

Empirical substitution models of SARS-CoV-2 protein evolution for phylogenetic inference

David Ferreira

Universidade de Vigo

Luis Daniel González-Vázquez

Universidade de Vigo

Miguel Arenas

marenas@uvigo.es

Universidade de Vigo

Research Article

Keywords: Substitution model, SARS-CoV-2 proteins, Protein evolution, Model selection, Protein folding stability, Phylogenetics

Posted Date: July 21st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9790437/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Probabilistic phylogenetic inference requires substitution models of molecular evolution, which should be as realistic as possible to yield accurate predictions. At the protein level, empirical substitution models are well established in phylogenetics. However, the number of available empirical substitution models remains limited, and only a few focus on rapidly evolving viruses. Indeed, no substitution models have yet been developed specifically for SARS-CoV-2 proteins, despite their potential to improve evolutionary predictions. Here, we present a set of empirical substitution models for SARS-CoV-2 proteins, including the main protease, papain-like protease, spike protein, and the complete proteome. We estimated these models using maximum likelihood from thousands of protein sequences and validated with independent datasets. Overall, the estimated empirical substitution models outperformed currently available empirical models in terms of phylogenetic likelihood and revealed distinct evolutionary patterns among SARS-CoV-2 proteins. Additionally, using forward-in-time evolutionary simulations, we found that SARS-CoV-2 proteins evolved under these empirical substitution models generated protein variants with folding stabilities more realistic than those produced under other empirical substitution models. We conclude that evolutionary inferences based on protein-specific substitution models are generally more accurate and biologically realistic than those based on generalist models, and recommend the empirical substitution models presented here for evolutionary predictions of SARS-CoV-2 proteins.

Introduction

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) continues to circulate globally, causing significant morbidity and millions of deaths each year [1]. The evolution of this virus has shown the fixation of multiple mutations (substitutions) that confer fitness advantages such as altered transmissibility, increased virulence, and enhanced immune evasion capabilities [2–6]. Consequently, understanding the molecular evolution of SARS-CoV-2 is crucial for designing effective treatments that target circulating viral diversity [7, 8] and for predicting past and future evolutionary trajectories [9–13], which can also inform the design of therapeutic strategies against the virus [14].

In this regard, empirical substitution models of protein evolution, which capture the observed substitution patterns in terms of relative rates of change among amino acids and amino acid frequencies at the equilibrium, can provide valuable insights for evolutionary predictions [15–18]. Actually, these models are widely used in evolutionary biology to infer phylogenetic trees, ancestral sequences, evolutionary parameters and selection pressures, as well as to simulate molecular evolution, among other applications [15–21]. Thus, empirical substitution models of protein evolution have been broadly implemented in popular phylogenetic frameworks (e.g. *RAxML-NG* [22], *ModelTest-NG* [23], *IQ-TREE* [24] and *MEGA* [25], among others). However, the number of available empirical substitution models remains limited considering the vast variety of existing protein families [26], which can lead to unrealistic modeling in many cases [e.g., 27]. Since evolutionary patterns can vary among proteins [26] and the choice of substitution model can affect the accuracy of phylogenetic trees and ancestral

sequence reconstructions [28–30], the development of empirical substitution models for proteins whose evolutionary patterns are not represented by existing models is important. In the case of SARS-CoV-2, the virus encodes a variety of proteins (including nonstructural, structural, and accessory proteins [6, 31, 32]) required for virus replication, but no empirical substitution models have yet been developed to mimic the evolution of these proteins.

Here, we present empirical substitution models of evolution for several key SARS-CoV-2 proteins, including spike, main protease (Mpro, nsp5), and papain-like protease (PLpro, nsp3), as well as a model for the entire SARS-CoV-2 proteome. The models were inferred using the maximum likelihood (ML) approach applied to thousands of protein sequences, and their fit to real SARS-CoV-2 data was evaluated by comparison with currently available empirical models, using phylogenetic likelihood and the folding stability of modeled proteins.

Material and Methods

Estimation of empirical substitution models of protein evolution for SARS-CoV-2

The extensive sequencing of SARS-CoV-2 enables comprehensive evolutionary analyses of the virus. Sequences were obtained from the GISAID and NCBI databases, and we applied a filtering step to include only sequences without ambiguities and without indels of 10 or more amino acids with respect to the reference protein sequence (NCBI Protein ID: YP_009724389.1), as well as sequences sharing at least 75% amino acid identity with the reference to exclude poor-quality annotations and select realistic variants [33]. Sequences from each dataset were aligned using *MAFFT* [34]. For each of the spike, Mpro, and PLpro SARS-CoV-2 proteins (with sequence lengths of 1273, 306 and 316 amino acids, respectively), we constructed a protein sequence dataset of 100,000 randomly sampled sequences for estimating the corresponding substitution model, hereafter referred to as the *method dataset*. For the entire proteome (sequence length of 34,386 amino acids), the method dataset included 20,000 randomly sampled sequences. We also constructed *test datasets* to evaluate the estimated substitution model for each study protein and for the proteome. In particular, we generated 10 independent test datasets for each protein and for the proteome, each comprising 1,000 randomly sampled sequences (not included in the method dataset), except for the proteome datasets, which included 100 sequences each. Larger sample sizes were not employed due to computational limitations of ML calculations.

We estimated the empirical substitution models for each SARS-CoV-2 protein and for the proteome using ML with the *QMaker* method [35] implemented in *IQ-TREE2* [24], based on the corresponding method dataset and a phylogenetic tree. This phylogenetic tree was previously reconstructed with ML under the available best-fitting empirical substitution model selected for the corresponding dataset with *ModelFinder* [36] (Table S1; Supplementary Material). Accordingly, we estimated the general time-reversible 20·20 exchangeability matrix (*Q*), which includes the relative rates of change between amino acids and the amino acid frequencies at the equilibrium (Equation I).

$$Q_{ij} = M_{ij}f_{ij}(l)$$

The resulting exchangeability matrices for each SARS-CoV-2 protein and for the proteome (hereafter named SC2-Mpro, SC2-PLpro, SC2-Spike, and SC2-Proteome) were included in a public database of empirical substitution models of protein evolution, *eModelDB* (<https://emodldb.streamlit.app/> [26]), to facilitate accessibility and comparison with other substitution models.

Evaluation of the empirical substitution models of SARS-CoV-2 protein evolution

We evaluated the estimated empirical substitution models using phylogenetic likelihood and the folding stability of modeled proteins. In particular, we compared the fit of the empirical substitution models, including the models estimated here and other currently available empirical models, with the test datasets.

To evaluate the models through phylogenetic likelihood, we calculated the phylogenetic likelihood of the SARS-CoV-2 substitution models and the existing best-fitting models selected with *ModelFinder* (Kalyaanamoorthy et al. 2017; Minh et al. 2020) for each test dataset. All substitution models were evaluated both considering and ignoring their amino acid frequencies at the equilibrium (+ F). We then used the Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC) [37–39], as implemented in *ModelFinder*, to identify the substitution model providing the best phylogenetic fit to each test dataset. Additionally, we found it interesting to assess the similarity between each SARS-CoV-2 substitution model and widely used empirical substitution models. For this purpose, we normalized the instantaneous rate matrices (scaling values between 0 and 1) and the stationary amino acid frequencies across all substitution models, enabling direct comparison among their elements. Subsequently, we performed a principal component analysis (PCA) using the *Python* scikit-learn library (Pedregosa et al. 2011) to identify the relative distances among substitution models.

Next, we evaluated the SC2-Mpro and SC2-PLpro substitution models through the folding stability of modeled proteins, as the structures of these proteins can be computationally modeled in intensive simulations of protein evolution along evolutionary histories. In particular, we performed forward-in-time simulations of SARS-CoV-2 Mpro and PLpro sequences under the estimated SARS-CoV-2 substitution models, under the currently available substitution models that previously produced the best phylogenetic fit and, under a structurally constrained substitution (SCS) model [40, 41]. For this purpose, we simulated evolutionary histories using a standard population birth-death model [13], which assigns birth and death rates according to fitness effects derived from protein stability, while modeling protein evolution under each studied substitution model using *ProteinEvolver2* [13]. The Wuhan strain sequence of Mpro (YP_009725301.1) and PLpro (YP_009742610.1) was assigned to the root node and evolved forward in time (along the birth-death evolutionary history and under the substitution model), generating multiple sequence alignments of 1,000 sequences (see Table S2, Supplementary Material, for detailed information on the parameters used for the birth-death and substitution models). A total of 100

simulated multiple sequence alignments with sequence identity similar (± 0.01) to that observed in the real test dataset were generated. Finally, we estimated the protein folding stability (ΔG) of both real and simulated protein sequences under each substitution model using *DeltaGREM* [41, 42], using the Wuhan strain as the reference structural template.

Results

Substitution patterns in SARS-CoV-2 proteins

The empirical substitution models estimated for SARS-CoV-2 proteins revealed both general and protein-specific evolutionary patterns. Overall, the SC2-Mpro and SC2-PLpro models showed multiple amino acid changes with low relative substitution rates [e.g., histidine-valine (H-V) and serine-glutamine (S-Q)], whereas the substitution models estimated for the spike and the entire proteome were less restrictive (Table 1 and Fig. 1). Indeed, the substitution models showed some recurrent evolutionary patterns (Table 1) associated with the physicochemical properties of the amino acids involved in the substitution events, as detailed below.

Table 1

Frequent amino acid replacements identified in SARS-CoV-2 protein substitution models. Amino acid replacements with the highest relative rates of change (above 0.9 in normalized rates ranging from 0 to 1) in the SARS-CoV-2 empirical substitution models are shown. The table lists the amino acids involved in each type of substitution, the normalized relative substitution rate, and a classification based on the physicochemical properties of the involved amino acids (shown in parentheses). This classification includes (1) conservation of size and hydrophobicity, (2) diversification of physicochemical properties, and (3) conservation of amino acid charge.

SARS-CoV-2 substitution model	Frequent amino acid replacement					
SC2-Mpro	Ala-Val (1)	Lys-Arg (3)	Pro-His (2)	Thr-Ile (1)	Val-Ile ⁽¹⁾	Leu-Phe (1)
	0.911	0.944	1.000	0.996	0.951	0.947
SC2-PLpro	Ala-Asp (3)	Ala-Val (1)	Ser-Gly (2)	Thr-Ile (1)	Leu-Pro (2)	Pro-Ser (2)
	0.996	0.978	0.907	1.000	1.000	1.000
SC2-Spike	His-Gln (3)	Gln-E ⁽³⁾	Gly-Asp (2)	Arg-Lys (3)	Tyr-His (3)	-
	1.000	0.952	0.992	0.953	1.000	
SC2-Proteome	Val-Ala (1)	His-Tyr (1)	Ile-Thr ⁽¹⁾	Pro-Ser (2)	-	-
	0.950	1.000	1.000	0.915		

1. *Conservation of amino acid size and hydrophobicity in certain substitutions.* Substitutions between amino acids of similar size (and often hydrophobicity) were commonly observed in the estimated models. For example, the alanine-valine (A-V) substitution, which preserves both hydrophobicity and size, showed a high substitution rate in all estimated models. Substitutions between amino acids of comparable size but variable polarity were also frequent. For instance, threonine-isoleucine (T-I), valine-isoleucine (V-I) and histidine-tyrosine (H-Y) generally showed high substitution rates in the SC2-Mpro, SC2-PLpro and SC2-Proteome models (Fig. 1). Similarly, the leucine-phenylalanine (L-F) substitution, which preserves hydrophobicity, also showed a high substitution rate.
2. *Diversification of physicochemical properties in certain amino acid substitutions.* Some substitutions between amino acids with distinct physicochemical properties were also detected in the estimated models. Notably, proline-histidine (P-H) substitutions involve the replacement of a non-polar, conformationally restricted residue (P) with a polar residue that can acquire a positive charge (H). Indeed, proline-serine (P-S) and proline-leucine (P-L) substitutions [involving non-polar and structural torsional effects (P) toward other residues] were relatively frequent in the studied proteins. Additional recurrent substitutions involved glycine-serine (G-S) and glycine-aspartate (G-D), which replace the smallest non-polar residue with either a small polar or a charged residue.
3. *Conservation of charge and related physicochemical properties in certain amino acid substitutions.* Amino acid substitutions that preserve residue charge were frequently observed in proteins that must modulate surface properties involved in host recognition and immune escape. Most of these substitutions were detected in the SC2-Spike model. For example, the arginine-lysine (R-K) substitution preserves positive charge, whereas histidine-glutamine (H-Q) and histidine-tyrosine (H-Y) substitutions conserve polarity and related properties. In SC2-Mpro, the high lysine-arginine (K-R) substitution rate reflects a conservative exchange between two positively charged residues, suggesting maintenance of local electrostatic interactions while allowing modest flexibility in side-chain geometry. This pattern is consistent with constraints related to structural stability or substrate recognition. The glutamine-glutamate (Q-E) substitution, which replaces a polar residue with a negatively charged residue, was frequent in spike. Similarly, in SC2-PLpro, alanine-aspartate (A-D) substitutions, involving exchanges between a small non-polar residue and a negatively charged residue, were also frequent.

Regarding stationary amino acid frequencies, which influence exchangeability rates among amino acids, we observed elevated frequencies of proline (P, 0.123) in the SC2-Mpro model, leucine (L, 0.131) and isoleucine (I, 0.086) in the SC2-PLpro model, phenylalanine (F, 0.124) and valine (V, 0.080) in the SC2-Spike model, and lysine (K, 0.247) in the SC2-Proteome model.

Evaluation of empirical substitution models of SARS-CoV-2 proteins through phylogenetic likelihood

The estimated empirical substitution models for SARS-CoV-2 proteins generally fit the test datasets better, yielding lower BIC and AIC scores, than existing empirical substitution models in almost all test

datasets (Fig. 2 and Tables S3-S6; Supplementary Material). Notably, considering the observed amino acid frequencies (+ F) for SC2-Mpro and SC2-PLpro models resulted in worse fitting with the test datasets than the currently available empirical substitution models (Fig. 2AB). The SC2-Spike model also showed improved fit relative to existing empirical substitution models. However, in this case, incorporating observed amino acid frequencies from the test datasets further enhanced model performance (Fig. 2C). Similarly, the SC2-Proteome model showed its best fit when evaluated using the observed amino acid frequencies (Fig. 2D). We next assessed the performance of the SC2-Proteome model relative to the SARS-CoV-2 protein-specific models (SC2-Mpro, SC2-PLpro, and SC2-Spike) and standard empirical substitution models for the corresponding test datasets. For the Mpro and PLpro test data, the SC2-Proteome model ranked in an intermediate position based on BIC and AIC scores, with slightly improved performance when using observed amino acid frequencies (+ F) (Tables S3-S6). Interestingly, for the spike data, the SC2-Proteome + F model ranked second or third in terms of fit to the several proteome test datasets, alternating with the SC2-Spike model and just behind the SC2-Spike + F model (Tables S3-S6).

Similarity between substitution models for SARS-CoV-2 proteins and other currently available empirical substitution models

We compared the estimated SARS-CoV-2 substitution models with other existing empirical substitution models that previously showed high fit to the SARS-CoV-2 test datasets. Principal component analysis (PCA) based on the exchangeability matrices revealed a clear separation among empirical models inferred from general data (i.e., JTT [43]), lineage-specific data (i.e., HIVw [44], FLU [45], FLAVI [46], and Q.Bird [35]), and nuclear generalist data (i.e., VT [47] and PMB [48]) (Fig. 3), in agreement with previous studies [26]. With respect to the SARS-CoV-2-specific models, the SC2-Spike and SC2-Proteome models clustered closely, whereas the SC2-Mpro and SC2-PLpro models were more distantly related to other models (Fig. 3). Notably, the alanine-valine (A-V) substitution is strongly emphasized in SARS-CoV-2-specific models, with extremely high rates (above 0.75). In contrast, standard empirical models assign substantially lower rates to this replacement, including JTT, PMB, and VT models (A-V rates below 0.35), as well as the FLU model (A-V rate of 0.23), whereas other virus-specific models, such as HIVw and FLAVI, show intermediate A-V rates (0.52 and 0.41, respectively). An opposite pattern is observed for the threonine-serine (T-S) substitution. This amino acid replacement shows high rates in general empirical models, reaching 0.49 in JTT, 0.56 in PMB, and 0.59 in VT, but is comparatively reduced in SARS-CoV-2-specific models and in some other models such as Q.Bird, HIVw, FLU, and FLAVI (below 0.25). The valine-methionine (V-M) replacement further distinguishes the Q.Bird model from other empirical models, showing a high rate (above 0.57). A similarly elevated rate is observed in the SC2-PLpro model (0.63), whereas this rate remains below 0.33 in both general empirical models and the remaining SARS-CoV-2-specific models. Finally, although the HIVw, FLU, and FLAVI models do not display elevated rates for any particular substitution type, they share a general pattern of reduced rates across multiple substitution types. This pattern is shared with the SC2-Mpro and SC2-PLpro models but contrasts with

the other SARS-CoV-2-specific models. These differences show how SARS-CoV-2-specific substitution models capture variable evolutionary constraints relative to both general and lineage-specific empirical models.

Evaluation of empirical substitution models of SARS-CoV-2 proteins through protein folding stability

We compared the folding stability of protein sequences simulated under three models (the estimated SARS-CoV-2-specific models, the likelihood-based best-fitting traditional empirical model, and the SCS model) with that of the real sequences, using 100 diversity-matched simulated alignments for Mpro and PLpro. For Mpro simulated data, all the substitution models simulated proteins with only a slight decrease in folding stability relative to the real sequences (Fig. 4). The SC2-Mpro model performed significantly better than the traditional likelihood-based best-fitting empirical model, although it did not outperform the SCS model. For PLpro, simulated protein sequences under the SC2-PLpro model showed folding stabilities closer to those observed in the real data than sequences generated with the SCS model, which produced overly stable proteins, or the traditional likelihood-based best-fitting empirical model (FLAVI), which produced less stable proteins.

Discussion

Probabilistic phylogenetics requires substitution models of molecular evolution to determine which states are likely to be present over evolutionary time [49, 50]. In this regard, empirical substitution models of protein evolution remain widely used in the field due to their computational efficiency and implementation in popular software [22–25]. An important factor is that different models can mimic the evolution of a given dataset with varying accuracy, potentially influencing evolutionary inferences derived from them [29, 30, 51]. The current collection of empirical substitution models of protein evolution remains limited relative to the diversity of protein families and taxa that exist [26] and, therefore, expanding this set with additional models is important to achieve appropriate model fitting for newly generated data [27]. An example is the evolution of SARS-CoV-2 proteins, which is central for reconstructing the phenotypic past of the virus, projecting its future evolutionary trajectories, and informing the development of effective therapeutic strategies. Motivated by the lack of empirical substitution models for SARS-CoV-2 proteins, we developed empirical substitution models for fundamental SARS-CoV-2 proteins as well as for the entire viral proteome. We found that SARS-CoV-2-specific substitution models generally fit SARS-CoV-2 protein data better than standard empirical models, indicating that they more accurately capture the evolutionary patterns of SARS-CoV-2 proteins.

Interestingly, we identified distinct substitution patterns among the SARS-CoV-2-specific substitution models, with the viral proteases showing stronger evolutionary constraints than the spike protein and the proteome. This is expected, as proteases perform specific catalytic functions that usually require tight structural packing [52, 53] and the conservation of key catalytic residues [54]. Accordingly, substitutions

in these enzymes are often disfavored when they do not preserve side-chain size or structural compatibility [55, 56]. Despite these constraints, the SC2-Mpro model showed tolerance to substitutions between lysine and leucine, likely occurring at solvent-exposed or structurally flexible regions. Similarly, frequent proline-leucine substitutions observed in the SC2-PLpro model can reflect a combination of local structural tolerance and compositional bias. Although proline is a relatively infrequent amino acid, its replacement by leucine suggests that these positions are not deeply buried or critically constrained.

Regarding the spike protein, crucial for host cell entry and immune evasion [2], its model (SC2-Spike) exhibited high rates of amino acid replacement, including adjustments in charge and polarity. Such substitutions are consistent with its predominantly surface-exposed role and the need for adaptive flexibility. For example, frequent tyrosine-histidine substitutions can facilitate interface plasticity by enabling rapid modulation of binding affinity or antibody recognition through changes between aromatic and polar/charged states [57–59]. Tyrosine residues are often enriched in antibody-binding sites, where they contribute through aromatic and hydrogen-bonding interactions [60–63], whereas histidine substitutions are commonly used to introduce pH-dependent and plastic binding behavior that can modulate affinity and antigen recognition [64, 65]. Additional substitutions between basic residues, particularly lysine and arginine, likely occur at solvent-exposed or membrane-proximal regions, where flexible or loop-like regions can accommodate changes in side-chain chemistry [66, 67]. Such lysine-arginine replacements are known to remodel surface electrostatics, enhance membrane affinity, and modulate membrane fusion efficiency [66–69]. Likewise, polar residues such as histidine and glutamine are often enriched in flexible, interaction-prone regions that support adaptable surface and phase-separation interactions [62, 63, 68, 70]. Finally, the SC2-Proteome substitution model captures general evolutionary patterns shared across SARS-CoV-2 proteins, emphasizing broad constraints such as hydrophobic core conservation and side-chain size compatibility, while lacking the protein-specific selective pressures observed in the models based on individual proteins. Interestingly, the high lysine frequency observed in this model can reflect an overall compositional bias toward basic residues in SARS-CoV-2 proteins.

The estimated SARS-CoV-2 substitution models generally provided a better likelihood-based fit to independent SARS-CoV-2 datasets than the currently available standard empirical substitution models. In particular, the SARS-CoV-2 protein-specific substitution models generally outperformed standard empirical models, reflecting protein-specific selective pressures acting on SARS-CoV-2 protein evolution. The SC2-Proteome substitution model ranked after the protein-specific models, suggesting that it effectively captures substitution patterns shared across diverse SARS-CoV-2 proteins. Notably, the performance of SARS-CoV-2 substitution models depended on whether the estimated amino acid frequencies at the equilibrium or the amino acid frequencies observed in the study test data were considered. Models based on estimated amino acid frequencies at the equilibrium were more accurate in mimicking the evolution of small SARS-CoV-2 proteins, Mpro and PLpro, which we believe can reflect the importance of certain amino acids under strong functional constraints [54, 71]. In contrast, substitution models derived from large proteins (spike and the proteome) showed improved performance when using amino acid frequencies observed in the study test data, possibly because

these frequencies provide a more general instantaneous perspective, whereas background compositional information can become noisier due to the presence of multiple embedded regions with distinct evolutionary patterns. Consistent with this interpretation, principal component analysis revealed that SC2-Mpro and SC2-PLpro cluster separately from SC2-Spike and SC2-Proteome, reflecting distinct protein- and lineage-specific evolutionary constraints. A clear example is the alanine-valine substitution, which is strongly favored in SARS-CoV-2-specific models (particularly SC2-Mpro, SC2-PLpro, and SC2-proteome) but occurs at substantially lower rates in standard empirical models and at intermediate levels in other virus-specific models. This pattern supports a tendency toward larger and more hydrophobic residues at certain positions of SARS-CoV-2 proteins [72]. In contrast, the threonine-serine substitution shows an opposite trend, being common in general empirical models but consistently reduced in SARS-CoV-2-specific and other viral models, suggesting tighter constraints on substitutions among polar residues. Additional model-specific signatures, such as elevated valine-methionine substitution rates in SC2-PLpro and Q.Bird, further highlight functional or structural specialization. Together, these patterns indicate that empirical substitution models capture biologically meaningful differences in selective pressures rather than universal substitution tendencies.

Besides, we evaluated the performance of the SC2-Mpro and SC2-PLpro models by examining the folding stability of simulated proteins. Overall, sequences evolved under these SARS-CoV-2-specific models showed folding stabilities closer to those observed in real proteins than did sequences simulated under other empirical models or even under an SCS model (Fig. 4). The latter finding was partially unexpected, as SCS models have often outperformed empirical models in similar contexts [28, 29, 41, 54, 73–75], and empirical models have frequently failed to generate proteins with realistic stability profiles [28, 76, 77]. We suspect that the strong performance of the substitution models for SARS-CoV-2 proteins likely reflects the importance of using empirical models based on protein-specific data, which implicitly capture evolutionary constraints while allowing their straightforward integration into probabilistic methods for evolutionary inference.

The empirical substitution models of SARS-CoV-2 protein evolution developed in this study have been incorporated into the *EModelDB* empirical model database. Our results indicate that these models capture protein-specific evolutionary patterns in SARS-CoV-2 and can improve the accuracy of phylogenetic inferences, enabling more reliable predictions of past and future evolutionary trajectories. Accurate substitution models are generally needed to model protein evolution across diverse organisms, and although inferring them requires large data, the continuing expansion of sequence databases and advances in computational methods are helping to overcome this limitation. Altogether, the SARS-CoV-2 substitution models presented here can enhance evolutionary inferences for this virus and motivate the development of additional protein-specific substitution models. We also highlight that widely used phylogenetic frameworks should expand their set of implemented substitution models [26] to facilitate the use of appropriate substitution models for specific datasets.

Declarations

Data Availability

The estimated substitution models for SARS-CoV-2 proteins, the method and test datasets, and the data simulated under different models, along with the estimates, are available from the Zenodo repository at the <https://doi.org/10.5281/zenodo.18763423>.

Acknowledgments

This study was supported by the Grant CNS2023-144363 funded by MICIU/AEI/10.13039/501100011033 and by European Union NextGenerationEU/PRTR. DF was contracted by the Grant CNS2023-144363 funded by MICIU/AEI/10.13039/501100011033 and by European Union NextGenerationEU/PRTR. LDGV was funded by a fellowship from Xunta de Galicia ED481A-2023/089, programa de axudas á etapa predoutoral da Xunta de Galicia (Consellería de Cultura, Educación, Formación Profesional e Universidades) cofinanciado pola Unión Europea no marco do Programa FSE+ Galicia 2021-2027. Funding for open access charge: Universidade de Vigo/CISUG. We thank the Centro de Supercomputación de Galicia (CESGA) for providing the computational resources.

Conflicts of Interest

The authors declare no conflicts of interest.

Author contributions

DF, LDGV and MA conceived and designed the study. LDGV prepared data. DF prepared data and performed the analyses. DF and MA interpreted the results. All the authors wrote the manuscript and approved its final version.

Supplementary Material

The supplementary material is available at the journal online.

References

1. WHO (2025) WHO Coronavirus (COVID-19) dashboard. In. <https://data.who.int/dashboards/covid19/cases>
2. Harvey WT, Carabelli AM, Jackson B et al (2021) SARS-CoV-2 variants, spike mutations and immune escape. *Nat Rev Microbiol* 19(7):409–424
3. Cao Y, Wang J, Jian F et al (2022) Omicron escapes the majority of existing SARS-CoV-2 neutralizing antibodies. *Nature* 602(7898):657–663
4. Carabelli AM, Peacock TP, Thorne LG et al (2023) SARS-CoV-2 variant biology: immune escape, transmission and fitness. *Nat Rev Microbiol* 21(3):162–177
5. Raharinarina NA, Gubela N, Bornigen D et al (2025) SARS-CoV-2 evolution on a dynamic immune landscape. *Nature* 639(8053):196–204

6. González-Vázquez LD, Arenas M (2023) Molecular Evolution of SARS-CoV-2 during the COVID-19 Pandemic. *Genes* 14(2):407
7. Markov PV, Ghafari M, Beer M et al (2023) The evolution of SARS-CoV-2. *Nat Rev Microbiol* 21(6):361–379
8. Meijers M, Ruchnewitz D, Eberhardt J et al (2023) Population immunity predicts evolutionary trajectories of SARS-CoV-2. *Cell* 186(23):5151–5164e5113
9. Neher RA, Russell CA, Shraiman BI (2014) Predicting evolution from the shape of genealogical trees. *Elife* 3:e03568
10. Lässig M, Mustonen V, Walczak AM (2017) Predicting evolution. *Nat Ecol Evol* 1(3):0077
11. Levinstein Hallak K, Rosset S (2022) Statistical modeling of SARS-CoV-2 substitution processes: predicting the next variant. *Commun Biol* 5(1):285
12. Wortel MT, Agashe D, Bailey SF et al (2023) Towards evolutionary predictions: Current promises and challenges. *Evol Appl* 16(1):3–21
13. Ferreira D, Gonzalez-Vazquez LD, Prado-Comesana A et al (2025) Forecasting protein evolution by integrating birth-death population models with structurally constrained substitution models. *Elife* 14:RP106365
14. Rouzine IM, Rozhnova G (2023) Evolutionary implications of SARS-CoV-2 vaccination for the future design of vaccination strategies. *Commun Med (Lond)* 3(1):86
15. Thorne JL (2000) Models of protein sequence evolution and their applications. *Curr Opin Genet Dev* 10:602–605
16. Yang Z (2006) *Computational Molecular Evolution*. Oxford University Press, Oxford, England.
17. Arenas M (2015) Trends in substitution models of molecular evolution. *Front Genet* 6:319
18. Ferreira D, Pazos E, Arenas M (2026) Trends in substitution models of protein evolution for phylogenetic inference. *Mol Phylogenet Evol* 214:108473
19. Arenas M (2012) Simulation of Molecular Data under Diverse Evolutionary Scenarios. *PLoS Comput Biol* 8(5):e1002495
20. Shapiro B, Rambaut A, Drummond AJ (2006) Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Mol Biol Evol* 23(1):7–9
21. Ayuso-Fernandez I, Molpeceres G, Camarero S et al (2022) Ancestral sequence reconstruction as a tool to study the evolution of wood decaying fungi. *Front Fungal Biol* 3:1003489
22. Kozlov AM, Darriba D, Flouri T et al (2019) RAXML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* 35(21):4453–4455
23. Darriba D, Posada D, Kozlov AM et al (2020) ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. *Mol Biol Evol* 37(1):291–294
24. Minh BQ, Schmidt HA, Chernomor O et al (2020) IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* 37(5):1530–1534

25. Tamura K, Stecher G, Kumar S (2021) MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Mol Biol Evol* 38(7):3022–3027
26. Iglesias-Rivas P, Del Amparo R, Cabaleiro JA et al (2025) Empirical substitution models of protein evolution: database, relationships, and modeling considerations. *Database* 2025:baaf052
27. Keane TM, Creevey CJ, Pentony MM et al (2006) Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evol Biol* 6:29
28. Arenas M, Bastolla U (2020) ProtASR2: Ancestral reconstruction of protein sequences accounting for folding stability. *Methods Ecol Evol* 11(2):248–257
29. Del Amparo R, Arenas M (2022) Consequences of Substitution Model Selection on Protein Ancestral Sequence Reconstruction. *Mol Biol Evol* 39(7):msac144
30. Del Amparo R, Arenas M (2023) Influence of substitution model selection on protein phylogenetic tree reconstruction. *Gene* 865:147336
31. Gonzalez-Vazquez LD, Iglesias-Rivas P, Ferreiro D et al (2025) Heterogeneous Evolution Among SARS-CoV-2 Genes and Variants of Concern. *J Med Virol* 97(9):e70604
32. Chung C, Irudayaraj P, Lallow E et al (2024) An overview of SARS-CoV-2 viral proteins with relevance to improved diagnostic and therapeutic platforms. *Front Virol* Volume 4–2024
33. Gayvert K, McKay S, Lim WK et al (2023) Evolutionary trajectory of SARS-CoV-2 genome shifts during widespread vaccination and emergence of Omicron variant. *Npj Viruses* 1(1):5
34. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 30(4):772–780
35. Minh BQ, Dang CC, Vinh LS et al (2021) QMaker: Fast and Accurate Method to Estimate Empirical Models of Protein Evolution. *Syst Biol* 70(5):1046–1060
36. Kalyaanamoorthy S, Minh BQ, Wong TKF et al (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* 14(6):587–589
37. Schwarz G (1978) Estimating the dimension of a model. *Ann Stat* 6(2):461–464
38. Luo A, Qiao H, Zhang Y et al (2010) Performance of criteria for selecting evolutionary models in phylogenetics: a comprehensive study based on simulated datasets. *BMC Evol Biol* 10:242
39. Akaike H (1981) Likelihood of a model and information criteria. *J Econometrics* 16:3–14
40. Arenas M, Dos Santos HG, Posada D et al (2013) Protein evolution along phylogenetic histories under structurally constrained substitution models. *Bioinformatics* 29(23):3020–3028
41. Arenas M, Sanchez-Cobos A, Bastolla U (2015) Maximum likelihood phylogenetic inference with selection on protein folding stability. *Mol Biol Evol* 32(8):2195–2207
42. Minning J, Porto M, Bastolla U (2013) Detecting selection for negative design in proteins through an improved model of the misfolded state. *Proteins* 81(7):1102–1112
43. Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* 8(3):275–282

44. Nickle DC, Heath L, Jensen MA et al (2007) HIV-specific probabilistic models of protein evolution. *PLoS ONE* 2(6):e503
45. Dang CC, Le QS, Gascuel O et al (2010) FLU, an amino acid substitution model for influenza proteins. *BMC Evol Biol* 10:99
46. Le TK, Vinh LS (2020) FLAVI: An amino acid substitution model for flaviviruses. *J Mol Evol* 88:445–452
47. Muller T, Vingron M (2000) Modeling amino acid replacement. *J Comput Biol* 7(6):761–776
48. Veerassamy S, Smith A, Tillier ER (2003) A transition probability model for amino acid substitutions from blocks. *J Comput Biol* 10(6):997–1010
49. Minin V, Abdo Z, Joyce P et al (2003) Performance-based selection of likelihood models for phylogeny estimation. *Syst Biol* 52(5):674–683
50. Lemmon AR, Moriarty EC (2004) The importance of proper model assumption in bayesian phylogenetics. *Syst Biol* 53(2):265–277
51. Duchene S, Di Giallonardo F, Holmes EC (2016) Substitution Model Adequacy and Assessing the Reliability of Estimates of Virus Evolutionary Rates and Time Scales. *Mol Biol Evol* 33(1):255–267
52. Jack BR, Meyer AG, Echave J et al (2016) Functional Sites Induce Long-Range Evolutionary Constraints in Enzymes. *PLoS Biol* 14(5):e1002452
53. Ribeiro AJM, Tyzack JD, Borkakoti N et al (2020) A global analysis of function and conservation of catalytic residues in enzymes. *J Biol Chem* 295(2):314–324
54. Ferreira D, Khalil R, Sousa SF et al (2024) Substitution Models of Protein Evolution with Selection on Enzymatic Activity. *Mol Biol Evol* 41(2):msae026
55. Abriata LA, Palzkill T, Dal Peraro M (2015) How structural and physicochemical determinants shape sequence constraints in a functional enzyme. *PLoS ONE* 10(2):e0118684
56. Ladunga I, Smith RF (1997) Amino acid substitutions preserve protein folding by conserving steric and hydrophobicity properties. *Protein Eng* 10(3):187–196
57. Fellouse FA, Barthelemy PA, Kelley RF et al (2006) Tyrosine plays a dominant functional role in the paratope of a synthetic antibody derived from a four amino acid code. *J Mol Biol* 357(1):100–114
58. Koide S, Sidhu SS (2009) The importance of being tyrosine: lessons in molecular recognition from minimalist synthetic binding proteins. *ACS Chem Biol* 4(5):325–334
59. Akiba H, Tsumoto K (2015) Thermodynamics of antibody-antigen interaction revealed by mutation analysis of antibody variable regions. *J Biochem* 158(1):1–13
60. Nakayama T, Mizohata E, Yamashita T et al (2015) Structural features of interfacial tyrosine residue in ROBO1 fibronectin domain-antibody complex: Crystallographic, thermodynamic, and molecular dynamic analyses. *Protein Sci* 24(3):328–340
61. Peng HP, Lee KH, Jian JW et al (2014) Origins of specificity and affinity in antibody-protein interactions. *Proc Natl Acad Sci U S A* 111(26):E2656–2665

62. Mejias-Gomez O, Madsen AV, Skovgaard K et al (2023) A window into the human immune system: comprehensive characterization of the complexity of antibody complementary-determining regions in functional antibodies. *MAbs* 15(1):2268255
63. Qu L, Qiao X, Qi F et al (2021) Analysis of Binding Modes of Antigen-Antibody Complexes by Molecular Mechanics Calculation. *J Chem Inf Model* 61(5):2396–2406
64. Murtaugh ML, Fanning SW, Sharma TM et al (2011) A combinatorial histidine scanning library approach to engineer highly pH-dependent protein switches. *Protein Sci* 20(9):1619–1631
65. Schroter C, Gunther R, Rhiel L et al (2015) A generic approach to engineer antibody pH-switches using combinatorial histidine scanning libraries and yeast display. *MAbs* 7(1):138–151
66. Sokalingam S, Raghunathan G, Soundrarajan N et al (2012) A study on the effect of surface lysine to arginine mutagenesis on protein stability and structure using green fluorescent protein. *PLoS ONE* 7(7):e40410
67. Clarke RJ (2023) Electrostatic switch mechanisms of membrane protein trafficking and regulation. *Biophys Rev* 15(6):1967–1985
68. Li L, Vorobyov I, Allen TW (2013) The different interactions of lysine and arginine side chains with lipid membranes. *J Phys Chem B* 117(40):11906–11920
69. Rice A, Wereszczynski J (2017) Probing the disparate effects of arginine and lysine residues on antimicrobial peptide/bilayer association. *Biochim Biophys Acta Biomembr* 1859(10):1941–1950
70. Long Y, Meng F, Kondo N et al (2011) Conserved arginine residue in the membrane-spanning domain of HIV-1 gp41 is required for efficient membrane fusion. *Protein Cell* 2(5):369–376
71. Wu G (2009) Amino acids: metabolism, functions, and nutrition. *Amino Acids* 37(1):1–17
72. Cross TJ, Takahashi GR, Diessner EM et al (2020) Sequence Characterization and Molecular Modeling of Clinically Relevant Variants of the SARS-CoV-2 Main Protease. *Biochemistry* 59(39):3741–3756
73. Ferreira D, Branco C, Arenas M (2024) Selection among site-dependent structurally constrained substitution models of protein evolution by approximate Bayesian computation. *Bioinformatics* 40(3):btae096
74. Lorca-Alonso I, Otero-de-Navascues F, Arenas M et al (2026) Site-Specific Structure and Stability Constrained Substitution Models Improve Phylogenetic Inference. *Syst Biol* 75(3):581–604
75. Ferreira D, Khalil R, Gallego MJ et al (2022) The evolution of the HIV-1 protease folding stability. *Virus Evol* 8(2):veac115
76. Arenas M, Weber CC, Liberles DA et al (2017) ProtASR: An Evolutionary Framework for Ancestral Protein Reconstruction with Selection on Folding Stability. *Syst Biol* 66(6):1054–1064
77. Bordner AJ, Mittelman HD (2014) A new formulation of protein evolutionary models that account for structural constraints. *Mol Biol Evol* 31(3):736–749

Figures

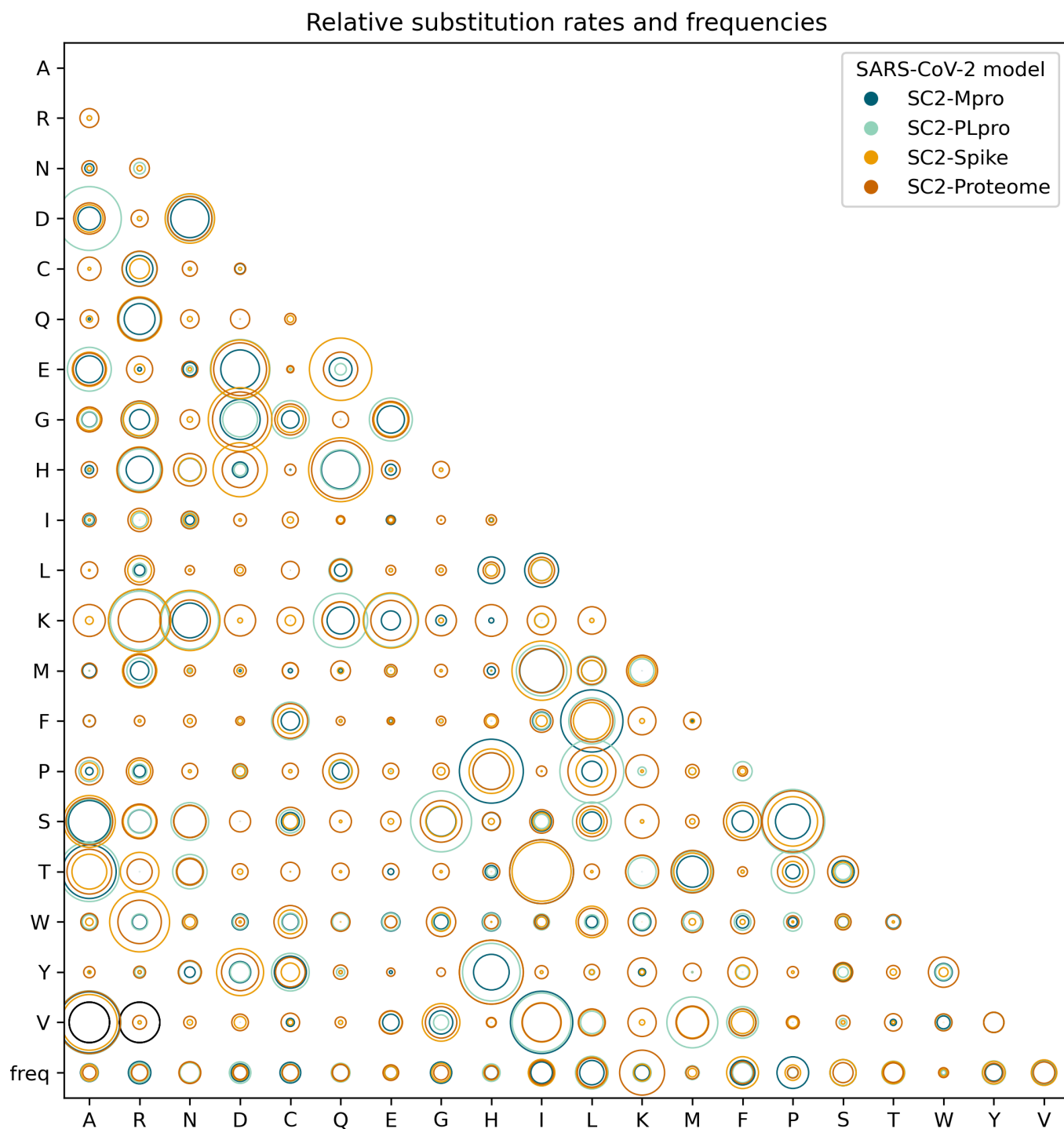


Figure 1

Relative substitution rates among amino acids and amino acid frequencies at the equilibrium in the estimated SARS-CoV-2 substitution models. Relative amino acid substitution rates are shown above, and amino acid frequencies at the equilibrium are shown below. Estimates for each SARS-CoV-2 substitution model are displayed in colors as indicated in the figure legend.

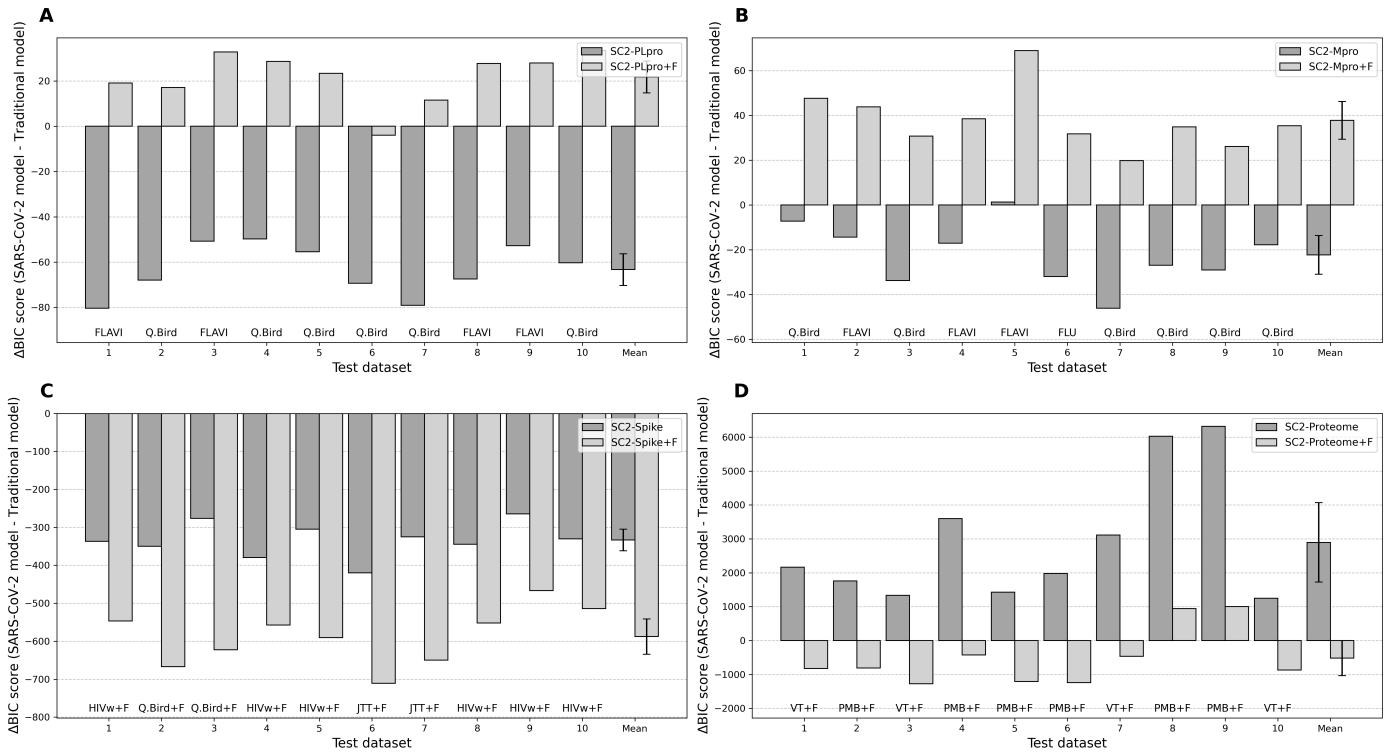


Figure 2

Evaluation of the SARS-CoV-2 substitution models through phylogenetic likelihood in comparison with standard empirical substitution models. The panel shows the difference in Bayesian Information Criterion scores (ΔBIC) between each SARS-CoV-2 substitution model and the best-fitting standard empirical model for each test dataset. Negative ΔBIC scores indicate that the SARS-CoV-2 substitution model provides a better fit to the corresponding test dataset than the standard empirical model. Each plot corresponds to a specific SARS-CoV-2 protein or the proteome. For each protein, we evaluated two SARS-CoV-2 substitution models: the base model, which includes the estimated amino acid frequencies at the equilibrium (dark gray), and a model using the amino acid frequencies observed in the test dataset (+F, light gray). Results are shown for 10 independent test datasets (1-10), and the best-fitting standard model used as the baseline is indicated below. The last bar provides the mean ΔBIC across all test datasets, with error bars representing the 95% confidence interval.

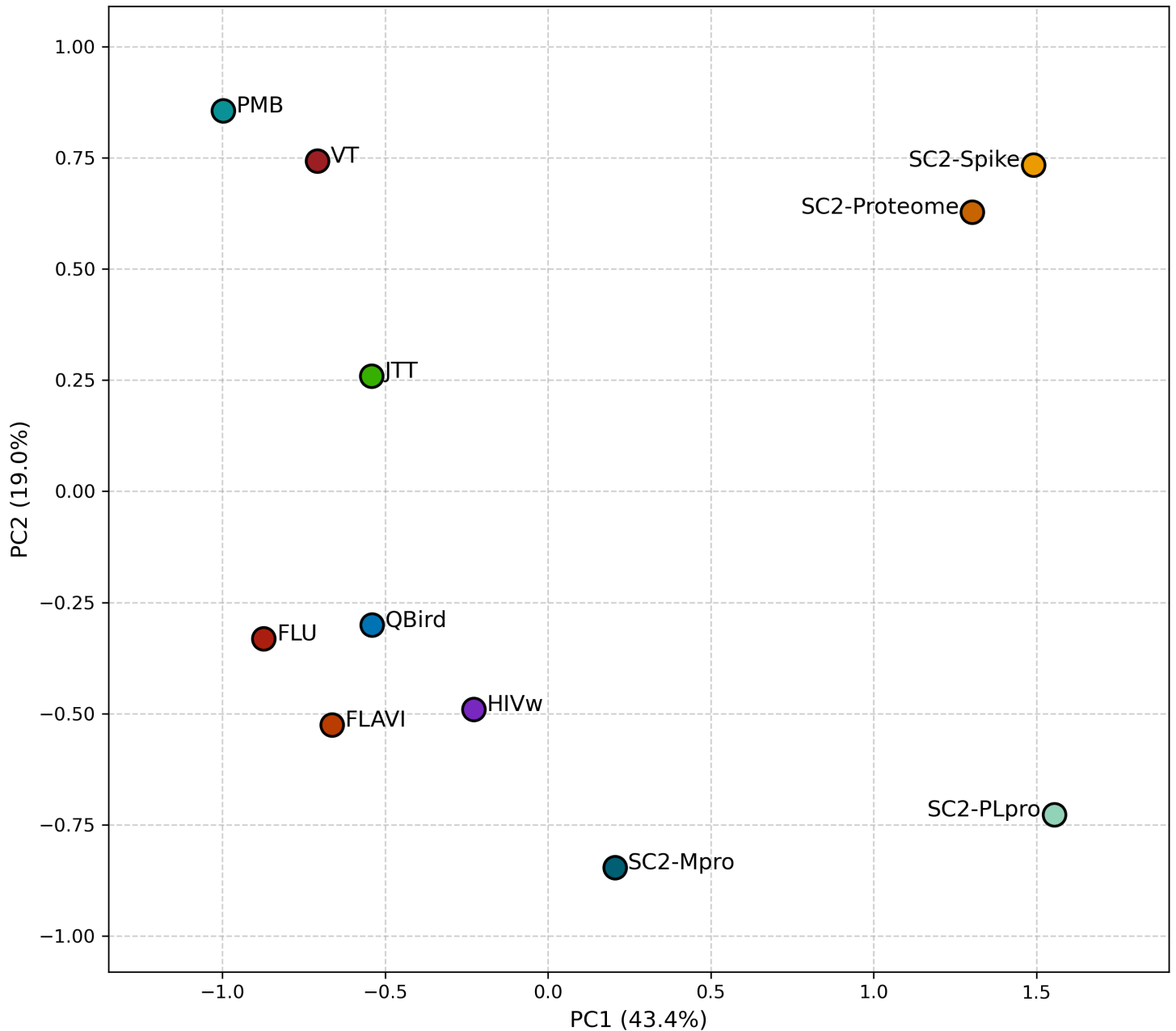


Figure 3

Principal component analysis of SARS-CoV-2 and standard empirical substitution models. The principal component analysis shows the positions of the SARS-CoV-2 substitution models relative to standard empirical substitution models that provided high phylogenetic likelihoods with the test data. Each point represents a substitution model. Distances between substitution models reflect dissimilarities among their substitution matrices and revealed several clusters (e.g., PMB and VT models, or SC2-Spike and SC2-Proteome models). The first principal component (PC1) explains 43.4% of the variance (horizontal axis), and the second principal component (PC2) explains 19.0% (vertical axis).

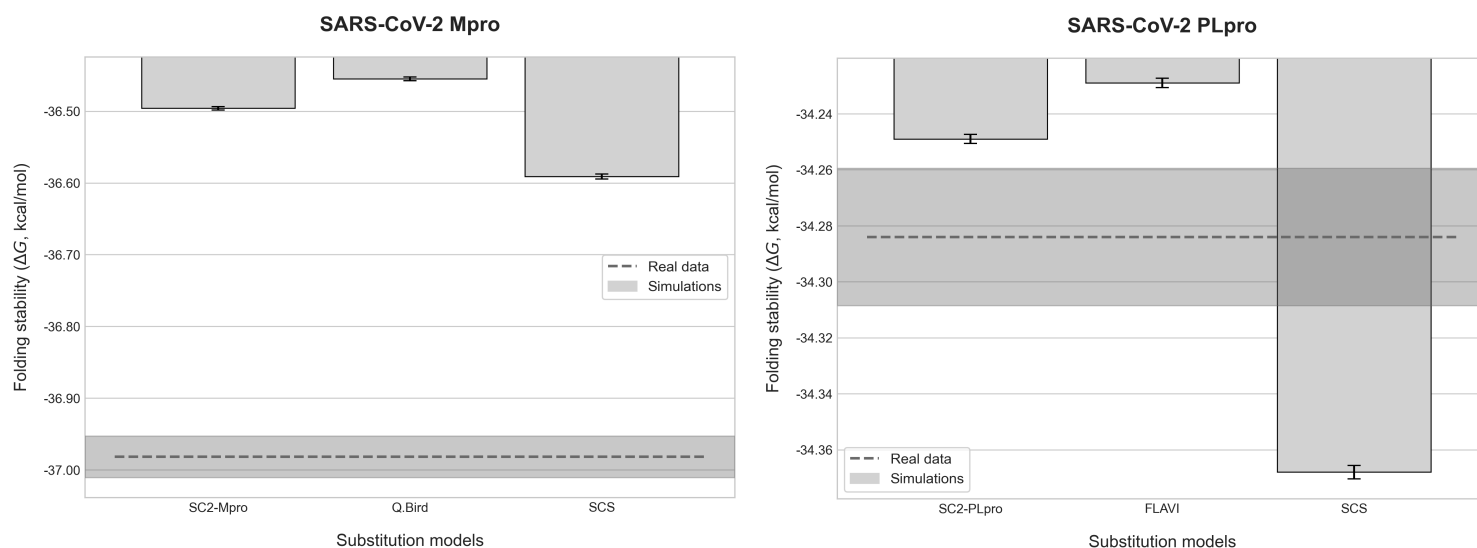


Figure 4

Protein folding stability of real and simulated Mpro and PLpro sequences under diverse substitution models. Bars show the folding stability (ΔG , kcal/mol) for Mpro (left) and PLpro (right) sequences simulated under three different substitution models: the SARS-CoV-2-specific model, the standard likelihood-based best-fitting empirical model, and the structurally constrained substitution (SCS) model. Gray bars represent the mean ΔG of sequences from 100 simulated multiple sequence alignments, with error bars indicating the 95% confidence interval. The dark gray dashed horizontal line shows the mean ΔG of the real protein variants, with the surrounding gray shaded band indicating the 95% confidence interval.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryMaterial.pdf](#)