PCCP

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d5cp04811g

Deciphering the temporal and spatial mutation dynamics of the SARS-CoV-2 spike glycoprotein

Muhammad Hasan,†[a] Shihong Chen, †[a] Mengqi Jia, †[a] Cheuk Fung Alvin Leung,†[a] Wentao Xu, [a] Ka Pui Chang, [a] Chi Ming Kan,[a] Shanqi Yap,[a] Yuan Bing, [b] Kaicheng Zhu, *[a] Xiakun Chu*[c] and Haibin Su *[ade]

Received 11th December 2025,
Accepted 22nd May 2026

DOI: 10.1039/d5cp04811g

rsc.li/pccp

We present a statistical pipeline with two parallel procedures to analyze SARS-CoV-2 spike evolution: (1) *probability sequence density* analysis for probing its sequence space, and (2) *leading mutations by the composite metric*. This metric integrates mutation eigenvector information with pairwise couplings to outline evolutionarily significant mutations, coined *leading mutations*, from massive data sets. Our results reveal progressive increase in sequence mutation rates over time, alongside scaling behaviors predictive of variant emergence and evolutionary trends in spike mutation patterns. These findings characterize the mechanisms by which the spike glycoprotein acquires new mutations, offering insights into its evolutionary dynamics.

1 Introduction

Since the initial emergence of SARS-CoV-2, countless mutational events have occurred to shape its evolutionary trajectory.[1] Central to the evolution of SARS-CoV-2 is its spike glycoprotein. Being the primary antigen and fusion protein of the virus,[2,3] strong selective pressure exerted on the spike renders it one of the most mutationally active regions in the SARS-CoV-2 genome.[4,5] To track these rapid changes, global collaborative efforts have leveraged databases like the Global Initiative on Sharing All Influenza Data (GISAID) to curate vast repositories of spike sequence data.[6] These resources not only enable real-time epidemiological surveillance,[7] but also serve as records that offer critical insights into the dynamics of SARS-CoV-2 spike evolution.

Studies of organismic or viral mutations often focus on their immediate functional effects. It is important to elucidate the phenotypic changes due to genotypic alterations, but this superficial level of investigation is insufficient to unravel the mechanisms underpinning adaptive evolution—this is where studies of mutation rates come into play.[8,9] While maintaining high mutation rates increases the likelihood for viruses to accrue beneficial mutations,[10] it may reduce their overall fitness because most mutations are either neutral or slightly deleterious.[11–13] This dichotomy drives the evolution of mutation rates in both cellular and acellular life,[14,15] and greatly influences the rate at which adaption occurs. Arising naturally from this delicate balance, previous studies have reported a correlation between mutation rates and population sizes across various organisms.[8] Extending this notion to viral systems such as the SARS-CoV-2 spike glycoprotein can help uncover the determinants of their evolutionary trajectories.

Another collective feature determined by the emergence of multiple mutations is the statistics of their distributions. Many-body systems, in which many interacting components give rise to emergent behaviors, are omnipresent across the sciences. Viral evolution exemplifies such a system, with carriers of different mutations not acting in isolation but in interweaved manners to drive the population's evolution akin to chemical reaction networks.[15–17] Time-resolved analysis of the mutation distributions in viruses allows our understanding of their evolutionary dynamics to be deepened. By incorporating phylogenetic methods, variant clusters can be identified from large-scale sequence datasets, through which the mutation distance between a given progeny and its nearest ancestor can be established.[18,19] This approach enables the characterization of scaling behaviors in viral evolution—relationships that quantitatively capture the principles governing the organization and dynamics of complex biological systems.[20–22]

While a thorough understanding of both the mutation rate and distribution of SARS-CoV-2 paints a global picture of its evolution, it remains essential to study the details of its amino

[a] *Department of Chemistry, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China. E-mail: kczhu@ust.hk*
[b] *Songshan Lake Materials Laboratory, Dongguan City, Guangdong, China*
[c] *Advanced Materials Thrust, Function Hub, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China. E-mail: xiakunchu@hkust-gz.edu.cn*
[d] *IAS Center for AI for Scientific Discoveries, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China*
[e] *AI-X Open Lab., HKUST Shenzhen Hong Kong Collaborative Innovation Research Institute, China. E-mail: haibinsu@ust.hk*
† These authors contributed equally to this work.

acid mutations. We previously proposed a method to score the evolutionary significance of mutable sites based on their individual mutation strengths, such that a representative set of key mutations can be identified from endless combinations.[23] However, higher-order structures exist within complex networks:[24] as the spike continues to acquire mutations in large numbers, accounting for their epistatic interactions has become necessary to clarify its evolution.[25–27] A natural way to systematically organize such high-dimensional data is to pack it into multi-dimensional arrays known as tensors; decomposing these tensors into simpler orthogonal factor matrices often makes them more interpretable.[28–31] With their marked usefulness in analyzing physical, chemical, and biological data, we believe tensor decomposition serves as the key to extract the leading factors of SARS-CoV-2 evolution.

The need to monitor the interactions between SARS-CoV-2 mutations in a higher order perspective is reinforced by the spike glycoprotein's intricate multi-constrainted biophysical landscape.[27,32,33] The spike is a class I fusion homotrimer whose receptor-binding domain (RBD) exists in a metastable pre-fusion state that transitions dynamically between a receptor-accessible up state and a receptor-inaccessible down state, creating a fundamental tension between human angiotensin-converting enzyme 2 (hACE2) binding affinity and thermodynamic stability.[2,32,34] Deep mutational scanning has shown that many mutations that enhance hACE2 affinity simultaneously destabilize the protein, forcing the virus to navigate a narrow fitness ridge.[32,35] In variants such as B.1.617.2 (Delta, δ) and Omicron (o), experimental structural dynamics studies have revealed that mutations directly alter RBD motion and equilibrium distribution between conformational states,[36] while molecular dynamics simulations have further shown that these effects are modulated by synergistic interactions with surface glycans.[37] Consequently, variant fitness is not a simple sum of individual mutational effects but an emergent property of the spike's global physicochemical profile, where epistatic interactions stabilize otherwise deleterious but functionally advantageous mutations.[27]

Here, we introduce a bifurcated statistical pipeline for deciphering the temporal mutation pattern of the SARS-CoV-2 spike glycoprotein (Fig. 1). To begin with, our *probability sequence density* (PSD) $P(\boldsymbol{d})$ analysis characterizes the scaling behaviors in viral evolution. In this process, $\log P(\boldsymbol{d}) - \log \boldsymbol{d}$ slope of each monthly data set is computed, and its time gradient can serve as an early warning signal for variant emergence. On the other hand, our *leading mutations by the composite metric* ($\mathrm{LM^{CM}}$) method integrates sequence, site, and amino acid data from Tucker decomposition of a mutation tensor, and pair correlation information of mutations, to outline a set of evolutionarily important mutations, coined *leading mutations*, every month. Each set of $\mathrm{LM^{CM}}$-derived leading mutations is publicly accessible on our online platform at https://hbsulab.github.io/deLemus/. From these results, we present a comprehensive analysis of the evolutionary dynamics of the SARS-CoV-2 spike glycoprotein by studying temporal heterogeneities within its mutation patterns, shedding light on key mutation trends of the spike glycoprotein.

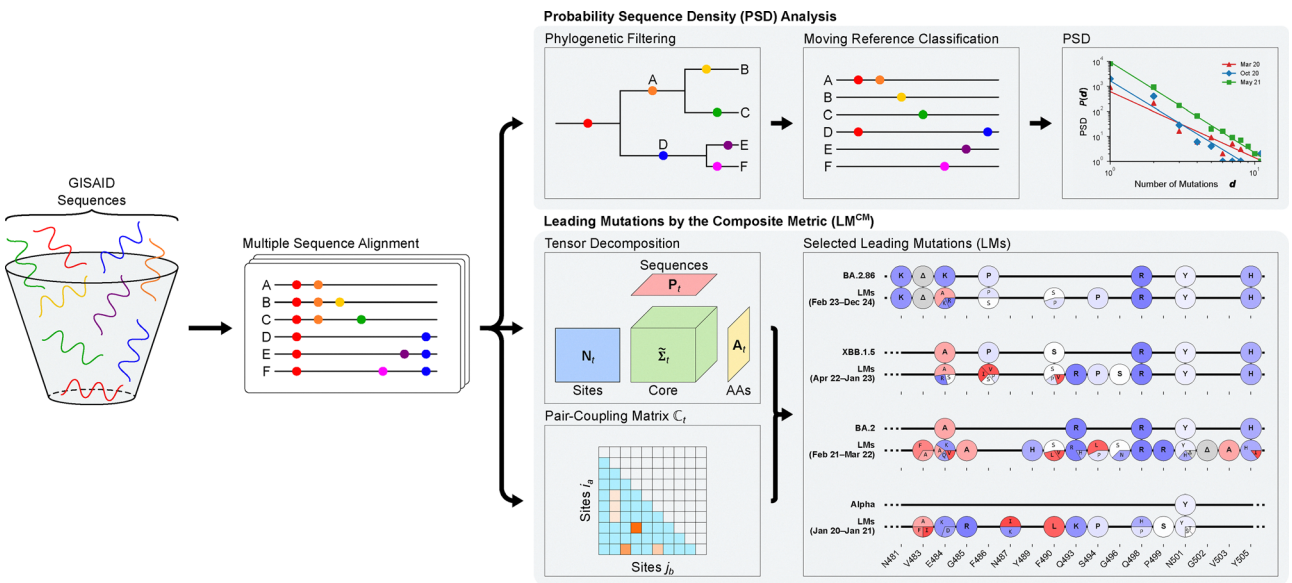

Fig. 1 A statistical pipeline for deciphering the temporal mutation pattern of the SARS-CoV-2 spike glycoprotein. First, spike glycoprotein amino acid sequences downloaded from the Global Initiative on Sharing All Influenza Data (GISAID) hCoV-19 database are filtered and aligned to generate sets of non-degenerate sequences grouped by month. Then, these data sets are processed in two parallel steps. For the probability sequence density (PSD) $P(\boldsymbol{d})$ analysis, the $\log P(\boldsymbol{d}) - \log \boldsymbol{d}$ slope of each monthly data set is computed, and its time gradient can serve as an early warning signal for the emergence of new variants. On the other hand, our leading mutations by the composite metric ($\mathrm{LM^{CM}}$) method integrates sequence, site, and amino acid (AA) data from Tucker decomposition of a mutation tensor, and pair correlation information of mutations, to outline a set of evolutionarily important mutations every month. The diagram of selected leading mutations (LMs) displays a subset of top LMs outlined across four intervals, juxtaposed with the reported mutations carried by representative variants that emerged by the end of each period. Circles or wedges colored blue correspond to mutations that increase hydrophilicity, whereas those colored red indicate the opposite effect. The complete set of RBD LMs outlined each month can be found in Fig. S6.

2 Numerical methods

To elucidate the emergent evolutionary dynamics of the SARS-CoV-2 spike glycoprotein, we implemented a bifurcated statistical pipeline that integrates global scaling analysis with higher-order mutation interaction modeling (Fig. 1). Inspired by low-rank tensor decomposition techniques that revolutionized modern condensed matter physics,[29,38,39] we treated viral populations as many-body systems in which mutations interact in a network-like manner. We first applied *probability sequence density* (PSD) analysis, $P(\boldsymbol{d})$, to characterize their scaling behaviors and monitor temporal changes in the $\log P(\boldsymbol{d}) - \log \boldsymbol{d}$ slope indicative of evolutionary regime shifts. Complementarily, to account for epistatic interactions among rapidly accumulating mutations, we constructed monthly mutation tensors from multiple sequence alignments and performed Tucker decomposition to extract dominant latent features. Using our *leading mutations by the composite metric* ($\mathrm{LM^{CM}}$) method, these tensor-derived mutation strengths were then integrated with pair-coupling correlations to identify *leading mutations* on a monthly basis.

To track the evolutionary trajectory of the SARS-CoV-2 spike glycoprotein, we compiled a longitudinal dataset spanning four years of the pandemic. By December 2023, more than 15 million amino acid sequences were retrieved from the GISAID hCoV-19 database.[6] We employed the EPI_ISL_402124 sequence as the baseline reference for alignment.[40] To isolate distinct evolutionary signals, minimize computational redundancy, and lessen sampling biases,[41,42] we implemented a rigorous filtering protocol to remove identical sequences. This resulted in a refined dataset of 667 213 unique, distinguishable sequences for analysis. Sequences were further partitioned into temporal cohorts based on their submission month, to which multiple sequence alignment was applied using Clustal Omega,[43] and phylogenetic lineages were assigned using the Nextstrain framework.[19] To distinguish novel evolutionary events from baseline noise, we identified substitutions and deletions by comparing each sequence to its most recent common ancestor.

We characterized the evolutionary dynamics using two primary quantitative measures of mutation rate. The first measure is the sequence-level mutation rate Ξ in the units of āā $\mathrm{seq^{-1}\ mo^{-1}}$, calculated as the mean number of novel mutations per sequence for each variant in a given month. Novel mutations were identified through a comparative analysis with the genetically closest ancestor from preceding months. To ensure statistical rigor, we calculated the corresponding 95% confidence intervals for each major variant utilizing the Student's *t*-distribution to account for sample variance. The second measure is the site mutation rate in the units of āā $\mathrm{site^{-1}\ mo^{-1}}$, which evaluates selective pressure at the residue level. For each variant, this was determined by calculating the mean novel mutation count per site in the variant's final month of observation, subtracting it from the variant's first month of emergence, and normalizing it by the duration of the variant's circulation. This site-specific rate was subsequently compared against the total number of unique sequences to assess the relationship between population-level reporting and residue-level divergence.

Furthermore, we investigated the global distribution of mutations through the PSD analysis, $P(\boldsymbol{d})$, where $\boldsymbol{d}$ represents the number of novel mutated sites across each sequence. By examining the relationship between $P(\boldsymbol{d})$ and $\boldsymbol{d}$ in a log–log coordinate system, we identified a persistent power-law scaling relationship (Fig. S1). The slope of this scaling—derived from the $\log P(\boldsymbol{d}) - \log \boldsymbol{d}$ relationship—was monitored on a monthly basis to detect shifts in the virus's evolutionary regime. Details regarding the statistical implementation and the goodness-of-fit assessments of the PSD analysis are provided in Section I of the SI.

To capture the complex interdependency between mutation sites and amino acid types, we represented the multiple sequence alignment data as a high-dimensional mutation tensor. For each month t, we constructed an $m \times l \times a_0$ tensor $\tilde{\mathbf{H}}_t = (H_t^{ijk})$, where m is the number of non-degenerate sequences, l is the sequence length fixed at 1273 sites, and a_0 is the number of amino acid types including deletions. An entry $H_t^{ijk} = 1$ signifies the presence of mutation of type k at site j in sequence i. We factorized this high-dimensional structure by Tucker decomposition, defined algebraically as

$$\tilde{\mathbf{H}} = \mathbf{P}_t \otimes \tilde{\Sigma}_t \otimes \mathbf{N}_t \otimes \mathbf{A}_t, \tag{1}$$

to extract its latent features, where $\mathbf{P}_t$, $\mathbf{N}_t$, and $\mathbf{A}_t$ represent the characteristic matrices for sequences, sites, and amino acid types, respectively. The core tensor $\tilde{\Sigma}_t$ contains the weight coefficients, or eigenvalues, for these interactions. From these components, we constructed a mutation strength matrix $\mathbf{M}_t$ whose entry identifies the amino acid substitution of a specific site that exerts the strongest influence on sequence divergence.

Evolutionary success is often driven by the cooperation of mutations rather than isolated ones.[26] To quantify these dependencies, we constructed a pair-coupling matrix $\mathbf{C}_t = (C_t^{ij})$ based on the correlation of mutation frequencies:

$$C_t^{ij} = \left[\sum_{a,b} \left(f_t^{i_a j_b} - f_t^{i_a} f_t^{j_b} \right) \right]^{1/2}, \tag{2}$$

where $f_t^{i_a j_b}$ denotes the joint frequency of mutation types a and b at sites i and j, and $f_t^{i_a}$, $f_t^{j_b}$ denote their independent frequencies respectively. Through singular value decomposition, we isolated the primary pair-coupling matrix $\mathbb{C}_t$ from $\mathbf{C}_t$, which represents the most significant signals of cooperations between mutation pairs.

Finally, we established the $\mathrm{LM^{CM}}$ framework which ranks mutations based on a composite score $L_t^{i_a}$ that integrates individual mutation strength $M_t^{i_a}$ with its network of couplings $\mathbb{C}_t^{i_a j_b}$:

$$L_t^{i_a} = M_t^{i_a} \sum_{j,b} \mathbb{C}_t^{i_a j_b} M_t^{j_b}. \tag{3}$$

Detailed operational construction of $\mathbf{M}_t$, the reduced pair-coupling matrix $\mathbb{C}_t$, and the sensitivity of the Tucker decomposition to rank choice are described in Section II of the SI (Fig. S2–S4). By identifying mutations with the highest $L_t^{i_a}$ scores, we can forecast *leading mutations* that are biophysically primed for selection in future variants. This dynamic ranking is updated monthly and served *via*

our open-access platform at https://hbsulab.github.io/deLemus/ to assist the global research community in early variant detection.

3 Results and discussion

3.1 Temporal dynamics of mutation rates

Mutations are the primary drivers of viral evolution. To understand how novel SARS-CoV-2 variants persistently emerge, it is necessary to decipher the underlying rate at which new mutations appear on its spike glycoprotein.[1] For instance, the drift-barrier hypothesis posits that genetic drift imposes an intrinsic lower bound on mutation rates that cannot be bypassed by selection, which can be verified by a positive correlation between base substitutional mutation rates and strengths of drift experienced by organisms.[8,14] To see if similar relationships can be observed within the diverse SARS-CoV-2 population, we analyzed the behavior of spike mutation rates across its sublineages.

We observed an increase in the sequence-level mutation rate Ξ of the spike glycoprotein over the course of the pandemic (Fig. 2). Early pre-Omicron variants demonstrated relatively low and stable mutation rates, with mean values for D614G (1.18 ± 0.05 āā seq^{-1} mo^{-1}) and Alpha (1.18 ± 0.09 āā seq^{-1} mo^{-1}) highlighting the limited sequence divergence at the onset of viral transmission.[45] Subsequent variants of the Omicron family exhibited elevated Ξ values, with mean rates ranging from 1.41 ± 0.26 āā seq^{-1} mo^{-1} for BA.4&5 to 1.55 ± 0.47 āā seq^{-1} mo^{-1} for BA.1, indicative of accelerated evolutionary dynamics and increased mutational variance.

Furthermore, we observed a robust positive correlation between the site-specific mutation rates and the genetic diversity of major SARS-CoV-2 variants. These include B.1.1.7 (Alpha, α), B.1.351 (Beta, β), P.1 (Gamma, γ), B.1.617.2 (Delta, δ), and Omicron (o) (Fig. 2 Inset). Direct frequency-based analyses are often susceptible to sampling biases, and certain sequences can be either heavily underrepresented or overrepresented due to discrepancies in sequencing intensities.[41,42] We therefore quantified genetic diversity using the count of unique sequences, which mitigates oversampling by tallying each distinct sequence only once.

Notably, highly transmissible variants like Delta and Omicron exhibited higher site mutation rates of 0.29 āā site^{-1} mo^{-1} and 0.47 āā site^{-1} mo^{-1} respectively, as well as greater within-variant genetic diversity.[46,47] Positive coupling between these traits is expected but mechanistically and causally nontrivial. One interpretation is that a high substitution rate helps maintain genetic diversity within a viral population, which promotes adaptation to optimize its fitness characteristics such as transmissibility.[48,49] On the other hand, more transmissible variants could lead to greater generational turnover and thus greater substitution rate and genetic diversity *via* rapid parallel infection of the host population.[50] In contrast, the geographically constrained Beta and Gamma variants displayed comparatively lower site mutation rates and genetic diversities.[51]

3.2 Evolution in the sequence space

Complementing the temporal aspect of viral evolution in terms of rate heterogeneity is the spatial information embedded in its sequence space. To uncover the patterns hidden within such a vast ensemble of sequences, we focused on the scaling behaviors of spike mutations. The intriguing scaling relationship between physical parameters often encapsulates the underlying complex organization processes, particularly in the field of evolutionary biology.[20,21,52,53] In this study, we defined the term *probability sequence density* (PSD) $P(\boldsymbol{d})$, which denotes a probability distribution over the integral displacement $\boldsymbol{d}$ of a given sequence from its nearest ancestor in the sequence space.

From the PSDs of the monthly non-degenerate sequence sets, we observed a linear scaling relationship between the quantities $P(\boldsymbol{d})$ and $\boldsymbol{d}$, which manifested itself across different time frames of the SARS-CoV-2 pandemic (Fig. 3a). This scaling was highly robust, and was confirmed by the systematic examination of the $\log P(\boldsymbol{d}) - \boldsymbol{d}$ plots for every month from January 2020 to December 2023 (Fig. S5). March 2020, October 2020, and May 2021 were selected as representative snapshots because they illustrate the temporal evolution of the scaling slope during the first major saltation event: transitioning from an early random-mutation regime through the emergence of

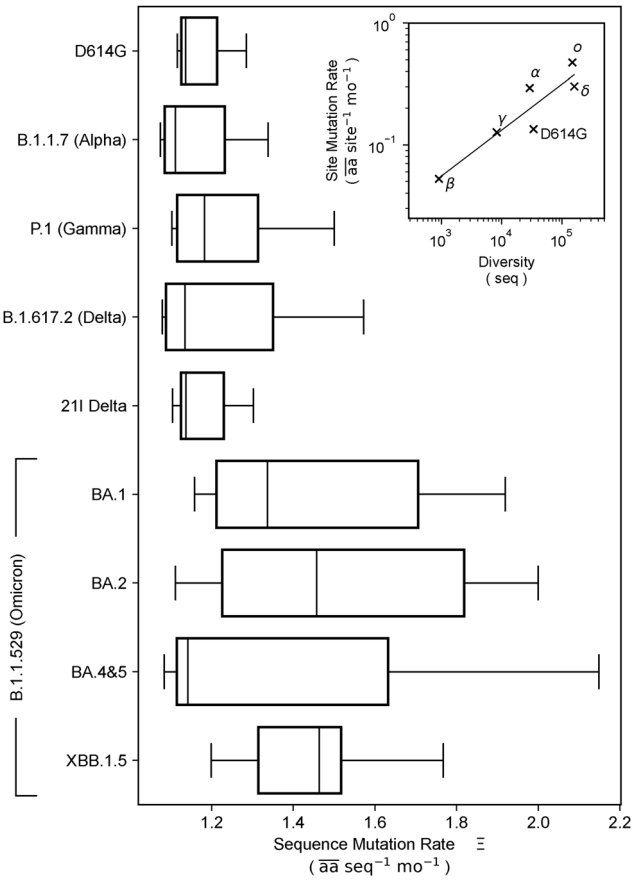

Fig. 2 Mutation rates Ξ of SARS-CoV-2 spike glycoproteins across major variants, expressed in units of āā seq^{-1} mo^{-1}. Variants are ordered top to bottom according to their designation date by the World Health Organization.[44] Inset: Correlation between site mutation rates and effective population size for variants of concern ($R^2 = 0.853$, $n = 6$), expressed in āā site^{-1} mo^{-1}. Other abbreviations: α – B.1.1.7 (Alpha), β – B.1.351 (Beta), γ – P.1 (Gamma), δ – B.1.617.2 (Delta), o – Omicron.

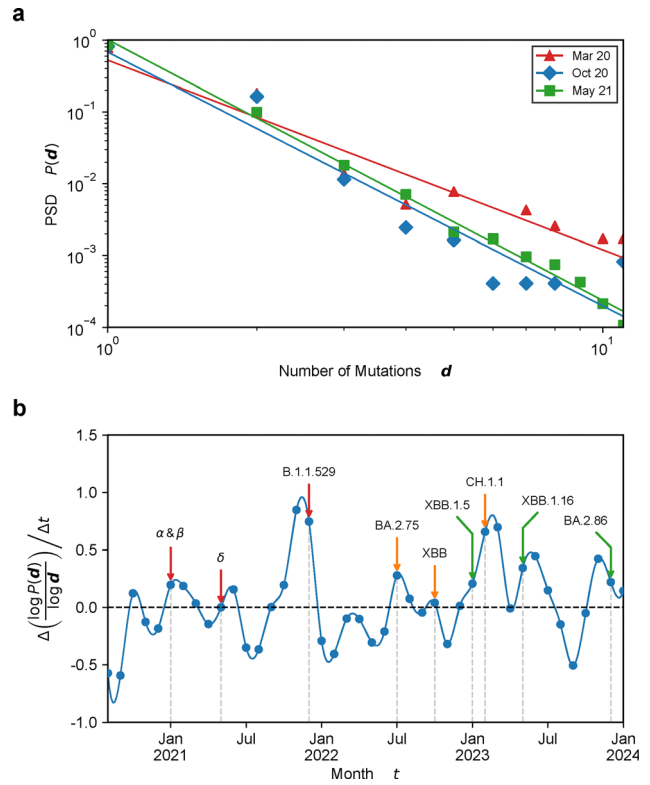

Fig. 3 (a) Scaling relationship between the probability sequence density (PSD) $P(\boldsymbol{d})$ and number of mutations d in selected months ($R_{Mar20}^2 = 0.903$, $n = 10$; $R_{Oct20}^2 = 0.899$, $n = 9$; $R_{May21}^2 = 0.992$, $n = 11$). (b) Evolution of the time gradient $\Delta(\log P(\boldsymbol{d})/\log \boldsymbol{d})/\Delta t$. Months wherein notable SARS-CoV-2 strains were designated by the World Health Organization as variants of concern, interest, or under monitoring are annotated with red, green, and orange arrows respectively.[44] Other abbreviations: α – B.1.1.7 (Alpha), β – B.1.351 (Beta), δ – B.1.617.2 (Delta).

Alpha and Beta variants and culminating in the rapid global dominance of Delta shortly thereafter.[54]

The effects of this scaling are twofold. First, the consistently negative slope of the $\log P(\boldsymbol{d}) - \boldsymbol{d}$ plot indicates that spike sequences with lower mutation loads prevailed over those with more mutations in each monthly ensemble. Specifically, most SARS-CoV-2 spike glycoproteins were seen to harbor around one to three mutations per month (Fig. 3a), corroborating previous estimates;[55] spike sequences surpassing this range are exceedingly rare. Second, such a scaling relationship can be described as a power law, from which network-like interactions between individual SARS-CoV-2 variants can be inferred.[20,52] Being a type of heavy-tailed distribution, a power law distribution creates a heightened likelihood of large-valued events than a normal distribution.[22,53,56] This suggests that the evolution of the SARS-CoV-2 spike glycoprotein is a process highly susceptible to abrupt and massive mutational events, likely coming from immunocompromised patients with prolonged infections.[57,58] The three major evolutionary leaps sustained by the virus, from Alpha to Delta, from Delta to BA.1, and from XBB to BA.2.86, serve as prime examples that demonstrate how a few extreme events could disproportionately impact the global SARS-CoV-2 population.[54,59,60] Studies have suggested that such a burst of mutations equips saltation variants with enhanced immune evasion capabilities,[60,61] or helps them navigate through entrapped regions of the fitness landscape epistatically,[54,58,62] which may have led to their global predominance.

We further examined the dynamics of the scaling relationships between $P(\boldsymbol{d})$ and $\boldsymbol{d}$, from which we discovered an intriguing link between the time gradient of the $\log P(\boldsymbol{d}) - \log \boldsymbol{d}$ slope, $\Delta(\log P(\boldsymbol{d})/\log \boldsymbol{d})/\Delta t$, and SARS-CoV-2 spike evolution. In general, $\log P(\boldsymbol{d})/\log \boldsymbol{d}$ can be viewed as a quantity describing the shape of the PSD, so its time derivative enables us to track time-dependent changes in the distribution of spike sequences based on their $\boldsymbol{d}$ values. For example, a reduction in this ratio corresponds to a shift toward a sequence ensemble dominated by low-$\boldsymbol{d}$ sequences, whereas an increment corresponds to the opposite. Under this interpretation, large positive peaks in a $\Delta(\log P(\boldsymbol{d})/\log \boldsymbol{d})/\Delta t$-time plot would represent a sudden influx of heavily mutated spike sequences within the virus population, which would signify the emergence of a new variant. In alignment with our hypothesis, while multiple peaks and troughs can be observed from Fig. 3b, the emergence of most major SARS-CoV-2 variants is accompanied by a positive peak in the time gradient. Therefore, this finding suggests the importance of PSD analysis in understanding the evolution of viruses, showing it can be used for the early detection of variant displacement events.

3.3 Evolution of leading mutations

To characterize how mutations shape the evolutionary trajectory of SARS-CoV-2 in greater detail, we analyzed the mutation patterns of each amino acid site across the spike glycoprotein. Given the large number of existing spike mutations, identifying those that exert the strongest influence on viral evolution is critical for forecasting future genomic shifts. Addressing this challenge, we developed the LM^{CM} method to effectively outline leading mutations in the spike glycoprotein, providing a predictive window into which residues are likely to drive upcoming variant emergence. These leading mutations are tracked and publicly accessible *via* our online platform at https://hbsulab.github.io/deLemus/. To exemplify the predictive power of our LM^{CM} method and the evolutionary significance of its outlined leading mutations, a representative selection of these located in the receptor-binding domain (RBD) of the spike glycoprotein is displayed in Fig. 4.

Several key RBD mutations in SARS-CoV-2 have emerged across different variants, significantly impacting hACE2 binding affinity; notably, the LM^{CM} method consistently identified these dangerous substitutions months before they achieved global dominance. The N501Y mutation, first identified as a leading mutation in June 2020 using the LM^{CM} method—well before the surge of Alpha—is present in the major variants Alpha, Beta, Gamma, and Omicron. While N501Y has been shown to enhance RBD stability and hACE2 binding from Alpha through Gamma *via* mechanisms like π–π stacking with the receptor's Y41 residue,[63–65] its advantageous effects are context-dependent and less pronounced in Omicron,[66] underscoring the importance of epistatic interactions between N501Y and its surrounding residues.[36] Similarly, the

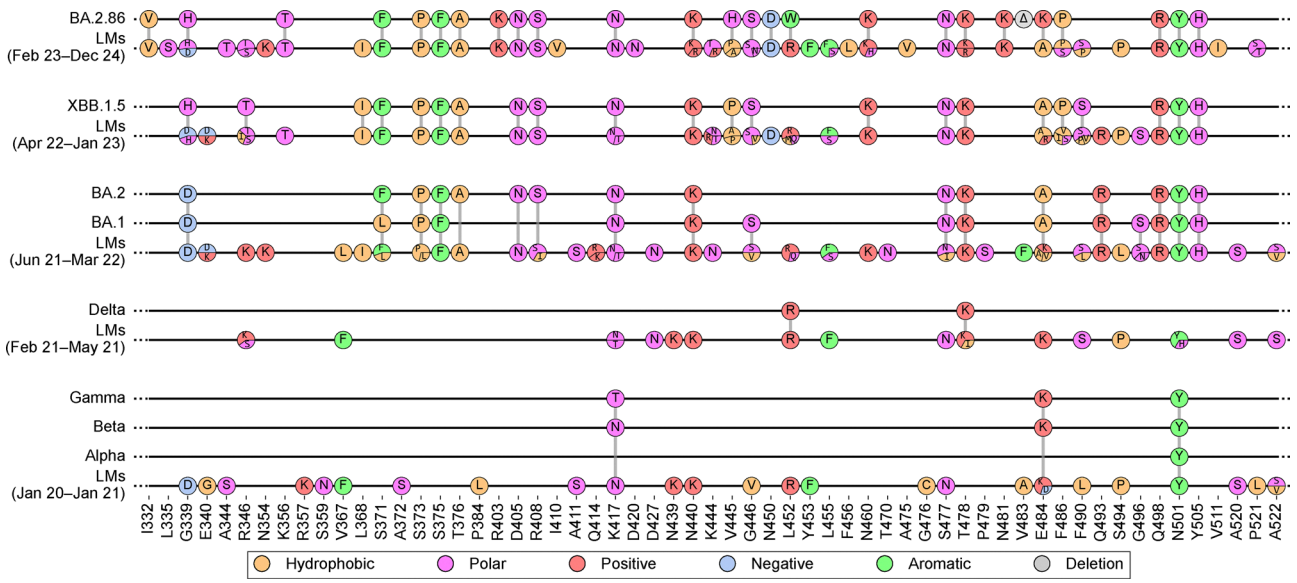

Fig. 4 Top-ranked LM^{CM}-outlined leading mutations (LMs) in the receptor-binding domain (RBD) of the spike. Variants designated as variants of concern or interest in the same month by the World Health Organization are grouped together.[44] Each pie chart represents the distribution of top-ranked LMs for a given month, where each colored sector corresponds to a specific mutation and its relative frequency within that month's top-ranked set. Vertical lines connect LMs with reported mutations if they coincide. The color scheme corresponds to the side-chain property of the residue being mutated to: Hydrophobic (orange; A, I, L, V, G, P, C, M), polar (magenta; H, N, Q, S, T), positive (red; K, R), negative (blue; D, E), aromatic (lime; F, W, Y), and deletion (gray; Δ). The complete set of RBD LMs can be found in Fig. S6 and S7.

E484K mutation, captured by our method as early as January 2020 and notably found in the Beta variant, promotes stronger hACE2 binding *via* local structural rearrangements when combined with N501Y and D614G.[67] The T478K mutation, highlighted by LM^{CM} in November 2020 as a future mutation of concern and later found characteristic of the Delta variant, facilitates tighter receptor binding by forming additional hydrogen bonds at the RBD-hACE2 interface.[68] Another mutation associated with Delta, L452R, was identified by our pipeline in April 2020, and contributes to increased hACE2 affinity through secondary structure rearrangements with the aforementioned E484K and N501Y.[67,69] Finally, the N440K mutation, outlined by LM^{CM} as an evolutionarily significant site, enhances electrostatic complementarity with hACE2, aiding its high transmissibility.[70]

The time-resolved spectra of the LM^{CM}-derived leading mutations bridge the gap between the global evolutionary trends in sequence ensembles and local functional alterations in the spike protein, allowing for the early detection of fitness-enhancing amino acid substitutions. One notable trend we observed was the accumulation of basic or polar amino acids in the RBD of the spike (Fig. S6). This convergence in physicochemical properties of substitutions was particularly pronounced within the receptor-binding motif (RBM) of the RBD (Fig. S7 and S8), an essential stretch of amino acids that engages the hACE2 receptor to mediate viral entry into host cells.[71] While the enrichment of positively charged or polar residues in the RBM tends to improve spike-receptor interactions due to the predominantly negatively charged surface of the hACE2 receptor,[70,72] the convergence ratio—defined here as the monthly fraction of newly appearing leading mutations among all RBM-based leading mutations—revealed a plateau in this trend over time (Fig. S8 and S9). This saturation, identified through our leading mutation analysis, may suggest a potential evolutionary trade-off between the charge optimization of the RBM for better infectivity and the immune evasion capability of the virus.[73] In fact, recent studies have reported that newer Omicron subvariants exhibit diminishing gains in hACE2 binding affinity when compared to their predecessors,[74,75] a shift in evolutionary strategy that can be anticipated by monitoring the divergence in LM^{CM} spectral outputs. By identifying these evolutionarily significant sites months before they manifest in dominant global variants, the LM^{CM} framework demonstrates robust capability to predict the physicochemical trajectory of SARS-CoV-2 spike evolution.

4 Conclusions

Our statistical pipeline allowed us to characterize the collective effects arising from the mass acquisition of mutations by the SARS-CoV-2 spike glycoprotein. From our comprehensive analysis, several distinct patterns were revealed in the temporal and spatial domains of its evolution. For the former, mutation rate heterogeneity across variants, exemplified by its increase in recent Omicron sublineages, underscored the virus's capacity to perpetuate its continual adaptation. For the latter, the power-law scaling we uncovered highlights how a very few extreme mutational leaps can entirely reshape the evolutionary trajectory of the virus. In particular, this enabled us to devise the temporal gradient of the $\log P(\boldsymbol{d}) - \log \boldsymbol{d}$ slope, capable of quantifying these abrupt shifts, as an early indicator for the emergence of evolutionarily important variants. Moreover, at

the single amino acid level, our LM^{CM} method bridges the global sequence ensemble dynamics with site-specific functional impacts of individual mutations by outlining crucial leading mutations within massive sequence data sets; its results are posted in our publicly accessible platform at https://hbsulab.github.io/deLemus/. Overall, the findings advance our understanding of viral evolution and demonstrates the need for extensive disease surveillance to better understand the evolutionary dynamics of circulating viruses.

Conflicts of interest

There are no conflicts to declare.

Data availability

The codes and data are available at https://github.com/hbsulab/Deciphering-the-Temporal-Mutation-Pattern-of-the-SARS-CoV-2-Spike-Glycoprotein.

Supplementary information (SI) include the tensor decomposition method and pair-coupling matrices; evolutionary trend of leading mutations outlined by the composite metric; physiochemical properties of leading mutations in reported variants, *etc*. See DOI: https://doi.org/10.1039/d5cp04811g.

Acknowledgements

We acknowledge the data contributors and the GISAID Initiative for sharing the genetic sequences. We are grateful to Prof. Prabal Maiti for fruitful discussions and advice. This work is supported in part by the Research Grants Council of Hong Kong (16305623), HKUST 20-20 Grant (VP2020S25SC04), Guangdong S&T Program (2025A0505000027), and Society of Interdisciplinary Research (SOIRÉE).

References

1 P. V. Markov, M. Ghafari and M. Beer, *et al.*, The evolution of SARS-CoV-2, *Nat. Rev. Microbiol.*, 2023, **21**, 361–379.
2 A. C. Walls, Y.-J. Park and M. A. Tortorici, *et al.*, Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein, *Cell*, 2020, **181**, 281–292.e6.
3 Y. Huang, C. Yang and X.-F. Xu, *et al.*, Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19, *Acta Pharmacol. Sin.*, 2020, **41**, 1141–1149.
4 M. Amicone, V. Borges and M. J. Alves, *et al.*, Mutation rate of SARS-CoV-2 and emergence of mutators during experimental evolution, *Evol. Med. Public Health*, 2022, **10**, 142–155.
5 A. M. Carabelli, T. P. Peacock and L. G. Thorne, *et al.*, SARS-CoV-2 variant biology: immune escape, transmission and fitness, *Nat. Rev. Microbiol.*, 2023, **21**, 162.
6 Y. Shu and J. McCauley, GISAID: Global initiative on sharing all influenza data - from vision to reality, *Eurosurveillance*, 2017, **22**, 30494.
7 A. Rambaut, E. C. Holmes and Á. O'Toole, *et al.*, A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology, *Nat. Microbiol.*, 2020, **5**, 1403–1407.
8 M. Lynch, Evolution of the mutation rate, *Trends Genet.*, 2010, **26**, 345–352.
9 K. M. Peck and A. S. Lauring, Complexities of Viral Mutation Rates, *J. Virol.*, 2018, **92**, e01031-17.
10 S. Duffy, Why are RNA virus mutation rates so damn high?, *PLoS Biol.*, 2018, **16**, e3000003.
11 M. Kimura, Evolutionary Rate at the Molecular Level, *Nature*, 1968, **217**, 624–626.
12 T. Ohta, Slightly Deleterious Mutant Substitutions in Evolution, *Nature*, 1973, **246**, 96–98.
13 K. Sprouffske, J. Aguilar-Rodríguez and P. Sniegowski, *et al.*, High mutation rates limit evolutionary adaptation in Escherichia coli, *PLOS Genet.*, 2018, **14**, e1007324.
14 M. Lynch, M. S. Ackerman and J.-F. Gout, *et al.*, Genetic drift, selection and the evolution of the mutation rate, *Nat. Rev. Genet.*, 2016, **17**, 704–714.
15 A. S. Lauring and R. Andino, Quasispecies Theory and the Behavior of RNA Viruses, *PLoS Pathog.*, 2010, **6**, 1–8.
16 M. Eigen, Selforganization of matter and the evolution of biological macromolecules, *Naturwissenschaften*, 1971, **58**, 465–523.
17 E. Domingo, J. Sheldon and C. Perales, Viral Quasispecies Evolution, *Microbiol. Mol. Biol. Rev.*, 2012, **76**, 159–216.
18 Á. O'Toole, E. Scher and A. Underwood, *et al.*, Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool, *Virus Evol.*, 2021, **7**, veab064.
19 J. Hadfield, C. Megill and S. M. Bell, *et al.*, NextStrain: Real-time tracking of pathogen evolution, *Bioinformatics*, 2018, **34**, 4121–4123.
20 J. H. Brown, V. K. Gupta and B.-L. Li, *et al.*, The fractal nature of nature: power laws, ecological complexity and biodiversity, *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 2002, **357**, 619–626.
21 G. B. West and J. H. Brown, The origin of allometric scaling laws in biology from genomes to ecosystems: towards a quantitative unifying theory of biological structure and organization, *J. Exp. Biol.*, 2005, **208**, 1575–1592.
22 M. E. J. Newman, Power laws, Pareto distributions and Zipf's law, *Contemp. Phys.*, 2005, **46**, 323–351.
23 M. Hasan, Z. He and M. Jia, *et al.*, Dynamic expedition of leading mutations in SARS-CoV-2 spike glycoproteins, *Comput. Struct. Biotechnol. J.*, 2024, **23**, 2407–2417.
24 A. R. Benson, D. F. Gleich and J. Leskovec, Higher-order organization of complex networks, *Science*, 2016, **353**, 163–166.
25 D. M. Weinreich, Y. Lan and C. S. Wylie, *et al.*, Should evolutionary geneticists worry about higher-order epistasis?, *Curr. Opin. Genet. Dev.*, 2013, **23**, 700–707.
26 T. N. Starr and J. W. Thornton, Epistasis in protein evolution, *Protein Sci.*, 2016, **25**, 1204–1218.
27 T. N. Starr, A. J. Greaney and W. W. Hannon, *et al.*, Shifting mutational constraints in the SARS-CoV-2 receptorbinding domain during viral evolution, *Science*, 2022, **424**, eabo7896.
28 T. G. Kolda and B. W. Bader, Tensor Decompositions and Applications, *SIAM Rev.*, 2009, **51**, 455–500.

29 L. Grasedyck, D. Kressner and C. Tobler, A literature survey of low-rank tensor approximation techniques, *GAMM-Mitt*, 2013, **36**, 53–78.
30 X. Bi, X. Tang and Y. Yuan, *et al.*, Tensors in Statistics, *Annu. Rev. Stat. Appl.*, 2021, **8**, 345–368.
31 A. Auddy, D. Xia and M. Yuan, Tensors in High-Dimensional Data Analysis: Methodological Opportunities and Theoretical Challenges, *Annu. Rev. Stat. Appl.*, 2025, **12**, 527–551.
32 T. N. Starr, A. J. Greaney and S. K. Hilton, *et al.*, Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding, *Cell*, 2020, **182**, 1295–1310.e20.
33 L. Jaroszewski, M. Iyer and A. Alisoltani, *et al.*, The interplay of SARS-CoV-2 evolution and constraints imposed by the structure and functionality of its proteins, *PLoS Comput. Biol.*, 2021, **17**, 1–22.
34 D. Wrapp, N. Wang and K. S. Corbett, *et al.*, Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation, *Science*, 2020, **367**, 1260–1263.
35 A. J. Greaney, T. N. Starr and C. O. Barnes, *et al.*, Mapping mutations to the SARS-CoV-2 RBD that escape binding by different classes of antibodies, *Nat. Commun.*, 2021, **12**, 4196.
36 V. Calvaresi, A. G. Wrobel and J. Toporowska, *et al.*, Structural dynamics in the evolution of SARS-CoV-2 spike glycoprotein, *Nat. Commun.*, 2023, **14**, 1421.
37 W. Xu, T. Guo and H. Su, Evolutionary aspect of spike glycoprotein's conformational dynamics, *Phys. Chem. Chem. Phys.*, 2026, **28**, 5645–5655.
38 R. Orús, A practical introduction to tensor networks: Matrix product states and projected entangled pair states, *Ann. Phys.*, 2014, **349**, 117–158.
39 M. C. Bañuls, Tensor Network Algorithms: A Route Map, *Annu. Rev. Condens. Matter Phys.*, 2023, **14**, 173–191.
40 P. Zhou, X.-L. Yang and X.-G. Wang, *et al.*, A pneumonia outbreak associated with a new coronavirus of probable bat origin, *Nature*, 2020, **579**, 270–273.
41 Z. Chen, A. S. Azman and X. Chen, *et al.*, Global landscape of SARS-CoV-2 genomic surveillance and data sharing, *Nat. Genet.*, 2022, **54**, 499–507.
42 A. F. Brito, E. Semenova and G. Dudas, *et al.*, Global disparities in SARS-CoV-2 genomic surveillance, *Nat. Commun.*, 2022, **13**, 7003.
43 F. Sievers, A. Wilm and D. Dineen, *et al.*, Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega, *Mol. Syst. Biol.*, 2011, **7**, 539.
44 WHO, WHO COVID-19 Dashboard. https://data.who.int/dashboards/covid19/variants. 2024.
45 S. Duchene, L. Featherstone and M. Haritopoulou-Sinanidou, *et al.*, Temporal signal and the phylodynamic threshold of SARS-CoV-2, *Virus Evol.*, 2020, **6**, veaa061.
46 P. Mlcochova, S. A. Kemp and M. S. Dhar, *et al.*, SARSCoV-2 B.1.617.2 Delta variant replication and immune evasion, *Nature*, 2021, **599**, 114–119.
47 Y. Fan, X. Li and L. Zhang, *et al.*, SARS-CoV-2 Omicron variant: recent progress and future perspectives, *Signal Transduction Targeted Ther.*, 2022, **7**, 141.
48 M. Vignuzzi, J. K. Stone and J. J. Arnold, *et al.*, Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population, *Nature*, 2006, **439**, 344–348.
49 Y. Xiao, I. Rouzine and S. Bianco, *et al.*, RNA Recombination Enhances Adaptability and Is Required for Virus Spread and Virulence, *Cell Host Microbe*, 2016, **19**, 493–503.
50 K. Hanada, Y. Suzuki and T. Gojobori, A Large Variation in the Rates of Synonymous Substitution for RNA Viruses and Its Relationship to a Diversity of Viral Infection and Transmission Modes, *Mol. Biol. Evol.*, 2004, **21**, 1074–1080.
51 H. Tegally, E. Wilkinson and J. L.-H. Tsui, *et al.*, Dispersal patterns and influence of air travel during the global expansion of SARS-CoV-2 variants of concern, *Cell*, 2023, **186**, 3277–3290.e16.
52 R. V. Solé, S. C. Manrubia and M. Benton, *et al.*, Criticality and scaling in evolutionary ecology, *Trends Ecol. Evol.*, 1999, **14**, 156–160.
53 G. U. Yule, A mathematical theory of evolution, based on the conclusions of Dr J. C. Willis, F. R. S, *Philos. Trans. R. Soc., B*, 1925, **213**, 21–87.
54 B. F. Nielsen, C. M. Saad-Roy and Y. Li, *et al.*, Host heterogeneity and epistasis explain punctuated evolution of SARS-CoV-2, *PLoS Comput. Biol.*, 2023, **19**, 1–24.
55 R. Sender, Y. M. Bar-On and S. Gleizer, *et al.*, The total number and mass of SARS-CoV-2 virions, *Proc. Natl. Acad. Sci.*, 2021, **118**, e2024815118.
56 J. M. Halley, Ecology, evolution and 1/f-noise, *Trends Ecol. Evol.*, 1996, **11**, 33–37.
57 L. Corey, C. Beyrer and M. S. Cohen, *et al.*, SARS-CoV-2 Variants in Patients with Immunosuppression, *N. Engl. J. Med.*, 2021, **385**, 562–566.
58 C. A. Smith and B. Ashby, Antigenic evolution of SARSCoV-2 in immunocompromised hosts, *Evol. Med. Public Health*, 2023, **11**, 90–100.
59 R. Viana, S. Moyo and D. G. Amoako, *et al.*, Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa, *Nature*, 2022, **603**, 679–686.
60 D. Planas, I. Staropoli and V. Michel, *et al.*, Distinct evolution of SARS-CoV-2 Omicron XBB and BA.2.86/JN.1 lineages combining increased fitness and antibody evasion, *Nat. Commun.*, 2024, **15**, 2254.
61 L. B. Snell, S. Pickering and A. Alcolea-Medina, *et al.*, Antibody escape drives emergence of diverse spike haplotypes resembling variants of concern in persistent SARSCoV-2 infections, *Cell Rep. Med.*, 2026, **7**, 102587.
62 A. L. Taylor and T. N. Starr, Deep mutational scanning of SARS-CoV-2 Omicron BA.2.86 and epistatic emergence of the KP.3 variant, *Virus Evol.*, 2024, **10**, veae067.
63 B. Luan, H. Wang and T. Huynh, Enhanced binding of the N501Y-mutated SARS-CoV-2 spike protein to the human ACE2 receptor: insights from molecular dynamics simulations, *FEBS Lett.*, 2021, **595**, 1454–1461.
64 W. Dejnirattisai, D. Zhou and P. Supasa, *et al.*, Antibody evasion by the P.1 strain of SARS-CoV-2, *Cell*, 2021, **184**, 2939–2954.e9.

65 F. Tian, B. Tong and L. Sun, *et al.*, N501Y mutation of spike protein in SARS-CoV-2 strengthens its binding to receptor ACE2, *eLife*, 2021, **10**, e69091.
66 L. Wu, L. Zhou and M. Mo, *et al.*, SARS-CoV-2 Omicron RBD shows weaker binding affinity than the currently dominant Delta variant to human ACE2, *Signal Transduction Targeted Ther.*, 2022, **7**, 8.
67 D. Mannar, J. W. Saville and X. Zhu, *et al.*, Structural analysis of receptor binding domain mutations in SARSCoV-2 variants of concern that modulate ACE2 and antibody binding, *Cell Rep.*, 2021, 110156.
68 Y. Wang, C. Liu and C. Zhang, *et al.*, Structural basis for SARS-CoV-2 Delta variant recognition of ACE2 receptor and broadly neutralizing antibodies, *Nat. Commun.*, 2022, **13**, 871.
69 A. Aggarwal, S. Naskar and N. Maroli, *et al.*, Mechanistic insights into the effects of key mutations on SARS-CoV-2 RBD-ACE2 binding, *Phys. Chem. Chem. Phys.*, 2021, **23**, 26451–26458.
70 C. Nie, A. K. Sahoo and R. R. Netz, *et al.*, Charge Matters: Mutations in Omicron Variant Favor Binding to Cells, *ChemBioChem*, 2022, **23**, e202100681.
71 C. B. Jackson, M. Farzan and B. Chen, *et al.*, Mechanisms of SARS-CoV-2 entry into cells, *Nat. Rev. Mol. Cell Biol.*, 2022, **23**, 3–20.
72 B. Jawad, P. Adhikari and R. Podgornik, *et al.*, Key Interacting Residues between RBD of SARS-CoV-2 and ACE2 Receptor: Combination of Molecular Dynamics Simulation and Density Functional Calculation, *J. Chem. Inf. Model.*, 2021, **61**, 4425–4441.
73 S. Xue, Y. Han and F. Wu, *et al.*, Mutations in the SARSCoV-2 spike receptor binding domain and their delicate balance between ACE2 affinity and antibody evasion, *Protein Cell*, 2024, **15**, 403–418.
74 H. L. Nguyen, T. Q. Nguyen and M. S. Li, SARS-CoV-2 Omicron Subvariants Do Not Differ Much in Binding Affinity to Human ACE2: A Molecular Dynamics Study, *J. Phys. Chem. B.*, 2024, **128**, 3340–3349.
75 Q. Wang, I. A. Mellis and J. Ho, *et al.*, Recurrent SARSCoV-2 spike mutations confer growth advantages to select JN.1 sublineages, *Emerging Microbes Infect.*, 2024, **13**, 2402880.