Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Immunology

Review

SARS-CoV-2: lessons in virus mutation prediction and pandemic preparedness

Weiyi Tang[1], Jenna Kim[1], Raphael TC Lee[2,3], Sebastian Maurer-Stroh[1,2,3,4,5,6], Laurent Renia[1,7,8] and Matthew Z Tay[1,9]

The COVID-19 pandemic has prompted an unprecedented global response. In particular, extraordinary efforts have been dedicated toward monitoring and predicting variant emergence due to its huge impact, particularly for vaccine escape. Broadly, we classify such methods into two categories: forward mutation prediction, where phenotypes are first observed and the responsible genotypes traced, and reverse mutation prediction, which starts with selected pathogen genetic profiles and characterizes their associated phenotypes. Reverse mutation prediction strategies have advantages in being able to sample a more complete evolutionary space since sequences that do not yet exist can be sampled. The rapid improvement in the maturity and scale of reverse mutation prediction strategies, such as deep mutational scanning, has led to significant amounts of data for machine learning, with concomitant improvement in the prediction results from computational tools. Such integrated prediction approaches are generalizable and offer significant opportunities for anticipating viral evolution and for pandemic preparedness.

Addresses
[1] A*STAR Infectious Diseases Labs (AIDL), Agency for Science, Technology and Research (A*STAR), Singapore
[2] Bioinformatics Institute (BII), Agency for Science, Technology and Research (A*STAR), Singapore
[3] GISAID Global Data Science Initiative (GISAID), Munich, Germany
[4] National Public Health Laboratory, Singapore, Singapore
[5] Department of Biological Sciences, National University of Singapore, Singapore, Singapore
[6] Human Potential Translational Research Programme, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore
[7] Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore
[8] School of Biological Sciences, Nanyang Technological University, Singapore
[9] Department of Biochemistry, National University of Singapore, Singapore

Corresponding author: Tay, Matthew Z (matthew_tay@IDLabs.a-star.edu.sg)

Current Opinion in Immunology 2025, **95**:102560

This review comes from a themed issue on **Immunity and COVID**

Edited by **Benedetta Terziroli Beretta-Piccoli**

For complete overview of the section, please refer to the article collection, "Immunity and COVID (2025)"

Available online 15 May 2025

https://doi.org/10.1016/j.coi.2025.102560

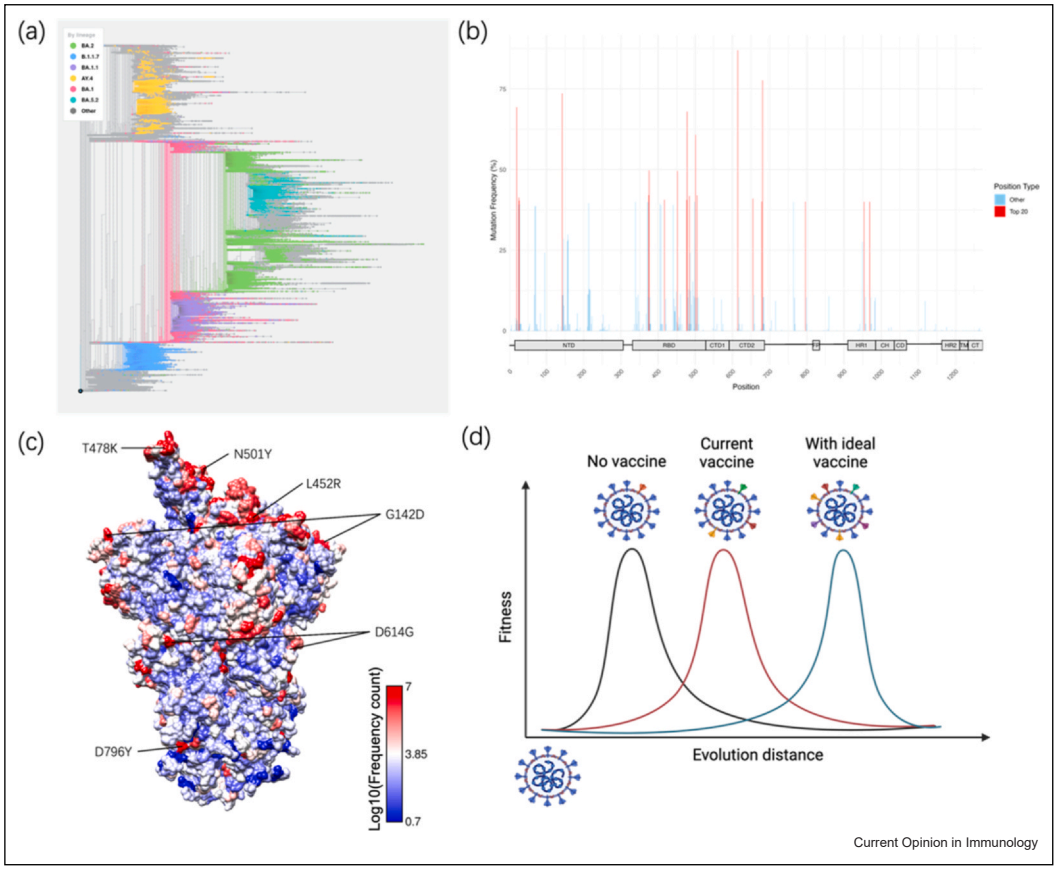
0952–7915/© 2025 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Introduction

Since its emergence in Wuhan, China, in December 2019, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has caused over 700 million cases worldwide [1]. Five years later, the virus continues to mutate and evolve (Figure 1a). As of December 2024, the World Health Organization has designated JN.1 as the current variant of interest, which carries over 90 mutations compared to the ancestral Wuhan strain. Analysis of over 10 million SARS-CoV-2 sequences reveals that the receptor-binding domain (RBD) is the most frequently mutated region of the spike protein, followed by the N-terminal domain and C-terminal domain (Figure 1b and c). Mutations such as D614G and N501Y are particularly notable, as they enhance ACE2 receptor affinity and promote immune escape [2,3]. These evolutionary adaptations underscore the challenges faced in controlling SARS-CoV-2 and similar viruses.

Vaccines remain a key countermeasure against viral infections. Numerous vaccines and antibody therapies have been licensed around the world to combat COVID-19,

Figure 1

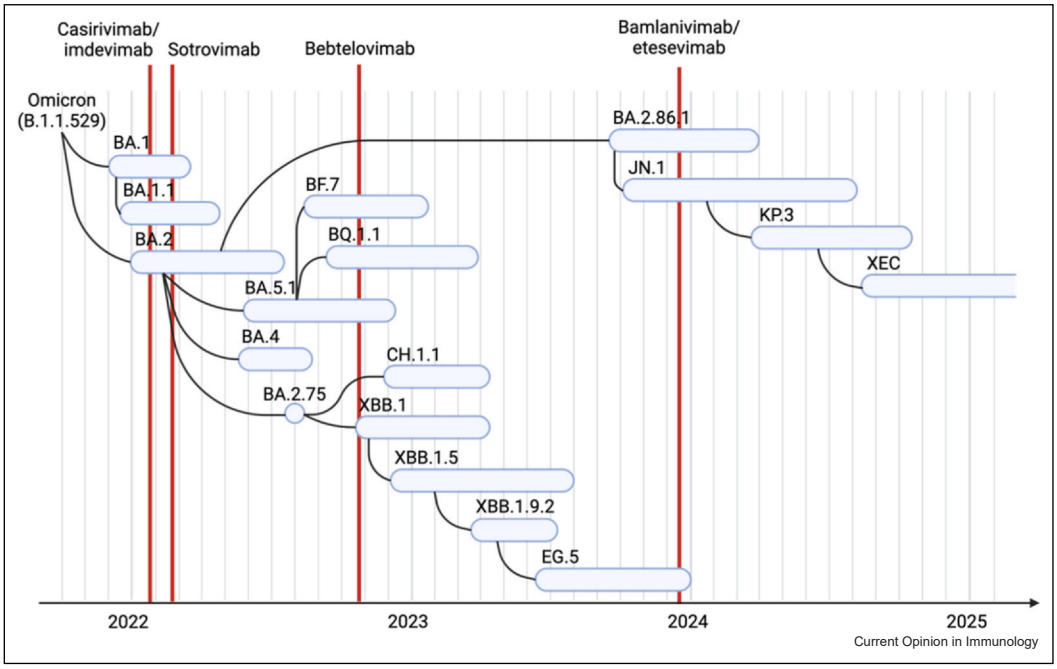


SARS-CoV-2 evolutionary history. **(a)** SARS-CoV-2 phylogenetic tree containing 12 649 996 high-quality genomes from GISAID (updated on November 29, 2024). **(b)** The mutation frequency of each position on the SARS-CoV-2 spike protein. Sequence data were obtained from GISAID, and 5 291 257 sequences were selected with no unidentified amino acids and length over 1260 amino acids. All sequences were compared to the original Wuhan strain by DIAMOND to identify mutations [5]. The 20 most frequently mutated positions are colored in red. **(c)** The mutation frequency of each position on SARS-CoV-2 spike protein mapped to the 3D structure. The structure is obtained from PDB (#7DWZ). **(d)** Ideal vaccines can apply strong selection pressure to the virus, forcing it to undergo more rounds of mutations to escape and/or gain fitness. The code utilized for mutation frequency analysis is openly available on Github (https://github.com/tangweiyi123/MutationFrequencyAnalysis). The SARS-CoV-2 genomic sequences analyzed in this study were obtained from the GISAID database. The findings of this study are based on metadata associated with 5 291 257 sequences available on GISAID up to April 23, 2025, and accessible at https://doi.org/10.55876/gis8.250422nv. GISAID, Global Initiative on Sharing All Influenza Data.

but their efficacy has waned over time due to the emergence of different SARS-CoV-2 strains [4]. The ability to predict potential mutation trajectories and their functional impact is essential for informing vaccine development and public health policies as viruses continue to evolve. By identifying likely future variants, researchers can design vaccines that are more resistant to immune escape — either by blocking viral evasion altogether or by forcing the virus to accumulate additional mutations, which may reduce its binding efficiency to host cells (Figure 1d). This approach not only enhances the durability of vaccine protection but also helps extend the time between necessary updates, thereby reducing the logistical and financial burdens of large-scale vaccine distribution, particularly in resource-limited settings. Moreover, by maintaining higher levels of immunity within the population, these strategies can contribute to slowing viral transmission, facilitating herd immunity, and mitigating the long-term impact of a pandemic.

Here, we review current strategies for predicting virus mutation and suggest a framework for classifying major approaches. Similar to the idea of forward and reverse genetics, we define 'forward mutation prediction' as the process of identifying specific phenotypes, such as immune escape or fitness, then tracing the genotypes responsible. Viral surveillance and experimental evolution fall under this category. More recently, a new class of predictive methods has emerged, which we term 'reverse mutation prediction'. These begin with the selection of predefined genetic profiles of pathogens, followed by the characterization of their associated

Figure 2



Timeline of Omicron descendant circulation and FDA revocations of emergency use authorization for monoclonal antibody treatments [11]. Each Omicron descendant represents its circulation time when its prevalence is > 1% (except BA.2.75) [7,12]. FDA, Food and Drug Administration.

phenotypic traits. Deep mutational scanning (DMS), immunological profiling, and computational modeling exemplify this approach. Just as reverse genetics offered a complement to forward genetics to enable more comprehensive and mechanistic studies, we believe that reverse mutation prediction strategies will enhance forward mutation prediction, providing improved forecasting capabilities and deeper insights into the mechanisms driving pathogen evolution.

'Forward mutation prediction'

Global surveillance

Global virus surveillance operates as an interconnected network in which health care facilities, research laboratories, and environmental monitoring stations collect samples from patients, wastewater, and animals to track viral evolution [6]. These samples undergo genomic sequencing to identify genetic changes, and the results are uploaded to international databases, such as the platform of the Global Initiative on Sharing All Influenza Data data science initiative [7,8]. By compiling genomic sequences, researchers can detect mutation patterns and predict how a virus may evolve, including shifts in variant dominance, immune escape potential, and disease severity (Figure 2).

An example of large-scale surveillance in action is the United States Centers for Disease Control and Prevention's (CDC's) monitoring program, which tracked nationally circulating SARS-CoV-2 variants from May 2023 to September 2024. During this period, the predominant variants were identified as descendants of Omicron XBB and JN.1, with JN.1 lineages becoming increasingly prevalent. Data from this report informed the selection of subsequent antigenic targets as well as provided insight into potential impacts on vaccine efficacy [9]. Real-time genomic tracking of viruses in this manner not only facilitates the early identification of emerging variants but also enables public health officials to anticipate infection surges and tailor response strategies accordingly.

In addition to human cases, surveillance efforts extend to genomic analyses of samples from wild animal populations. Studies have documented instances of SARS-CoV-2 transmission between humans and animals, raising concerns that certain species could serve as reservoirs for new variants [6]. While definitive evidence for SARS-CoV-2 adaptation in wild animals remains limited, collecting and sequencing animal samples aids in assessing the risk of cross-species transmission events that could introduce novel viral variants into human populations. By comparing the evolutionary trajectories of viruses across different geographical regions, researchers can also gain insight into how selective pressures — such as vaccination coverage and population immunity — shape viral spread and variant selection. Integrating these regional analyses into global

surveillance can improve predictions of viral evolution and guide vaccine design strategies [10].

Despite advancements in virus tracking, achieving truly comprehensive global surveillance remains a challenge. Differences in protocols for sample collection, sequencing, and data sharing can introduce inconsistencies that complicate comparative analyses across regions, undermining the accuracy and reliability of mutation predictions. Furthermore, rapid and widespread sampling is essential for real-time monitoring, yet many developing countries do not have the necessary infrastructure for systematic genomic sequencing. These geographic disparities result in the underrepresentation of viral evolution in regions with fewer resources, creating biases that reduce the generalizability of predictive models and impair early detection and response efforts.

Addressing these limitations is critical toward establishing an integrated and responsive global virus surveillance system. Efforts to develop universally recognized data standards, such as the CDC's Data Modernization Initiative [13], can enhance quality, interoperability, and access. Strategic investments in sequencing infrastructure also have the potential to expand global coverage. International collaborations, technology transfers, and targeted funding programs focused on genomic sequencing capacity are particularly impactful, as demonstrated by the effectiveness of programs like the Africa CDC's Pathogen Genomics Initiative [14]. Additionally, innovations in sequencing technology — such as portable sequencing platforms and accessible bioinformatics tools — play a pivotal role in democratizing genomic surveillance. Making these technologies more broadly available across diverse geographic and economic settings can strengthen both the robustness and equity of global mutation prediction.

Experimental evolution

Another strategy for predicting new variants is passaging the virus *in vitro* or *in vivo* under controlled experimental conditions and applying whole-genome sequencing to identify mutations. Unlike global surveillance and immunological profiling, which focus on variants that have already emerged in natural settings, experimental evolution aims to observe mutations that arise spontaneously under experimental conditions. These mutations initially occur at random but may disappear or become fixed in the system. Mutations that consistently appear in both laboratory and natural viral populations may indicate broader evolutionary trends, offering valuable insights for vaccine and therapeutic development [15,16]. Conversely, some mutations may only be observed in natural settings, shaped by stochasticity, host-specific immune pressures, or environmental factors not easily replicated in the lab.

Chronic HIV-1 infection provides a natural case study for viral evolution. Within a single individual, successive waves of viral escape variants arise in response to the evolving neutralizing antibody response [17]. The patterns of escape, primarily through a shifting glycan 'shield', reveal an effectively infinite capacity for HIV-1 — and potentially other viruses — to evolve resistance against antibodies while maintaining fitness. These insights from patient-derived samples have shaped our understanding of immune evasion and guided the design of broadly neutralizing antibodies and immunogens for vaccine research.

In acute viral infections such as with SARS-CoV-2, however, opportunities to study within-host evolution are more limited. Instead, studies must rely on animal or cell-based models to simulate selective pressures and observe mutation trajectories. For example, a direct passage of early SARS-CoV-2 isolates in multiple human cell lines led to the identification of mutations consistent with those observed in Alpha, Gamma, and Omicron variants [18]. A study with long-term serial passaging of SARS-CoV-2 in VeroE6 cells showed independent acquisition of the same mutations across different lines, suggesting a pattern of convergent evolution [15]. *In vivo* passaging of B.1.617.2 (Delta) in K18-ACE2 mice revealed *de novo* mutations that are associated with disease severity and antibody resistance in late passages [19]. Collectively, these models provide a method for anticipating evolutionary paths that may arise before they become widespread in human populations.

Compared to *in vitro* and *in silico* approaches, *in vivo* models provide a more comprehensive representation of natural infection by accounting for complex physiological interactions. As a result, they have been indispensable in characterizing current and potential SARS-CoV-2 variants. However, animal models rarely recapitulate human physiology in its entirety. This makes the selection of appropriate models for specific applications an important consideration. The incorporation of *in vivo* techniques as part of a multi-pronged strategy will best inform our prediction and understanding of the evolution of future SARS-CoV-2 variants.

Evolutionary passaging can also be leveraged to investigate the robustness of therapies against emerging variants. For instance, long-term passaging of SARS-CoV-2 with various antibody cocktails has been used to assess viral resistance to therapeutic antibodies [20]. Additionally, adaptation studies using SARS-CoV-2 in ACE2 low-expressing cell lines identified mutations associated with increased receptor affinity and entry efficiency [21]. Passaging SARS-CoV-2 in the presence of the antiviral drugs nirmatrelvir and ensitrelvir identified the M49L mutation as a driver of drug resistance, both *in vitro* and *in vivo* [22]. Similarly, exposing SARS-CoV-2 to

increasing concentrations of GS-441524, the parent nucleoside of remdesivir, led to the identification of the S759A and V792I mutations conferring drug resistance [23]. An efficient way to induce mutations and study drug resistance is using mutagenesis compounds like N4-hydroxycytidine [24]. However, such studies must carefully navigate ethical considerations and potential public health risks.

A key advantage of experimental evolution is the ability to track virus evolution in a controlled setting. However, this approach faces three challenges that cannot be fully addressed simultaneously: biological relevance, throughput, and cost. *In vitro* cell lines offer a high-throughput and cost-effective model but fail to capture mutations influenced by host immunity, tissue-specific infection dynamics, and other complex biological factors. More sophisticated models, such as organoids and humanized mice, better mimic human physiology and may therefore yield more clinically relevant findings. Still, these models require significant labor and resources, particularly for long-term passaging and maintaining multiple passage lineages. Moreover, the inherently limited scale of such experiments introduces stochasticity, meaning only a small subset of possible mutations will emerge. This constrains the ability to comprehensively analyze all mutations that might contribute to a particular phenotype and thus limits the breadth of insights that can be obtained.

Expanding predictive horizons

Global surveillance and experimental evolution are central tools in forward mutation prediction and have provided important insights into viral evolution. However, these approaches heavily depend on observing and extrapolating from variants that have already emerged. This limits their ability to anticipate entirely novel or unprecedented variants. The Omicron variant, which featured multiple novel mutations without clear evolutionary intermediates, is a prime example. Closing this gap in predictive capability thus requires complementary strategies that proactively explore the broader mutational landscape.

'Reverse mutation prediction'

Deep mutational scanning

DMS has emerged as a transformative technique, revolutionizing our ability to understand protein structure–function relationships. In virology contexts, DMS provides a prospective approach to measure the effects of a large number of mutations on the spike protein, even before they emerge in circulating viral variants. Traditional DMS is performed using cell-surface display systems in yeast or mammalian cells to measure protein binding. For example, site-saturation mutagenesis of the BA.2.86 RBD spanning 200 amino acids assessed the effects of all single mutations on ACE2 affinity and RBD expression. This study revealed that while the Q493E mutation (present from BA.2.86 to KP.3) alone impairs ACE2 affinity, the strong positive epistasis among L455S, F456L, and Q493E allows KP.3 to maintain ACE2 binding and enhances immune evasiveness [25]. Traditional DMS methods, however, are typically restricted to specific protein regions and primarily measure binding affinity without assessing broader functional effects.

Recent advances have expanded DMS capabilities. The application of a genotype–phenotype-linked pseudovirus system by Jesse Bloom's lab enables evaluation of the entire protein and includes assessment of antibody neutralization [26]. The extended DMS accurately predicted non-RBD mutations with escape potential against various antibodies [26]. Additionally, DMS analyses of full XBB.1.5 and BA.2 spikes identified non-RBD mutations that improve ACE2 binding, such as Q52H and T572I. Other non-RBD mutations were found to improve serum antibody escape while reducing ACE2 binding ability, suggesting that they act by altering RBD conformation. By estimating mutation effects on viral growth and serum escape, they also partially predicted the evolutionary success of SARS-CoV-2 clades, demonstrating the potential of DMS to recapitulate real-world evolutionary trajectories [27]. Notably, the use of a pseudovirus system — where the SARS-CoV-2 spike protein is displayed on a lentiviral pseudovirus particle that cannot self-replicate without additional helper plasmids — avoids the risk of generating fitter, potentially hazardous live viral strains.

Expanding DMS from single domains to full proteins demands vast resources for constructing and testing all possible mutations, and the complexity of testing combinatorial mutations increases exponentially. High-throughput methods and automation are essential for comprehensive DMS studies [28]. Another limitation is that the protein-level evaluations may not always accurately reflect viral fitness, necessitating further validation through additional functional assays. Nevertheless, DMS is already being used for the identification of epitopes for the development of resistance-resilient inhibitors. For instance, DMS of Zika virus NS2B-NS3 protease identified that key residues in an important allosteric site are actually mutable, discouraging further pursuit of this site and suggesting other epitopes where virus escape is less likely to occur [29].

Immunological profiling

Immunological profiling provides a real-world corollary to DMS, leveraging historical human population-level data to understand virus–host interactions and the evolution of new variants. Two common methods are antigenic cartography, which provides a broad mapping of

population immunity, and epitope mapping, which focuses on detailed molecular interactions. By analyzing immune responses, researchers can identify current immune pressure points and potential viral escape routes, allowing predictions of mutations most likely to emerge and become fixed in populations.

Antigenic cartography is a powerful technique for visualizing the distances of antigens and antisera [30,31]. Originally developed using hemagglutination inhibition data for influenza virus, it can also be applied to neutralization assay data for other viruses. Several antigenic mappings have been constructed for SARS-CoV-2 in different studies. For example, one study analyzing the panel of 21 SARS-CoV-2 variants and 15 groups of human sera identified substitutions at spike protein positions 417, 452, 484, and 501 as key determinants of antigenic differences, though sensitivity to these substitutions varied greatly between individuals [32]. Another antigenic cartography study on a broad collection of pre-Omicron and Omicron variants indicated that cross-neutralization improved following hybrid immunity from bivalent boosting, supporting future vaccine updates [33]. Additionally, antigenic mapping of Omicron sublineages found substantial cross-neutralization of BA.5 antiserum against BQ1.1, suggesting immunological relationships between these variants [34].

Epitope mapping identifies antibody or T cell receptor binding sites on viral antigens. By characterizing heavily targeted epitopes, researchers can pinpoint antigen regions under strong selective pressure and predict mutations that are most feasible for the virus. Combined with population immunity data from antigenic cartography, epitope mapping enables prediction of which variants are most likely to emerge and spread widely. Linear epitope mapping uses overlapping peptide libraries, while conformational epitope mapping requires more advanced methods such as X-ray crystallography, peptide arrays, alanine scanning, domain exchange, hydrogen-deuterium exchange, chemical cross-linking, and hydroxyl radical footprinting [35]. A recent study mapping T cell epitopes in SARS-CoV-2 vaccine recipients and convalescent patients identified conserved and immunogenic nonspike epitopes, which could contribute to immune surveillance against emerging variants [36].

One limitation of immunological profiling is its retrospective nature. Unlike some other methods, it relies on recruiting patient cohorts, which can take months or even years. This delay between sample collection and analysis makes it difficult to capture real-time viral evolution, hindering immediate public health responses. Although efforts have been made to increase throughput via automation [37,38], this inherent delay remains a challenge. Another avenue for improving the efficiency of immunological profiling is developing artificial intelligence models to predict binding epitopes. However, training these models requires substantial data sets from preliminary studies, limiting their application to newly emerging viruses [39].

Computational modeling

The different sources of information described above can be utilized either individually or in an integrative manner via various computational approaches to predict viral evolution and the emergence of variants.

Rich epidemiological data provide a wealth of information that can be used for the prediction of viral evolution. Several methods have emerged that make use of such epidemiological data. For instance, the use of multinomial logistic regression to describe variant frequency changes [40] showed strong results in estimating viral fitness. A natural language–based model, the Constrained Semantic Change Search model, achieved 95–99% clustering accuracy for viral subtypes and identified immune escape mutations with area under the curve values ranging from 0.57 to 0.85 across different viral proteins [41]. Breaking sequences down to evaluate the impact of individual mutations, an analytical epidemiological model has also been employed to predict the transmissibility of mutations [42]. The model was able to predict that variants of concern, such as Alpha, Delta, and Omicron, have significant transmission advantages when they were less than 2% of sampled sequences.

In the absence of surveillance or epidemiological data, biophysical information alone can provide a rich source of information for predicting future viral variants. EVEscape calculated escape probability through a novel framework combining prepandemic data with biophysical information, identifying 50% of its top predictions in actual variants by May 2023, with 66% accuracy for high-frequency mutations [43]. This method was validated against experimental scans for SARS-CoV-2, influenza, and HIV, demonstrating its utility in early pandemic preparedness.

Still, algorithms can benefit greatly from epidemiological data. PandoGen [44], a model pretrained on general protein sequences and fine-tuned on SARS-CoV-2 spike protein sequences from Global Initiative on Sharing All Influenza Data, demonstrated strong predictive power by combining protein language modeling with reward-based learning. It outperformed larger models like ProGen2 [45] despite being 30× smaller, and successfully predicted Delta variant sequences 1 month before their emergence, as well as BA.5 sequences 10 days in advance.

The advent of 'reverse mutation prediction' techniques such as DMS and high-throughput immunological profiling has resulted in sufficiently large data sets that can be utilized by computational approaches for training or

validation of viral escape predictions. Chi et al. [46] combined molecular dynamics analyses with FoldX [47] stability calculations to assess the potential for SARS-CoV-2 mutations to escape antibody neutralization by simulating changes in binding affinity against a panel of monoclonal antibodies. This method showed mean Pearson correlations of 0.52 between predicted binding affinity changes and experimental escape fractions across 19 antibody-antigen complexes. Similarly, the structure-based model SpikePro, which estimates the binding affinity to ACE2 and other neutralizing antibodies, could also be used to predict viral transmissibility, infectivity, immune escape, and basic reproductive number [48].

Epidemiological and biophysical data can also be integrated with DMS and immunological profiling to generate more comprehensive fitness models. As detailed in Meijers et al., antigenic fitness can be calculated using cross-immunity matrices derived from neutralization titers, while intrinsic fitness incorporates protein stability and receptor binding metrics [49]. This model successfully identified temporal windows for antigenic escape by tracking when high population immunity coincided with strong antigenic selection. Another such comprehensive framework accurately predicted the fitness of emerging SARS-CoV-2 variants and helped explain the reversion of the Q493R mutation by accounting for epistatic interactions [50]. Several other models estimate site-specific mutation fitness by incorporating phenotypic measures of antigenic drifts [51] or DMS data [52,53]. One recent success has been the incorporation of regional genomic surveillance data with DMS through mechanistic modeling, which produced predicted trends that closely matched historical dynamics of SARS-CoV-2 variant fitness, including region-specific differences in viral fitness driven by prior variant exposure in local populations [54].

Across the studies discussed here and in the broader literature, computational models have demonstrated reasonably strong predictive accuracy in identifying viral mutations, estimating fitness, and forecasting immune escape. Performance metrics such as area under the curve values, correlation coefficients, and classification accuracies have generally fallen within moderate to high ranges, supporting the applicability of these models in pandemic preparedness. Still, the potential for false-positive and false-negative predictions is a key concern. False positives, which predict consequential mutations that never arise, can divert resources toward unnecessary interventions. False negatives, where impactful mutations go undetected, may delay critical public health responses.

Quantitative assessments of false-positive and false-negative rates are beginning to emerge but remain limited. A major challenge lies in the heterogeneity of model validation pipelines, as differing training sets, prediction targets, and evaluation metrics make direct comparisons difficult. Interpretation is further complicated by the temporal dynamics of viral evolution. For example, a study analyzing an artificial neural network for SARS-CoV-2 mutation forecasting found that several mutations initially deemed false positives later appeared in independent data sets collected months after the model was trained [55]. This time lag blurs the boundary between accurate prediction and overprediction, especially in rapidly evolving viruses like SARS-CoV-2.

This underscores the benefit of integrating diverse data types — including immunological profiling, DMS, epidemiology, and biophysical data — to enrich model training and validation. Machine learning approaches that incorporate epistatic effects, immune dynamics, and host variability further enhance biological realism. Together, these integrative and iterative strategies are critical for improving the predictive accuracy and real-world utility of computational models in mutation prediction.

Scaling and validating reverse mutation prediction

While reverse mutation prediction offers a proactive framework for anticipating viral evolution, it is heavily dependent on high-throughput experimental infrastructure. Techniques such as DMS and large-scale immunological profiling require specialized platforms, extensive reagent libraries, automation pipelines, and considerable technical expertise. Scaling these methods to cover full-length viral proteins and combinations of mutations demands substantial resources in terms of time, labor, and cost. As a result, their implementation is often confined to well-equipped research centers, limiting broader accessibility and slowing widespread adoption.

Furthermore, despite their predictive power, experimental validation of reverse prediction outputs remains an ongoing challenge. Many mutations identified through DMS or computational modelling as potentially significant — whether in terms of ACE2 binding, antibody escape, or viral fitness — have yet to be confirmed *in vivo* or in clinical settings [26]. For example, some mutations predicted to enhance immune escape are supported by pseudovirus assays but lack corroboration in full viral replication systems or epidemiological surveillance. Similarly, antigenic cartography and epitope mapping offer retrospective insights into immune pressure but may lag behind real-time variant emergence.

This deficit reflects both the complexity of translating *in vitro* findings to real-world viral behavior and the logistical limitations of high-throughput confirmation. Comprehensive validation requires integrating multiple experimental systems, such as live virus assays, animal models, and longitudinal cohort studies, which each

have their own constraints. As a result, many reverse prediction findings remain preliminary, emphasizing the need for continued investment in experimental infrastructure and coordinated efforts to accelerate validation workflows. Bridging this gap will be essential to maximize the utility of reverse mutation prediction for timely and actionable pandemic preparedness.

Mutation prediction in the pandemic preparedness framework

While forward and reverse mutation prediction provide valuable foresight into the evolutionary trajectories of viruses, they represent only one component of a comprehensive pandemic preparedness strategy. The predictive utility of these approaches is enhanced — and their limitations mitigated — when integrated with other key domains of infectious disease control, including host–pathogen biology, behavioral modeling, and health care system readiness.

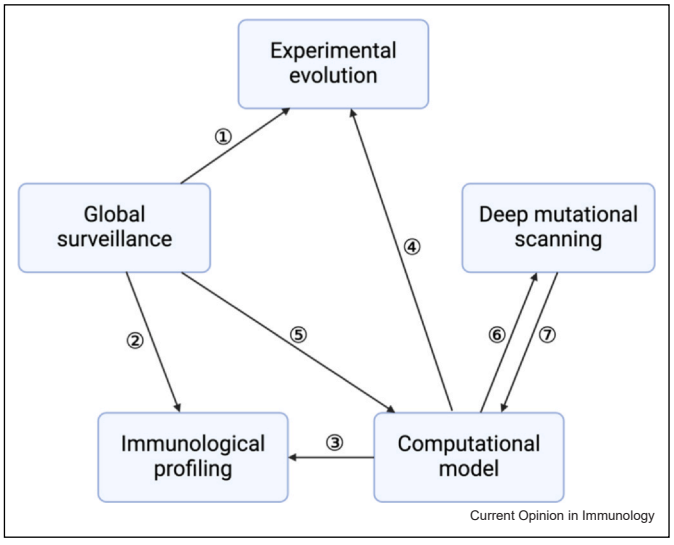
A fundamental complement to mutation prediction is the study of host–pathogen interactions, which governs not only viral infectivity and immune escape but also transmission dynamics and disease severity. The fitness of a given mutation cannot be fully interpreted in isolation from the host environment in which it evolves, as factors such as tissue tropism, host immune history, genetic susceptibility, and even microbiome composition can influence whether a mutation confers a true advantage *in vivo* [56]. Consequently, prediction models often benefit from concurrent research on immunological correlates of protection, viral entry mechanisms, and the like to infer which mutations are likely to enhance fitness or escape immunity. At the same time, reverse prediction techniques such as DMS can feed back into basic virology by identifying regions of the viral proteome under strong functional constraint or immune pressure. Without accounting for these biological variables, predictions risk misrepresenting the real-world impact of emerging mutations.

Integrated models that incorporate viral evolution alongside human behavior additionally offer a more holistic approach to forecasting outbreaks. These approaches can, for instance, project the geographic spread of a predicted immune escape variant based on travel patterns and local seroprevalence or estimate the likelihood of emergence under varying levels of vaccination coverage. Mutation prediction enhances these models by identifying candidate variants of concern, while the broader models offer contextual filters such as prior infection rates and demographics to assess which mutations are most likely to have population-level impact.

Mutation prediction also interfaces with health care system and community readiness. For instance, identifying mutations that alter viral epitopes may inform monoclonal antibody development, but such advances are only effective when supported by responsive health care infrastructure and robust supply chains. At the community level, timely insights into variant evolution enable early implementation of targeted public health messaging and nonpharmaceutical interventions, such as masking or social distancing, which can mitigate transmission before widespread clinical measures are available. In turn, well-prepared communities with effective surveillance, rapid testing, and consistent reporting contribute critical real-time data that refine mutation prediction models.

The full value of forward and reverse mutation prediction is realized when these approaches embedded within a broad, multidisciplinary framework that spans host–pathogen studies, human behavior, public health infrastructure, and more. Strengthening the feedback loops between mutation prediction and these interconnected domains will be essential not only for timely

Figure 3



How each strategy may offset the limitation of the other. (1) Immune responses vary significantly among individuals due to genetic diversity, pre-existing immunity, and other factors, which are not reflected under controlled experimental conditions. Population-based data from global surveillance can provide insights into how host factors influence viral evolution. (2) New mutations identified through global surveillance allow researchers to study their effects on epitope changes and immune evasion. (3) Mutations predicted by computational modeling can be further verified by immunological experiments, ensuring accuracy in real-world scenarios. (4) Computational modeling may help explain why certain mutations arise under experimental conditions. (5) The massive numbers of variant sequences and associated metadata from global surveillance can be utilized for sequence-based computational modeling and prediction. (6) Structure-based computational modeling can be used for *in silico* DMS, reducing financial costs [57]. (7) The high-throughput nature of DMS experiments enables the generation of huge data sets for computational modeling.

detection of emerging threats but also for informed, equitable, and effective responses in future pandemics.

Conclusion

SARS-CoV-2 continues to evolve more than 5 years after its emergence, and a multifaceted approach is key to keep pace with its escape of vaccines and therapeutics. Forward mutation prediction through global surveillance and experimental evolution identifies mutations that are actively spreading or may arise under selective pressures. At the same time, reverse mutation prediction through DMS, immunological profiling, and computational modeling enables researchers to anticipate immune escape pathways and gauge potential fitness impacts of new variants. Integrating and validating findings across these predictive methods, alongside robust vaccination strategies, can help reduce the frequency of vaccine updates while preserving their long-term efficacy (Figure 3). Furthermore, embedding mutation prediction within a multidisciplinary approach with insights from host–pathogen biology, behavioral and epidemiological modeling, and health care system readiness enhances our global health resilience and ability to stay ahead of future pandemics.

While new mutation prediction methods continue to improve, comparing models has not been straightforward due to different metrics and learning/test sets used. Head-to-head evaluation of the different components within integrated models will be essential to focus and prioritize research efforts. Nonetheless, the strategies pioneered during the COVID-19 pandemic offer a scalable and generalizable foundation for confronting other rapidly evolving pathogens and developing future-resistant vaccines and therapeutics. Going forward, improving integration and coordination between predictive science, clinical systems, and public health infrastructure will be critical to realizing the full potential of mutation prediction as a cornerstone of pandemic preparedness.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We gratefully acknowledge all data contributors, that is, the authors and their originating laboratories responsible for obtaining the specimens, and their submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based. This work has been supported in part by funding from the Singapore Ministry of Health (MOH) Programme for Research in Epidemic Preparedness and REsponse (PREPARE) fund (PREPARE-CS1-2022-004, PREPARE-OC-VT-2022-001, PREPARE-OC-VT-2024-003). TWY was supported by the Singapore International Graduate Award (SINGA). MZT was supported by a grant from the National Medical Research Council of Singapore (NMRC/OF-YIRG19nov-0059). LR was supported by a Start-up University Grant from the Ministry of Education (#022388-00001).

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. World Health Organization. **COVID-19 cases | WHO COVID-19 dashboard** [Internet]. 2025 [updated 2025; cited 2025 Mar 11]. Available from: ⟨https://data.who.int/dashboards/covid19/cases⟩.

2. Starr TN, Greaney AJ, Hilton SK, Ellis D, Crawford KH, Dingens AS, Navarro MJ, Bowen JE, Tortorici MA, Walls AC, *et al.*: **Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding**. *Cell* 2020, **182**:1295.

3. Plante JA, Liu Y, Liu J, Xia H, Johnson BA, Lokugamage KG, Zhang X, Muruato AE, Zou J, Fontes-Garfias CR, *et al.*: **Spike mutation D614G alters SARS-CoV-2 fitness**. *Nature* 2021, **592**:116-121.

4. Chen RE, Zhang X, Case JB, Winkler ES, Liu Y, VanBlargan LA, Liu J, Errico JM, Xie X, Suryadevara N, *et al.*: **Resistance of SARS-CoV-2 variants to neutralization by monoclonal and serum-derived polyclonal antibodies**. *Nat Med* 2021, **27**:717-726.

5. Buchfink B, Reuter K, Drost H-G: **Sensitive protein alignments at tree-of-life scale using DIAMOND**. *Nat Methods* 2021, **18**:366-368.

6. Tosta S, Moreno K, Schuab G, Fonseca V, Segovia FMC, Kashima S, Elias MC, Sampaio SC, Ciccozzi M, Alcantara LCJ, *et al.*: **Global SARS-CoV-2 genomic surveillance: what we have learned (so far)**. *Infect Genet Evol* 2023, **108**:105405.

7. Khare S, Gurry C, Freitas L, B Schultz M, Bach G, Diallo A, Akite N, Ho J, Tc Lee R, Yeo W, *et al.*: **GISAID's role in pandemic response**. *China CDC Wkly* 2021, **3**:1049-1051.

8. Uludağ M, Incitti R, Gao X, Heeney J, Gojobori T, Alam I: **Global surveillance of emerging SARS-CoV-2 variants**. *Bull World Health Organ* 2024, **102**:374 374A.

9. Ma KC: **Genomic surveillance for SARS-CoV-2 variants: circulation of omicron XBB and JN.1 lineages — United States, May 2023–September 2024**. *MMWR Morb Mortal Wkly Rep* 2024, **73**:938-945.

10. Meijers M, Ruchnewitz D, Eberhardt J, Łuksza M, Lässig M: **Population immunity predicts evolutionary trajectories of SARS-CoV-2**. *Cell* 2023, **186**:5151-5164.e13.

11. Centers for Medicare & Medicaid Services. COVID-19 Monoclonal Antibodies | CMS [Internet]. 2024 [updated 2024; cited 2025 Mar 11]. Available from: ⟨https://www.cms.gov/monoclonal⟩.

12. Gangavarapu K, Latif AA, Mullen JL, Alkuzweny M, Hufbauer E, Tsueng G, Haag E, Zeller M, Aceves CM, Zaiets K, *et al.*: **Outbreak.info genomic reports: scalable and dynamic surveillance of SARS-CoV-2 variants and mutations**. *Nat Methods* 2023, **20**:512-522.

13. Centers for Disease Control and Prevention: Data Modernization Initiative (DMI) [Internet]. 2024 [updated 2024; cited 2025 Apr 20]. Available from: ⟨https://www.cdc.gov/data-modernization/php/about/dmi.html⟩.

14. Onywera H, Ondoa P, Nfii F, Ogwell A, Kebede Y, Christoffels A, Tessema SK: **Boosting pathogen genomics and bioinformatics workforce in Africa**. *Lancet Infect Dis* 2024, **24**:e106-e112.

15. Foster CSP, Walker GJ, Jean T, Wong M, Brassil L, Isaacs S, Lyu Y, Turville S, Kelleher A, Rawlinson WD: Long-Term Serial Passaging of SARS-CoV-2 Reveals Signatures of Convergent Evolution. 2023, doi:⟨10.1101/2023.11.02.565396⟩.

16. Sonnleitner ST, Sonnleitner S, Hinterbichler E, Halbfurter H, Kopecky DBC, Koblmüller S, Sturmbauer C, Posch W, Walder G: **The mutational dynamics of the SARS-CoV-2 virus in serial passages in vitro**. *Virol Sin* 2022, **37**:198-207.

17. • Wei X, Decker JM, Wang S, Hui H, Kappes JC, Wu X, Salazar-Gonzalez JF, Salazar MG, Kilby JM, Saag MS, *et al.*: **Antibody neutralization and escape by HIV-1**. *Nature* 2003, **422**:307-312.

This seminal paper showcases the sequential selection of HIV-1 viral quasispecies within a single individual, driven predominantly by antibody escape. This same theme has now come to be applied to other viruses at a population level.

18. Chung H, Noh JY, Koo B-S, Hong JJ, Kim HK: **SARS-CoV-2 mutations acquired during serial passage in human cell lines are consistent with several of those found in recent natural SARS-CoV-2 variants**. *Comput Struct Biotechnol J* 2022, **20**:1925-1934.

19. Willett JDS, Gravel A, Dubuc I, Gudimard L, dos Santos Pereira Andrade AC, Lacasse É, Fortin P, Liu J-L, Cervantes JA, Galvez JH, *et al*.: **SARS-CoV-2 rapidly evolves lineage-specific phenotypic differences when passaged repeatedly in immune-naïve mice**. *Commun Biol* 2024, **7**:1-13.

20. Ma H, Zong H-F, Liu J-J, Yue Y-L, Ke Y, Liao Y-J, Tang H-N, Wang L, Wang S-S, Yuan Y-S, *et al*.: **Long-term passaging of pseudotyped SARS-CoV-2 reveals the breadth of monoclonal and bispecific antibody cocktails**. *Acta Pharm Sin* 2023, **44**:1455-1463.

21. Carrascosa-Sàez M, Marqués M-C, Geller R, Elena SF, Rahmeh A, Dufloo J, Sanjuán R: **Cell type-specific adaptation of the SARS-CoV-2 spike**. *Virus Evol* 2024, **10**:veae032.

22. Bouzidi HS, Driouich J-S, Klitting R, Bernadin O, Piorkowski G, Amaral R, Fraisse L, Mowbray CE, Scandale I, Escudié F, *et al*.: **Generation and evaluation of protease inhibitor-resistant SARS-CoV-2 strains**. *Antivir Res* 2024, **222**:105814.

23. Stevens LJ, Pruijssers AJ, Lee HW, Gordon CJ, Tchesnokov EP, Gribble J, George AS, Hughes TM, Lu X, Li J, *et al*.: **Mutations in the SARS-CoV-2 RNA-dependent RNA polymerase confer resistance to remdesivir by distinct mechanisms**. *Sci Transl Med* 2022, **14**:eabo0718.

24. Kumar P, Zhang X, Shaha R, Kschischo M, Dobbelstein M: **Identification of antibody-resistant SARS-CoV-2 mutants via N4-Hydroxycytidine mutagenesis**. *Antivir Res* 2024, **231**:106006.

25. Taylor AL, Starr TN: **Deep mutational scanning of SARS-CoV-2 Omicron BA.2.86 and epistatic emergence of the KP.3 variant**. *Virus Evol* 2024, **10**:veae067.

26. Dadonaite B, Crawford KHD, Radford CE, Farrell AG, Yu TC, Hannon WW, Zhou P, Andrabi R, Burton DR, Liu L, *et al*.: **A pseudovirus system enables deep mutational scanning of the full SARS-CoV-2 spike**. *Cell* 2023, **186**:1263-1278.e20.

27. •• Dadonaite B, Brown J, McMahon TE, Farrell AG, Figgins MD, Asarnow D, Stewart C, Lee J, Logue J, Bedford T, *et al*.: **Spike deep mutational scanning helps predict success of SARS-CoV-2 clades**. *Nature* 2024, **631**:617-626.
This paper utilizes pseudovirus-based deep mutational scanning to show that clade growth in the real world could be predicted by identification of mutations that all2ow improved immune escape, receptor binding, or cell entry *in vitro*, thus forming one of the first direct mechanistic links between *in vitro* phenotypes and real-world epidemiology.

28. Starita LM, Fields S: **Deep mutational scanning: a highly parallel method to measure the effects of mutation on protein function**. *Cold Spring Harb Protoc* 2015, **2015**:pdb.top077503.

29. • von Delft F, Ni X, Richardson R, Godoy A, Ferla M, Kikawa C, Scheen J, Hannon W, Capkin E, Lahav N, *et al*.: **Crystallographic fragment screening and deep mutational scanning of Zika virus NS2B-NS3 protease enable development of resistance-resilient inhibitors**. *Res Sq* 2025, https://doi.org/10.21203/rs.3.rs-5876218/v1 rs.3.rs-5876218.
This paper utilizes DMS to map the mutability of epitopes, identifying stable epitopes suitable for targeting by resistance-resilient therapeutics, offering a pathway for usage of a reverse mutation prediction strategy for pre-emptive therapeutics development.

30. Smith DJ, Lapedes AS, de Jong JC, Bestebroer TM, Rimmelzwaan GF, Osterhaus ADME, Fouchier RAM: **Mapping the antigenic and genetic evolution of influenza virus**. *Science* 2004, **305**:371-376.

31. Sitaras I: **Antigenic cartography: overview and current developments**. In *Animal Influenza Virus: Methods and Protocols*. Edited by Spackman E. Springer; 2020:61-68.

32. Wilks SH, Mühlemann B, Shen X, Türeli S, LeGresley EB, Netzl A, Caniza MA, Chacaltana-Huarcaya JN, Corman VM, Daniell X, *et al*.: **Mapping SARS-CoV-2 antigenic relationships and serological responses**. *Science* 2023, **382**:eadj0070.

33. Rössler A, Netzl A, Knabl L, Bante D, Wilks SH, Borena W, von Laer D, Smith DJ, Kimpel J: **Characterizing SARS-CoV-2 neutralization profiles after bivalent boosting using antigenic cartography**. *Nat Commun* 2023, **14**:5224.

34. Mykytyn AZ, Rosu ME, Kok A, Rissmann M, van Amerongen G, Geurtsvankessel C, Vries RD, de Munnink BBO, Smith DJ, Koopmans MPG, *et al*.: **Antigenic mapping of emerging SARS-CoV-2 omicron variants BM.1.1.1, BQ.1.1, and XBB.1**. *Lancet Microbe* 2023, **4**:e294-e295.

35. Dang X, Guelen L, Hulsik DL, Ermakov G, Hsieh EJ, Kreijtz J, Stammen-Vogelzangs J, Lodewijks I, Bertens A, Bramer A, *et al*.: **Epitope mapping of monoclonal antibodies: a comprehensive comparison of different technologies**. *mAbs* 2023, **15**:2285285.

36. Bozkus CC, Brown M, Velazquez L, Thomas M, Wilson EA, O'Donnell T, Ruchnewitz D, Geertz D, Bykov Y, Kodysh J, et al.: T Cell Epitope Mapping Reveals Immunodominance of Evolutionarily Conserved Regions Within SARS-CoV-2 Proteome. 2024, doi:⟨10.1101/2024.10.23.619918⟩.

37. Lomakin YA, Ovchinnikova LA, Terekhov SS, Dzhelad SS, Yaroshevich I, Mamedov I, Smirnova A, Grigoreva T, Eliseev IE, Filimonova IN, *et al*.: **Two-dimensional high-throughput on-cell screening of immunoglobulins against broad antigen repertoires**. *Commun Biol* 2024, **7**:1-13.

38. Thalén NB, Karlander M, Lundqvist M, Persson H, Hofström C, Turunen SP, Godzwon M, Volk A-L, Malm M, Ohlin M, *et al*.: **Mammalian cell display with automated oligo design and library assembly allows for rapid residue level conformational epitope mapping**. *Commun Biol* 2024, **7**:1-11.

39. Wang C, Wang J, Song W, Luo G, Jiang T: **EpiScan: accurate high-throughput mapping of antibody-specific epitopes using sequence information**. *npj Syst Biol Appl* 2024, **10**:1-16.

40. Abousamra E, Figgins M, Bedford T: **Fitness models provide accurate short-term forecasts of SARS-CoV-2 variant frequency**. *PLoS Comput Biol* 2024, **20**:e1012443.

41. • Hie B, Zhong ED, Berger B, Bryson B: **Learning the language of viral evolution and escape**. *Science* 2021, **371**:284-288.
This paper represents a novel language-based approach for codifying viral evolution, using both 'grammar' and 'semantics frameworks from language models to separately represent virus evolutionary fitness and immune escape, respectively.

42. Lee B: **Inferring effects of mutations on SARS-CoV-2 transmission from genomic surveillance data**. *Nat Commun* 2025, **16**:441.

43. Thadani NN, Gurev S, Notin P, Youssef N, Rollins NJ, Ritter D, Sander C, Gal Y, Marks DS: **Learning from prepandemic data to forecast viral escape**. *Nature* 2023, **622**:818-825.

44. Ramachandran A, Lumetta SS, Chen D: **PandoGen: generating complete instances of future SARS-CoV-2 sequences using Deep Learning**. *PLoS Comput Biol* 2024, **20**:e1011790.

45. Nijkamp E, Ruffolo JA, Weinstein EN, Naik N, Madani A: **ProGen2: exploring the boundaries of protein language models**. *Cell Syst* 2023, **14**:968-978 e3..

46. Chi LA, Barnes JE, Patel JS, Ytreberg FM: **Exploring the ability of the MD+FoldX method to predict SARS-CoV-2 antibody escape mutations using large-scale data**. *Sci Rep* 2024, **14**:23122.

47. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L: **The FoldX web server: an online force field**. *Nucleic Acids Res* 2005, **33**:W382-W388.

48. Cia G, Kwasigroch JM, Rooman M, Pucci F: **SpikePro: a webserver to predict the fitness of SARS-CoV-2 variants**. *Bioinformatics* 2022, **38**:4418-4419.

49. Meijers M, Ruchnewitz D, Eberhardt J, Karmakar M, Łuksza M, Lässig M: **Concepts and Methods for Predicting Viral Evolution**. 2024, doi:⟨10.48550/arXiv.2403.12684⟩. This paper presents a set of methods for data-driven predictive analysis of viral evolution,

providing a framework for integration of multiple data sources that is generalizable across viruses.

50. Wang D, Huot M, Mohanty V, Shakhnovich EI: **Biophysical principles predict fitness of SARS-CoV-2 variants**. *Proc Natl Acad Sci USA* 2024, **121**:e2314518121.

51. Lou J, Liang W, Cao L, Hu I, Zhao S, Chen Z, Chan RWY, Cheung PPH, Zheng H, Liu C, *et al*.: **Predictive evolutionary modelling for influenza virus by site-based dynamics of mutations**. *Nat Commun* 2024, **15**:2546.

52. Huddleston J, Barnes JR, Rowe T, Xu X, Kondor R, Wentworth DE, Whittaker L, Ermetal B, Daniels RS, McCauley JW, *et al*.: **Integrating genotypes and phenotypes improves long-term forecasts of seasonal influenza A/H3N2 evolution**. *eLife* 2020, **9**:e60067.

53. Lee JM, Huddleston J, Doud MB, Hooper KA, Wu NC, Bedford T, Bloom JD: **Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human H3N2 influenza variants**. *Proc Natl Acad Sci USA* 2018, **115**:E8276-E8285.

54. •• Raharinirina NA, Gubela N, Börnigen D, Smith MR, Oh D-Y, Budt M, Mache C, Schillings C, Fuchs S, Dürrwald R, *et al*.: **SARS-CoV-2 evolution on a dynamic immune landscape**. *Nature* 2025, **639**:196-204.

This paper synthesized a comprehensive mechanistic model incorporating deep mutational scanning and regional genomic surveillance data to accurately predict shifts in variant dynamics at a regional level, thus showcasing that given sufficient training input, computational models can achieve very fine accuracy and relevance to real world situations.

55. Saldivar-Espinoza B, Macip G, Garcia-Segura P, Mestres-Truyol J, Puigbò P, Cereto-Massagué A, Pujadas G, Garcia-Vallve S: **Prediction of recurrent mutations in SARS-CoV-2 using artificial neural networks**. *IJMS* 2022, **23**:14683.

56. Zhang Y, Chen S, Tian Y, Fu X: **Host factors of SARS-CoV-2 in infection, pathogenesis, and long-term effects**. *Front Cell Infect Microbiol* 2024, **14**:1407261.

57. Tsai K-C, Lee Y-C, Tseng T-S: **Comprehensive deep mutational scanning reveals the immune-escaping hotspots of SARS-CoV-2 receptor-binding domain targeting neutralizing antibodies**. *Front Microbiol* 2021, **12**:698365.