Title: Learning virus genotype-fitness landscape in embedding space

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

Affiliations:

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

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

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

[4]Beijing Volcano Engine Technology Co., Ltd., Beijing, 100098, China

[5]BYHEALTH Institute of Nutrition & Health, Guangzhou, 510663, 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, Shanghai, 200030, China

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

[9]Shanghai Institute for Biomedical and Pharmaceutical Technologies, Shanghai, 200032, China

†These authors contributed equally to this work.

*Corresponding author. zhangxin.zhangx@bytedance.com (X.Z.); zhangxg2@by-health.com (X.G.Z.); li_yixue@gzlab.ac.cn (Y.X.L.)

Abstract: Predicting the SARS-CoV-2 epidemic and "immune escape" mutations remain crucial problems. We present a theoretical framework called Phenotype-Embedding (P-E) theorem and prove that the virus fitness can calculate by selecting appropriate sequence embedding under the VAE framework. Starting from the P-E theorem and based on a modified Transformer model, we obtain a calculable quantitative relationship between "immune escape" mutations and the fitness of the virus lineage and plot a genotype-fitness landscape in the embedded space. We accurately calculated the viral fitness and the basic reproduction number (R_0) using only the sequence data of SARS-CoV-2 spike protein. In addition, our model can simulate viral neutral evolution and spatio-temporal selection, decipher the effects of epistasis and recombination, and more accurately predict viral mutations associated with immune escape. Our work provides a theoretical framework for constructing genotype-phenotype landscapes and a paradigm for the interpretability of deep learning in virus evolution research.

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

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

Introduction

Evolutionary mutations in viruses often lead to vaccine failure and escape of neutralizing antibody recognition and induce a weakened or ineffective drug effect, resulting in a substantial increase in the infection rate. Virus mutations cause phenotypic changes that enhance fitness and become fixed in the population. Therefore, people hope to detect the evolutionary mutations that drive virus "immune escape" through various high-throughput experimental techniques (1, 2), given an empirical description of the genotype-fitness landscape whose topological properties have decisive effects on the evolutionary process and the predictability of evolution.

However, experimental techniques have always faced challenges when considering the possibility of epistasis and recombination across virus strains and directly giving the association between "immune escape" mutations and virus fitness.

Since the global COVID-19 outbreak, the GISAID database [https://gisaid.org] has collected more than 20 million SARS-CoV-2 genome sequences with the Spatio-temporal information. Such massive data gives us an advantage in discovering the biological mechanisms behind it. We can use virus genome sequence information to decipher the dynamic mechanism driving virus evolution test the role of epistatic and recombination constraints (*3*), identify and predict virus "immune escape" mutations, obtain virus epidemic-related fitness information, and provide evidence for further experimental verification and verification.

In this regard, B. Hie et al. have done pioneering work (*4*). Using a natural language model, they have constructed a language "embedding" for a given virus sequence, building a quantitative correlation between mutated sequence structures and language regulations, and correlating it with immune evasion ability. The model can predict whether a mutation will occur given a sequence context.

To obtain the per-lineage fitness of SARS-CoV-2, F. Obermeyer et al. (*5*) estimated virus growth as a linear combination of the effects of individual mutation and developed a hierarchical Bayesian multinomial logistic regression model PyR$_0$. The model can infer the relative prevalence of all virus lineages within a geographic area and detects fitness-related mutations and lineages with increased prevalence rates. The model can identify amino acid substitutions most significantly associated with increased fitness and define significance as the posterior mean or posterior standard deviation.

M. C. Maher et al. (*6*) developed statistical models incorporating epidemiological covariates, taking into account the effects of drivers and the related fitness of different viruses.

Our work focuses on virus fitness calculations by constructing a computable quantitative representation of the genotype-fitness landscape. For this purpose, we set up a theoretical framework called the P-E theorem. It proves that the fitness of the virus population can calculate precisely by selecting appropriate embedding expressions for virus genome mutation under the Variational Auto-Encoder (VAE) framework. Then, following the basic idea learning language of virus evolution (4), through our developed deep learning natural language model CoT2G-F, we realized the accurate calculation of virus population fitness without introducing and artificially coupling macro-epidemiological information and obtained a precise mathematical representation of the virus genotype-fitness landscape in embedding space.

Based on the CoT2G-F model and following the P-E theorem, we derive a quantitative relationship between the fitness of the SARS-CoV-2 lineages and "immune escape" mutations. The fitness as a macroscopic observable biological variable can be accurately described as the mathematical expectation of a latent variable function according to a hidden state distribution.

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

The formulation of the G-P principle and its strict mathematical formulation provide a new valuable perspective and interpretability to apply deep learning methods in the research of the virus evolution. Our research paradigm of directly calculating R_0 through "immune escape" mutations provides a feasible way to construct genotype-phenotype landscapes.

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

Results

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

Understanding the evolutionary causes and consequences of the genotype-fitness landscape is fundamental and challenging work. The features of the genotype-fitness landscape has a decisive effect on the evolutionary process and the predictability of evolution. But, the impact of the genotype-fitness landscape concept on evolutionary biology has been limited owing to the lack of empirical information about the topography of real genotype-fitness landscapes and the precise and quantified relationship between genomic variation and fitness. Or because this relationship is too complicated to calculate (10). We attempt to solve this problem from the basic statistical theory combined with a deep learning model to give a computable quantitative description of the genotype fitness landscape of the virus based on the genome sequence of the SARS-CoV-2 alone.

Starting from the definitions of evolutionary biology (10，14), we obtain a formal expression for the fitness of the virus population, which is consistent with the mathematical expression given in (5). Under the Variational Autoencoder (VAE) framework, 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: “An observable macrobiological phenotype can be calculated under the VAE framework if we can find a reasonable embedded representation of the related microscopic genotype”. The P-E theorem provides a way to establish a deep learning model to calculate virus fitness (see and supplemental note 1, formulas (1-11)). It means that under the VAE framework with a reasonable embedding representation, we can accurately calculate the fitness of the virus population as an observable macrobiological phenotype by using only virus genome sequence data and construct the genotype-fitness landscape of virus population (Box 1 and Fig. 1A).

Modelling of virus evolution mechanisms based on virus sequences

Based on the P-E theorem, we try to set up a natural language model under the VAE framework and calculate the fitness of the virus population. In the natural language that determines the structure of a virus genome sequence, each DNA or amino acid sequence of a virus must conform to the "grammar" of the biological world. By learning and understanding these grammars, we can grasp the evolutionary laws and the composition rules of virus genomic sequences, and then find out the driver mutations that induce "immune escape", obtaining essential information related to virus fitness. In recent years, significant progress has been made in learning the composition rules of DNA or amino acid sequences through natural language models and studying the evolution mechanism of biological species (*4*, *7*–*9*) .

Referring to B. Hie's work (4), we construct a natural language model CoT2G-F, which is a Co-attention-based Transformer model to bridge Genotype and Fitness: (i) Introducing co-attention and self-attention mechanisms to extract the Spatio-temporal correlation and long-range site interaction information within and across virus sequences. These mechanisms enable the model to extract epistatic and recombination signals that promote the occurrence of immune escape mutations. (ii) Dividing training process into two stages: pre-training, simulating the neutral evolution of the virus by randomly masking one or more continuous bases at any position of the input virus sequence, learning the "grammar" rules of sequence composition; and fine-tuning, in chronological order construct the evolutionary map of the virus sequences dynamically, perform supervised fine-tuning according to the Spatial-temporal correlation constraints and target sequences, reproduce the role of environmental selection, bringing the model closer to the virus evolution process (fig. S4). (iii) Using Transformer to replace BiLSTM (4) as the kernel of the natural language pre-training model. When constructing the "semantic" representation or "embedding" of the input virus sequences, the Transformer model with self-attention mechanism can further extract the upstream and downstream long-range related mutation information of the sequence itself.

As of November 2022, 14.96 million SARS-CoV-2 spike proteins in GISAID (www.gisaid.org) after strict quality control were selected to train the CoT2G-F model (table S1). All the data is for pre-training, but the data used for fine-tuning carries out in stages to verify the model's predictive ability. According to the model CoT2G-F, the hidden state distribution of DNA bases or amino acids can be obtained, and a sequence "embedding" that satisfies this hidden state distribution is guaranteed to conform to the "grammar" rules even if there are "semantic" changes. Thus, given an "embedded" representation of input mutated sequences, the "immune escape" mutations can be identified and inferred from the degree of "semantic" change (Fig. 1B, 1C and 1B and see supplemental notes 2-4). Further work is to start from the P-E theorem (Formula (11)) and use the CoT2G-F model to determine the micro-state function of a latent variable corresponding to the virus fitness in the hidden space.

Identifying and predicting "immune escape" mutations

Following the CSCS decision rule for semantics and syntax changes by B. Hie (*4*), we tested the ability of the model CoT2G-F. We compare the performance of our model with the other two related mainstream models from the framework design in three technical dimensions, and our model demonstrates superior performance (see supplementary materials). It can be seen from Fig. 2A that our model performs better than BiLSTM and Vanilla Transformer models, and the ability to identify and predict immune escape mutations has significantly improved. Since we consider the Spatio-temporal dynamics of virus evolution during training processes, the model should have the ability both identify existing and predict future "immune escape" mutations. We proved this using more than 1 million SARS-CoV-2 genomic sequence data from the United Kingdom as an example and got better and more comprehensive prediction results (Fig. 2B, 2C and 2D). This lays a foundation for the subsequent fitness calculation.

We take SARS-CoV-2 genome data collected and submitted by the UK to the GISAID database (the Omicron is not yet pandemic) from August 2021 to October 2021 to perform the prediction (table S1). The model has good reproducibility for already emerged immune escape mutations, and the Omicron strains BA.1 and BA.2.10.4 have been predicted (fig. S5). To verify the predictive ability, we took the data from the United Kingdom submitted to GISAID after March and before October

2022 as input and step-by-step to predict the future occurrence of immune escape mutations. Almost all of the immune escape mutations carried by the Omicron lineage were forecasted, which emerged after March 2022 (Fig. 2D). Considering that our model training use data in GISAID as of March 31, 2022, this result is undoubtedly exciting. It shows that our model can give the correct prediction six months in advance, forecasting the subsequently emerged virus lineages, even without dynamic fine-tuning (see supplementary materials, fig. S4).

Since we randomly mask a sequence span with an average amino acid length of 3 during pre-training, our model can not only simulate various patterns of continuous evolution-related genetic drift, but also can reproduce and predict the occurrence of discrete evolutionary events caused by so-called "catastrophes" events. (see supplemental notes 2-4). In our prediction results, there are a large number of predicted indels, which contain considerable indels that occur in the Omicron virus lineage (see supplementary materials, table S2). Due to the limited space and the topic of this paper, we have not started the detailed analysis, but it is worthy of further study.

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

The immune escape ability of a virus lineage determines its fitness or the basic reproduction number (R_0). The problem is how to establish the quantitative relationship between the "immune escape" mutation and the fitness of the virus conforming to the P-E theorem under the framework of the natural language model. B.

Hie et al. propose in their natural language model that the degree of "semantic" change is related to the immune evasion ability, and the "grammatical" fitness is related to the fitness of the virus (*4*). Meanwhile, F. Obermeyer et al. use a hierarchical Bayesian regression model to fit the relative growth of virus lineages through multinomial likelihood (*5*). They point out that the regression coefficient of their model is equivalent to the per-lineage fitness of the virus.

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

If evolution by natural selection is to occur, there must be a genotype change in the virus population across generations. Therefore, the fitness of the population is a concept of change. If the genotype doesn't change, there's no fitness change. The syntactic nature of the stable genome composition does not directly contribute to the fitness of the virus. If the genome composition already conforms to its syntactic nature, then only the "semantic" changes that affect the gain of genomic function and contribute to the virus fitness change. The "grammar" rules of the natural language model determine the global composition of the sequence to ensure virus survival, and the "semantic" changes correspond to the changes in the "immune escape" ability. Therefore, "semantics" and "grammar" together determine the fitness of viruses. Under the theoretical framework of deep learning, above property can be expressed naturally as the convolution of "semantic" and "grammatical" terms of CSCS criterion (*4*). And when multiplied by the coefficient λ, the final mathematical representation is consistent with the P-E theorem (see supplemental note 1 and note 4, formulas (11) and (31)). This results suggest that the term related to the "semantics" change of the model CoT2G-F and the CSCS criterion (*4*) may be the micro-state function of the latent variable corresponding to the virus 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 VAE model or its extension model. The mathematical expectation of this function in terms of the hidden state distribution is the corresponding macroscopic observable variable that we hope to obtain. We can extend and define a conditional semantic change score (CSC score) from the CSCS score presented by B. Hie et al. (*4*). The conditional semantic change score CSC, as a function of latent variables describing "semantic" changes, is used to measure the "immune escape" ability related to virus sequence mutation. Thinking that the absolute mean of the CSC score is an index of the immune escape ability of a virus lineage, we want to see how this index changes across different virus lineages (see supplemental notes 2-4, formula (27)). We selected 3.7 million SARS-CoV-2 spike protein sequence data from the UK to calculate the absolute mean distribution of the CSC score according to the hidden state probability distribution (Fig. 3).

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

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

It has always been a long-term goal in biology to create genotype-fitness landscapes by mapping DNA sequences to mutation combinations observed in phylogeny or evolutionary experiments. Due to consideration of the epistatic effects of mutations, when the mutation number is too large, the model will become very complicated and difficult to calculate and verify experimentally (10–12) . E. D.

Vaishnav et al. (13) argue that the complete fitness landscape defined by a fitness function maps every sequence (including mutations) in the sequence space to every fitness associated with it. But so far, no theoretical model can explicitly give a concise computational model and framework for the genotype-fitness landscape while fully considering epistasis and recombination potential. Under the framework of the P-E theorem, we try to solve this problem through our deep learning natural language model CoT2G-F combined with CSC scores.

Based on the functional relationship between CSC score and the latent variable of CoT2G-F model (see supplemental note 4, formula (22)), we know that the CSC score is a relative semantic change corresponding to the relative immune escape ability of the virus. The CSC score measures the genomic mutation degree of emerging virus strains relative to wild-type virus strains. Furthermore, the "semantic" mutation corresponding to genomic mutation - CSC score and the "grammatical" coincidence degree corresponding to mutation state distribution - hidden state distribution together determine a measure of the virus 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 obvious and verified (Fig. 3A and 3B, Fig. 4). This result and the P-E theorem provide a mathematical basis for redefining macro-phenotypes such as R_0 and the resulting computability.

The genotype-fitness landscape

The calculation formula of the R_0 has a whole novel meaning now (see supplemental note 4, formula (31), (32) and (33)): The R_0 is the convolution of the semantic change function and the syntactic state distribution function in the embedding space. The meaning of convolution here is to give the average of the cumulative effects of the contribution of all sampled virus variants to the R_0 according to the reference sequence (we select Wuhan-Hu-1 as a reference sequence, see supplementary materials). More importantly, the R_0 can precisely rewrite as a mathematical expectation of a semantic change function according to a hidden state probability distribution giving the quantitative relationship between the "immune escape" ability of the virus and the fitness of the virus lineage. Finally, we present a comprehensive and precise formulation of the virus genotype-fitness landscape. The discovery of consistency between these two different representations provides a solid mathematical basis and research paradigm for applying deep learning models to biological problems and gaining interpretability (see supplemental notes 1-4).

Studying the intrinsic relationship between statistical mechanics and deep learning has always been an intriguing subject (*15–17*). We know that the core concept of statistical mechanics is: "The macroscopic physical observables can characterize as the ensemble average of the corresponding microstate variable functions". Inspired by this, the inference is naturally drawn: "Under the VAE framework, an observable macro-biological variable can express as the mathematical expectation of a function of a latent variable according to a hidden state distribution in decoder space". We call this framework bridging Genotype-Phenotype landscapes the G-P principle, which is an important corollary of the P-E theorem and gives a novel interpretable application of deep learning theory to life science (Fig. 5, supplemental note 1 and 4, formulas (11), (34) and (35)).

Finally, we can plot the genotype-fitness landscape in the embedded space based on the variation state of the viral genome sequence according to formula (30). It is a two-dimensional hypersurface, and we can obtain the fitness of the virus lineage by integrating the surface density of the specific region corresponding to the virus lineage on this hypersurface, get and define the immune escape force of the virus (see supplemental note 4).

Inferring the R_0 of the virus lineage

According to our theoretical framework, computing the fitness of each virus lineage and the fold increase in relative fitness to obtain R_0 requires adequate sampling for each virus lineage in a specific region and time interval. In this region and time interval where and when the virus begins to emerge, sustaining spreads, increases in the number and eventually reaches a plateau. Based on the biological definition of fitness, the selected sampling time interval needs to ensure the virus is transmitted for enough generations to reduce the violent fluctuation of the signal (*5*). It is also crucial for determining the occurrence frequency of the largest cluster of the different lineages and simultaneously calculating the absolute mean of the CSC score (see supplemental note 4, formula (30), fig. S3A and 3B). Currently, the GISAID (www.gisaid.org) contains over 4 million genome sequences of SARS-CoV-2 from the United Kingdom [http://gisaid.org], and the Spatio-temporal distribution of the virus lineages reflects the major trend of the world epidemic of SARS-CoV-2, which can serve as a model system for a reliable and feasible G-P principle.

When computing the virus per-lineage fitness, we used 3.7 million SARS-CoV-2 spike proteins from the GISAID database (www.gisaid.org) and fully adopted the Pango lineage designation and assignment. For acquiring the factor λ_k in the calculation of the per-lineage fitness (see supplemental note 4, formula (30)), lineage assignment is performed with Pangolin firstly. Then, the $\mathrm{k_{th}}$ lineage and its nearest sublineages were selected to form a population, sampling for a given time interval $\mathrm{t_{bin}}$, and λ_k is the occurrence frequency of the $\mathrm{k_{th}}$ lineage in this small population. The calculation result has shown in Figure 4, where the time interval we choose is 10 days (see fig. S6). Our model correctly infers the WHO classification variant Omicron (Pango lineage BA.2.37) to have a very high relative fitness in line with current official monitoring results: ~18 times higher than the original Wuhan lineage (Fig. 4, [95% confidence interval (CI) 17.29 to 18.00], fig. S7), accurately predicts its rise in the spread regions.

The general trends of the R_0 are consistent with the results of (*5*), but the R_0 value is higher overall. It is because when calculating the contribution of immune escape mutation sites to immune escape abilities and R_0, our model considers all possible interactions between all sequence sites of the virus and mutations and takes into account tempo-spatial correlations and recombination effects. This result validates our proposed P-E theorem and G-P principle. Now the relative fitness of the virus lineage and R_0 is precisely computable.

Discussion and conclusion

Applicating AI and deep learning methods to biology have two key points. One is how to make the model more in line with biological logic, and the other is to go deep into the connotation of the model and give the interpretability. This study explores these two aspects and presents an initial research paradigm. We first came up with a general theorem, the P-E theorem, and in the second step, we construct a model CoT2G-F that more closely matches the evolutionary biology scenario. The calculation of R_0 realizes by the model under the guidance of the P-E theorem.

In our model the R_0 is a mathematical expectation of the "semantic" change according to a hidden state probability distribution. It is also a convolution of the "semantic" change function and a hidden state distribution function. The hidden state probability distribution as a prior probability distribution determines the "grammatical" term in the CSCS criterion (Fig. 5, supplementary notes 3 and 4). Both directly gives quantitative relationships among "semantics" change and "grammatical " fitness, "immune escape" mutation and the fitness of virus lineages (see supplementary materials, supplemental note 4, formula (30) and (33)). This result can be the correspondence in the biology of the core concept in statistical mechanics that "macroscopic observable physical quantities are the ensemble average of the corresponding microscopic state variable functions". We call this the G-P (Genotype-Phenotype) principle and it can be regarded as a corollary of the P-E theorem. The G-P principle provides a new valuable perspective and interpretability for applying deep learning to virus evolution. It reveals the intrinsic correlation between deep learning theory and statistical mechanics studied by many researchers (*15–17*).

Currently, most deep learning algorithms for biological studies attempt to establish the relationship between microscopic molecular biology characteristics and macroscopic biology variables to obtain the interpretability of the models (*18–24*). However, it is always challenging to establish a direct, quantitative and computable correlation between them. The practice and results of the R_0 calculation in the present work face the question and give a feasible paradigm for the computational modelling of genotype-phenotype landscapes, which is a concrete and successful example of the G-P principle or the P-E theorem. We believe that according to our research path proposed in the present work, many research examples that follow the G-P principle or the P-E theorem will emerge future. Our model, the derived mathematical representation of R_0, and computational results provide a beautiful commentary on the first principle of "selection imprinting recorded in viral genomic mutation."

In addition, the model uses the Transformer as a kernel and further introduces the co-attention and continuous span masking mechanism considering the Spatial-temporal related virus evolution. The model has advantages in extracting the upstream and downstream long-range correlation mutation information from the sequence itself and the ability to get the Spatio-temporal features related to virus evolution. With the introduction of a continuous mask mechanism, the model can mimic genetic drift and determine what kind of mutations will be fixed by selection. All together, inputting the virus genome sequence alone, the model can more accurately identify and predict possible future "immune escape" mutations by calculating the conditional "semantic" change related to the latent variables. The model's power has demonstrated in (Fig. 2A, 2B, 2C and 2D).

Due to the introduction of a continuous masking mechanism, our model is able to reproduce and predict the occurrence of discontinuous evolutionary events caused by so-called "catastrophes" to some extent. In our prediction results, there is a large number of predicted indels, including quite of few amino acid insertions of the spike protein presented by the Omicron lineage. These insertions conform to the "semantics" plus "grammar" constraints of the sequence composition (table S2). It is worthy of further in-depth research and experimental verification.

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

Finally, it is worth pointing out that the proposed P-E theorem is universal in biological research. There are many fundamental states of biology, such as brain homeostasis, metabolic homeostasis, cell fate state, tumor microenvironment, advanced ageing state, etc. The maintenance, imbalance and remodeling of these macroscopically observable phenotypic states can effectively study based on the P-E theorem. The P-E theorem provides a solid mathematical foundation for the deterministic description of biological homeostasis and deciphering its stability mechanism behind it.

References and Notes

1. D. Mathew *et al.*, *Science*. **369**, eabc8511 (2020).
2. T. N. Starr *et al.*, *Cell*. **182**, 1295-1310.e20 (2020).
3. J. A. G. M. de Visser, S. F. Elena, I. Fragata, S. Matuszewski, *Heredity*. **121**, 401–405 (2018).
4. B. Hie, E. D. Zhong, B. Berger, B. Bryson, *Science*. **371**, 284–288 (2021).
5. F. Obermeyer *et al.*, *Science*. **376**, 1327–1332 (2022).
6. M. C. Maher *et al.*, *Sci. Transl. Med.* **14**, eabk3445 (2022).
7. E. C. Alley, G. Khimulya, S. Biswas, M. AlQuraishi, G. M. Church, *Nat Methods*. **16**, 1315–1322 (2019).
8. T. Bepler, B. Berger, Learning protein sequence embeddings using information from structure (2019), , doi:10.48550/arXiv.1902.08661.
9. R. Rao *et al.*, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2019; https://papers.nips.cc/paper/2019/hash/37f65c068b7723cd7809ee2d31d7861c-Abstract.html), vol. 32.
10. J. A. G. M. de Visser, J. Krug, *Nature Reviews Genetics*. **15**, 480–490 (2014).
11. D. A. Kondrashov, F. A. Kondrashov, *Trends in Genetics*. **31**, 24–33 (2015).
12. I. Fragata, A. Blanckaert, M. A. Dias Louro, D. A. Liberles, C. Bank, *Trends in Ecology & Evolution*. **34**, 69–82 (2019).
13. E. D. Vaishnav *et al.*, *Nature*. **603**, 455–463 (2022).
14. D. J. Futuyma, M. Kirkpatrick, in *Evolution* (Sinauer Associates, Inc,

Sunderland, Massachusetts, Fourth., 2017), pp. 103–134.

15. Y. Bahri *et al.*, *Annual Review of Condensed Matter Physics*. **11**, 501–528 (2020).
16. G. Carleo *et al.*, *Rev. Mod. Phys.* **91**, 045002 (2019).
17. Y. Liu, X. Yao, *Neural Networks*. **12**, 1399–1404 (1999).
18. L. Qiu, L. Lin, V. M. Chinchilli, Mutational Interpretable Learning from Multi-view Data (2022), , doi:10.48550/arXiv.2202.13503.
19. K. M. Boehm, P. Khosravi, R. Vanguri, J. Gao, S. P. Shah, *Nat Rev Cancer*. **22**, 114–126 (2022).
20. I. Cetin, O. Camara, M. A. G. Ballester, Attri-VAE: attribute-based, disentangled and interpretable representations of medical images with mutational autoencoders (2022), , doi:10.48550/arXiv.2203.10417.
21. L. Seninge, I. Anastopoulos, H. Ding, J. Stuart, *Nat Commun*. **12**, 5684 (2021).
22. H. A. Elmarakeby *et al.*, *Nature*. **598**, 348–352 (2021).
23. A. Cheerla, O. Gevaert, *Bioinformatics*. **35**, i446–i454 (2019).
24. S. Ainsworth, N. Foti, A. K. Lee, E. Fox, Interpretable VAEs for nonlinear group factor analysis (2018), , doi:10.48550/arXiv.1802.06765.

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

Funding: Y. X. Li is Supported by the CAS Research Fund, Grant No. XDB38050200 and the Self-supporting Program of Guangzhou Laboratory, Grant No. SRPG22-001 and SRPG22-007.

Author contributions: All authors conceived the project and methodology. Y. X. Liu led the completion of all data preparation and preprocessing. Y. Luo, Y. X. Liu, X. Lu and H. Gao performed the computational experiments and wrote the software, R. K. He contributed to conceptual design and data preparation. X. Zhang, X. G. Zhang and Y.X. Li outlined the overall theoretical framework.Y. X. Li laid the foundations of mathematics. All authors interpreted the results and wrote the manuscript.

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

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

Supplementary Materials

Materials and Methods

Supplementary Text

Figs. S1 to S5

Tables S1 to S2

References

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

Population genetics (*14*) : the average fitness ϖ of the virus population:

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

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

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

We define a function $P(\boldsymbol{x})$ as a probability state distribution quantifying the virus lineages' fitness changes caused by sequence variation states, and $P(\boldsymbol{x})$ can have a expand form using a Gaussian Mixture Model (GMM).

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

$\pi_k N(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ 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:

$$\varpi = E(\omega) = \int \boldsymbol{\omega}(\boldsymbol{x})P(\boldsymbol{x})d\boldsymbol{x} = \sum_{k=1}^{K}\pi_k\varpi_k \qquad (4)$$

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

$$\varpi_k \overset{\text{def}}{=} \frac{\lambda_k}{n_k}\sum_{i=1}^{n_k}\omega_i^k \qquad (5)$$

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

$$\varpi_k = \varpi_{|P(\boldsymbol{x})} = \int \boldsymbol{\omega}(\boldsymbol{x})\, P(\boldsymbol{x})\, d\boldsymbol{x} \qquad (6)$$

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

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

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

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

Under the VAE framework together with Bayes' theorem, because $\boldsymbol{x} = W\boldsymbol{z} + B$, and $\frac{dx}{dz} = \mathrm{P}_{\varphi}(\mathbf{x}\,|\,\mathbf{z})/\mathrm{Q}_{\theta}(\mathbf{z}\,|\,\mathbf{x}) = W$, we have

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

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

Considering that the average of states is by virus lineage, the fitness of a virus lineage is determined by the dominant virus cluster in the lineage， and then by doing a simple matrix operation we can get $\mathrm{W}\boldsymbol{\omega}(\mathrm{z}) \rightarrow \lambda\boldsymbol{\omega}(\mathrm{z})$. λ is a diagonal matrix. Then, we have

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

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

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

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 VAE 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 ω(**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 "ω(**z**)", "$\mathrm{P}(\boldsymbol{z})$", and “time” as axes. The stable points on the hypersurface correspond to the genomic variation states that contribute the most to the "immune escape" ability. We can obtain the fitness of the virus lineage by integrating the surface density of the specific region corresponding to the virus lineage on this hypersurface. Now, if we take the gradient of the ω(**z**) along the time and the $\mathrm{P}(\boldsymbol{z})$ axis on a two-dimensional hypersurface and we got and defined the immune escape force of the virus.

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

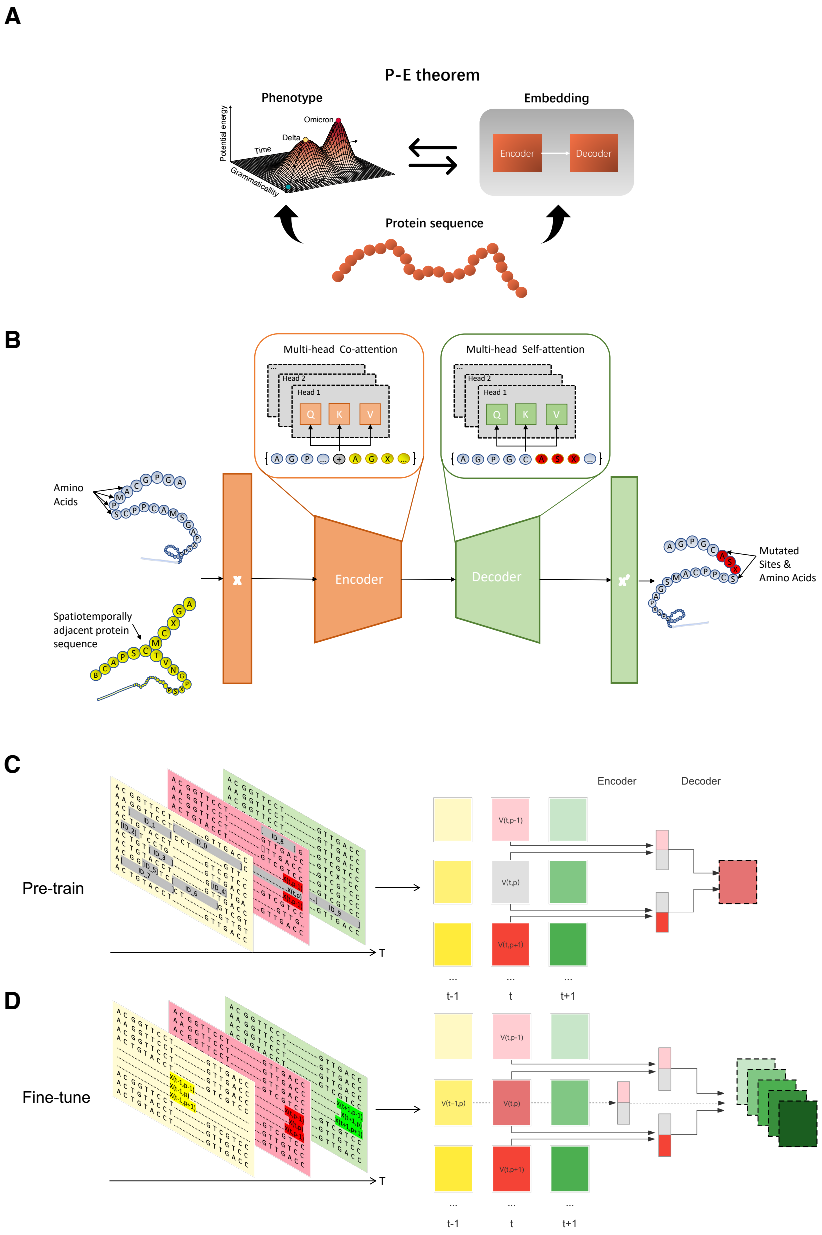

A
P-E theorem
Phenotype
Potential energy
Time
Grammaticality
Delta
Omicron
Embedding
Encoder
Decoder
Protein sequence
B
Multi-head Co-attention
Head 2
Head 1
Q
K
V
Multi-head Self-attention
Amino Acids
Spatiotemporally adjacent protein sequence
Encoder
Decoder
Mutated Sites & Amino Acids
C
Pre-train
T
Encoder
Decoder
V(t,p-1)
V(t,p)
V(t,p+1)
t-1
t
t+1
D
Fine-tune
T
V(t,p-1)
V(t−1,p)
V(t,p)
V(t,p+1)
t-1
t
t+1

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

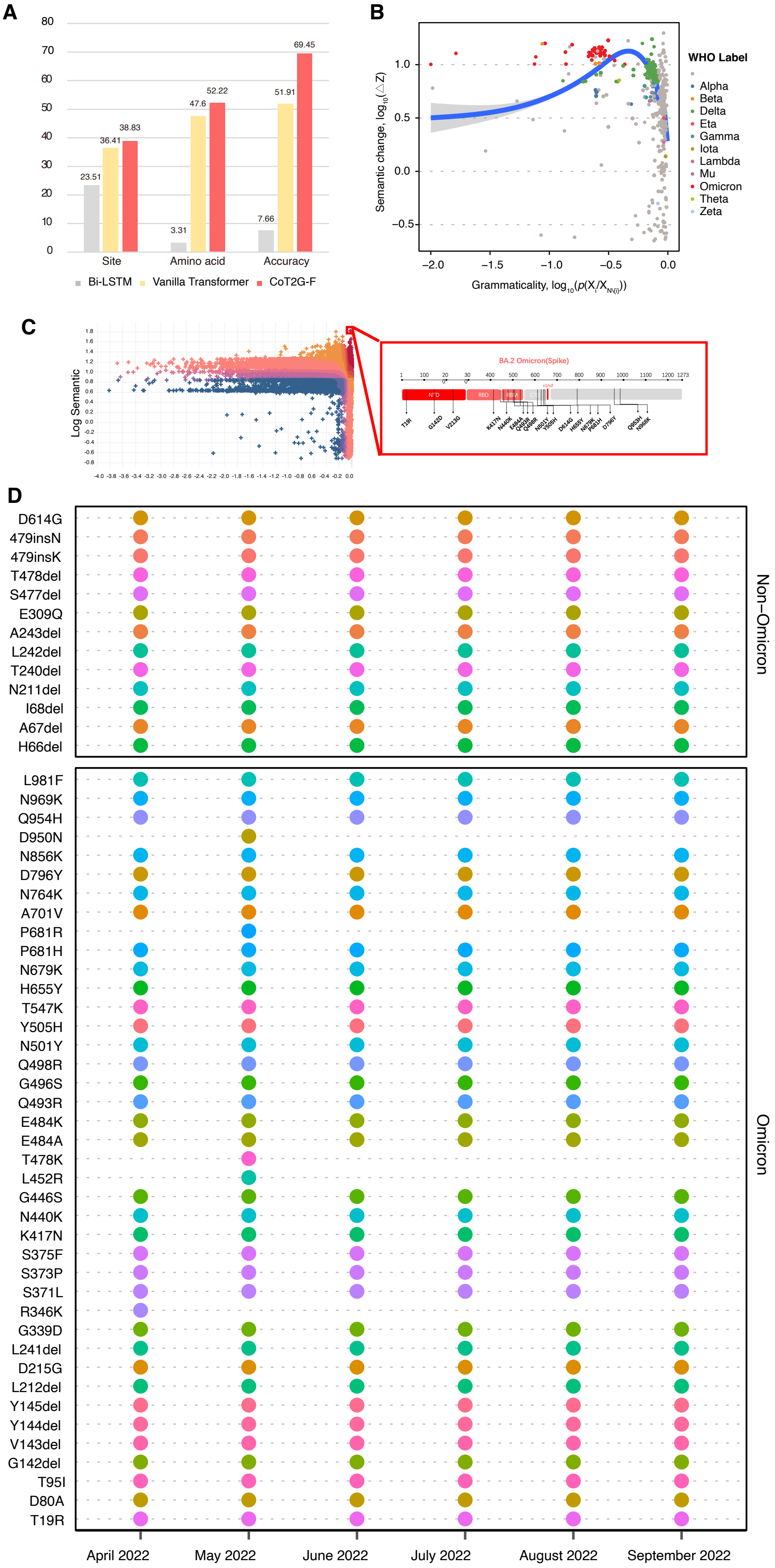
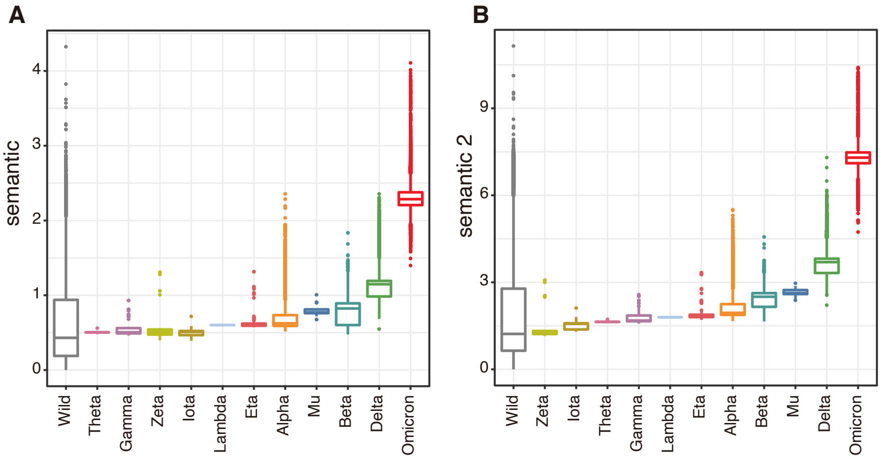

A
80
70
60
50
40
30
20
10
0
23.51
36.41
38.83
3.31
47.6
52.22
7.66
51.91
69.45
Site
Amino acid
Accuracy
Bi-LSTM
Vanilla Transformer
CoT2G-F
B
Semantic change, $\log_{10}(\triangle Z)$
Grammaticality, $\log_{10}(p(X_i/X_{N\setminus\{i\}}))$
1.0
0.5
0.0
−0.5
−2.0
−1.5
−1.0
−0.5
0.0
WHO Label
Alpha
Beta
Delta
Eta
Gamma
Iota
Lambda
Mu
Omicron
Theta
Zeta
C
Log Semantic
BA.2 Omicron(Spike)
NTD
RBD
RBM
CTD
s1/s2
T19I
G142D
V213G
K417N
N440K
E484A
Q493R
Q498R
N501Y
Y505H
D614G
H655Y
N679K
P681H
D796Y
Q953H
N968K
D
D614G
479insN
479insK
T478del
S477del
E309Q
A243del
L242del
T240del
N211del
I68del
A67del
H66del
Non−Omicron
L981F
N969K
Q954H
D950N
N856K
D796Y
N764K
A701V
P681R
P681H
N679K
H655Y
T547K
Y505H
N501Y
Q498R
G496S
Q493R
E484K
E484A
T478K
L452R
G446S
N440K
K417N
S375F
S373P
S371L
R346K
G339D
L241del
D215G
L212del
Y145del
Y144del
V143del
G142del
T95I
D80A
T19R
Omicron
April 2022
May 2022
June 2022
July 2022
August 2022
September 2022

Fig. 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 $\mathrm{k_{th}}$ virus lineage, which is defined as the immune escape capacity of the $\mathrm{k_{th}}$ virus lineage. Figures (A) and (B) both clearly show that with the progress of the global epidemic, from the Wuhan virus lineage to the Delta lineage and then to the Omicron lineage, the immune evasion ability continues to increase and even accelerates. Figures (A) and (B) show the same absolute mean distribution trend of the CSC_i^k, but figure (B) changes is more smoothly.



615 **Fig. 4. Relative fitness derived from the natural language model CoT2G-F versus date of lineage emergence.** Circle size is proportional to the sampling number for different lineage in the time interval. This figure uses a time interval of 10 days. Referring to the Pango lineage designation and assignment, the R_0, which is the fold increase in relative fitness of the virus lineages according to the Wuhan lineage, is 620 plotted in different colours. The results for 5, and 20-day time intervals have shown in (fig. S6). All the results are almost consistent, reflecting the robustness of the model.

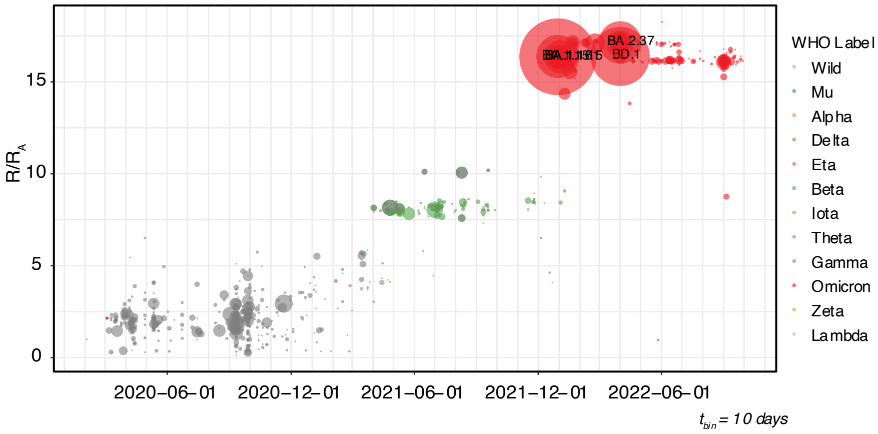
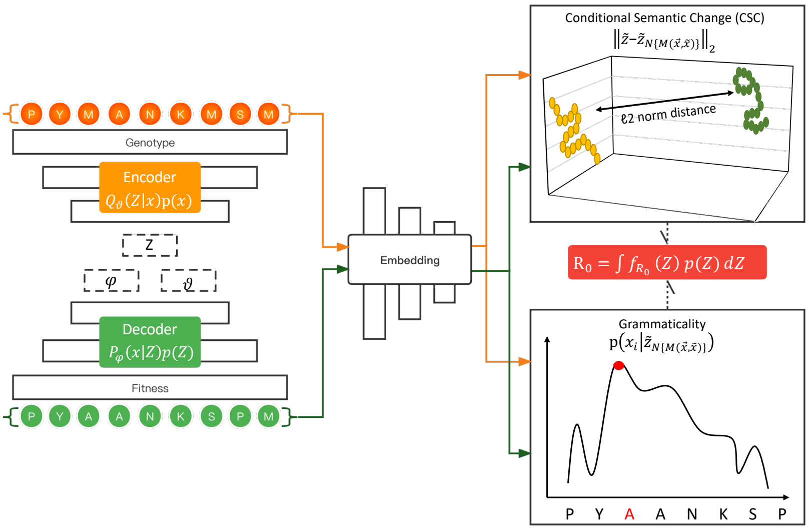


Fig. 5. CoT2G-F framework for building genotype-phenotypic landscapes.

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 two hidden state probability distributions $Q_\theta(\mathbf{z}|\mathbf{x})$ and $P(\mathbf{z})$ in the encoder-decoder space, the prior probability $P(\mathbf{x})$ and the posterior probability distribution $P_\varphi(\mathbf{x}|\mathbf{z})$ by 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 probability distribution $P(\mathbf{z})$ in the decoder space, the model can express the R_0 as the mathematical expectation of a specific latent variables function related to viral sequence mutation according to a hidden state probability distribution, building a genotype-fitness landscape.


P Y M A N K M S M
Genotype
Encoder
$Q_\theta(Z|x)p(x)$
Z
φ
ϑ
Decoder
$P_\varphi(x|Z)p(Z)$
Fitness
P Y A A N K S P M
Embedding
Conditional Semantic Change (CSC)
$\left\|\tilde{z}-\tilde{z}_{N\{M(\vec{x},\tilde{x})\}}\right\|_2$
ℓ2 norm distance
$R_0 = \int f_{R_0}(Z)\, p(Z)\, dZ$
Grammaticality
$p\left(x_i|\tilde{z}_{N\{M(\vec{x},\tilde{x})\}}\right)$
P Y A A N K S P