Title: Genotypic-phenotypic landscape computation based on first principle and deep learning

Authors: Yuexing Liu[1,†], Yao Luo[4,†], Xin Lu[1,†], Hao Gao[7,†], Ruikun He[7], Xin Zhang[7,*], Xuguang Zhang[5,*], Yixue 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.

*Correspondence: xzhang@picb.ac.cn (Xin Zhang); fxgzhang@mengniu.cn (Xuguang Zhang); li_yixue@gzlab.ac.cn (Yixue Li)

Abstract

The relationship between genotype and fitness is fundamental to evolution, but quantitatively mapping genotypes to fitness has remained challenging. We propose the Phenotypic-Embedding theorem (P-E theorem) that bridges genotype-phenotype through an encoder-decoder deep learning framework. Inspired by this, we proposed a more general first principle for correlating genotype-phenotype, and the Phenotypic-Embedding theorem provides a computable basis for the application of first principle. As an application example of the P-E theorem, we developed the Co-attention based Transformer model to bridge Genotype and Fitness (CoT2G-F) model, a Transformer-based pre-train foundation model with downstream supervised fine-tuning (SFT) that can accurately simulate the neutral evolution of viruses and predict immune escape mutations. Accordingly, following the calculation path of the P-E theorem, we accurately obtained the basic reproduction number (R_0) of SARS-CoV-2 from first principles, quantitatively linked immune escape to viral fitness, and plotted the genotype-fitness landscape. The theoretical system we established provides a general and interpretable method to construct genotype-phenotype landscapes, providing a new paradigm for studying theoretical and computational biology.

Keywords

Genotype-fitness landscape; interpretability; deep learning; immune escape; SARS-CoV-2; the relative basic reproduction number (R_0).

Main

Due to the large amount of scientific data left to us, the SARS-CoV-2 virus has become a model virus for basic research on the transmission mechanism of virus-related epidemics, despite the Covid-19 pandemic having ended. 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. Now the GISAID database [https://gisaid.org] has collected more than 15 million SARS-CoV-2 genome sequences with spatio-temporal information. This massive data set provides an opportunity to uncover 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 [3] by utilizing a natural language model. F. Obermeyer et al. [4] developed a hierarchical Bayesian multinomial logistic regression model, PyR0, to estimate the per-lineage fitness of SARS-CoV-2. M. C. Maher et al. [5] developed statistical models incorporating epidemiological covariates to account for the effects of driver mutations and the associated fitness of different viruses.

Our work focuses on constructing a computable representation of the genotype-fitness landscape by deciphering the effects of epistasis and recombination, identifying and predicting viral mutations associated with immune escape and quantifying virus fitness. To do this, following the fitness definition in population genetics, we first

developed the P-E theorem, a theoretical framework that allows us theoretically to quantitatively calculate the fitness of a virus population by selecting appropriate embedding expressions for the virus genome under the Encoder-Decoder Seq2Seq framework. Then, on the basis of the P-E theorem, a more universal first principle of Genotype-Phenotype for landscape computing was proposed, and on the basis of the P-E theorem, we provide a computable basis for Genotype-Phenotype landscape computation under the deep learning framework. By calculating the transmissibility R_0 of SARS-CoV-2, a macrobiological phenotype determined by the genotype of the virus, we give an example of constructing a real genotype-fitness landscape following first principles by using the P-E theorem. We then constructed a Transformer-based pre-train foundation model with downstream supervised fine-tuning (SFT) to calculate and predict virus immune escape mutations and the fitness without introducing and artificially coupling macro-epidemiological information, resulting in a precise mathematical representation of the virus genotype-fitness landscape in embedding space. In our work, the fitness, a macroscopic biological variable, can be accurately described mathematically as the expectation of a latent variable function according to a hidden state distribution.

Our model can also be used as a generative model to predict the likely occurrence of "immune escape" mutations. Our prospective and retrospective calculations (Figure 3B) have confirmed these results.

Results

Phenotype-Embedding (P-E) theorem, first principles and constructing genotype-fitness landscape

Unravelling the evolutionary causes and consequences of the genotype-fitness landscape is a fundamental yet challenging task. The features of the genotype-fitness landscape profoundly influence 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 information about the topography of natural genotype-fitness landscapes and the precise and quantified relationship between genomic variation and fitness [6]. 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.

Starting from the definitions of evolutionary biology [6,7], we derive a formal expression for the fitness of a virus population, which is consistent with the mathematical expression given by F. Obermeyer et al. [4]. Under the Encoder-Decoder Seq2Seq framework that satisfies the universality theorem of neural networks [8], we introduce the Bayes theorem and combine it with Gaussian Mixed Model (*GMM*) expansion and Expectation-Maximum (*EM*) algorithm to deduce the P-E theorem under the condition of linear expansion according to the virus lineage (see Box 1 and supplemental note 1, formulas (1-11)): “An observable macrobiological phenotype can be calculated under the Encoder-Decoder framework if we can find a reasonable

embedded representation of the related microscopic genotype".

The P-E theorem provides a mathematical basis to establish a deep learning model to calculate virus fitness. This means that under an Encoder-Decoder framework with reasonable embedded representations, we can use only viral genomic sequence data to accurately compute viral population fitness as an observable macroscopic biological phenotype, and construct the genotype-fitness landscape of viral populations (Figure 1A).

Inspired by the P-E theorem and the formal logic of fitness calculation in the proof of the P-E theorem, we propose a more general first principle of genotype-phenotype: "A macroscopically observable biology variable can express as a mathematical expectation of a microscopic state function according to a state probability distribution" (see supplemental note 4, Formula (35) and (35)'). This "first principles" provides a mathematical basis for precisely defining biological phenotypes, allowing us to deeply understand and solve computational problems of biological phenotypes, and at the same time, the P-E theorem provides a computable basis for the application of first principles. On the basis of this work, we further deduce how to start from the first principles of universality and combine the P-E theorem to achieve a more general quantitative calculation framework of biological phenotypes (see Figure 1A). With the first principles and the P-E theorem, we can then design a suitable deep computational model to calculate the biological phenotypes of interest.

Modelling of virus evolution mechanisms based on virus sequences

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 [3,9–11]. To effectively utilize the P-E theorem for fitness calculation, we need to devise a natural language model implemented within the Encoder-Decoder framework.

Different from B. Hie's work [3], we construct a pre-trained + supervised fine-tuning(SFT) foundation model, a Co-attention-based Transformer pre-train model CoT2G-F to bridge Genotype and Fitness (see Figure 1): (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 (Supplementary Figure 4).

Our approach stands out from other NLP models used in the study of viral sequence evolution primarily due to our optimized supervised fine-tuning (SFT)

153 strategy. Specifically, we first start from the pre-training model, combine the P-E
154 theorem to determine the mutation that contributes the most to the fitness of the virus
155 as the target sequence, and then use the time-dependent dynamic fine-tuning mode to
156 obtain the spatio-temporal selection signal of virus evolution that may not be directly
157 attainable through other models (Figure 1C and 1D, Supplementary Figure 4A and 4B).

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

170 **Identifying and predicting "immune escape" mutations**

171 Following the CSCS decision rule for semantics and syntax changes proposed by
172 B. Hie [3], we tested the performance of our model CoT2G-F against two related
173 mainstream models from the framework design in three technical dimensions. Our
174 model demonstrated good performance (see supplementary materials). As shown in

175 Figure 2A, our model outperformed the BiLSTM and Vanilla Transformer models, and
176 the ability to identify and predict immune escape mutations was improved. By
177 considering the spatio-temporal dynamics of virus evolution during training processes,
178 the model was able to identify existing and predict future "immune escape" mutations.
179 We tested this using more than 2 million SARS-CoV-2 genomic sequence data from the
180 United Kingdom as an example (Figure 2B, 2C and 2D, Supplementary Figure 4, Figure
181 5C and 5D).

182 Convergent evolution is a widespread phenomenon observed throughout the
183 evolutionary process, including the later stages of viral epidemics. It refers to the
184 independent emergence of similar traits or adaptations in different virus lineages, often
185 in response to similar environmental pressures or functional requirements. This
186 evolution typically occurs under strong environmental selection constraints [13].
187 Results showed that our model can predict and reproduce the convergent evolutionary
188 mutations. For instance, nine of the 18 convergent mutation sites in the representative
189 lineages of pre-Omicron and Omicron [13–15] were predicted with the model
190 (Supplementary Figure 6).

191 Notably, our model was fine-tuning trained using data from GISAID as of March
192 31, 2022, demonstrating its ability to give accurate predictions up to six months in
193 advance without dynamic fine-tuning. Besides, our model can also be used as a
194 generative model to generate possible immune escape mutations for screening and early
195 warning study (Supplementary Figure 6).

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 relative basic reproduction number (R_0). The problem is establishing a 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. If evolution by natural selection is to occur, there must be a genotypic change in the virus population across generations. This change dependent on the genotype; if the genotype does not change, there is no change in fitness. 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 [3]. 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)). Based on the above reasoning, we can argue that F. Obermeyer et al. [4] defined virus per-lineage fitness as precisely a special case of the P-E theorem. These results suggest that the term related to the "semantics" change of the model CoT2G-F and the CSCS criterion [3] 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. [3]. 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 (Figure 3).

Figure 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 "immune escape" ability of the virus lineage is a calculable and quantifiable parameter directly related to the virus genome sequence mutation, which can be used to assess the cumulative collective effects of the immune escape mutations in an emerging virus lineage. The absolute mean of the *CSC* score accurately describes the changing trend

237 of the “immune escape” ability of virus lineages and thus reveals the epidemic potency
238 of the virus. This result further suggests that the *CSC* score is the microstate function
239 corresponding to the virus fitness indicated by the P-E theorem.

240 **Expressing the fitness of a virus lineage as a convolution of “semantics” and**
241 **“grammar” and the mathematical expectation of a semantic change function**
242 **according to a hidden state probability distribution**

243 It has long been a goal in biology to create genotype-fitness landscapes by
244 mapping DNA sequences to mutation combinations observed in phylogeny or
245 evolutionary experiments. However, due to the consideration of the epistatic effects of
246 mutations, when the mutation number is large, the model becomes complicated and
247 difficult to calculate and verify experimentally [6,7,16] . E. D. Vaishnav et al. [17] argue
248 that the complete fitness landscape defined by a fitness function maps every sequence
249 (including mutations) in the sequence space to its associated fitness. However, no
250 theoretical model has yet been able to provide a concise computational model and
251 framework for the genotype-fitness landscape while fully considering epistasis and
252 recombination potential. Under the framework of the P-E theorem, we attempt to solve
253 this problem through our deep learning natural language model CoT2G-F combined
254 with *CSC* scores.

255 Based on the functional relationship between the *CSC* score and the latent variable
256 of the CoT2G-F model (see Supplemental Note 4, Formula (22)), we know that the *CSC*
257 score is a relative measure of semantic change corresponding to the relative immune
258 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 (Figure 3A and 3B, Figure 4). This result and the P-E theorem provide a mathematical basis for redefining macro-phenotypes such as R_0 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

281 according to a hidden state probability distribution, giving the quantitative relationship
282 between the "immune escape" ability of the virus and the fitness of the virus lineage.
283 Finally, we present a comprehensive and precise formulation of the virus genotype-
284 fitness landscape. The consistency between these two different representations provides
285 a solid mathematical basis and research paradigm for applying deep learning models to
286 biological problems and gaining interpretability (see Supplemental Notes 1-4).

287 Studying the intrinsic relationship between statistical mechanics and deep learning
288 has always been an intriguing subject [18–20]. We know that the core concept of
289 statistical mechanics is: "The macroscopic physical observables can be characterized
290 as the ensemble average of the corresponding microstate variable functions". Inspired
291 by this, the inference is naturally drawn: "Under the Encoder-Decoder framework, a
292 macro-biological observable can be expressed as the mathematical expectation of a
293 function of a latent variable according to a hidden state distribution in the decoder
294 space". We call this framework bridging genotype-phenotype landscapes is an
295 important corollary of the P-E theorem and gives a novel interpretable application of
296 deep learning theory to life science (Figure 5, Supplemental Note 1 and 4, equations
297 (11), (35) and (36)). Finally, we can plot the genotype-fitness landscape in the
298 embedded space based on the variation state of the viral genome sequence according to
299 equation (30). It is a two-dimensional hypersurface, and we can obtain the fitness of the
300 virus lineage by integrating the surface density of the specific region corresponding to
301 the virus lineage on this hypersurface, thereby defining the immune escape force of the
302 virus (see Supplemental Note 4, Figure 5B and Supplementary Figure 9).

303 **Inferring the R_0 of the virus lineage**

304 According to our theoretical framework, computing the fitness of each virus
305 lineage and the fold increase in relative fitness to obtain the transmissibility potential，
306 ie, basic reproduction number (R_0) of SARS-CoV-2 requires adequate sampling for each
307 virus lineage in a specific region and time interval. In this region and time interval, the
308 virus begins to emerge, sustain its spread, increase in number, and eventually reach a
309 plateau. Based on the biological definition of fitness, the selected sampling time interval
310 needs to ensure the virus is transmitted for enough generations to reduce the violent
311 fluctuation of the signal [4]. It is also crucial for determining the occurrence frequency
312 of the dominant cluster of the different lineages and simultaneously calculating the
313 absolute mean of the *CSC* score (see Supplemental Note 4, Formula (30), Figures S3A
314 and S3B). Currently, the GISAID (www.gisaid.org) contains over 2 million genome
315 sequences of SARS-CoV-2 from the United Kingdom [http://gisaid.org], and the spatio-
316 temporal distribution of the virus lineages reflects the major trend of the world epidemic
317 of SARS-CoV-2, which can serve as a model system for a reliable and feasible G-P
318 principle.

319 We computed the virus per-lineage fitness using 2.7 million SARS-CoV-2 spike
320 proteins from the GISAID database (www.gisaid.org) and adopted the Pango lineage
321 designation and assignment. To calculate the factor λ_k in the per-lineage fitness (see
322 Box 1, Formula (9) and Supplemental Note 4, Formula (30)), we first performed lineage
323 assignment with Pangolin. We then selected the k_{th} lineage and its nearest sublineages
324 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 Figure 4, where the time interval we chose was 10 days (see Supplementary Figure 6). Our model correctly inferred that the WHO classification variant Omicron lineages (Pango lineage BA.2.X) had very high relative fitness in a pandemic. It was approximately: ~ 10-12 times higher than the original Wuhan and lambda lineages (Figure 4A, [95% confidence interval (CI) 16.78 to 16.84], Supplementary Figure 7), accurately predicting its rise in the spread regions.

In our study, we propose a simple calibration process to obtain the absolute basic reproduction number ($\hat{R}_0$) of virus lineages by using an observable actual number of the absolute basic reproduction number ($\hat{R}_0$) for the Delta lineage of SARS-CoV-2. (Figure 4A, Supplemental Note 4, Formula (34)). Our calculated absolute basic reproduction numbers ($\hat{R}_0$) align with the results obtained by F. Obermeyer *et al.* [4] (Figure 4C) and are also in agreement with the actual and official monitoring results (Figure 4D). This result validates our proposed P-E theorem, allowing for the precise computation of the relative fitness of the virus lineage and R_0.

Discussion

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 a mathematical basis for interpretability. This study explores these two aspects and presents an initial research paradigm. In the first step, we propose a general theorem, the P-E theorem, which is a computational path of macrobiological phenotypes under the framework of deep learning. Then in the

second step, we constructed an encoder-decoder Seq2Seq deep learning model CoT2G-F that is more in line with the evolutionary biology scenario. The model can accurately predict the immune escape mutation of SARS-CoV-2 and calculate the immune escape ability of the virus lineage. In particular, because the CoT2G-F is a basic pre-trained foundation model for virus evolution research, the model can perform different downstream analysis tasks by various supervised fine-tuning (SFT).

In our work, The calculation of R_0 was realized by the model CoT2G-F under the guidance of the P-E theorem, which gives the first principle-based interpretability of the model. The R_0 is a mathematical expectation of the "semantic" change according to a hidden state distribution, which 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", and is a corollary of the P-E theorem, which 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 [18–20].

The practice and results of the R_0 calculation in the present work provide a feasible paradigm for the computational modelling of genotype-phenotype landscapes and a compelling commentary on the first principle of "selection imprinting recorded in viral

genomic mutation", which is a concrete and successful example of the P-E theorem. We believe that, following the research path proposed in the present work, many research examples that adhere to the P-E theorem will emerge in the future.

In our study, we employ the Transformer as a kernel in our model, incorporating co-attention、self-attention and continuous span masking mechanisms to mimic the spatial-temporal dynamical associations of virus evolution. This comprehensive approach that combines the main deep learning algorithms provides several noteworthy benefits. Firstly, it enables us to extract long-range correlation mutation information from both upstream and downstream regions of the viral sequence, thereby capturing crucial spatio-temporal features relevant to virus evolution. Secondly, our model can effectively simulate genetic drift and discerns which mutations are more likely to be advantageous and subsequently fixed through natural selection, and exhibits significantly improved accuracy in the identification and prediction of potential future "immune escape" mutations. The efficacy of our model is clearly demonstrated in Figures 2A, 2B, 2C, and 2D, which provide compelling evidence of its predictive capabilities.

Importantly, we have constructed a formal three-dimensional embedding space to visualize the virus genotype-fitness landscape, utilizing the quantitative correlation between "immune escape" mutations and viral lineage fitness. This pioneering approach represents the first precise mathematical representation of a virus' genotype-fitness landscape, enabling us to understand the topological properties that drive virus evolution and immune escape. The development of this fundamental methodology

establishes a revolutionary theoretical and technical framework for future research, with potential applications across diverse contexts.

The proposed P-E theorem may demonstrate its universality in biological research by effectively studying fundamental phenotypic states in biology, such as brain homeostasis, metabolic homeostasis, cell fate state, ‘tumour microenvironment, survival capability, and advanced ageing state [21]. This theorem establishes a robust mathematical foundation for deterministically describing biological homeostasis and deciphering its stability mechanisms.

Acknowledgements

We thank Guoping Zhao, Li Jin, Jiarui. Wu, Peiji Zeng, Ruibin Liu and Zhijian Lang for their deep and helpful discussions. We thank Tao Zeng and Junwei Liu for their assistance with the manuscript. We are also particularly grateful to Brian Hie for his constructive responses to our questions about the details of his related excellent models.

Funding

This work was supported by Grant No. XDB38050200, R&D Program of Guangzhou Laboratory, Grant No. SRPG22-001, and R&D Program of Guangzhou Laboratory, Grant No. SRPG22-007.

Author contributions

Conceptualization: Yixue Li. Methodology: Yao. Luo, Yuexing Liu, Xin Lu, Hao Gao, Ruikun He, Xin. Zhang, Xuguan Zhang and Yixue. Li. Investigation: Yao Luo, Yuexing Liu, Xin Lu, Hao Gao and Ruikun He. Resources: Yixue Li. Data curation: Yuexing Liu, Yao. Luo, Xin Lu, Hao Gao and Ruikun He. Writing, original draft: Yixue Li, Yuexing Liu, Yao Luo, Xin Lu, Hao Gao. Writing, review and editing: Yixue Li, Yuexing Liu. Supervision: Yixue Li. Funding acquisition: Yixue Li.

Competing interests

All the authors have no conflicts of interest to declare.

Methods

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 (Extended Data Table 1). Lineage assignment was performed with Pangolin (version 4.1.2) [22] subsequently. To get the spike protein sequences, genome sequences were firstly aligned to SARS-CoV-2 reference genome (Wuhan-Hu-1, NC_045512.2) [23] using the MAFFT7 (version 7.505) [24]. 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 (Supplementary Figure 1). 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 (Supplementary Figure 4). In each month, sequences were arranged by their evolution distance relative to Wuhan-Hu-1 strain (NC_045512.2) of SARS-CoV-

2 (Supplementary Figure 5). 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 a pre-train foundation model CoT2G-F with downstream supervised fine-tuning(SFT). It is an Encoder-Decoder Seq2Seq architecture natural language model 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

Figure2A 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. Hie [3]. to capture the semantic and syntactic variations of protein sequences. It is worth noting that B. Hie (4) and our method (CoT2G-F) solve different problems: B. Hie [3]. 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 (Figure 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.

484 The first step is the pre-training process. First, from the outbreak of the virus to the
485 present, the time cross-section is divided into months, and the global SARS-CoV-2
486 genome sequences in one month are arranged according to the evolutionary distance to
487 form a two-dimensional genome sequence matrix $\boldsymbol{X_t} = (\boldsymbol{x}_{(t,1)}, \dots, \boldsymbol{x}_{(t,p)}, \dots, \boldsymbol{x}_{(t,N)})$,
488 the row of the submatrix $\boldsymbol{x_{t,p}}$ represents the genome sequence of a virus around spatial
489 position *p*, and *t* is the corresponding time of a certain time cross-section. During pre-
490 training, the co-attention module of CoT2G-F starts from the first row of the matrix $\boldsymbol{X_t}$,
491 and sequentially takes three spatially connected Spike protein sequences
492 $(\boldsymbol{x}_{(t,p-1)}, \boldsymbol{x}_{(t,p)}, \boldsymbol{x}_{(t,p+1)})$, get $(\boldsymbol{x}_{(t,p-1)}, \boldsymbol{x}_{(t,p)})$ and $(\boldsymbol{x}_{(t,p)}, \boldsymbol{x}_{(t,p+1)})$ pairwise co-
493 attention signals, and then the Transformer module randomly masks any amino acid
494 sites of the Spike protein with a proportion under 15%, which is completely different
495 from [4]. We adopt such a pre-training process that can not only replicate the "neutral
496 evolution" source of virus variation but also consider the long-range associations
497 between sites in the viral genome sequence composition rules, as well as possible
498 recombination effects. Therefore, our pre-training model that reflects the "semantics"
499 and "grammar" of viral sequence composition is more reasonable. After completing the
500 first step of training, we can already calculate the "semantic" changes of the Spike
501 protein of each input virus genome sequence as defined in (see supplemental note 4)
502 and select the target sequences according to the *CSCS* defined in [4] for the second step
503 training.

504 The second step of training is the dynamic fine-tuning process. The occurrence of
505 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 $\boldsymbol{X_{t-1}}$, $\boldsymbol{X_t}$, $\boldsymbol{X_{t+1}}$, the five selected sequences in $\boldsymbol{X_{t+1}}$ with the largest *CSCS* scores used as the training targets, and then the three spatially correlated sequences （$\boldsymbol{x}_{(t,p-1)}$, $\boldsymbol{x}_{(t,p)}$, $\boldsymbol{x}_{(t,p+1)}$） are selected according to the same pre-training mode in $\boldsymbol{X_t}$, and finally in $\boldsymbol{X_{t-1}}$, select a sequence $\boldsymbol{x}_{(t-1,p)}$ with the closest evolution distance to the sequence $\boldsymbol{x}_{(t,p)}$ in $\boldsymbol{X_t}$, when calculating the pairwise co-attention mechanism between （$\boldsymbol{x}_{(t,p-1)}$, $\boldsymbol{x}_{(t,p)}$, $\boldsymbol{x}_{(t,p+1)}$）, while computing the co-attention mechanism between $\boldsymbol{x}_{(t-1,p)}$ and $\boldsymbol{x}_{(t,p)}$. 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 [3], 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 relative 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 $CSC=\|\hat{\boldsymbol{z}}-\hat{\boldsymbol{z}}[\tilde{x}_i]\|_{\ell_x}$, $\|\cdot\|_{\ell_x}$ is the ℓ_x norm, $x = 1, 2$ (see supplemental note 1), in which $\hat{\boldsymbol{z}}=\frac{1}{N}\sum_{i=1}^{N}\hat{\boldsymbol{z}}_{\boldsymbol{i}}$. 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 $\hat{\boldsymbol{z}}$ to align and normalize the CSC score (see supplemental note 4).

Data Availability

GISAID sequence data is publicly available at https://gisaid.org. 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/record/7388491. 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/record/7388491.

Code Availability

The source code for data preprocessing and modelling and available at https://github.com/martyLY/CoT2G-F.

557 **References**

558 1. Mathew D, Giles JR, Baxter AE, et al. Deep immune profiling of COVID-19 patients
559 reveals distinct immunotypes with therapeutic implications. Science 2020;
560 369:eabc8511
561 2. Starr TN, Greaney AJ, Hilton SK, et al. Deep Mutational Scanning of SARS-CoV-2
562 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding.
563 Cell 2020; 182:1295-1310.e20
564 3. Hie B, Zhong ED, Berger B, et al. Learning the language of viral evolution and escape.
565 Science 2021; 371:284–288
566 4. Obermeyer F, Jankowiak M, Barkas N, et al. Analysis of 6.4 million SARS-CoV-2
567 genomes identifies mutations associated with fitness. Science 2022; 376:1327–
568 1332
569 5. Maher MC, Bartha I, Weaver S, et al. Predicting the mutational drivers of future
570 SARS-CoV-2 variants of concern. Sci. Transl. Med. 2022; 14:eabk3445
571 6. de Visser JAGM, Krug J. Empirical fitness landscapes and the predictability of
572 evolution. Nature Reviews Genetics 2014; 15:480–490
573 7. Fragata I, Blanckaert A, Dias Louro MA, et al. Evolution in the light of fitness
574 landscape theory. Trends in Ecology & Evolution 2019; 34:69–82
575 8. Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal
576 approximators. Neural Networks 1989; 2:359–366
577 9. Alley EC, Khimulya G, Biswas S, et al. Unified rational protein engineering with
578 sequence-based deep representation learning. Nat Methods 2019; 16:1315–1322
579 10. Bepler T, Berger B. Learning protein sequence embeddings using information from
580 structure. 2019;
581 11. Rao R, Bhattacharya N, Thomas N, et al. Evaluating Protein Transfer Learning with
582 TAPE. Advances in Neural Information Processing Systems 2019; 32:
583 12. Vaswani A, Shazeer N, Parmar N, et al. Attention is All you Need. Advances in
584 Neural Information Processing Systems 2017; 30:
585 13. Cao Y, Jian F, Wang J, et al. Imprinted SARS-CoV-2 humoral immunity induces

586 convergent Omicron RBD evolution. Nature 2022;
587 14. Focosi D, Quiroga R, McConnell S, et al. Convergent Evolution in SARS-CoV-2
588 Spike Creates a Variant Soup from Which New COVID-19 Waves Emerge.
589 International Journal of Molecular Sciences 2023; 24:2264
590 15. Ito J, Suzuki R, Uriu K, et al. Convergent evolution of SARS-CoV-2 Omicron
591 subvariants leading to the emergence of BQ.1.1 variant. Nat Commun 2023;
592 14:2671
593 16. Kondrashov DA, Kondrashov FA. Topological features of rugged fitness landscapes
594 in sequence space. Trends in Genetics 2015; 31:24–33
595 17. Vaishnav ED, de Boer CG, Molinet J, et al. The evolution, evolvability and
596 engineering of gene regulatory DNA. Nature 2022; 603:455–463
597 18. Bahri Y, Kadmon J, Pennington J, et al. Statistical Mechanics of Deep Learning.
598 Annual Review of Condensed Matter Physics 2020; 11:501–528
599 19. Carleo G, Cirac I, Cranmer K, et al. Machine learning and the physical sciences.
600 Rev. Mod. Phys. 2019; 91:045002
601 20. Liu Y, Yao X. Ensemble learning via negative correlation. Neural Networks 1999;
602 12:1399–1404
603 21. Wang F, Liu J, Gao F, et al. Exploring multi-omics latent embedding spaces for
604 characterizing tumor heterogeneity and tumoral fitness effects. 2023;
605 2023.07.05.547886
606 22. Rambaut A, Holmes EC, O'Toole Á, et al. A dynamic nomenclature proposal for
607 SARS-CoV-2 lineages to assist genomic epidemiology. Nature Microbiology 2020;
608 5:1403–1407
609 23. Wu F, Zhao S, Yu B, et al. A new coronavirus associated with human respiratory
610 disease in China. Nature 2020; 579:265–269
611 24. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7:
612 Improvements in Performance and Usability. Molecular Biology and Evolution
613 2013; 30:772–780
614
615

616 **Key points**

617 We present a precise formulation of the fundamental principles governing the
618 calculation of biological phenotypes. Specifically, macroscopically observable
619 biological variables (biological phenotypes) can be mathematically represented by the
620 expectations of their corresponding microstate functions within their microstate
621 distribution.

622 We introduce and substantiate the genotype-phenotype embedding theorem (P-E
623 theorem). This theorem, established for the first time, demonstrates that the
624 foundational principles of biological phenotype calculation can be accurately computed
625 within the framework of deep learning.

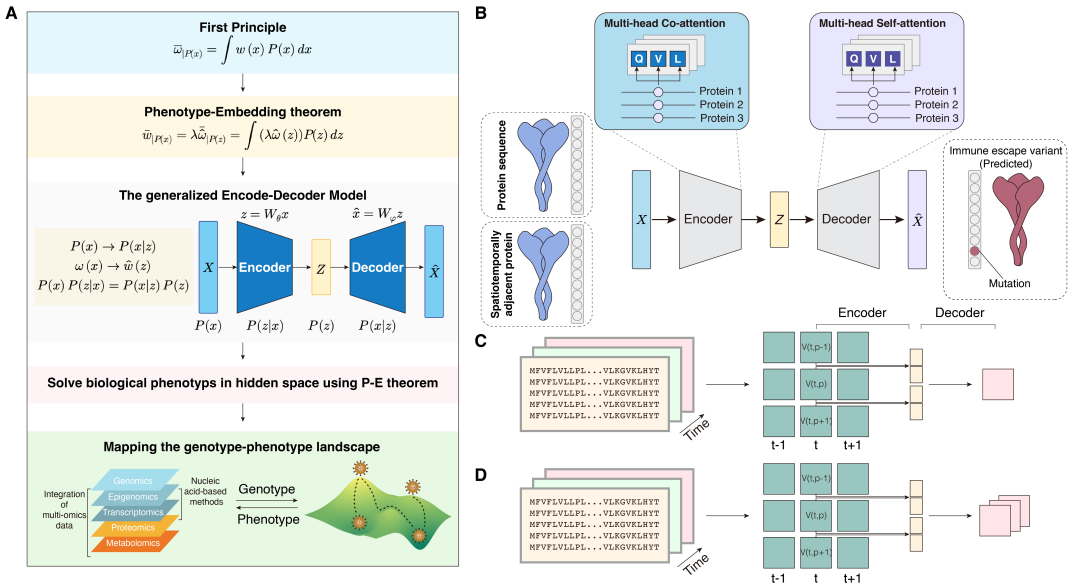
626 Employing a deep learning model, we successfully calculate the fitness (R_0) of
627 SARS-CoV-2, an observable macro-phenotype. We provide a concrete example of
628 biological phenotype computation.

629 The deep learning model we devised can effectively simulate the genetic drift and
630 environmental selection involved in virus evolution. Furthermore, it accurately predicts
631 the immune escape mutations of SARS-CoV-2.

632

Figure legends

Figure 1. The P-E theorem, the framework of the CoT2G_F model and simulating the real evolution scenario of viruses. (A) Biological phenotype calculation processes. The process has five steps: 1. First principle, 2. The P-E theorem, 3. The generalized encoder-decoder deep learning model, 4. Solve biological phenotyps inhidden space using P-E theorem, 5. Mapping the genotype-phenotype landscape. **(B)** CoT2G_F is a Transformer-based pre-train foundation model with downstream supervised fine-tuning(SFT), it is a standard encoder-decoder architecture that introduces a co-attention and continuous span masking mechanism. **(C** and **D)** Shows the co-attention combined pre-training and fine-tuning steps of the model (see Supplementary Materials, Figure Extended Data 4).



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

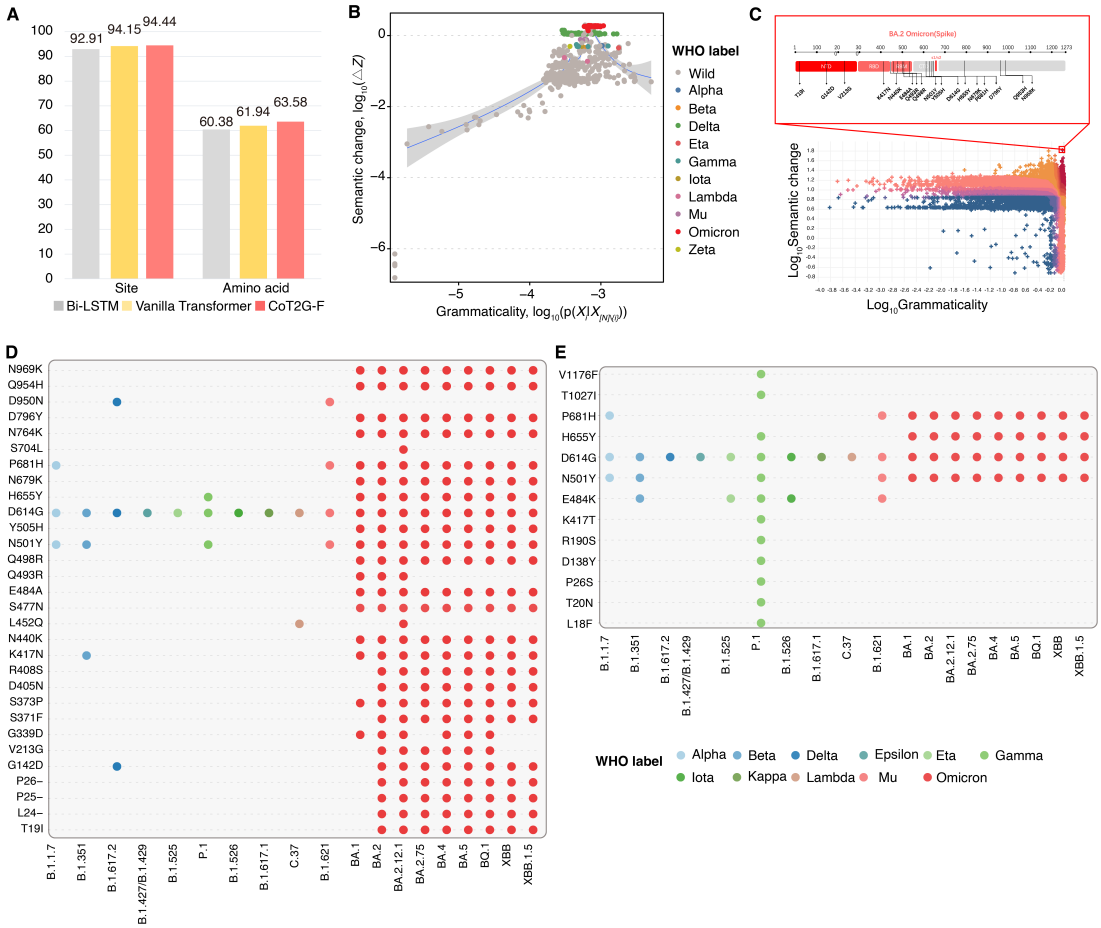
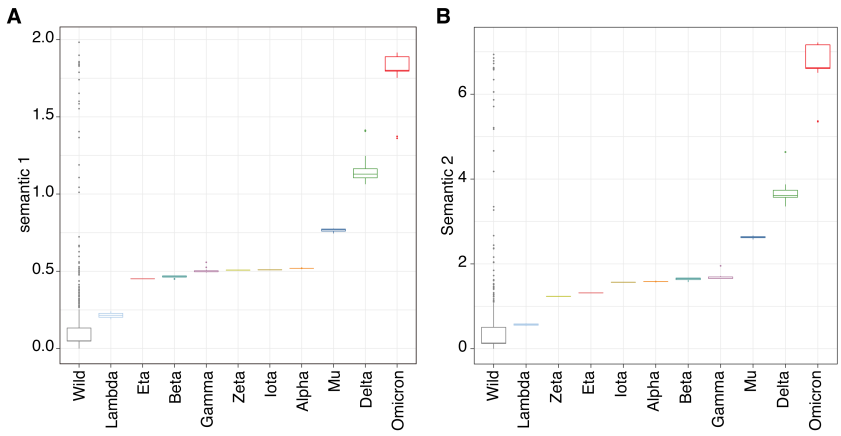
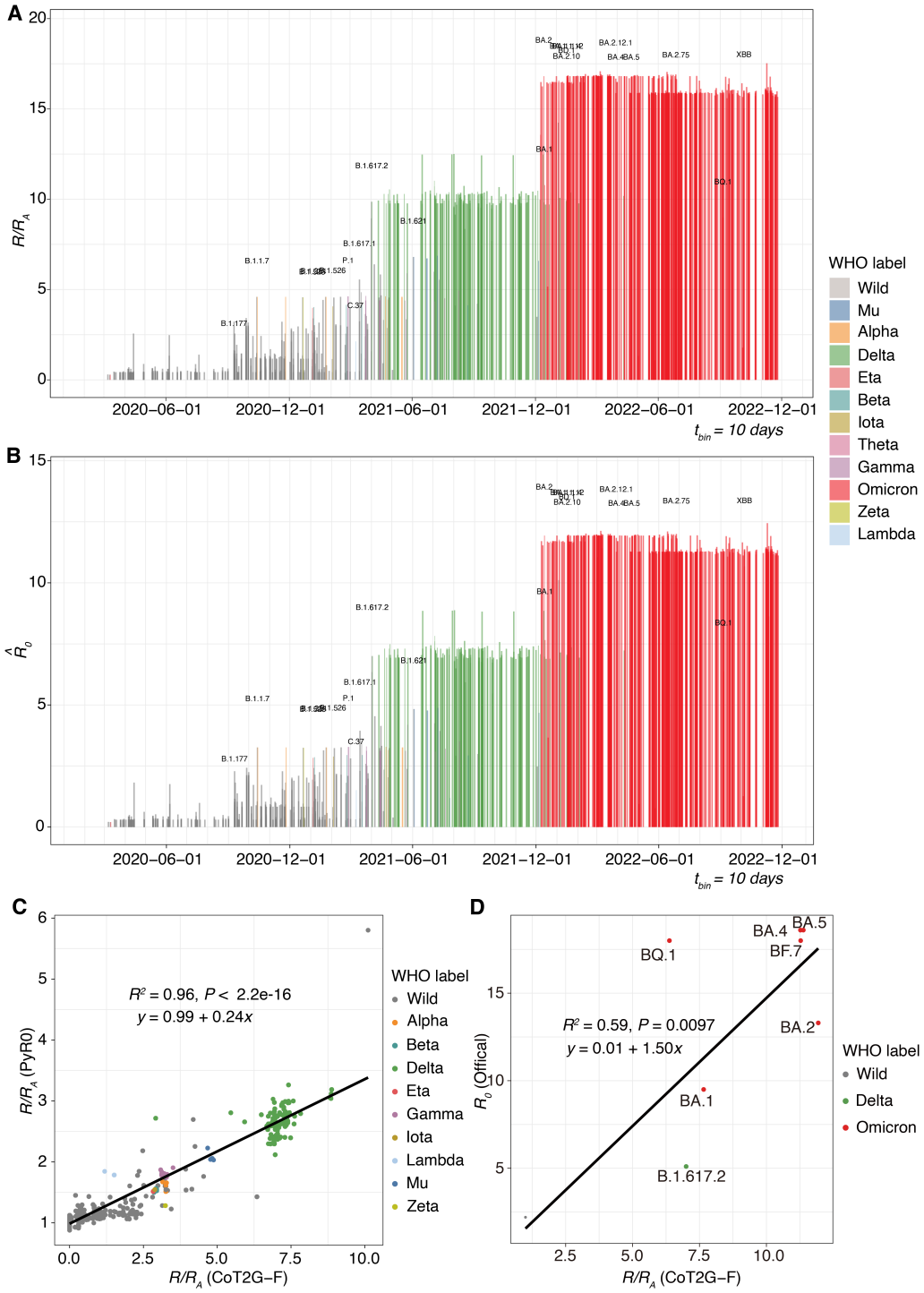

A
92.91
94.15
94.44
60.38
61.94
63.58
Site
Amino acid
Bi-LSTM
Vanilla Transformer
CoT2G-F
B
Semantic change, log10(△Z)
Grammaticality, log10(p(X|X(N|0)))
WHO label
Wild
Alpha
Beta
Delta
Eta
Gamma
Iota
Lambda
Mu
Omicron
Zeta
C
BA.2 Omicron(Spike)
Log10 Semantic change
Log10 Grammaticality
D
E
WHO label
Alpha
Beta
Delta
Epsilon
Eta
Gamma
Iota
Kappa
Lambda
Mu
Omicron

Figure 3. Absolute mean distribution plot of the semantic change parameter 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 CSC_i^k of the k_{th} virus lineage, which is defined as the immune escape capacity of the 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 CSC_i^k, but figure (B) changes is more smoothly.



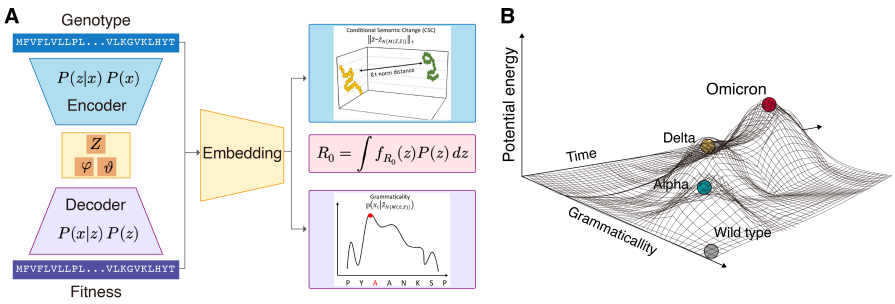
680 **Figure 4. Relative fitness derived from the natural language model CoT2G-F**
681 **versus date of lineage emergence.** (A) This figure uses a time interval of 10 days.
682 Referring to the Pango lineage designation and assignment, the relative basic
683 reproduction number (R_0), which is the fold increase in relative fitness of the virus
684 lineages according to the Wuhan lineage, is plotted in different colours. The results for
685 5, and 20-day time intervals have shown in (Supplementary Figure 6). All the results
686 are almost consistent, reflecting the robustness of the model. (B) The transmissibility
687 potential (ie, absolute basic reproduction number ($\hat{R}_0$)) of virus lineages of SARS-CoV-
688 2 by using an observable actual number of the absolute basic reproduction number ($\hat{R}_0$)
689 for the Wuhan lineage of SARS-CoV-2. (C) Comparison of the R/R_A calculated by the
690 Co2TG-F with the values calculated with PyR0. (D) Comparison of the R/R_A calculated
691 by the Co2TG-F with the official results.


A
B
C
D
R/R_A
R̂_0
t_bin = 10 days
2020–06–01
2020–12–01
2021–06–01
2021–12–01
2022–06–01
2022–12–01
BA.2
BA.2.10
BA.2.12.1
BA.4BA.5
BA.2.75
XBB
BA.1
BQ.1
B.1.617.2
B.1.621
B.1.617.1
P.1
B.1.1.7
B.1.526
C.37
B.1.177
WHO label
Wild
Mu
Alpha
Delta
Eta
Beta
Iota
Theta
Gamma
Omicron
Zeta
Lambda
R² = 0.96, P < 2.2e-16
y = 0.99 + 0.24x
R/R_A (PyR0)
R/R_A (CoT2G–F)
R² = 0.59, P = 0.0097
y = 0.01 + 1.50x
R_0 (Offical)
BA.4
BA.5
BF.7
BQ.1
BA.2
BA.1
B.1.617.2

692

693

Figure 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 $P(\boldsymbol{x})$ and hidden state distribution $P(\boldsymbol{z}|\boldsymbol{x})$ to the prior probability $P(\boldsymbol{z})$ and the posterior probability distribution $P(\boldsymbol{x}|\boldsymbol{z})$ 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 posterior probability distribution $P(\boldsymbol{z})$ 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 Genotype-Fitness landscape, in which only three virus lineages, including Omicron, were shown.



Tables

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:

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

$\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots\ldots, \lambda_K)$, 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:

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

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

$$\omega_k = \frac{1}{n_k}\sum_{i=1}^{n_k}\omega_i^k \quad (5)$$

ω_k is a mathematical expectation of the contribution of a single virus strain to virus lineage or virus population (if $\lambda_k = \pi_k$) fitness. Referring to formula (3) and (4),

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

The subscript $P(\cdot)$ represents the state probability distribution corresponding to virus population. As long as we find a suitable functional form of $\omega(\boldsymbol{x})$ 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 $\omega(\boldsymbol{x})$ 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 $P(\boldsymbol{z}|\boldsymbol{x})$ 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 $P(\boldsymbol{x}|\boldsymbol{z})$ be the prior and posterior probability related to the input and output of the model. The relationship among $P(\boldsymbol{x})$, $P(\boldsymbol{z}|\boldsymbol{x})$, $P(\boldsymbol{x}|\boldsymbol{z})$ and $P(\boldsymbol{z})$ is given by Bayes' theorem:

$$P(\boldsymbol{x})P(\boldsymbol{z}|\boldsymbol{x}) = P(\boldsymbol{x}|\boldsymbol{z})P(\boldsymbol{z}) \quad (7)$$

The function relationship defined by formula (1) still holds when $P(\boldsymbol{x})P(\boldsymbol{z}|\boldsymbol{x})$ maps to $P(\boldsymbol{x}|\boldsymbol{z})P(\boldsymbol{z})$. 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 $P(\boldsymbol{z}|\boldsymbol{x})$ and $P(\boldsymbol{z})$, the posterior probability distribution $P(\boldsymbol{x}|\boldsymbol{z})$ and the counterpart $\widetilde{\omega}(\boldsymbol{z})$ of $\omega(\boldsymbol{x})$.

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

$$\overline{\omega}_{|P(\boldsymbol{x})} = \int \omega(\boldsymbol{x})\, P(\boldsymbol{x})d\boldsymbol{x} = \int \omega(\boldsymbol{x})\, P(\boldsymbol{x})(P(\boldsymbol{x}\,|\,\boldsymbol{z})/P(\boldsymbol{z}\,|\,\boldsymbol{x}))d\boldsymbol{x} = \int \omega(\boldsymbol{x}')\, P(\boldsymbol{x}')d\boldsymbol{x}'$$

$$= \int \omega(W\boldsymbol{z}+B)\, P(W\boldsymbol{z}+B)|\frac{d\boldsymbol{x}'}{d\boldsymbol{z}}|d\boldsymbol{z} = \int \widetilde{\omega}(\boldsymbol{z})\, P(\boldsymbol{z})\, |W|d\boldsymbol{z}$$

and $\widetilde{\omega}(\boldsymbol{z}) = \omega(\boldsymbol{x}(\boldsymbol{z}))(P(\boldsymbol{x}(\boldsymbol{z})|\boldsymbol{z})/P(\boldsymbol{z}|\boldsymbol{x}(\boldsymbol{z}))), \frac{d\boldsymbol{x}}{d\boldsymbol{z}} = W$ is the Jacobian matrix, and $|W|$ is a determinant of Jacobian matrix that is the sum of combinations of products of elements in different positions of all the columns of matrix *W*. Considering the quasi-orthogonal property of base distribution in *GMM* expansion for ideal classification, let $\lambda_k = |W|_k$ is a mapping of the $|W|$ corresponding to the k_{th} virus lineages, and $\widetilde{\omega}_k$ is a hidden space state function corresponding to the fitness of the k_{th} virus lineage. We set $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)$ and $\widetilde{\boldsymbol{\omega}}(\boldsymbol{z}) = (\widetilde{\omega}_1, \ldots, \widetilde{\omega}_K)$ are vectors, then $|W|\widetilde{\omega}(\boldsymbol{z}) \to \boldsymbol{\lambda} * \widetilde{\boldsymbol{\omega}}(\boldsymbol{z})$. Finally, we have

$$\overline{\omega}_{|P(\boldsymbol{x})} = \int \omega(\boldsymbol{x})\, P(\boldsymbol{x})\, d\boldsymbol{x} = \int |W|\widetilde{\omega}(\boldsymbol{z})\, P(\boldsymbol{z})d\boldsymbol{z} = \int (\boldsymbol{\lambda} * \widetilde{\boldsymbol{\omega}}(\boldsymbol{z}))\, P(\boldsymbol{z})d\boldsymbol{z} \quad (8)$$

Obviously, formula (8) is the corresponding formal representation of formula (1) under the Encoder-Decoder Seq2Seq framework of deep learning, then we get a simple representation:

$$\overline{\omega}_{|P(\boldsymbol{x})} = \boldsymbol{\lambda} * \overline{\widetilde{\boldsymbol{\omega}}}_{|P(\boldsymbol{z})} \quad (9)$$

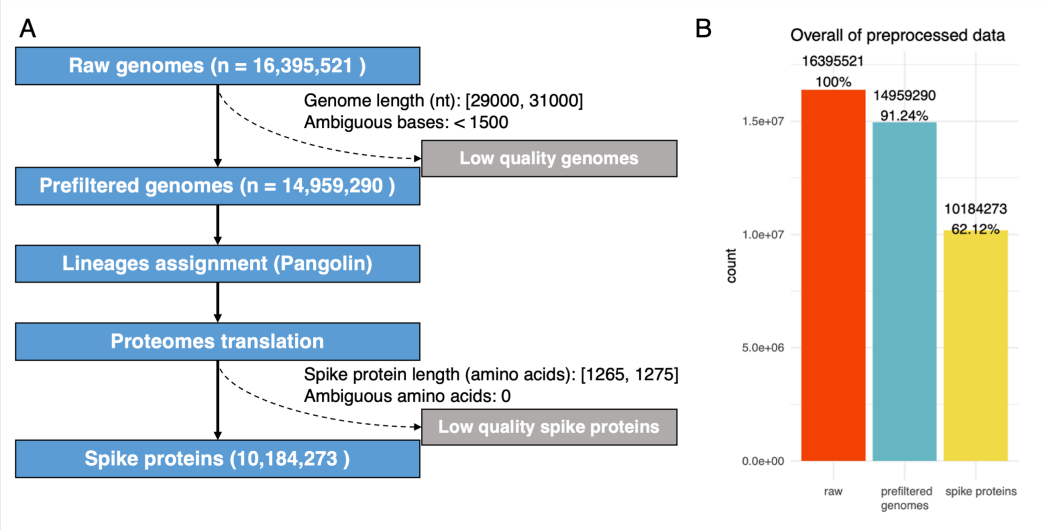
Because of the hierarchical structure of the population, lineage and cluster of the virus, and referring to the formula(1), (3), (4), (6) and (8), and based on the basic principles of population genetics (*14*), the elements of the vector $\boldsymbol{\lambda}$ are the corresponding elements of mixture 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 $\widetilde{\omega}(\boldsymbol{z})$ 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 "$\widetilde{\omega}(\boldsymbol{z})$", "$P(\boldsymbol{z})$", 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 $\widetilde{\omega}(\boldsymbol{z})$ along the time and the $P(\boldsymbol{z})$ axis on a two-dimensional hypersurface

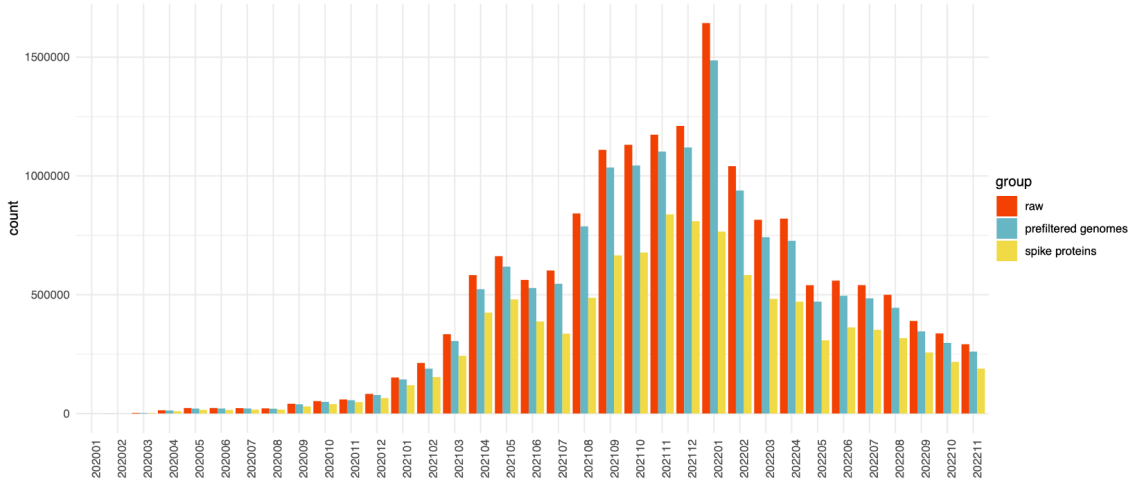
Supplementary Figures

Supplementary Figure 1 | Preprocessing of SARS-CoV-2 sequences

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



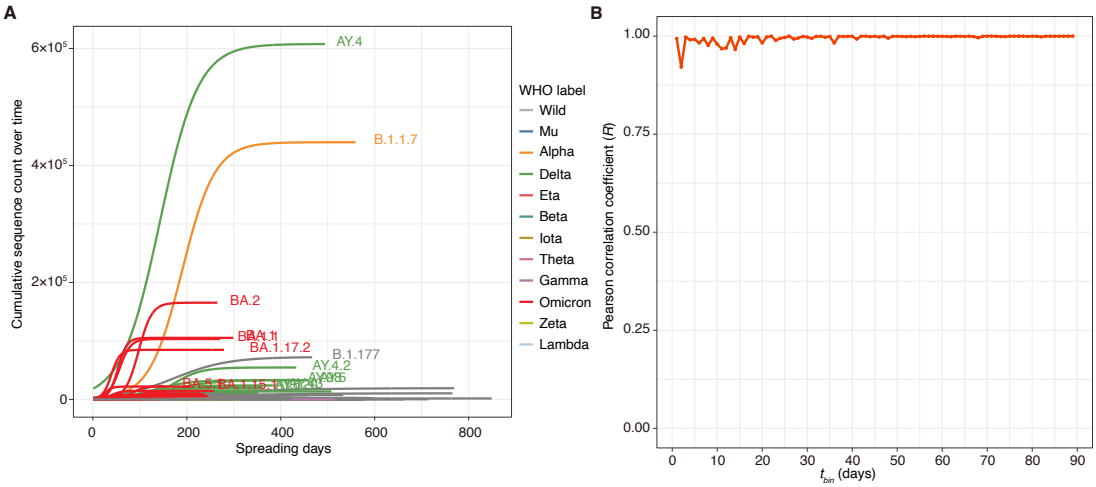
6 **Supplementary Figure 2 | Summary of the preprocessed SARS-CoV-2 sequences**



7

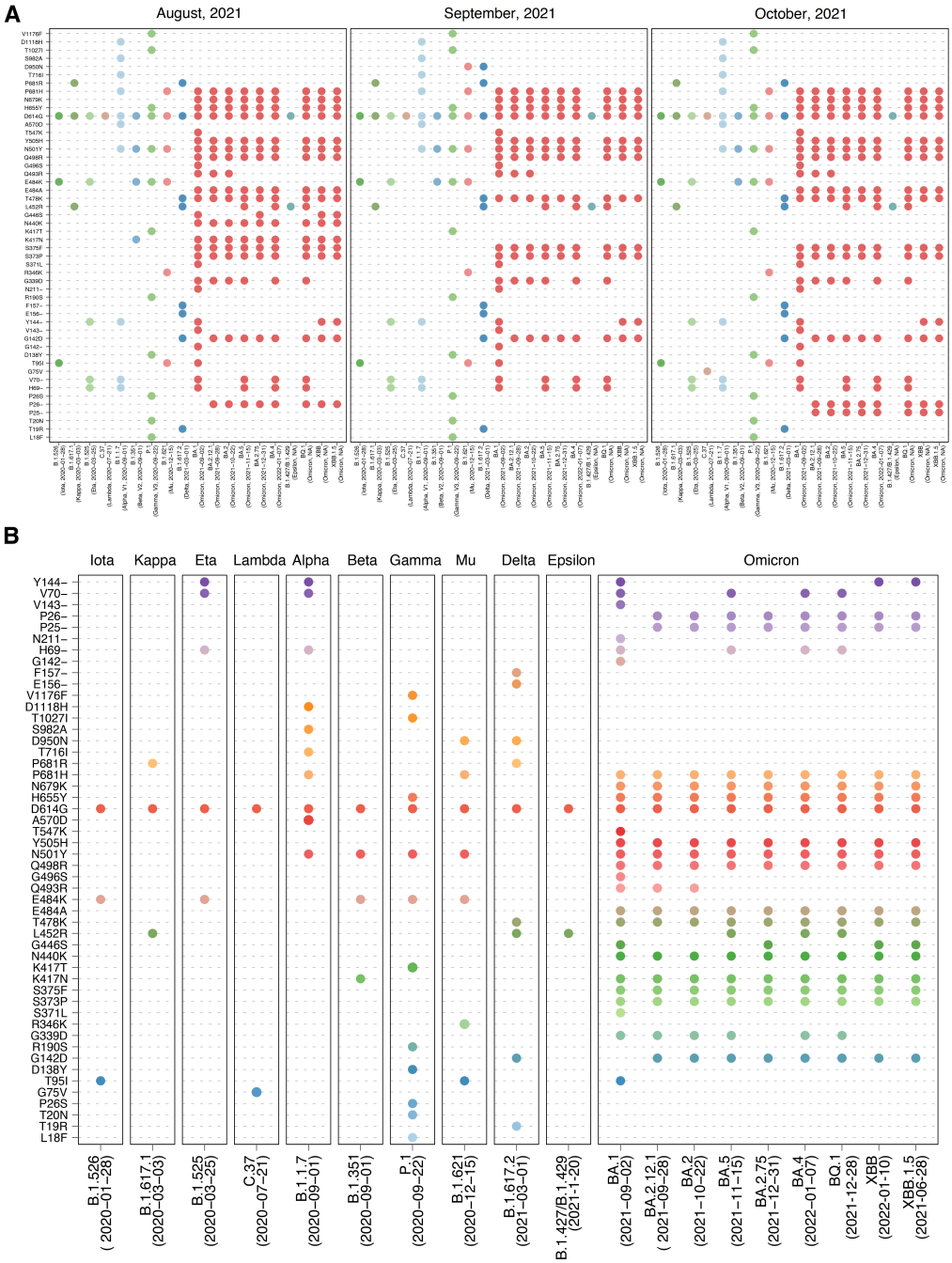
Supplementary Figure 3 | Growth rate of representative SARS-CoV-2 lineages in the UK

(A) Growth rates of SARS-CoV-2 lineages were fitted with a logistic model [1]. 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 dominant 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 dominant cluster in different lineages. (B) The different t_{bin} versus the correlations among the distributions of the occurrence frequencies of the dominant 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}.



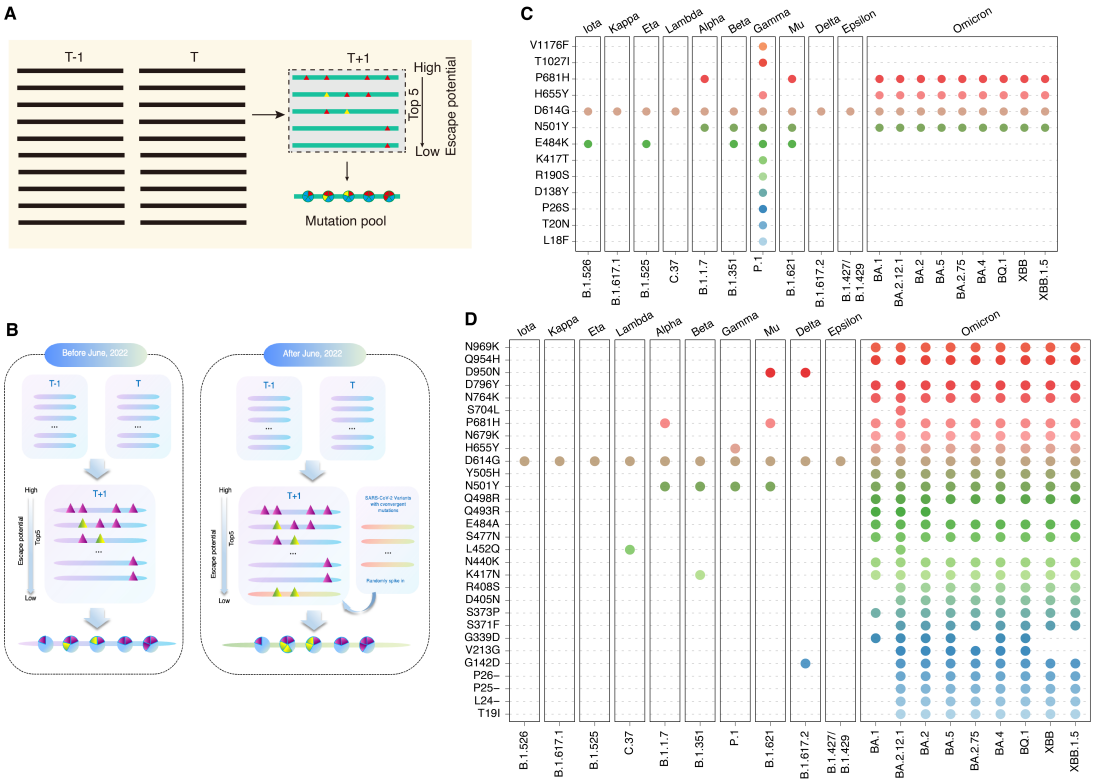
Supplementary Figure 4 | Reproducing known immune escape mutations

The prediction results show that the model has good reproducibility for already emerged "immune escape" mutations. We took data from August 2021 to October 2021 in GISAID (www.gisaid.org). The data were collected and submitted to the database by the United Kingdom.



Supplementary Figure 5 | The schematic diagram for the target sequences search of the dynamic fine-tuning step

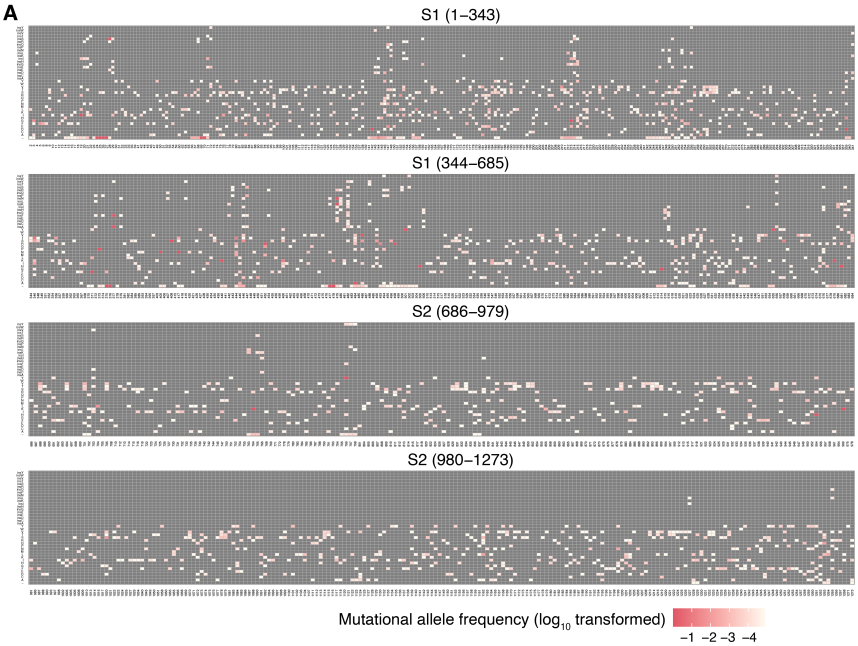
(**A**) The model selects virus sequences from three continuous cross-sections time, T-1, T, and T+1 time for the dynamic fine-tuning. Find the five virus sequences at T+1 time whose mutation sites are the key mutations that determine the virus lineage and whose sequences best meet the semantic and grammatical criteria. The five sequences are selected targets for one of the input sequence, and then supervised fine-tuning 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 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 were reproduced.

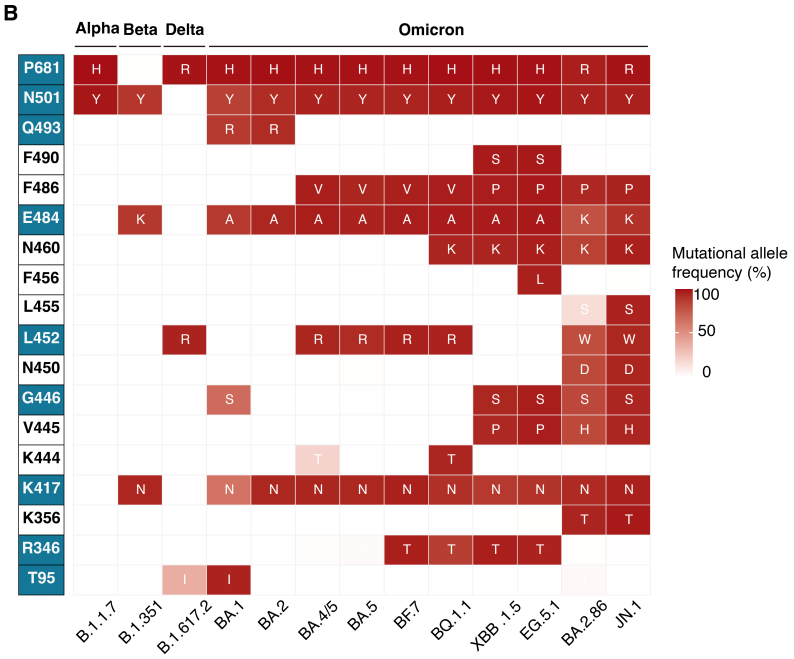

A
T-1
T
T+1
High
Top 5
Low
Escape potential
Mutation pool
B
Before June, 2022
After June, 2022
C
Iota
Kappa
Eta
Lambda
Alpha
Beta
Gamma
Mu
Delta
Epsilon
Omicron
V1176F
T1027I
P681H
H655Y
D614G
N501Y
E484K
K417T
R190S
D138Y
P26S
T20N
L18F
B.1.526
B.1.617.1
B.1.525
C.37
B.1.1.7
B.1.351
P.1
B.1.621
B.1.617.2
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
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

46

Supplementary Figure 6 | Model-Generated Spectrum of Immune Escape Mutants of SARS-CoV-2 and convergent evolutionary mutations

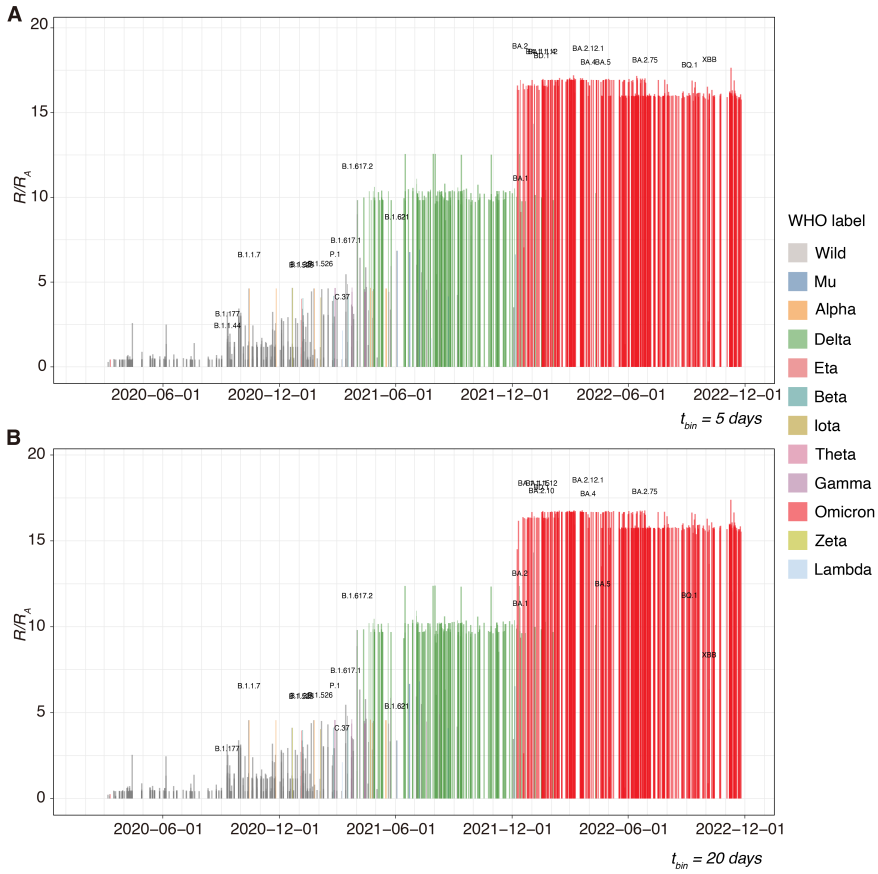
(**A**) The predicted immune escape mutant spectrum of SARS-CoV-2, including the three variants of the recently emerged virus lineage XBB.1.16. Heatmaps illustrating the predicted mutations across the Spike protein. The vertical coordinate stands for the possible mutations at each site, including amino acid substitute, insertions and deletions; while the horizontal coordinate represents for all positions (from 1 to 1273) in Spike protein. Squares are colored by mutational allele frequency ($\log_{10}$ transformed) according to the scale bar on the right. The model uses data from Britain from June to November. (**B**). The convergent evolutionary mutations predicted with the CoT2G-F. The heatmap illustrated the convergent mutations of the representative lineages of pre-Omicron and Omicron. The convergent mutation sites predicted with the CoT2G-F were highlighted in blue on the left side.


A
S1 (1–343)
S1 (344–685)
S2 (686–979)
S2 (980–1273)
Mutational allele frequency (log10 transformed)
−1 −2 −3 −4


B
Alpha Beta Delta
Omicron
P681 H R H H H H H H H R R
N501 Y Y Y Y Y Y Y Y Y Y Y
Q493 R R
F490 S S
F486 V V V V P P P P
E484 K A A A A A A A A K K
N460 K K K K K
F456 L
L455 S S
L452 R R R R R W W
N450 D D
G446 S S S S S S
V445 P P H H
K444 T T
K417 N N N N N N N N N N N
K356 T T
R346 T T T T
T95 I I
B.1.1.7 B.1.351 B.1.617.2 BA.1 BA.2 BA.4/5 BA.5 BF.7 BQ.1.1 XBB.1.5 EG.5.1 BA.2.86 JN.1
Mutational allele frequency (%)
100
50
0

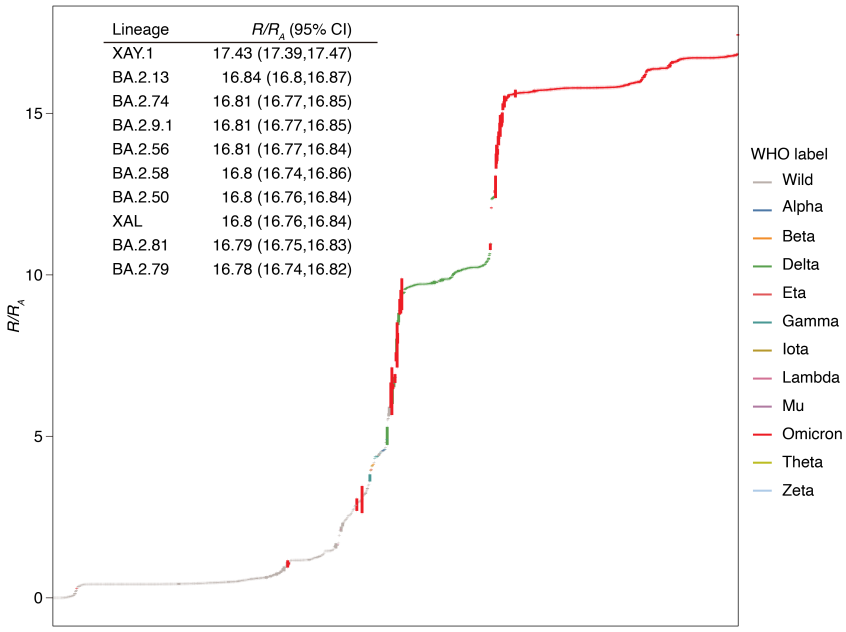
Supplementary Figure 7 | Relative fitness derived from the natural language model CoT2G-F versus date of lineage emergence

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.



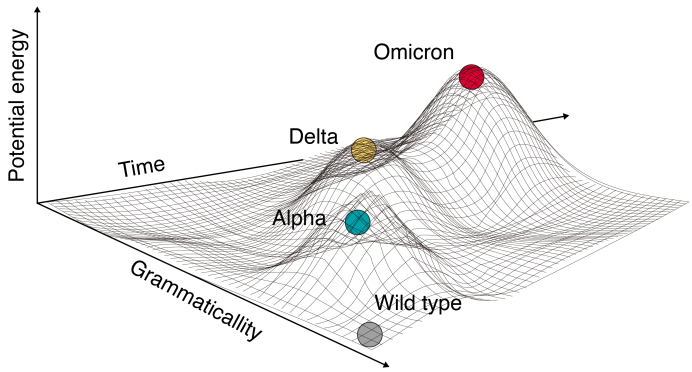
Supplementary Figure 8 | The confidence interval (CI) distribution for the R_0 calculation

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.



Supplementary Figure 9 | 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.



Extended Data Tables

Extended Data Table 1 | 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.

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		

Extended Data Table 2 | Indels predicted with the sequences from August 2021 and September 2021 (before the Omicron emerging)

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

Extended Data Table 3 | Key resources

REAGENT or RESOURCE	Source	IDENTIFIER
Software and algorithms		
Python (v3.8)	Rossum et al. [2]	https://www.python.org/
Biopython (v1.76)	Cock et al. [3]	https://biopython.org/
PyTorch (v1.12.1+cu102)	Paszke et al. [4]	https://pytorch.org/
Matplotlib (v3.4.1)	Hunter et. al. [5]	https://matplotlib.org
NLTK (v3.6.2)	Klein et al. [6]	https://www.nltk.org/
NumPy (v1.21.1)	Harris et al. [7]	https://numpy.org/
Pandas (v1.1.4)	McKinney et al. [8]	https://pandas.pydata.org/
Scipy (v1.3.1)	Virtanen et. al. [9]	https://scipy.org/
Scikit-learn (v0.21.3)	Pedregosa et al. [10]	https://scikit-learn.org
Fairseq (v0.12.2)	Ott et al. [11]	https://github.com/facebookresearch/fairseq
Flax (v0.5.2)	Heek et al. [12]	https://github.com/google/flax
JAX (v0.3.15)	Bradbury et al. [13]	https://github.com/google/jax
Optax (v0.1.3)	Babuschkin et al. [14]	https://github.com/deepmind/optax
Transformers (v4.20.1)	Wolf et al. [15]	https://github.com/huggingface/transformers
R (v4.2.2)	R Core Team [16]	https://cran.r-project.org
R package: tidyverse (v1.3.1)	Wickham et al. [17]	https://www.tidyverse.org/

Supplementary Information

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 [18], 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:

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

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 mean fitness of the k_{th} virus lineage. Apparently, the fitness of the virus population ω=$\omega(\boldsymbol{x})$ 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,..., x_i,..., x_L) represent L variation sites. Thus, each ω_k can involve the epistatic interaction of multiple mutation sites in the input sequence [19] 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 relative 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*).

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

π_k is the mixture coefficient, $N(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is the probability density function of a multivariate normal distribution with mean $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$. These coefficients π_k, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$ can obtain indirectly through the *EM* algorithm or deep neural network.

Now let's express the formula (1) in terms of the mathematical expectation of a probability density distribution function. According to formulas (3), the formal representation of the average fitness of the virus population can be obtained as follows:

$$\bar{\omega} = E(\omega) = \int \omega(\boldsymbol{x}) \, P(\boldsymbol{x}) d\boldsymbol{x}$$

$$= \int \sum_{k=1}^{K} \lambda_k \omega_k(\boldsymbol{x}) \, P(\boldsymbol{x}) d\boldsymbol{x} = \int \sum_{k=1}^{K} \lambda_k \omega_k(\boldsymbol{x}) \sum_{l=1}^{K} \pi_l \, N(\boldsymbol{x}|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l) d\boldsymbol{x}$$

Here, $\sum_{k=1}^{K} \lambda_k \omega_k(\boldsymbol{x})$ is the linear expansion result of the population fitness $\omega(\boldsymbol{x})$ according to virus lineage.

Given the independent and identically distributed (*IID*) traits of the virus lineage in the "ideal classification state", wherein the core of an "ideal classification state" is to attain

statistically minimal overlaps among each sub-distribution derived from the expansion of the standard normal distribution function, and the quasi-orthogonal attributes of the base distribution in *GMM* expansion during ideal classification, it is evident that:

$$\int \lambda_k \omega_k(\boldsymbol{x}) \pi_l N(\boldsymbol{x}|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l) d\boldsymbol{x} \cong$$

$$\cong \begin{cases} 0, & \text{when } k \neq l \\ \int \lambda_k \pi_k \omega_k(\boldsymbol{x}) N(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) d\boldsymbol{x} = \lambda_k \omega_k, & \text{when } k = l \end{cases}$$

$$\omega_k = \int \pi_k \omega_k(\boldsymbol{x}) N(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) d\boldsymbol{x}$$

In virus lineage level, ω_k is the fitness of the k_{th} virus lineage, and in virus population level, $\lambda_k = N_k/N$, N_k is the number of viruses in the k_{th} virus lineage, *N* is the total number of the virus population, so, $\sum_{k=1}^{K} \lambda_k = 1$, and under this condition conbining with the *EM* algorithm, we get formula (1) again.

$$\bar{\omega} = E(\omega) = \int \omega(\boldsymbol{x}) P(\boldsymbol{x}) d\boldsymbol{x}$$

$$= \sum_{k=1}^{K} \int \lambda_k \pi_k \omega_k(\boldsymbol{x}) N(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) d\boldsymbol{x} = \sum_{k=1}^{K} \lambda_k \omega_k \quad (4)$$

Formula (4) is a representation of the probability distribution of formula (1) in real space. The per-lineage fitness ω_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 and in the population leve,

$$\bar{\omega} = \sum_{k=1}^{K} \lambda_k \omega_k = \sum_{k=1}^{K} \lambda_k \left(\frac{1}{N_k} \sum_{i=1}^{N_k} \omega_i^k\right)$$

and per-lineage fitness:

$$\omega_k = \frac{1}{N_k} \sum_{i=1}^{N_k} \omega_i^k \quad (5)$$

Formula (5) is consistent with the definition by F. Obermeyer et al in their work [20]. According to the basic principles of population genetics [18] that the fitness of a virus lineage can be considered to be determined by the dominant virus cluster in the

lineage. Let ρ_k be the occurrence frequency of the dominant cluster in virus lineage when sampling virus strains, and $\rho_k = n_k/N_k$, n_k is the total number of samples taken from the dominant cluster in the k_{th} virus lineage. For the convenience of practical calculation, we can redefine the per-lineage fitness through the contribution of the dominant virus cluster of a lineage as follows:

$$\widehat{\omega}_k \overset{\text{def}}{=} \frac{\rho_k}{n_k}\sum_{i=1}^{n_k}\omega_i^k = \frac{1}{N_k}\sum_{i=1}^{n_k}\omega_i^k \tag{6}$$

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

$$\widehat{\omega}_k = \frac{1}{N_k}\sum_{i=1}^{n_k}\omega_i^k = \frac{1}{N_k}\sum_{i=1}^{n_k}|\omega_i^k| \tag{7}$$

$|\omega_i^k|$ 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 finding the dominant cluster in virus lineage and 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) and formula (5),

$$\omega_k = \frac{1}{N_k}\sum_{i=1}^{N_k}\omega_i^k \tag{8}$$

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), (4) and (5), and let $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots\dots, \lambda_K)$ and $\boldsymbol{\omega}(\boldsymbol{x}) = (\omega_1, \omega_2, \dots\dots, \omega_K)$ are vectors, then

$$\bar{\omega} = E(\omega) = \sum_{k=1}^{K}\lambda_k\omega_k = \int \omega(\boldsymbol{x})\,P(\boldsymbol{x})\,d\boldsymbol{x} = \bar{\omega}_{|P(\boldsymbol{x})} \tag{8'}$$

The subscript $P(\cdot)$ represents the state probability distribution corresponding to virus population. As long as we find a suitable functional form of $\omega(\boldsymbol{x})$ and a corresponding

probability distribution $P(\boldsymbol{x})$, we can obtain the fitness of viruses by calculating their mathematical expectations by the formula (8), then get the average fitness of the virus population and per-lineage fitness of virus lineages by formula (4) and (6). 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 $\omega(\boldsymbol{x})$ and $P(\boldsymbol{x})$ directly from real-world data with noise.

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

$$P(\boldsymbol{x}) = P(\boldsymbol{z})P(\boldsymbol{x}|\boldsymbol{z})/P(\boldsymbol{z}|\boldsymbol{x}) \tag{9}$$

The function relationship defined by formula (1) still holds when $P(\boldsymbol{x})$ maps to $P(\boldsymbol{z})P(\boldsymbol{x}|\boldsymbol{z})/P(\boldsymbol{z}|\boldsymbol{x})$. 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 $P(\boldsymbol{z}|\boldsymbol{x})$, $P(\boldsymbol{z})$, $P(\boldsymbol{x}|\boldsymbol{z})$ and the counterpart $\widetilde{\omega}(\boldsymbol{z})$ of $\omega(\boldsymbol{x})$.

Under the Encoder-Decoder Seq2Seq framework together with Bayes' theorem and Universality Theorem [21], because

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

we have

$$\int \omega(\boldsymbol{x})\, P(\boldsymbol{x})d\boldsymbol{x} = \int \omega(\boldsymbol{x})\, P(\boldsymbol{z})(P(\boldsymbol{x}|\boldsymbol{z})/P(\boldsymbol{z}|\boldsymbol{x}))d\boldsymbol{x}$$

$$= \int \omega(W\boldsymbol{z} + \boldsymbol{B})\, P(W\boldsymbol{z} + \boldsymbol{B})|W|d\boldsymbol{z}$$

Here,

$$W = \frac{d\boldsymbol{x}}{d\boldsymbol{z}}$$

is the Jacobian matrix, and $|W|=\sqrt{\det(W^T \cdot W)}$ is a determinant of Jacobian matrix that is the sum of combinations of products of elements in different positions of all the columns of matrix *W*. and through the encode-decode architecture's deep learning process, coupled with the Seq2Seq property, we can derive a parameterized representation of Bayes' theorem:

$$P(\boldsymbol{x}) = P(\boldsymbol{z})(P(\boldsymbol{x}|\boldsymbol{z})/P(\boldsymbol{z}|\boldsymbol{x})),$$

At this juncture, considering the Universality Theorem, convergence prerequisites of deep neural networks and the linear transformation characteristics of $\boldsymbol{x}=Wz+B$ in the output layer and the neighboring hidden layer, the equation

$$\int \omega(\boldsymbol{x})\, P(\boldsymbol{x})d\boldsymbol{x} = \int \omega(\boldsymbol{x})\, P(\boldsymbol{z})(P(\boldsymbol{x}|\boldsymbol{z})/P(\boldsymbol{z}|\boldsymbol{x}))d\boldsymbol{x}$$

holds. At this juncture, employing the linear transformation $\boldsymbol{x}=W\boldsymbol{z}+B$, the integral of x is translated into the integral of $\boldsymbol{z}$, yielding the equation

$$\int \omega(\boldsymbol{x})\, P(\boldsymbol{x})d\boldsymbol{x} = \int \omega\big(\boldsymbol{x}(\boldsymbol{z})\big)\, P(\boldsymbol{z})(P(\boldsymbol{x}(\boldsymbol{z})|\boldsymbol{z})/P(\boldsymbol{z}|\boldsymbol{x}(\boldsymbol{z})))|W|d\boldsymbol{z}$$

$$= \int \omega(W\boldsymbol{z}+B)\, P(W\boldsymbol{z}+B)|W|d\boldsymbol{z}.$$

if we set:

$$\omega(W\boldsymbol{z}+B) = \tilde{\omega}(\boldsymbol{z}) = \omega\big(\boldsymbol{x}(\boldsymbol{z})\big)(P(\boldsymbol{x}(\boldsymbol{z})|\boldsymbol{z})/P(\boldsymbol{z}|\boldsymbol{x}(\boldsymbol{z})) \qquad (9)'\text{-}1$$

This equation can be rewritten as:

$$\tilde{\omega}(\boldsymbol{z})P(\boldsymbol{z}|\boldsymbol{x}) = \omega(\boldsymbol{x})P(\boldsymbol{x}|\boldsymbol{z}) \qquad (9)'\text{-}2$$

Equation (9)'-2 implies that if $\tilde{\omega}(\boldsymbol{z})$ is a topological mapping of $\omega(\boldsymbol{x})$ in the hidden space within deep neural networks, then the product of the function $\omega(\boldsymbol{x})$ of the

microstate variable $\boldsymbol{x}$ and the conditional probability distribution of $\boldsymbol{x}$ in the real space equals the product of the function $\widetilde{\omega}(\boldsymbol{z})$ of the latent variable $\boldsymbol{z}$ and the conditional probability distribution of $\boldsymbol{z}$ in the hidden space. If we take into account the topological properties of encode-decode Seq-Seq deep neural networks, it is evident that this equation holds true. It follows naturally.

$$P(W\boldsymbol{z} + \boldsymbol{B}) = P(\boldsymbol{z}). \qquad (9)'\text{-}3$$

Consequently, we get finally

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

Given the quasi-orthogonal property inherent in the base distribution of Gaussian Mixture Model (*GMM*) expansion for optimal classification, let β_k represent the mapping of $|W|_k$ for the k_{th} virus lineage, and $\widetilde{\omega}_k$ denote a hidden space state function associated with the fitness of the k_{th} virus lineage. Forming vectors $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots\dots, \beta_K)$ and $\widetilde{\boldsymbol{\omega}}(\boldsymbol{z}) = (\widetilde{\omega}_1, \widetilde{\omega}_2, \dots, \widetilde{\omega}_K)$, the relationship

$$|W|\widetilde{\omega}(\boldsymbol{z}) \rightarrow \boldsymbol{\beta} * \widetilde{\boldsymbol{\omega}}(\boldsymbol{z}),$$

holds true. Then we get

$$\overline{\omega}_{|P(\boldsymbol{x})} = \int \omega(\boldsymbol{x})\, P(\boldsymbol{x})\, d\boldsymbol{x} = \int |W|\widetilde{\omega}(\boldsymbol{z})\, P(\boldsymbol{z}) d\boldsymbol{z} = \int (\boldsymbol{\beta} * \widetilde{\boldsymbol{\omega}}(\boldsymbol{z}))\, P(\boldsymbol{z}) d\boldsymbol{z} \qquad (10)$$

When a virus lineage contains enough virus strains, let $\omega(\boldsymbol{x}) = \omega_k(\boldsymbol{x})$, $P(\boldsymbol{x}) = P_k(\boldsymbol{x})$, and combining the formula (4), the following formula still holds,

$$\omega_{k_{|P_k(\boldsymbol{x})}} = \int \omega_k(\boldsymbol{x})\, P_k(\boldsymbol{x})\, d\boldsymbol{x}$$

$$= \int |W|_k \widetilde{\omega}_k(\boldsymbol{z})\, P_k(\boldsymbol{z}) d\boldsymbol{z} = \int \beta_k \widetilde{\omega}_k(\boldsymbol{z})\, P_k(\boldsymbol{z}) d\boldsymbol{z} = \beta_k \widetilde{\omega}_k$$

From formula (10) and see the logic for deriving formula (4), and combining formula (5), we have

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

$$= \int (\boldsymbol{\beta} * \tilde{\boldsymbol{\omega}}(\boldsymbol{z}))P(\boldsymbol{z})d\boldsymbol{z} = \sum_{k=1}^{K} \frac{\beta_k}{N_k} \sum_{i=1}^{N_k} \tilde{\omega}_i^k = \sum_{k=1}^{K} \beta_k \tilde{\omega}_k \qquad (10)'$$

Formula (10)' shares an identical structure with formulas (8) and (8)'. Hence, it is reasonable to equate $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots\dots, \beta_K) = \boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots\dots, \lambda_K)$, . Consequently,

$$\bar{\omega}_{|P(\boldsymbol{x})} = \int (\boldsymbol{\beta} * \tilde{\boldsymbol{\omega}}(\boldsymbol{z}))P(\boldsymbol{z})d\boldsymbol{z} = \int (\boldsymbol{\lambda} * \tilde{\boldsymbol{\omega}}(\boldsymbol{z}))P(\boldsymbol{z})d\boldsymbol{z} = \boldsymbol{\lambda} * \bar{\tilde{\boldsymbol{\omega}}}.$$

This leads to the final expression:

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

Formula (11) is a representation of the probability distribution of formula (1) in an embedded space. Because of the hierarchical structure of the population, lineage and cluster of the virus, referring to the formula(1), (3), (4), (6) and (10) and based on the basic principles of population genetics [18], the elements of the vector $\boldsymbol{\lambda}$ are the corresponding elements of mixture coefficient in formula (3). According to the formula (6) the element λ also associated with the occurrence frequency of a dominant cluster in virus lineage when sampling virus strains in a virus lineage and is a macroscopic parameter that links the hidden space with the actual space.

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

Based on formula (10), we know that $\omega(\boldsymbol{x})$ and $\tilde{\boldsymbol{\omega}}(\boldsymbol{z})$ 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 $\omega(\boldsymbol{x})$ 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 $\widetilde{\boldsymbol{\omega}}(\boldsymbol{z})$ 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 macroscopic phenotype, and genome sequence variation is related to microscopic genotype change, the meaning of equation (11) can be generalized, and the following phenotypic embedding (P-E) theorem in the biological field can be proposed:

Inference: 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."

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

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 [22], and it is paraphrased as follows:

For a virus sequence defined as $\boldsymbol{x} \stackrel{def}{=} (x_1, \dots, x_N)$ such that $x_i \in \mathcal{X}$, $i \in [N]$, where $\mathcal{X}$ is a finite set of characters, $\boldsymbol{x} \in V$, V is virus population set and $\boldsymbol{x}_k \in V_k$, 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 $\boldsymbol{x}[\tilde{x}_{\{i\}}] = \boldsymbol{x}[\tilde{x}_{i-\lambda+1}, \dots, \tilde{x}_{i+\lambda-1}] \stackrel{def}{=} (x_1, \dots, x_{i-\lambda}, \tilde{x}_{i-\lambda+1}, \dots, \tilde{x}_i, \dots, \tilde{x}_{i+\lambda-1}, x_{i+\lambda}, \dots, x_N)$, here, $1<=\lambda<=$Length(protein sequence) $*\,15\%$ and for SARS-CoV-2 spike protein $\bar{\lambda}$=3. For brevity, we will use $\boldsymbol{x}[\tilde{x}_{\{i\}}]$ instead of $\boldsymbol{x}[x_1, \dots, x_{i-\lambda}, \tilde{x}_{i-\lambda+1}, \dots, \tilde{x}_i, \dots, \tilde{x}_{i+\lambda-1}, x_{i+\lambda}, \dots, x_N]$ in the following. The syntax for a mutated sequence $\boldsymbol{x}$ is described as a probability distribution function:

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

Then, $p(\tilde{x}_{\{i\}}|\boldsymbol{x})$ is close to zero if $\boldsymbol{x}[\tilde{x}_{\{i\}}]$ does not conform to the grammar rules, and $p(\tilde{x}_{\{i\}}|\boldsymbol{x})$ is close to one if $\boldsymbol{x}[\tilde{x}_{\{i\}}]$ 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 $\hat{\boldsymbol{z}}_{\{i\}}$ is encoded from the context $\boldsymbol{x}_{[N]\setminus\{i\}} \stackrel{def}{=} (\dots, x_{i-\lambda}, x_{i+\lambda}, \dots)$, in which the viral genome

mutates at position $(x_{i-\lambda}, \ldots, x_{i+\lambda})$. In our model CoT2G-F, $\tilde{x}_{\{i\}}$ in $\boldsymbol{x}_{[N]\setminus\{i\}}$ 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

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

We define:

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

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

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

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

Here [22] defines

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

which is the average of semantic embeddings at all positions. Formula (2) shows that $\hat{\boldsymbol{z}}$ is a reference value for embedded "semantic" changes, and what virus lineage is chosen as a reference has biological significance. In our work, the Wuhan lineage was selected as a reference. $\hat{\boldsymbol{z}}[\tilde{x}_{\{i\}}]$ 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 $\hat{p}(\tilde{x}_{\{i\}}|\hat{\boldsymbol{z}}_{\{i\}})$ can be obtained through model training.

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

$$\hat{\boldsymbol{z}}_{\{i\}} \stackrel{def}{=} CoT2G - F_{\hat{f}_s}([\tilde{x}_{\{i\}}]|\boldsymbol{x})$$

When there are multiple mutations in the sequence, as defined

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

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

$$P(\boldsymbol{x}|\hat{\boldsymbol{z}}) = P(\boldsymbol{x}[\tilde{x}_{\{i\}}]|\hat{\boldsymbol{z}}) = \prod_{\tilde{x}_{\{i\}} \in \mathcal{M}(\boldsymbol{x},\tilde{\boldsymbol{x}})} \hat{p}(\tilde{x}_{\{i\}}|\hat{\boldsymbol{z}}) = \prod_{\tilde{x}_{\{i\}} \in \mathcal{M}(\boldsymbol{x},\tilde{\boldsymbol{x}})} p(\tilde{x}_{\{i\}}|\hat{\boldsymbol{z}})$$

In our CoT2G-F model, we can further get

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

Where $\boldsymbol{W}$ and $\boldsymbol{B}$ are both the parameters the model needs to learn and $\hat{\boldsymbol{z}}[\tilde{x}_{\{i\}}]$ 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 $\Delta\boldsymbol{z}[\tilde{x}_{\{i\}}]$ remains unchanged, and the rules from [22] still hold:

$$\alpha'(\tilde{x}_{\{i\}}; \boldsymbol{x}) \stackrel{def}{=} rank(\Delta\boldsymbol{z}[\tilde{x}_{\{i\}}]) + \beta rank(P(\boldsymbol{x}|\hat{\boldsymbol{z}})) \quad (15)$$

The distance $\Delta[\tilde{x}_{\{i\}}]$ in the embedding space describes the semantic change, and the hidden state probability distribution $P(\hat{\boldsymbol{z}})$ and the related posterior probability $P(\boldsymbol{x}|\hat{\boldsymbol{z}})$ 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 [22], we think that the first term $\Delta \boldsymbol{z}[\tilde{x}_{\{i\}}]$ 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. $\Delta[\tilde{x}_{\{i\}}]$ can also regard as a conditional "semantic" change（$CSC = \Delta \boldsymbol{z}[\tilde{x}_{\{i\}}]$）score. Late we will show that the per-lineage fitness can derive from the measure CSC with a hidden state distribution $P(\hat{\boldsymbol{z}})$ and $P(\boldsymbol{x}|\hat{\boldsymbol{z}})$ 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 [23], 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 [22,24,25] 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 20 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 [26] "masked language modelling"

objective and the "word dropout" regularization technique [27], 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:

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

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 $\mathrm{x}_{\mathrm{M/N}}$ given the masked sequence $\mathrm{x}_{\mathrm{N/M}}$ 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 $\boldsymbol{x}_{(t,p)}$ at t month slice，$\boldsymbol{x}_{(t,p-1)}$ and $\boldsymbol{x}_{(t,p+1)}$ is the spatial pre-order and follow-up contiguous sequences of $\boldsymbol{x}_{(\boldsymbol{t},\boldsymbol{p})}$, [:] means concatenate two sequences, so the

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

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

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.

$$\mathcal{L}_{\mathcal{P}re-train} = \alpha_{\mathcal{P}re-\mathcal{M}ask} * \mathcal{L}_{\mathcal{P}re-\mathcal{M}ask} + \alpha_{\mathcal{F}ollow-\mathcal{M}ask} * \mathcal{L}_{\mathcal{F}ollow-\mathcal{M}ask} \quad (18)$$

In the fine-tuning step, we follow the Constrained Semantic Change Search (CSCS) criterion proposed by B. Hi [22], 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 $\boldsymbol{x}$ and the mutated sequences as $\widetilde{\boldsymbol{x}} = (\tilde{x}_1, \ldots, \tilde{x}_N)$, we assume there are N mutant sequences, for one mutant sequence $\widetilde{\boldsymbol{x}}_j$, which consists of amino acid characters in $\widetilde{\boldsymbol{x}}_{ji}$ that disagree with those at the same position in the base sequence $\boldsymbol{x}$, which we denote as

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

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

$$\begin{aligned} \mathcal{CSCS} &= \mathcal{S}_{rank}(\widetilde{\boldsymbol{x}}_j; \boldsymbol{x}) + \beta * \mathcal{G}_{rank}(\widetilde{\boldsymbol{x}}_j; \boldsymbol{x}) \quad (19) \\ &= \mathcal{CSC} + \beta * \mathcal{G}_{rank}(\widetilde{\boldsymbol{x}}_j; \boldsymbol{x}) \end{aligned}$$

where $\mathcal{S}_{rank}(\widetilde{\boldsymbol{x}}_j; \boldsymbol{x})$ and $\mathcal{G}_{rank}(\widetilde{\boldsymbol{x}}_j; \boldsymbol{x})$ 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:

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

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:

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

where p is the mutation probability that the i site corresponds to the $\widetilde{\boldsymbol{x}}_{ji}$ is mutated to the amino acid corresponding to the i site in $\boldsymbol{x}$, 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:

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

For each sequence $\boldsymbol{x}$ at one month slice, the model needs to generate a list of target sequences, donated as $\widetilde{\boldsymbol{x}}$, which is the virus strain that may emerge "immune escape" mutations in the next month, the $|\widetilde{\boldsymbol{x}}|$ means all sites of $\widetilde{\boldsymbol{x}}$, and due to the autoregressive property of the decoder, when predicting $\widetilde{\boldsymbol{x}}_i$, the sites before the i site can also be introduced, i.e. $\widetilde{\boldsymbol{x}}_{<i}$. We need to independently minimize the negative log-likelihood of the amino acid $\widetilde{\boldsymbol{x}}_i$ given the $\boldsymbol{x}$ and $\widetilde{\boldsymbol{x}}_{<i}$ 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.

$$\mathcal{L}_{\mathcal{SP}} = \sum_{\mathrm{i}=1}^{|\tilde{\boldsymbol{x}}|} -\mathrm{logq}\left(\tilde{\boldsymbol{x}}_i \mid \left[\boldsymbol{x}_{(t,p)}: \boldsymbol{x}_{(t,p-1)}\right], \tilde{\boldsymbol{x}}_{<i}\right)$$

$$\mathcal{L}_{\mathcal{SF}} = \sum_{\mathrm{i}=1}^{|\tilde{\boldsymbol{x}}|} -\mathrm{logq}\left(\tilde{\boldsymbol{x}}_i \mid \left[\boldsymbol{x}_{(t,p)}: \boldsymbol{x}_{(t,p+1)}\right], \tilde{\boldsymbol{x}}_{<i}\right)$$

$$\mathcal{L}_{\mathcal{T}} = \sum_{\mathrm{i}=1}^{|\tilde{\boldsymbol{x}}|} -\mathrm{logq}\left(\tilde{\boldsymbol{x}}_i \mid \left[\boldsymbol{x}_{(t,p)}: \boldsymbol{x}_{(t-1,p)}\right], \tilde{\boldsymbol{x}}_{<i}\right)$$

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.

$$\mathcal{L}_{\mathcal{F}ine-tune} = \alpha_{\mathcal{SP}} * \mathcal{L}_{\mathcal{SP}} + \alpha_{\mathcal{SF}} * \mathcal{L}_{\mathcal{SF}} + \alpha_{\mathcal{T}} * \mathcal{L}_{\mathcal{T}} \quad (20)$$

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

<u>1. The "immune escape" ability score *CSC*</u>

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 [22]:

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

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

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

is the mean of all possible "semantic" changes in a sequence of N bases. If we select the Wuhan lineage as the reference [Wuhan-Hu-1, GenBank: MN908947.3], $\hat{\boldsymbol{z}}_{\boldsymbol{N}\backslash\{\mathcal{M}(\vec{\boldsymbol{X}},\tilde{\boldsymbol{X}})\}}$ is an absolute semantic change of a virus strain, and $\|\hat{\boldsymbol{z}} - \hat{\boldsymbol{z}}_{\boldsymbol{N}\backslash\{\mathcal{M}(\vec{\boldsymbol{X}},\tilde{\boldsymbol{X}})\}}\|_{\ell_x}$ is a relative semantic change of a virus strain according to the Wuhan virus lineage. In this way, a conditional semantic change score (*CSC*) under the condition that the sequence composition conforms to the grammar can be defined as:

$$CSC = \Delta\hat{\boldsymbol{z}}_{\boldsymbol{N}\backslash\{\mathcal{M}(\vec{\boldsymbol{X}},\tilde{\boldsymbol{X}})\}} = \|\hat{\boldsymbol{z}} - \hat{\boldsymbol{z}}_{\boldsymbol{N}\backslash\{\mathcal{M}(\vec{\boldsymbol{X}},\tilde{\boldsymbol{X}})\}}\|_{\ell_x} \tag{22}$$

and we can rewrite (21) as follows

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

CSC score is the measurement of the "immune escape" ability of the virus strain referring to the Wuhan lineage when the virus mutates. Formula (1) gives the fundamental definition of the mean fitness of the virus population and the fitness of the virus lineage. 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 ℓ_1 norm to define the *CSC* score. Because our model has better avoiding overfitting characteristics (Figure 3).

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

Based onur proposed Encoder-Decoder Seq2Seq architecture model, the encoder module created an embedded state distribution $P(\boldsymbol{z}|\boldsymbol{x})$ related to the virus fitness from

the input genome sequence $\boldsymbol{x}$. The decoder module can characterize the "immune escape" of the virus and decode a hidden state probability distribution $P(\boldsymbol{z})$ of a latent variable $\boldsymbol{z}$. $P(\boldsymbol{z})$ can also be formally expanded by the Gaussian Mixture Model (*GMM*) kernel function:

$$P(\boldsymbol{z}) = \sum_{k=1}^{K} \pi_k N(\boldsymbol{z} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{24}$$

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

Definition

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

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

$$\sum_{i}^{N} \Delta_i = \sum_{k}^{K} \Delta_k$$

N and K are the number of virus strains and the number of virus lineages in a virus population respectively. $\bar{\Delta}$ and $\bar{\Delta}_k$ are the expected values of Δ and Δ_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), and the quasi-orthogonal characteristics of the base distribution of *GMM* expansion in the ideal classification, the average "immune escape" ability of the virus population can express as:

$$\overline{CSC} = \bar{\Delta} = \sum_{k=1}^{K} \pi_k \int \Delta_k P^k(\boldsymbol{z}) d = \sum_{k=1}^{K} \pi_k \int \Delta_k N(\boldsymbol{z} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\, d\boldsymbol{z}$$

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

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:

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

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

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

$\widetilde{\omega}_i^k$ 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

$$\overline{CSC}^k = \frac{1}{N_k}\sum_{i=1}^{N_k} |\widetilde{\omega}_i^k| \tag{29}$$

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 k_{th} virus lineage. We define $\overline{CSC}^k$ as the average immune escape ability of the k_{th} virus lineage.

Let us look at the distribution of the absolute mean "semantic" changes computed by sampling different virus lineages (Figure 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. [20]. Let us now consider the following four facts: (i) The model determined that latent variable $\boldsymbol{z}$ reflected the semantic and grammatical changes of virus sequence site variation and the correlation between variation sites given by formula (2) and [22]; (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 [22]) (Figure 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 (Figure 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 [20] and general epidemiological observation.

Based on the above four facts, we believe that *CSC* = $\widetilde{\omega}(\boldsymbol{z})$ is the corresponding function of $\omega(\boldsymbol{x})$ 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), (6) , (22) and (29), and set $\beta_k = \hat{\lambda}_k$, we have

$$\widehat{\omega}_k = \hat{\lambda}_k \widetilde{\omega}_k \stackrel{\text{def}}{=} \frac{\hat{\lambda}_k}{n_k} \sum_{i=1}^{n_k} |\widetilde{\omega}_i^k| = \frac{\hat{\lambda}_k}{n_k} \sum_{i=1}^{n_k} CSC_i^k \tag{30}$$

Here, $\widehat{\omega}_k$ is the per-lineage fitness of the virus and it can be calculated directly under the framework of deep learning, $\hat{\lambda}_k$ is the occurrence frequency of the dominant cluster in the k_{th} virus lineage. Thus, formula (27) is the fitness of the virus population. In this way, we prove that the item of the degree of "semantic" change in the *CSCS* criterion proposed by B. Hie, et. al., [22] is related to the per-cluster fitness of the virus given by F. Obermeyer, et al. [20]. Although all the parameters involved in the calculation of the per-lineage fitness $\hat{\lambda}_k \widetilde{\omega}_k$ can be computed directly by the model, it is clear that the parameter $\hat{\lambda}_k$ and latent variable $\boldsymbol{z}$ 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 $P(\boldsymbol{z})$. 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 $\mathrm{k_{th}}$ virus lineage is deduced.

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

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

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

Here,

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

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

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:

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

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 dominant cluster in 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 dominant 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, Extended Data Supplementary Figure 3A and 3B). At this time, the problem of finding R_{0k} 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:

$$R_{0k} = \left(\frac{\hat{\lambda}_k}{n_k}\sum_{i=1}^{n_k} CSC_i^k\right)/(\hat{\lambda}_A\widetilde{\omega}_A) = \frac{\hat{\lambda}_k}{\hat{\lambda}_A}(\widetilde{\omega}_k/\widetilde{\omega}_A)$$

$$= \widehat{\omega}_k/\widehat{\omega}_A = \frac{\hat{\lambda}_k}{\hat{\lambda}_A}\int\left(\frac{CSC^k}{\widetilde{\omega}_A}\right)P^k(\boldsymbol{z})d\boldsymbol{z} = \int \hat{f}(\boldsymbol{z})P^k(\boldsymbol{z})d\boldsymbol{z} \quad (33)$$

Here, $\widehat{\omega}_A = \hat{\lambda}_A\widetilde{\omega}_A$ is the fitness of the Wuhan lineage, $\widehat{\omega}_k = \hat{\lambda}_k\widetilde{\omega}_k$ is the fitness of the k_{th} virus lineage, and the relative basic reproduction number R_{0k} is the fold increase of the per-lineage fitness $\widehat{\omega}_k$ over $\widehat{\omega}_A$. 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 relative basic reproduction number (R_0) of the virus lineage are precisely computable, which is proportional to the occurrence frequency of the dominant 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.

***R₀* 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 relative R_0.

In our study, we propose a simple calibration process to obtain the absolute basic reproduction number ($\hat{R}_o$) by using a selected observable actual number of the absolute basic reproduction number ($\hat{R}_o$) for any lineage of SARS-CoV-2. From formula (33), we can obtain the absolute basic reproduction number ($\hat{R}_o$) as below

$$\hat{R}_o \cong (\hat{\omega}_k/\hat{\omega}_C)\hat{R}_oc \quad (34)$$

Here, $\hat{\omega}_k$ and $\hat{\omega}_C$ are the relative per-lineage fitness of lineage k and C according to Wuhan strain, and $\hat{R}_oc$ is the actually observed absolute basic reproduction number of the reference lineage C. When we select Delta lineage as an observable reference [28,29], $\hat{R}_oc \sim 7$. (Figure 4B, Supplemental Note 4, Formula (30)). Our calculated absolute basic reproduction numbers ($\hat{R}_o$) align with the results obtained by F. Obermeyer *et al.* [20] and are also in agreement with the actual and official monitoring results. This result validates our proposed P-E theorem and G-P principle, allowing for the precise computation of the relative fitness and $\hat{R}_o$ of the virus lineages.

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

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

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 [22], 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 Supplementary Figure 9) . 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 (Figure 4A).

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

In equation (33), the R_{0k} on the left side is a macroscopically observable biological variable, and

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

on the right side is a mathematical expectation of a function $\hat{\lambda}_k CSC^k$ of the latent variable $\boldsymbol{z}$, which is related to genome sequence variation, according to the hidden state probability distribution $P(\boldsymbol{z})$. From equation (8) and (33), we can see that deep learning can help us build a bridge between Genotype-Phenotype, and we can draw a natural inference:

"A macroscopically observable biology variable $\bar{F}$ can express as a mathematical expectation of a microscopic state function $f(\boldsymbol{x})$ according to a state probability distribution $P(\boldsymbol{x})$".

This statement we call the G-P (Genotype-Phenotype) principle, it can represent by the following equation:

$$\bar{F} = \int f(\boldsymbol{x})\, P(\boldsymbol{x}) d\boldsymbol{x} \tag{35}$$

and according to the P-E theorem (11), the formula (35) is equivalent to

$$\bar{F} = \int f(\boldsymbol{x})\, P(\boldsymbol{x}) d\boldsymbol{x} = \lambda \int \tilde{f}(\boldsymbol{z})\, P(\boldsymbol{z}) d\boldsymbol{z} \tag{35'}$$

Variable $\boldsymbol{z}$ is a latent variable, λ is a macroscopic parameter that links the hidden space with the actual space. $\tilde{f}(\boldsymbol{z})$ is a function corresponding to $f(\boldsymbol{x})$ in hidden space and $P(\boldsymbol{z})$ is a prior probability distribution in hidden space, both of $\tilde{f}(\boldsymbol{z})$ and $P(\boldsymbol{z})$ can be calculated by deep learning.

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 [24,25]. 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:

$$\bar{f} = \int f(\Gamma) \rho(\Gamma) d\Gamma \tag{36}$$

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 (35) and (36) 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.

References

1. Tsoularis A, Wallace J. Analysis of logistic growth models. Mathematical Biosciences 2002; 179:21–55
2. Van Rossum G, Drake FL. Python 3 Reference Manual. 2009;
3. Cock PJA, Antao T, Chang JT, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 2009; 25:1422–1423
4. Paszke A, Gross S, Massa F, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. Advances in Neural Information Processing Systems 2019; 32:
5. Hunter JD. Matplotlib: A 2D graphics environment. Computing in Science & Engineering 2007; 9:90–95
6. Bird S, Klein E, Loper E. Natural language processing with Python: analyzing text with the natural language toolkit. 2009;
7. Harris CR, Millman KJ, van der Walt SJ, et al. Array programming with NumPy. Nature 2020; 585:357–362
8. McKinney W, others. Data structures for statistical computing in python. Proceedings of the 9th Python in Science Conference 2010; 445:51–56
9. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods 2020; 17:261–272
10. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 2011; 12:2825–2830
11. Ott M, Edunov S, Baevski A, et al. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations) 2019; 48–53
12. Heek J, Levskaya A, Oliver A, et al. Flax: A neural network library and ecosystem for JAX. 2023;
13. Bradbury J, Frostig R, Hawkins P, et al. JAX: composable transformations of Python+NumPy programs. 2018;

14. Babuschkin I, Baumli K, Bell A, et al. The DeepMind JAX Ecosystem. 2020;
15. Wolf T, Debut L, Sanh V, et al. Transformers: State-of-the-Art Natural Language Processing. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations 2020; 38–45
16. R Core Team. R: A Language and Environment for Statistical Computing. 2022;
17. Wickham H, Averick M, Bryan J, et al. Welcome to the Tidyverse. JOSS 2019; 4:1686
18. Futuyma DJ, Kirkpatrick M. The Evolution of a Population's Mean Fitness. Evolution 2017; 103–134
19. de Visser JAGM, Krug J. Empirical fitness landscapes and the predictability of evolution. Nature Reviews Genetics 2014; 15:480–490
20. Obermeyer F, Jankowiak M, Barkas N, et al. Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness. Science 2022; 376:1327–1332
21. Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. Neural Networks 1989; 2:359–366
22. Hie B, Zhong ED, Berger B, et al. Learning the language of viral evolution and escape. Science 2021; 371:284–288
23. Vaswani A, Shazeer N, Parmar N, et al. Attention is All you Need. Advances in Neural Information Processing Systems 2017; 30:
24. Rives A, Meier J, Sercu T, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences 2021; 118:e2016239118
25. Meier J, Rao R, Verkuil R, et al. Language models enable zero-shot prediction of the effects of mutations on protein function. Advances in Neural Information Processing Systems 2021; 34:29287–29303
26. Devlin J, Chang M-W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018;
27. Bowman SR, Vilnis L. Generating Sentences from a Continuous Space. the 20th SIGNLL Conference 2016; 10–21

774 28. Li Q, Guan X, Wu P, et al. Early Transmission Dynamics in Wuhan, China, of Novel
775 Coronavirus–Infected Pneumonia. New England Journal of Medicine 2020;
776 382:1199–1207
777 29. Majumder MS, Mandl KD. Early in the epidemic: impact of preprints on global
778 discourse about COVID-19 transmissibility. The Lancet Global Health 2020;
779 8:e627–e630
780
781