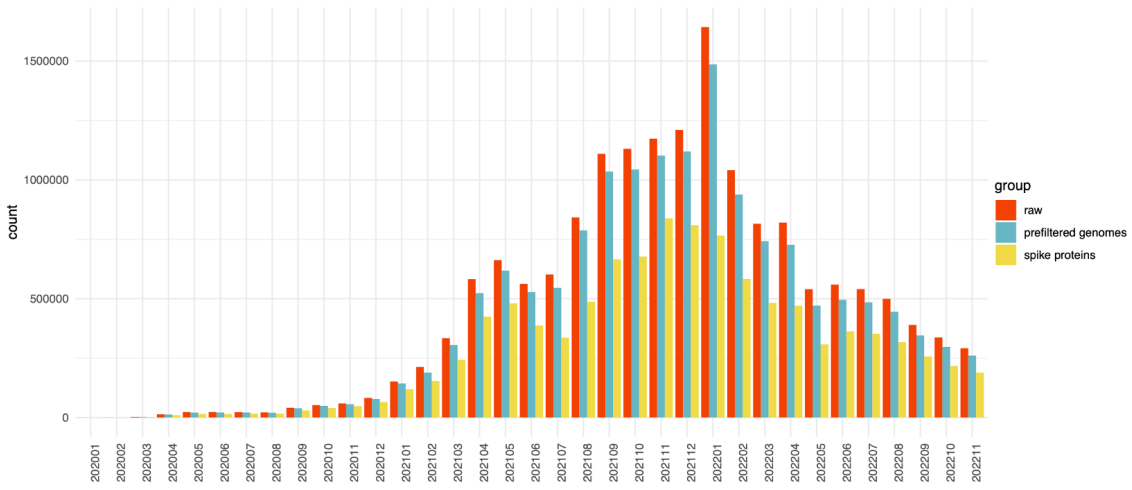
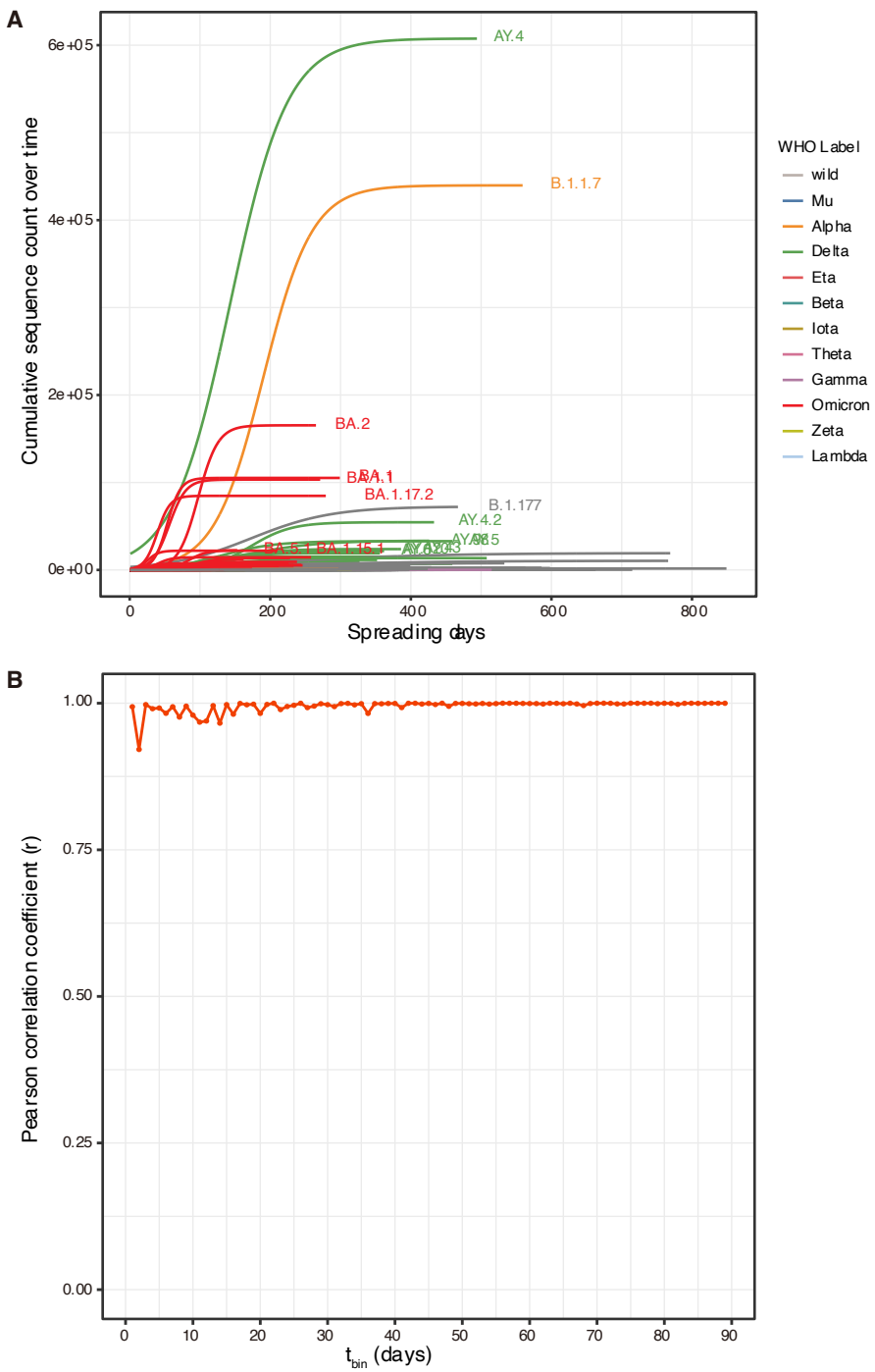


Fig. S2. Summary of the preprocessed SARS-CoV-2 sequences.



1395 **Fig. S3. Growth rate of representative SARS-CoV-2 lineages in the UK**

(A) Growth rates of SARS-CoV-2 lineages were fitted with a logistic model (*35*). Only representative lineages were shown. From the above logical growth curve of each virus lineage relative to the spread time in the figure, a reasonable sampling time interval t_{bin} can be selected for calculating the occurrence frequency of the largest

cluster in different lineages. As can be seen from (A), starting from the first day to the time less than 100 days appears to be a reasonable time interval for sampling virus strains and calculating the occurrence frequency of the largest cluster in different lineages. (B) The different t_{bin} versus the correlations among the distributions of the occurrence frequencies of the largest cluster in different lineages. It can be seen that after 5 days the correlation is over 0.94, and around 50 days the correlation is close to one and keep stable. Based on (A) and the observed robustness of the correlations from (B) we can easily select any day between 5-50 days as the end point of the time interval t_{bin}.

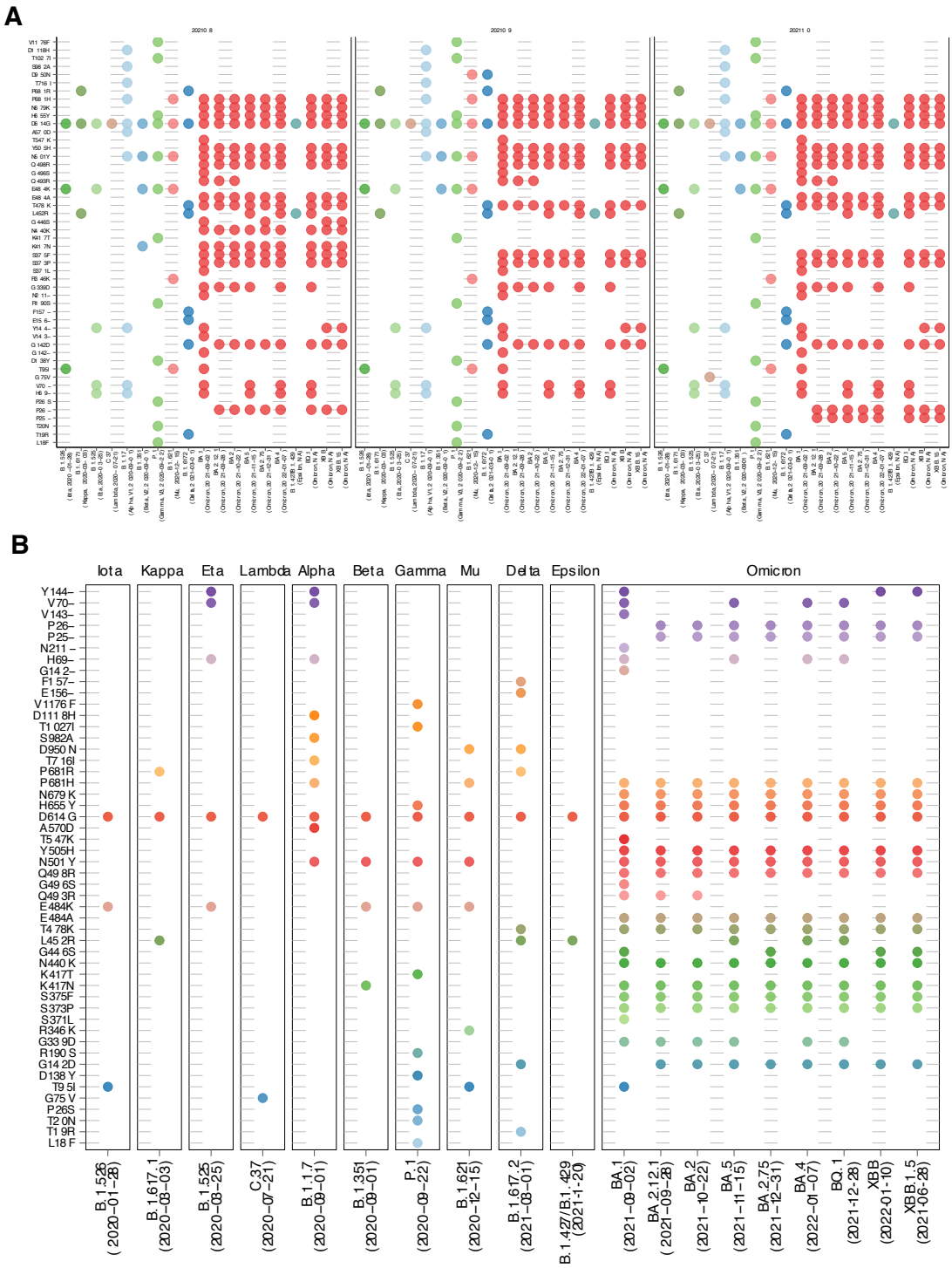


Fig. S4. Reproducing known immune escape mutations

The prediction results show that the model has good reproducibility for already emerged "immune escape" mutations. We took data from August 2021 and October

2021 in GISAID (www.gisaid.org) database. The data were collected and submitted to

1415 the database by the United Kingdom.

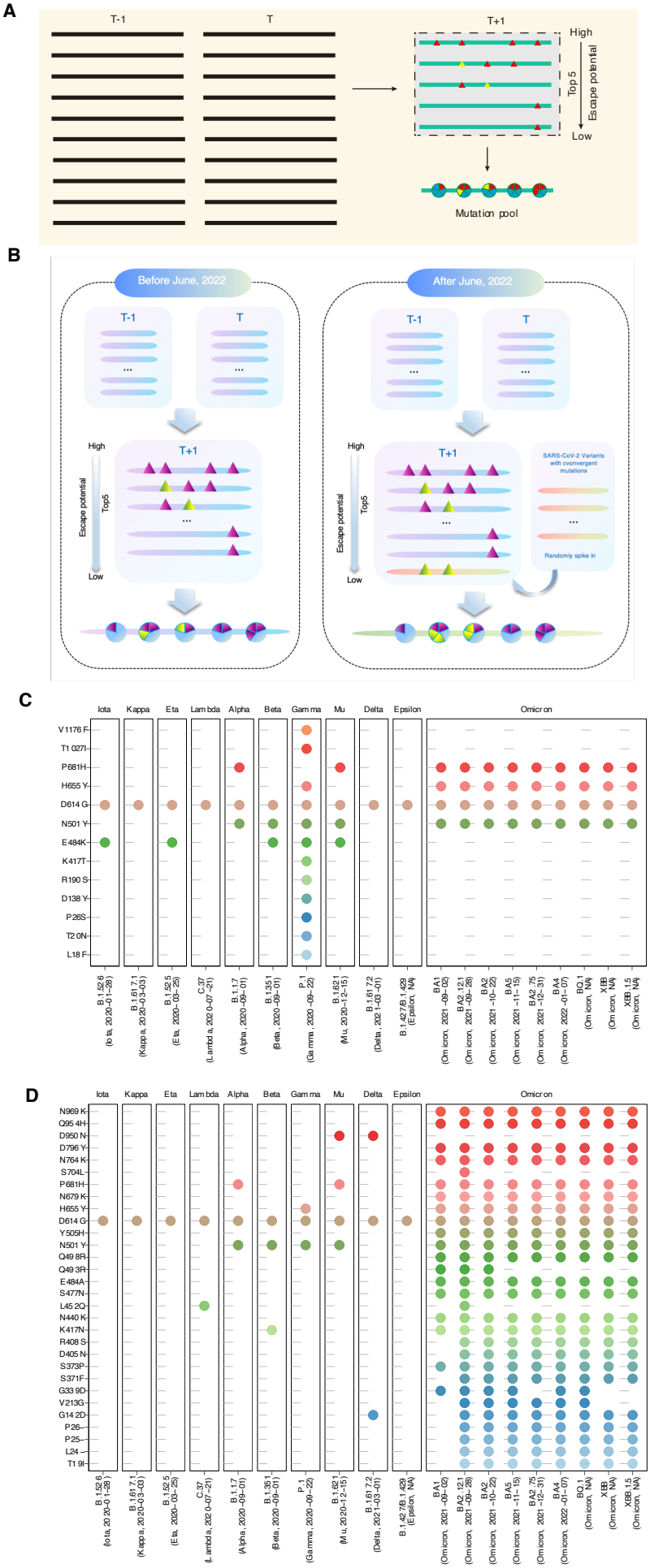

A
T-1
T
T+1
High
Top 5
Escape potential
Low
Mutation pool
B
Before June, 2022
After June, 2022
T-1
T
T+1
High
Escape potential
Top5
Low
SARS-CoV-2 Variants with cvonvergent mutations
Randomly spike in
C
Iota
Kappa
Eta
Lambda
Alpha
Beta
Gamma
Mu
Delta
Epsilon
Omicron
V1176F
T1027I
P681H
H655Y
D614G
N501Y
E484K
K417T
R190S
D138Y
P26S
T20N
L18F
B.1.526 (Iota, 2020-01-28)
B.1.617.1 (Kappa, 2020-03-03)
B.1.525 (Eta, 2020-03-25)
C.37 (Lambda, 2020-07-21)
B.1.1.7 (Alpha, 2020-09-01)
B.1.351 (Beta, 2020-09-01)
P.1 (Gamma, 2020-09-22)
B.1.621 (Mu, 2020-12-15)
B.1.617.2 (Delta, 2021-03-01)
B.1.427/B.1.429 (Epsilon, NA)
BA.1 (Omicron, 2021-09-02)
BA.2.12.1 (Omicron, 2021-09-28)
BA.2 (Omicron, 2021-10-22)
BA.5 (Omicron, 2021-11-15)
BA.2.75 (Omicron, 2021-12-31)
BA.4 (Omicron, 2022-01-07)
BQ.1 (Omicron, NA)
XBB (Omicron, NA)
XBB.1.5 (Omicron, NA)
D
N969K
Q954H
D950N
D796Y
N764K
S704L
P681H
N679K
H655Y
D614G
Y505H
N501Y
Q498R
Q493R
E484A
S477N
L452Q
N440K
K417N
R408S
D405N
S373P
S371F
G339D
V213G
G142D
P26--
P25--
L24--
T19I-

Fig. S5. The schematic diagram for the target sequences search of the dynamic fine-tuning step

1420 (**A**) The model selects virus sequences from three continuous cross-sections time, T-1, T, and T+1 time for the dynamic fine-tuning. Find the five virus sequences at T+1 time whose mutation sites are the key mutations that determine the virus lineage and whose sequences best meet the semantic and grammatical criteria. The five sequences are selected targets for one of the input sequence, and then supervised fine-tuning 1425 begins with the input virus sequence at T-1 and T time. In this fine-tuning process, at the same time, we have calculated the Spatiotemporal correlation among virus strains either.

(**B**) Fine-tuning of targeted hot spot mutation areas was conducted from March to June 2022. Randomly added sequences containing convergent evolutionary sites were 1430 incorporated into the pre-training model. From June to November 2022, input sequences from the UK were selected for predicting immune escape mutations.

(**C**) Prediction of immune escape mutations using data submitted to GISAID from the United Kingdom between March and June 2022.

(**D**) The immune escape mutations including the convergent evolutionary mutations 1435 were reproduced.

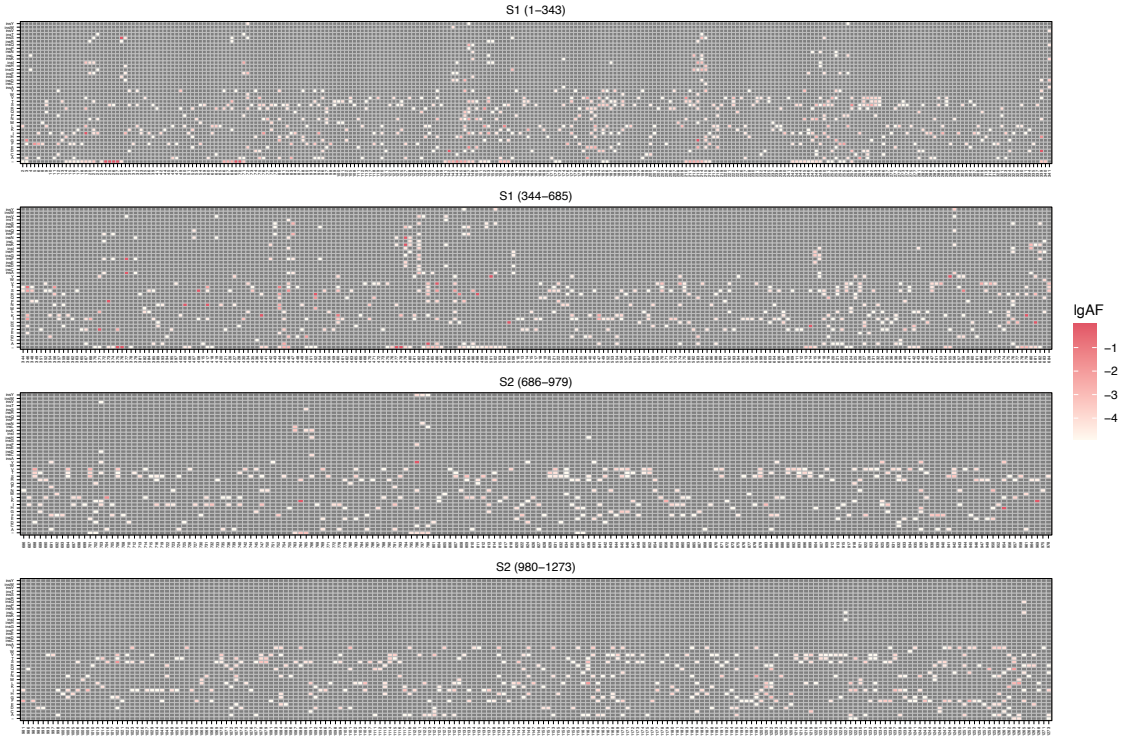


Fig. S6. Model-Generated Spectrum of Immune Escape Mutants of SARS-CoV-2

The predicted immune escape mutant spectrum of SARS-CoV-2, including the three variants of the recently emerged virus lineage XBB.1.16 and most of the existing convergent evolutionary mutations. The model uses data from Britain from June to November.

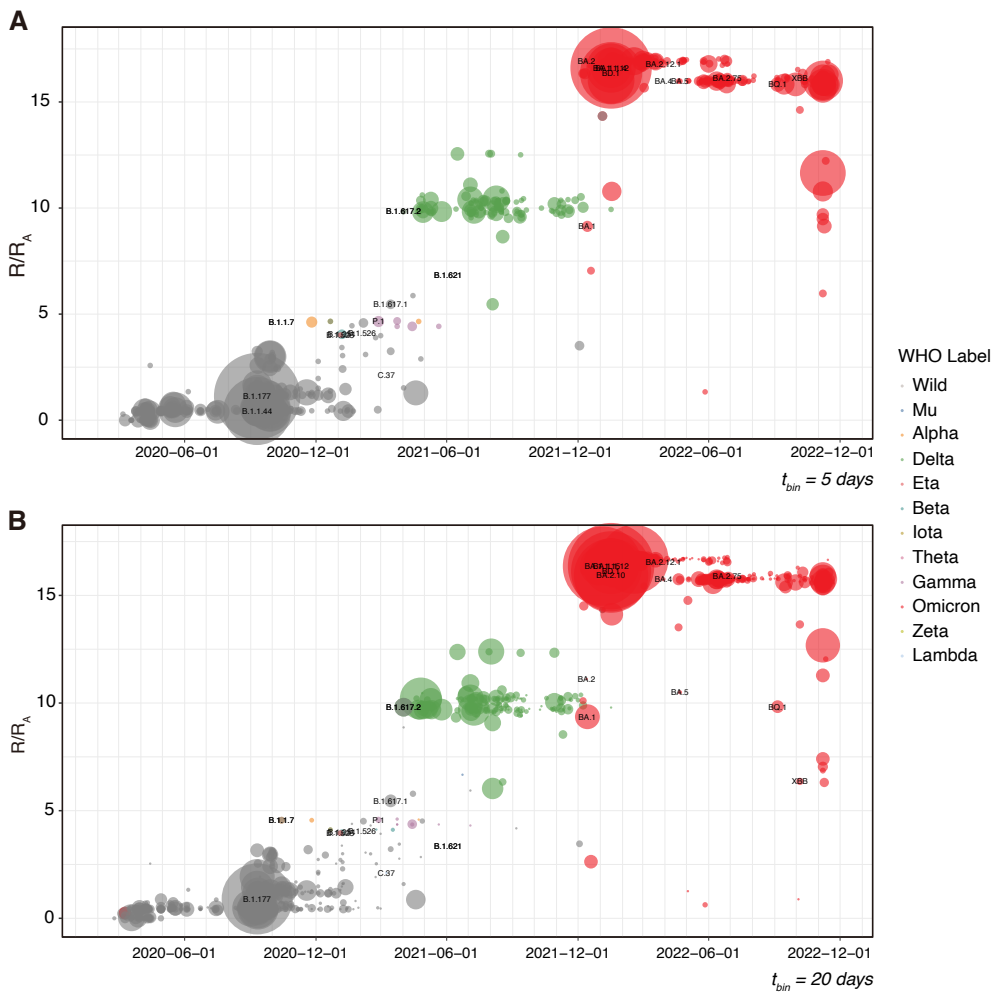


Fig. S7. Relative fitness derived from the natural language model CoT2G-F versus date of lineage emergence

Circle size is proportional to the sampling number for different lineage in certain time intervals (t_{bin}), which reflects the prevalence of the virus. Using the WHO classification, the R_0, ie. the fold increase in relative fitness of the k_{th} virus lineage according to the Wuhan lineage is plotted in different colours. Figures A and B use 5-day and 10-day sampling time intervals to calculate R_0, respectively. The results from A and B are almost completely consistent, reflecting the robustness of the model.

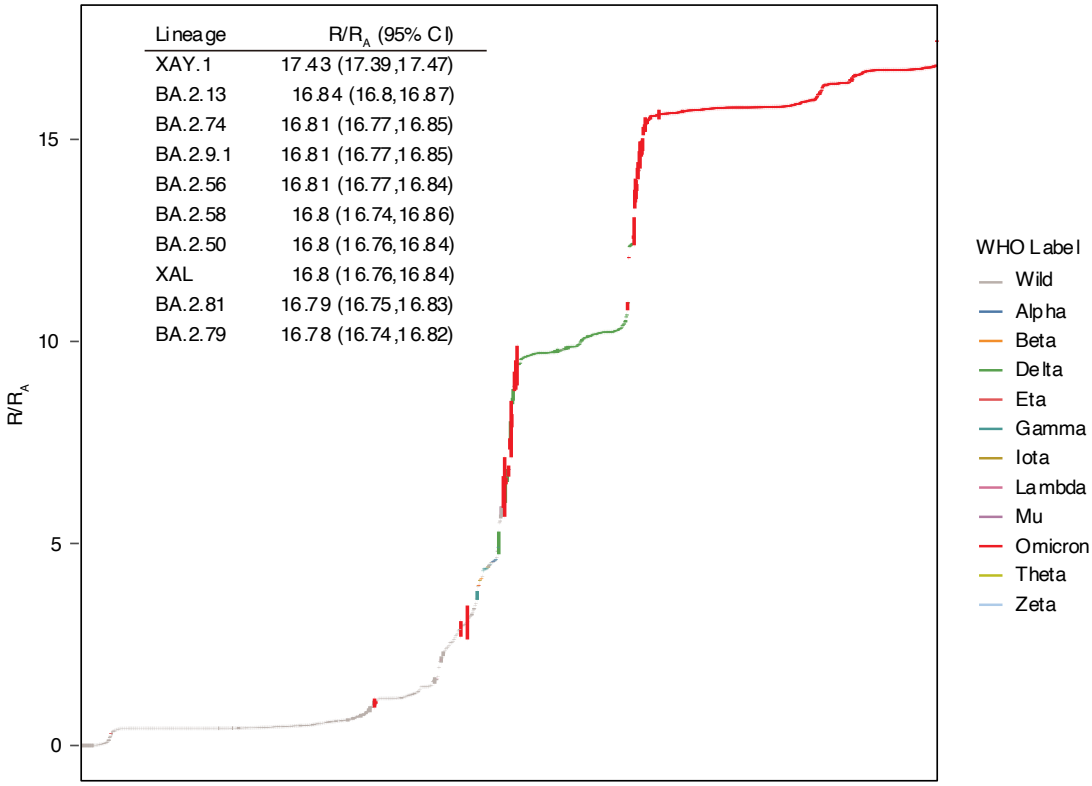


Fig. S8. The confidence interval (CI) distribution for the R_0 calculation

We select the sampling time interval (t_{bin}) from 1 day to 60 days to calculate the R_0 of the virus lineages, and the curves give the 60-day average of the R_0. This curve also reflects the robustness of the model to calculate R_0. The inset table lists the confidence interval (CI) of the ten fittest lineages of the pandemic inferred by our model. R/R_A is the fold increase in relative fitness over the Wuhan lineage in the UK.

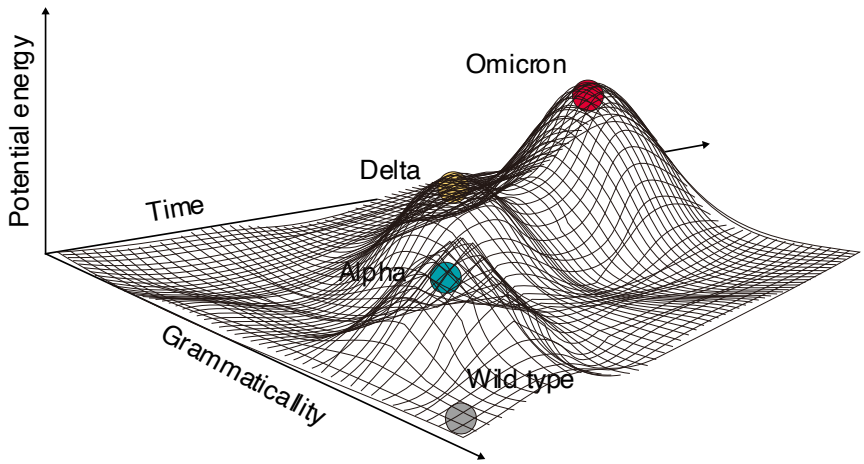


Fig. S9. Plot virus genotype fitness landscape

As a two-dimensional hypersurface, the genotype-fitness landscape can plot in a three-dimensional mapping space with “semantics”, “syntax”, and time as axes. Starting from the sampled data, we fitted a quadric surface function and drew a sketch of the Genityp-Fitness landscape, in which only three virus lineages, including Omicron, were shown.

Table S1. Summary of SARS-CoV-2 Spike protein sequences from the UK involved in this study

These data are used by the model to predict "immune escape" mutation and to calculate the R_0 of the virus lineage.

1475

Month	Genomes	Month	Genomes	Month	Genomes
Jan, 2020	0	Jan, 2021	50068	Jan, 2022	562692
Feb, 2020	3	Feb, 2021	65992	Feb, 2022	312710
Mar, 2020	395	Mar, 2021	71158	Mar, 2022	210240
Apr, 2020	4500	Apr, 2021	161156	Apr, 2022	154880
May, 2020	9337	May, 2021	103445	May, 2022	54265
Jun, 2020	9008	Jun, 2021	137734	Jun, 2022	23762
Jul, 2020	6763	Jul, 2021	245851	Jul, 2022	47680
Aug, 2020	1383	Aug, 2021	222611	Aug, 2022	28707
Sep, 2020	16974	Sep, 2021	273770	Sep, 2022	11597
Oct, 2020	20335	Oct, 2021	274571	Oct, 2022	22141
Nov, 2020	30958	Nov, 2021	291151	Nov, 2022	14283
Dec, 2020	26984	Dec, 2021	290254		

Table S2. Indels predicted with the sequences from August 2021 and September 2021 (before the Omicron emerging)

1480 With SARS-CoV-2 Spike protein sequences of September and October of 2021, a total of 7 mutations (referring to Wuhan-Hu-1) were predicted with the model. These mutations were subsequently identified in Omicron variants.

Indel	Classification	Omicron variants
H69del	Confirmed in Omicron variants	BA.1, BA.4, BA.5, BQ.1, B.1.525
V70del		BA.1, BA.4, BA.5, BQ.1, B.1.525
G142del		BA.1, BA.2, BA.4, BA.5, BA.2.12.1, BA.2.75, BQ.1, XBB
V143del		BA.1
Y144del		BA.1, XBB, B.1.525
N211del		BA.1
H66del	Novel	NA
A67del (next to H69del)		NA
I68del (next to H69del)		NA
Y145del (next to G142del)		NA
L212del		NA
S477del		NA
T478del		NA

A684del	NA
R685del	NA
S686del	NA
V687del	NA
A688del	NA

1485

Box 1. The mathematical foundation and derivation of P-E theorem

According to the basic definition of population genetics (*14*), the average fitness $\bar{\omega}$ of a virus population can be expressed mathematically as follows:

$$\bar{\omega} = \lambda_1\omega_1+\lambda_2\omega_2+\ldots\ldots+\lambda_K\omega_K = \sum_{k=1}^{K}\lambda_k\omega_k = \boldsymbol{\lambda}^T\boldsymbol{\omega} \quad (1)$$

$\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots\ldots, \lambda_K)$, λ_k is the occurrence frequency of the k_{th} virus lineage in the whole population, which is an observable measure related to amino acid substitution characteristics, and $\boldsymbol{\omega}=(\omega_1, \omega_2, \ldots\ldots, \omega_K)$, ω_k is the fitness of the k_{th} virus lineage, and $\boldsymbol{x} = (x_1, x_2, \ldots, x_i, \ldots, x_L)$ represent L variation sites. Thus, each ω_k can involve the epistatic interaction of multiple mutation sites in the input sequence (*12*) as following:

$$\omega_k(\boldsymbol{x})=a^{(0)}+\sum a_i^{(1)} x^k{}_i +\sum a_{ij}^{(2)} x^k{}_i x^k{}_j+\sum a_{ijk}^{(3)} x^k{}_i x^k{}_j x^k{}_k +\ldots+\sum a_{1,2,\ldots L}^{(L)} x^k{}_1 x^k{}_2\ldots x^k{}_L \quad (2)$$

We define a function $P(\boldsymbol{x})$ as a state probability distribution of the sequence's variation states in real space. This variation causes changes in viral fitness. The $P(\boldsymbol{x})$ can have an expanded form using a Gaussian Mixture Model (GMM).

$$P(\boldsymbol{x}) = \sum_{k=1}^{K}\pi_k N(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (3)$$

$\pi_k N(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is the k_{th} component in the mixture model, π_k is the mixture coefficient, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the mean and variance values of the relative fitness, and they can obtain indirectly through the EM algorithm or deep neural network. Combining formulas (1) and (3), the formal representation of the average fitness of the virus population can obtain as follows:

$$\bar{\omega} = E(\omega) = \int \boldsymbol{\omega}(\boldsymbol{x})P(\boldsymbol{x})d\boldsymbol{x} = \sum_{k=1}^{K}\pi_k\bar{\omega}_k \quad (4)$$

Let λ_k be the occurrence frequency of the dominant cluster in the k_{th} virus lineage when sampling virus strains, we can redefine the (dominant) per-lineage fitness as

$$\bar{\omega}_k \stackrel{\text{def}}{=} \frac{\lambda_k}{n_k}\sum_{i=1}^{n_k}\omega_i^k \quad (5)$$

Referring to formula (3), let $\boldsymbol{\omega}(\boldsymbol{x}) = \boldsymbol{\omega}_k(\boldsymbol{x})$, $P(\boldsymbol{x}) = N(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, the per-lineage fitness of the virus can express formally again as:

$$\bar{\omega}_k = \bar{\omega}_{|P(\boldsymbol{x})} = \int \boldsymbol{\omega}(\boldsymbol{x})\, P(\boldsymbol{x})\, d\boldsymbol{x} \quad (6)$$

The subscript P (·) represents different spaces corresponding to k. As long as we find a suitable functional form of $\boldsymbol{\omega}(\boldsymbol{x})$ and a corresponding probability distribution $P(\boldsymbol{x})$, we can obtain the per-lineage fitness of viruses by calculating their mathematical expectations by the formula (6), then get the average fitness of the virus population by formula (4). We can now try to do this within the context of deep learning, despite the fact that it is typically difficult to obtain and calculate $\boldsymbol{\omega}(\boldsymbol{x})$ and $P(\boldsymbol{x})$ directly from real-world data with noise.

Considering the independent and identically distributed (IID) properties of state random variables of different samples in general, we let $Q_\theta(\boldsymbol{z}|\boldsymbol{x})$ and $P(\boldsymbol{z})$ be the hidden state probability distribution in encoder-decoder spaces under the Encoder-Decoder Seq2Seq framework, and $P(\boldsymbol{x})$ and $P_\varphi(\boldsymbol{x}|\boldsymbol{z})$ be the prior and posterior probability related to the input and output of the model. The relationship among $P(\boldsymbol{x})$, $Q_\theta(\boldsymbol{z}|\boldsymbol{x})$, $P_\varphi(\boldsymbol{x}|\boldsymbol{z})$ and $P(\boldsymbol{z})$ is given by Bayes' theorem:

$$P(\boldsymbol{x})Q_\theta(\boldsymbol{z}|\boldsymbol{x}) = P_\varphi(\boldsymbol{x}|\boldsymbol{z})P(\boldsymbol{z}) \quad (7)$$

The function relationship defined by formula (1) still holds when $P(\boldsymbol{x})Q_\theta(\boldsymbol{z}|\boldsymbol{x})$ maps to $P_\varphi(\boldsymbol{x}|\boldsymbol{z})P(\boldsymbol{z})$. When the prior probability $P(\boldsymbol{x})$ is unknown, obtaining the fitness of virus lineages in the encoder or decoder space is transformed into finding the hidden state probability distributions $Q_\theta(\boldsymbol{z}|\boldsymbol{x})$ and $P(\boldsymbol{z})$, the posterior probability distribution $P_\varphi(\boldsymbol{x}|\boldsymbol{z})$

Under the Encoder-Decoder Seq2Seq framework together with Bayes' theorem and Universality Theorem of multilayer feedforward networks, because $\boldsymbol{x} \rightarrow \boldsymbol{x}' = W\boldsymbol{z} + \boldsymbol{B}$, we have

$$\bar{\omega}_{|\mathrm{P}(\boldsymbol{x})} = \int \boldsymbol{\omega}(\boldsymbol{x})\,\mathrm{P}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int \boldsymbol{\omega}(W\boldsymbol{z})\,\mathrm{P}(W\boldsymbol{z})\,\frac{d\boldsymbol{x}}{d\boldsymbol{z}}\,\mathrm{d}\boldsymbol{z}$$

$$= \int \boldsymbol{\omega}(\boldsymbol{z})\,\mathrm{P}(\boldsymbol{z})(\mathrm{P}_{\varphi}(\mathbf{x}\,|\,\mathbf{z})/\mathrm{Q}_{\theta}(\mathbf{z}\,|\,\mathbf{x}))\mathrm{d}\boldsymbol{z} = \int W\boldsymbol{\omega}(\boldsymbol{z})\,\mathrm{P}(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z}$$

and $\frac{d\boldsymbol{x}}{d\boldsymbol{z}} = \mathrm{P}_{\varphi}(\mathbf{x}\,|\,\mathbf{z})/\mathrm{Q}_{\theta}(\mathbf{z}\,|\,\mathbf{x}) = W$. Considering that the average of states is by virus lineage, the fitness of a virus lineage is determined by the dominant virus cluster in the lineage, and then by doing a simple matrix operation we can get $\mathrm{W}\boldsymbol{\omega}(\mathrm{z}) \rightarrow \lambda\boldsymbol{\omega}(\mathrm{z})$. λ is a diagonal matrix. Then, we have

$$\bar{\omega}_{|\mathrm{P}(\boldsymbol{x})} = \int \boldsymbol{\omega}(\boldsymbol{x})\,\mathrm{P}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int W\boldsymbol{\omega}(\boldsymbol{z})\,\mathrm{P}(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z} = \int \lambda\boldsymbol{\omega}(\boldsymbol{z})\,\mathrm{P}(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z} \qquad (8)$$

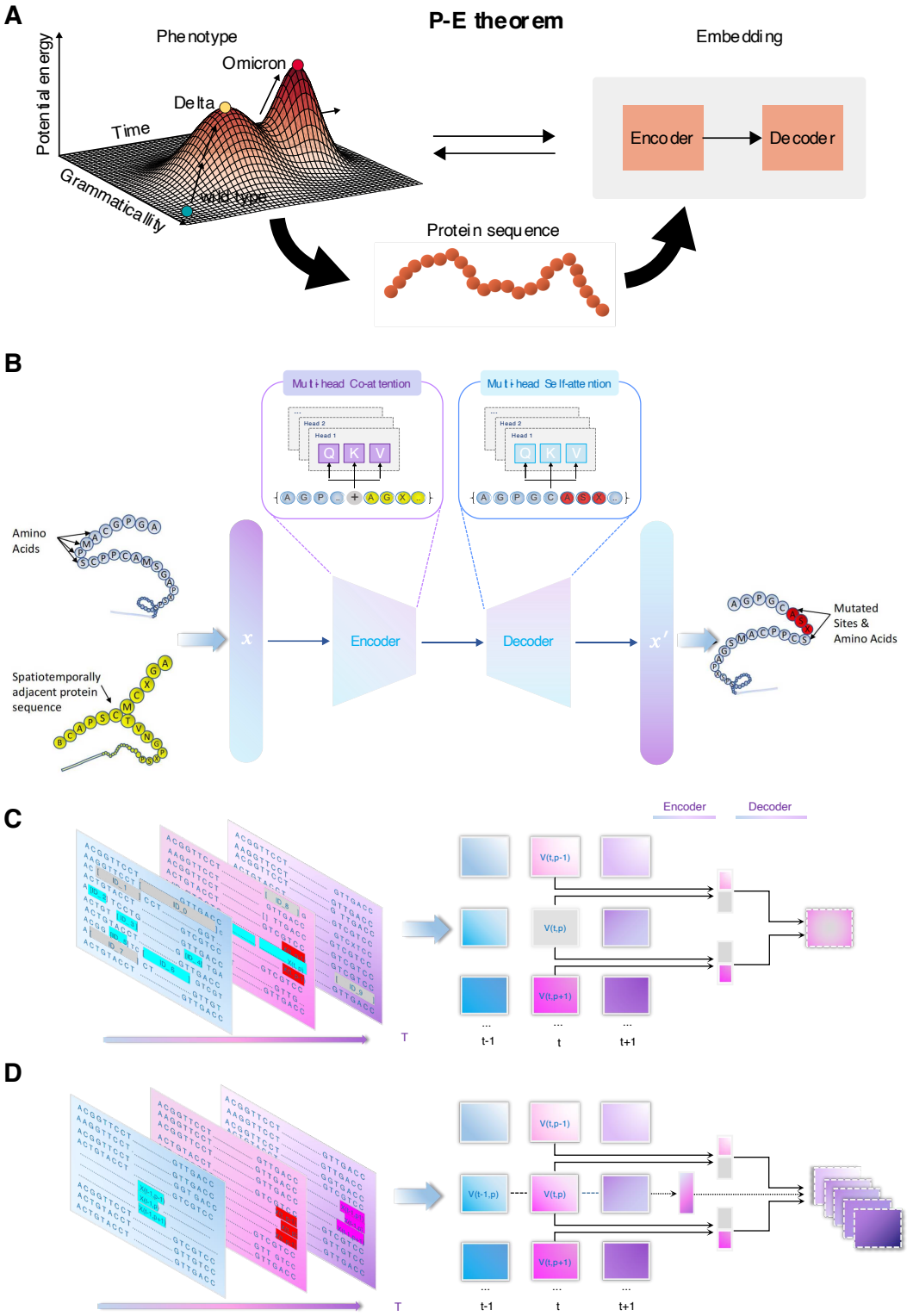
Obviously, formula (8) is the corresponding formal representation of formula (1) under the Encoder-Decoder Seq2Seq framework of deep learning, then, let $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots\ldots, \lambda_{\mathrm{K}})$ is a vector, which components are made up of diagonal elements of the diagonal matrix λ, we have:

$$\bar{\omega}_{|\mathrm{P}(\boldsymbol{x})} = \boldsymbol{\lambda}\bar{\boldsymbol{\omega}}_{|\mathrm{P}(\boldsymbol{z})} \qquad (9)$$

Because of the hierarchical structure of the population, lineage and cluster of the virus, and referring to the formula(1), (3), (4), (6) and (8), and based on the basic principles of population genetics (*14*), the elements of the vector $\boldsymbol{\lambda}$ are the corresponding elements of mixture coefficient in formula (4). The element λ is an occurrence frequency of a dominant cluster in virus lineage when sampling virus strains, and is a macroscopic parameter that links the hidden space with the actual space. Formula (9) gives a mathematical framework for representing phenotypes in embedded Spaces, leading to the Phenotype-Embedding (P-E) theorem: "An observable macro-biological phenotype can be computed under the Encoder-Decoder Seq2Seq framework if we can find a reasonable embedded representation of the related microscopic genotype."

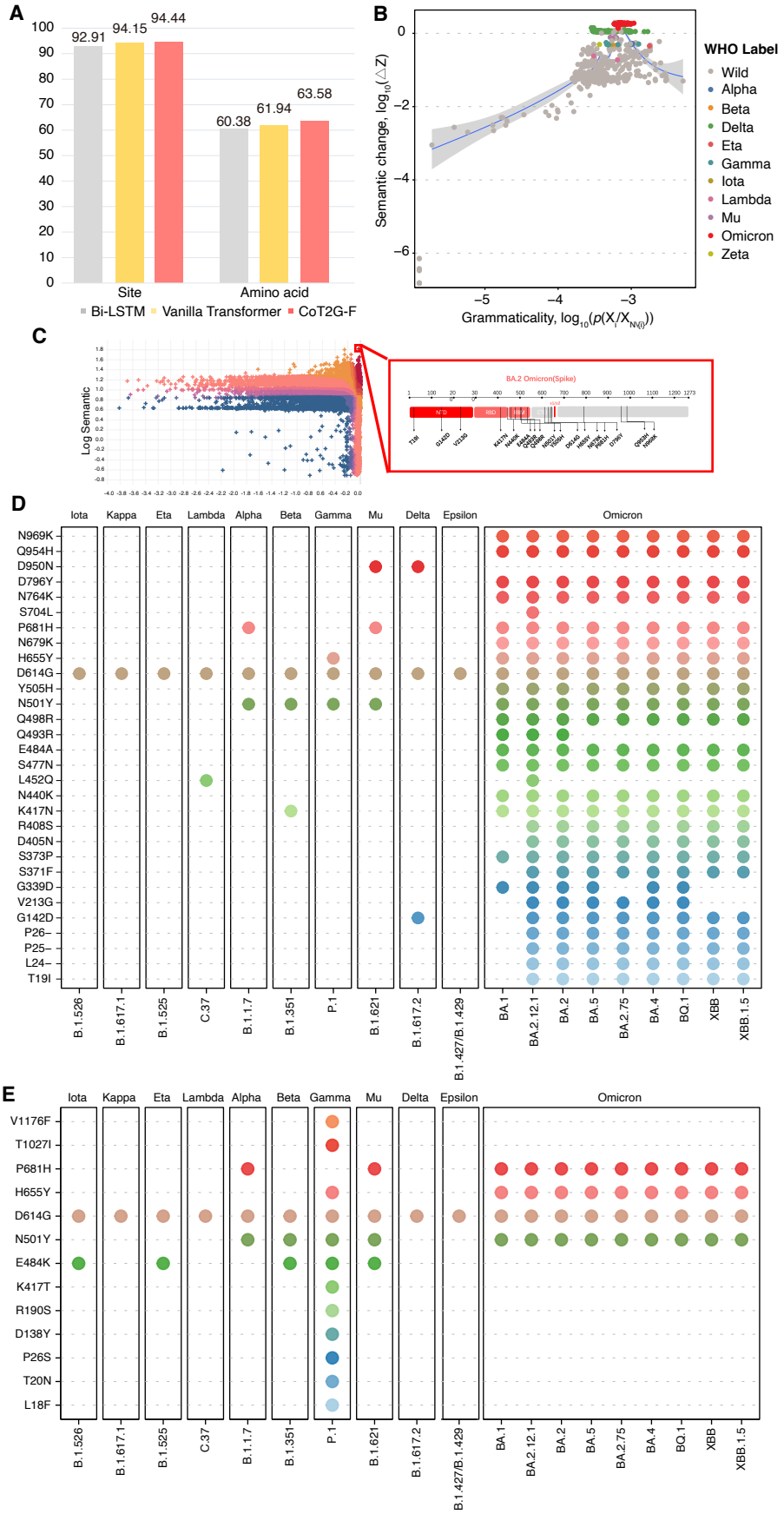
Genotype-fitness landscape: we can plot the genotype-fitness landscape in the embedded space based on the variation state of the viral genome sequence. Starting from P-E theorem, we can regard the $\omega(\mathbf{z})$ score as the "immune escape" potential of the virus, then, as a two-dimensional hypersurface, the genotype-fitness landscape can plot in a three-dimensional mapping space with "$\omega(\mathbf{z})$", "$\mathrm{P}(\boldsymbol{z})$", and "time" as axes. The stable points on the hypersurface correspond to the genomic variation states that contribute the most to the "immune escape" ability. We can obtain the fitness of the virus lineage by integrating the surface density of the specific region corresponding to the virus lineage on this hypersurface. Now, if we take the gradient of the $\omega(\mathbf{z})$ along the time and the $\mathrm{P}(\boldsymbol{z})$ axis on a two-dimensional hypersurface and we got and defined the immune escape force of the virus.

Fig. 1. The P-E theorem, the framework of the CoT2G_F model and simulating the real evolution scenario of viruses. (A) The P-E theorem. Based on Bayes theorem and the fully connected linear transformation under the Encoder-Decoder architecture, a quantitative relationship between virus immune escape mutations and its fitness and the genotype-fitness landscape can obtain in the embedding space. **(B)** CoT2G_F is a standard encoder-decoder architecture that introduces a co-attention and continuous span masking mechanism. **(C** and **D)** Shows the pre-training and fine-tuning steps of the model (see Supplementary Materials, Fig. S4).


A
P-E theorem
Phenotype
Omicron
Delta
Time
wild type
Potential energy
Grammaticality
Embedding
Encoder
Decoder
Protein sequence
B
Multi-head Co-attention
Multi-head Self-attention
Amino Acids
Spatiotemporally adjacent protein sequence
Encoder
Decoder
Mutated Sites & Amino Acids
C
Encoder
Decoder
V(t,p-1)
V(t,p)
V(t,p+1)
t-1
t
t+1
D
V(t-1,p)
V(t,p)
V(t,p+1)
t-1
t
t+1

1535

Fig. 2. Modeling of virus evolution mechanisms based on virus sequences. (A) The figure shows a comparative experiment result among our method (CoT2G-F), Vanilla Transformer and Bi-LSTM (see supplementary materials). (**B**) The virus lineage with a high prevalence rate has a significantly higher semantic change score (CSC) (see supplemental note 4), calculated by the UK-submitted spike protein sequence data for June 2020 to February 2022. The left side of the figure (**C**) visualizes the semantic changes and grammaticality output by our model CoT2G-F with the horizontal axis representing the grammaticality and the vertical axis representing the semantic changes. The upper right corner of figure (**C**) means that the bigger the semantic changes and the more likely the virus strain is to immune escape. The right side of the figure (**C**) shows the predicted mutation sites on the Spike protein sequence of the Omicron's B.A.2 virus lineage with high semantic changes and grammatical fitness, these mutation sites are real and marked at the bottom of the right figure. (**D**) Take the virus sequence data collected in the United Kingdom from August to October 2022 as input to identify and reproduce emergent "immune escape" mutations. (**E**) Input sequences collected from the United Kingdom between March 2022 and June 2022 were selected to predict immune escape mutations.


A
92.91
94.15
94.44
60.38
61.94
63.58
Site
Amino acid
Bi-LSTM
Vanilla Transformer
CoT2G-F
B
Semantic change, log10(△Z)
Grammaticality, log10(p(Xi/XN(i)))
WHO Label
Wild
Alpha
Beta
Delta
Eta
Gamma
Iota
Lambda
Mu
Omicron
Zeta
C
Log Semantic
BA.2 Omicron(Spike)
D
Iota
Kappa
Eta
Lambda
Alpha
Beta
Gamma
Mu
Delta
Epsilon
Omicron
N969K
Q954H
D950N
D796Y
N764K
S704L
P681H
N679K
H655Y
D614G
Y505H
N501Y
Q498R
Q493R
E484A
S477N
L452Q
N440K
K417N
R408S
D405N
S373P
S371F
G339D
V213G
G142D
P26–
P25–
L24–
T19I
B.1.526
B.1.617.1
B.1.525
C.37
B.1.1.7
B.1.351
P.1
B.1.621
B.1.617.2
B.1.427/B.1.429
BA.1
BA.2.12.1
BA.2
BA.5
BA.2.75
BA.4
BQ.1
XBB
XBB.1.5
E
V1176F
T1027I
P681H
H655Y
D614G
N501Y
E484K
K417T
R190S
D138Y
P26S
T20N
L18F

1555

Fig. 3. Absolute mean distribution plot of the semantic change parameter $\mathbf{CSC_i^k}$. Take the virus sequence data collected in the United Kingdom after March 2022. (A) Using the ℓ_1 norm. (B) Using the ℓ_2 norm. The horizontal axis in the figures is the different virus lineages according to the emerging chronological order. The vertical axis is the absolute mean of the semantic change $\mathrm{CSC_i^k}$ of the $\mathrm{k_{th}}$ virus lineage, which is defined as the immune escape capacity of the $\mathrm{k_{th}}$ virus lineage. Figures (A) and (B) both clearly show that with the progress of the global epidemic, from the Wuhan virus lineage to the Delta lineage and then to the Omicron lineage, the immune evasion ability continues to increase and even accelerates. Figures (A) and (B) show the same absolute mean distribution trend of the $\mathrm{CSC_i^k}$, but figure (B) changes is more smoothly.

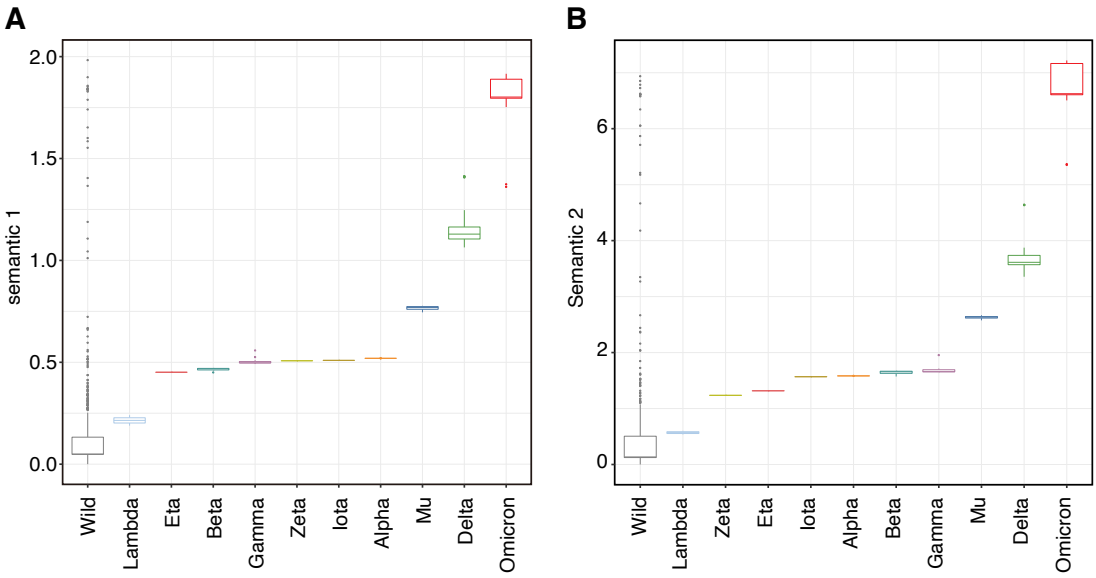


Fig. 4. Relative fitness derived from the natural language model CoT2G-F versus date of lineage emergence. Circle size is proportional to the sampling number for different lineage in the time interval. This figure uses a time interval of 10 days. Referring to the Pango lineage designation and assignment, the R_0, which is the fold increase in relative fitness of the virus lineages according to the Wuhan lineage, is plotted in different colours. The results for 5, and 20-day time intervals have shown in (fig. S6). All the results are almost consistent, reflecting the robustness of the model.

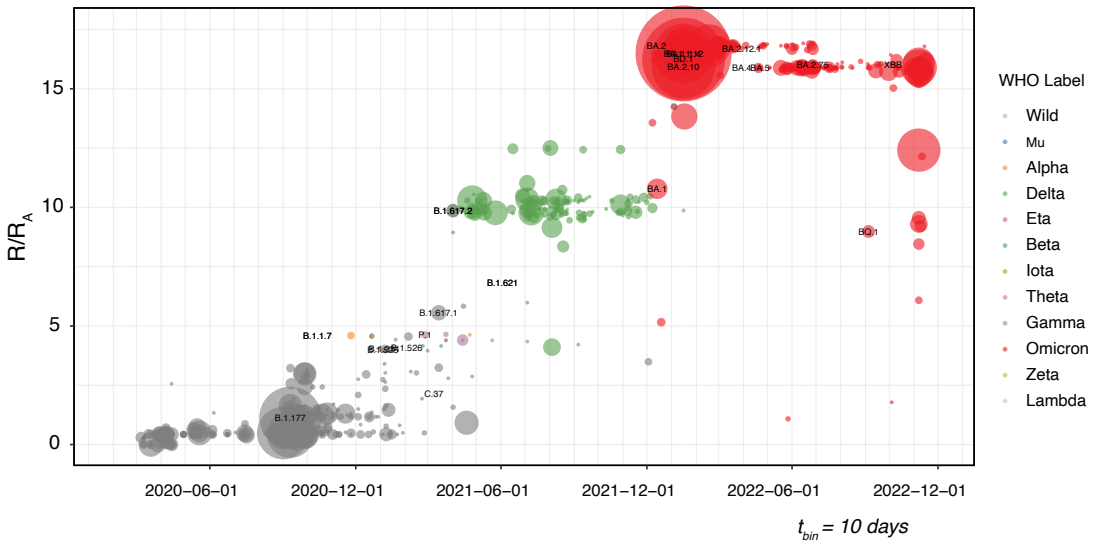
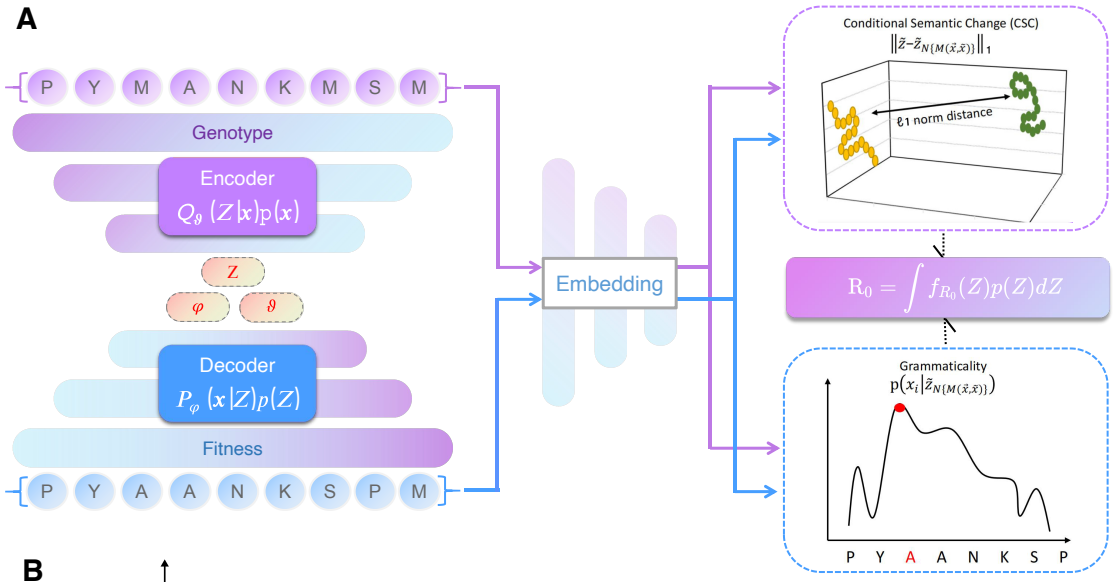
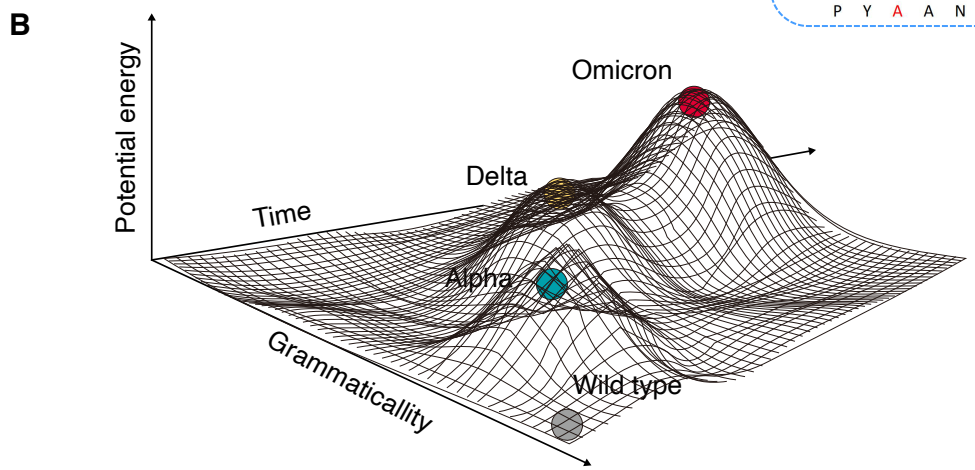


Fig. 5. CoT2G-F framework for building genotype-phenotypic landscapes. (A) CoT2G-F is a natural language deep learning model that introduces co-attention and continuous span masking mechanism and takes the Transformer as a kernel. The model links the prior probability $P(\mathbf{x})$ and hidden state distribution $Q_\theta(\mathbf{z}|\mathbf{x})$ to the prior probability $P(\mathbf{z})$ and the posterior probability distribution $P_\varphi(\mathbf{x}|\mathbf{z})$ in the encoder-decoder space via Bayes' theorem, mapping viral protein sequences as latent variables reflecting semantic and grammatical changes in the embedding space. Then refer to the semantic and grammatical changes to identify the virus sequence mutation. According to the hidden state prior probability distribution $P(\mathbf{z})$ in the decoder space, the model can express the R_0 as the mathematical expectation of a specific latent variables function related to viral sequence mutation according to a hidden state probability distribution, building a genotype-fitness landscape. (**B**) Plot virus genotype-fitness landscape. As a two-dimensional hypersurface, the genotype-fitness landscape can plot in a three-dimensional mapping space with "semantics", "syntax", and time as axes. Starting from the sampled data, we fitted a quadric surface function and drew a sketch of the Genityp-Fitness landscape, in which only three virus lineages, including Omicron, were shown.


A
P Y M A N K M S M
Genotype
Encoder
$Q_\vartheta(Z|x)p(x)$
Z
φ
ϑ
Decoder
$P_\varphi(x|Z)p(Z)$
Fitness
P Y A A N K S P M
Embedding
Conditional Semantic Change (CSC)
$\|\tilde{z}-\tilde{z}_{N(M(\tilde{x},\tilde{x}))}\|_1$
ℓ_1 norm distance
$R_0 = \int f_{R_0}(Z)p(Z)dZ$
Grammaticality
$p(x_i|\tilde{z}_{N(M(\tilde{x},\tilde{x}))})$
P Y A A N K S P


B
Potential energy
Omicron
Delta
Time
Alpha
Grammaticallity
Wild type