Computational and Structural Biotechnology Journal 27 (2025) 1370–1382

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Review Article

Predicting pathogen evolution and immune evasion in the age of artificial intelligence

D.J. Hamelin [a,b,c], M. Scicluna [a,b,c], I. Saadie [a,c], F. Mostefai [a,b,c], J.C. Grenier [a], C. Baron [a,b,c], E. Caron [d,e], J.G. Hussin [a,b,c,f,*]

[a] Montreal Heart Institute, Université de Montréal, Montréal, Quebec, Canada
[b] Mila - Quebec AI Institute, Montréal, Quebec, Canada
[c] Department of Biochemistry and Molecular Medicine, Faculty of Medicine, Université de Montréal, Montréal, Quebec, Canada
[d] CHU Sainte-Justine Research Center, Université de Montréal, Montréal, Quebec, Canada
[e] Yale Center for Immuno-Oncology, Yale Center for Systems and Engineering Immunology, Yale Center for Infection and Immunity, Yale School of Medicine, New Haven, CT, USA
[f] Department of Medicine, Faculty of Medicine, Université de Montréal, Montréal, Quebec, Canada

ARTICLE INFO

Keywords:
Viral evolution
Viral forecasting
Bioinformatics
Machine learning
Language models
Pandemic preparedness

ABSTRACT

The genomic diversification of viral pathogens during viral epidemics and pandemics represents a major adaptive route for infectious agents to circumvent therapeutic and public health initiatives. Historically, strategies to address viral evolution have relied on responding to emerging variants after their detection, leading to delays in effective public health responses. Because of this, a long-standing yet challenging objective has been to forecast viral evolution by predicting potentially harmful viral mutations prior to their emergence. The promises of artificial intelligence (AI) coupled with the exponential growth of viral data collection infrastructures spurred by the COVID-19 pandemic, have resulted in a research ecosystem highly conducive to this objective. Due to the COVID-19 pandemic accelerating the development of pandemic mitigation and preparedness strategies, many of the methods discussed here were designed in the context of SARS-CoV-2 evolution. However, most of these pipelines were intentionally designed to be adaptable across RNA viruses, with several strategies already applied to multiple viral species. In this review, we explore recent breakthroughs that have facilitated the forecasting of viral evolution in the context of an ongoing pandemic, with particular emphasis on deep learning architectures, including the promising potential of language models (LM). The approaches discussed here employ strategies that leverage genomic, epidemiologic, immunologic and biological information.

Contents

* Corresponding author at: Montreal Heart Institute, Université de Montréal, Montréal, Quebec, Canada.
E-mail address: julie.hussin@umontreal.ca (J.G. Hussin).

https://doi.org/10.1016/j.csbj.2025.03.044
Received 6 December 2024; Received in revised form 21 March 2025; Accepted 26 March 2025
Available online 28 March 2025

1. Forecasting viral evolution: a complex challenge

Viral pandemics have proven to be tremendous challenges responsible for significant loss of life, thus warranting extensive research into viral biology, therapeutics strategies, and public health interventions. The SARS-CoV-2 pandemic is the most thoroughly recorded pandemic in human history both genetically and epidemiologically, making it pivotal in facilitating our preparedness to future viral pandemics. One pandemic-preparedness strategy enabled by such extensive databases lies in developing frameworks to forecast the evolution of viruses and identify potentially harmful evolutionary events over the course of a pandemic. Such frameworks would be substantially beneficial as they would permit the anticipation of mutations and strains capable of circumventing human immunity and public health interventions. Nevertheless, anticipating viral evolution has proven to be an exceedingly challenging task, characterized by many considerations spanning biological, epidemiological, and social expertise.

One such consideration, made evident throughout the COVID-19 pandemic, lies in the breadth and constitution of available viral datasets. Sampling biases were found to heavily skew analyses and interpretations pertaining to SARS-COV-2 evolution and epidemiology, such as biases stemming from variations in sequencing capacities across countries [1]. Yet another limitation lies in data scarcity. Unlike SARS-CoV-2, the majority of viruses lack sufficient data to train adequately informed predictive models, although certain strategies sought to address this challenge by learning across viral families.

Beyond difficulties related to available data, the complexity of evolutionary dynamics poses nontrivial challenges. While it is feasible to investigate the evolutionary advantage of single mutations or combinations of mutations, the anticipation of multi-mutation variants remains challenging. For example, complex variants resulting from recombination events, extensive viral evolution within immunocompromised individuals, or animal reservoirs largely remain beyond the reach of current predictive methodologies. These constitute some of the limitations associated with predicting viral evolution. However, efforts have been made to overcome these challenges and narrow the gap in predictive capabilities.

In this review, we provide a state-of-the-art perspective on viral evolution and fitness prediction in the post-pandemic era, focusing on how big data and AI-driven approaches have transformed this field. Specifically, we discuss: (i) the various contributors to viral fitness as they relate to forecasting viral evolution; (ii) the essential role of abundant, high-quality data, bolstered by recent advancements in viral data collection and sharing; (iii) recent innovations in computational strategies for forecasting viral evolution, with a particular emphasis on phylogenetic-based frameworks, deep learning (DL), and protein language models; and (iv) the adaptability of these frameworks across viral species, with a focus on their translation into actionable insights to guide public health initiatives (Fig. 1). The unprecedented scale of sequencing and AI development during the COVID-19 pandemic has driven many of the methods discussed here, accelerating progress in viral forecasting. While many of these approaches were developed with SARS-CoV-2 evolution in mind, their design enables broader adaptability across RNA viruses, with several strategies already applied to multiple species. Looking forward, these advances set the stage for a more proactive and data-driven approach to pandemic preparedness.

2. Viral evolution and fitness

The primary task explored in this review pertains to forecasting viral evolution, a process characterized by genetic changes over time. The genetic diversification resulting from viral evolution is caused by mechanisms such as replication errors, recombination, genetic drift, bottleneck, natural selection and host pressures [2–4]. These crucial biological processes are responsible for the selection and fixation of viral mutations, and for viral variants. An integral component of forecasting viral evolution lies in achieving a thorough understanding of viral fitness, which encompasses the ability of a virus to survive, replicate, and transmit within a host population [5]. Although viral fitness is not synonymous with viral evolution, it consists of one of its key determinants. As such, while some of the methodologies discussed here solely predict viral fitness, others integrate it within larger schemes aimed at interrogating and predicting overall evolution. Developing a comprehensive understanding of the mutational landscape that drives viral fitness can provide insight into vaccine development. Furthermore, it enables the early flagging of Variants of Concern (VOC), which are viral strains with mutations that enhance transmissibility, virulence, or immune escape.

2.1. Defining viral fitness

The concept of fitness has been characterized by multiple definitions in the context of viruses. Wargo and Kurath [5] outlined three types of fitness pertaining to viruses: **replicative fitness**, which refers to the ability of a virus to produce infectious offspring in a specific environment; **transmission fitness**, which encompasses the ability of a virus to successfully spread across hosts in a population; and **epidemiologic fitness**, describing the capacity of a virus to become the dominant strain among competing variants in a population (Fig. 2A). Replicative and transmission fitness can be described using molecular mechanisms and are considered to be causal to pathogenicity. In contrast, epidemiological fitness results from the summation of replicative and transmission fitness, while heavily affected by social patterns across host populations. Its investigation is therefore largely observation-driven, relying on the geo-temporal distributions of viral strains. Importantly, these types of fitnesses reflect different stages in the life cycle and evolutionary process of a virus, and will therefore be regulated by distinctive selective pressures. Although replicative fitness is essential to the viral life cycle, excessive replication might increase susceptibility to immune responses, thereby driving immune evading adaptations. In contrast, transmission fitness is controlled by multiple factors that include host immunity and host cell entry. As such, mutations enabling immune evasion or enhancing host cell entry are known to contribute to transmission factors. Finally, epidemiological fitness relies upon a balance between both replicative fitness and transmission fitness, while being shaped by external factors such as host populational structure and behavior.

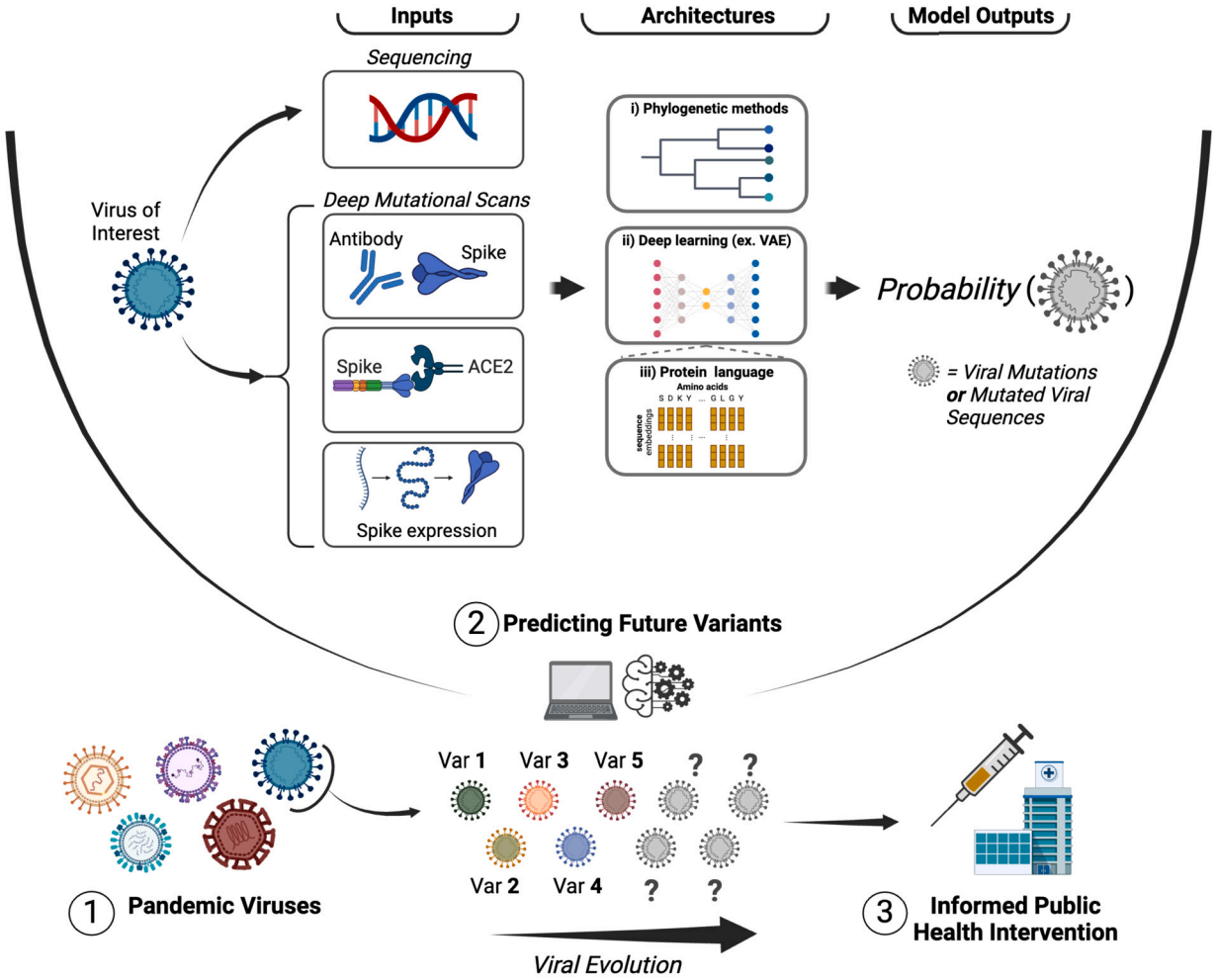

Fig. 1. Overview of key concepts and approaches. **(1)** The aim of this review consists of exploring recent frameworks aimed at anticipating the evolution of viruses posing pandemic threats. **(2)** The approaches discussed leverage diverse inputs which include genomic sequences as well as functional biological data acquired from high throughput Deep Mutational Scans. The examples shown here are specific to SARS-CoV-2 biology. A variety of computational architectures are discussed, including *i)* phylogenetic trees, *ii)* deep learning, and *iii)* specifically a subset of deep learning methods that utilize Protein Language Models. While varied in inputs and architectures, the methods discussed generally aim to predict the probability of a mutated sequence occurring. **(3)** The aim of the methods described in this review consists of generating insight pertaining to viral evolutionary trajectories to guide vaccine development and guide public health decision-making.

2.2. Main drivers of viral evolution

Developing a thorough understanding of the relationship between fitness and viral evolution is crucial to both identifying mutational drivers as well as forecasting viral evolution. All three types of fitness (replicative, transmission, and epidemiological) have contributed to the investigation of viral mutational drivers, and have been exploited by the predictive tools discussed in this review. Importantly, the development of high throughput experimental frameworks such as Deep Mutational Scans (DMS) have significantly facilitated the investigation of mutational drivers across various viruses and viral proteins [6]. DMS is a technique that systematically evaluates the functional impact of a large number of mutations across viral proteins (see section 3.2 for additional details). These regions often correspond to key viral functions such as host-cell entry, immune evasion, as well as viral replication efficiency (Fig. 2B), reflecting strong natural selection. The mutational drivers of SARS-CoV-2 have been extensively investigated during the pandemic, making it a well-documented case study.

Following SARS-CoV-2 entry into a host, the virus binds to the Angiotensin Converting Enzyme 2 (ACE2) receptor on the cell membrane of a host cell [7]. This binding occurs on the Receptor Binding Domain (RBD) region of the spike glycoprotein (Spike), and was found to be enhanced by factors such as the cleaving of the Furin site on the Spike protein as well as Spike conformational changes [8]. The appearance of prevalent mutations shown to improve these mechanisms established the importance of host cell entry as a prominent mutational driver in SARS-CoV-2 evolution [9]. Examples include Alpha mutations N501Y and P681H in the Spike protein, which improved RBD-ACE2 binding and Furin cleaving respectively [10,9]. Beyond cell entry, immune evasion has been identified as a strong driver of SARS-CoV-2 evolution [11]. Neutralizing antibodies were shown to play an important role in SARS-COV-2 evolution, with mutations enabling antibody evasion identified in most prevalent strains [11–15]. Examples include Spike mutations L452R and T478K found in the highly virulent Delta strain, as well as a large proportion of mutations characterizing the Omicron strain [11–13]. Importantly, one of the roles played by neutralizing antibodies consists of preventing cell entry through the disruption of the RBD-ACE2 binding interface. As such, host cell entry and antibody evasion were shown to be highly connected evolutionary processes, an interdependence described in other viruses such as Influenza A virus [16]. Other immune evasion mechanisms include the disruption of T cell response as well as the disruption of various immune cell signaling pathways [17–20]. While the evolutionary mechanisms described above primarily impact transmission fitness, mutations known to improve replicative

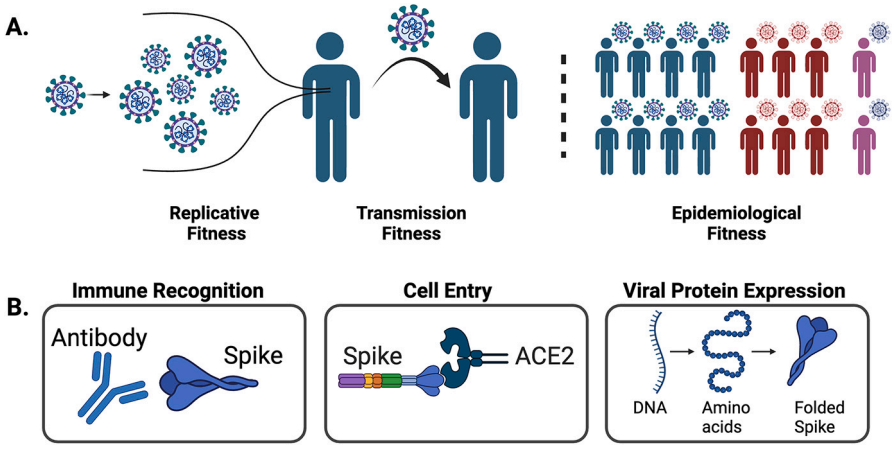

Fig. 2. Overview of viral fitness and evolutionary drivers. **A**. Viral fitness can be defined with respect to replicative fitness (left), transmission fitness (middle), and epidemiological fitness (right). **B**. Examples of evolutionary drivers specific to SARS-COV-2. These include Immune recognition of the SARS-CoV-2 Spike Glycoprotein by antibodies (left); entry of SAR-CoV-2 into host cells, mediated by the binding of the Spike Glycoprotein to the human ACE2 protein (middle); and the expression efficiency of viral proteins such as the Spike Glycoprotein.

fitness have been identified (Fig. 2B). These include the Spike D614G as well as the ORF1b P323L mutations found in SARS-CoV-2, both of which were associated with increased viral replication. While the examples described here were investigated in the context of SARS-CoV-2, mechanisms pertaining to viral entry, immune evasion, and viral replication have been identified as important selective pressures in numerous viruses [21,22]. Overall, these examples impose important constraints on viral evolution and therefore constitute essential considerations for frameworks predicting viral evolution.

3. The power of big data

Studies pertaining to viral evolution, fitness, and epidemiology have been expedited by advances in data collection and sharing, a trend befitting the age of Big Data. As demonstrated during the COVID-19 pandemic, a thorough understanding of viral evolution necessitates the collection of both viral genomic sequences as well as functional data. Viral genomic sequences, particularly when collected in a geographically and temporally distributed fashion, enable the investigation of viral evolution and epidemiology [23–25]. In contrast, functional data acquired through experimental assays provide information pertaining to the biological relevance of viral mutations [26–30]. Both categories of data-collection schemes have significantly evolved in recent years, enabling the rapid detection and response to novel virulent SARS-CoV-2 strains during the COVID-19 pandemic. As access to sufficient quantities of high-quality data constitutes a significant bottleneck to predicting viral evolution, we will discuss developments in such data-collection schemes. Specifically, we will focus on *i)* the collection and sharing of viral genomic data and *ii)* high-through functional assays such as DMS.

3.1. Data collection platforms

Viral genomic surveillance has proven to be a crucial component to the management of viral epidemics and outbreaks. Effective genomic surveillance relies on data-sharing platforms aimed at tracking viral evolution and epidemiology on a global setting. Prominent genomic data-sharing platforms include the National Center for Biotechnology Information (NCBI), as well as the Global Initiative on Sharing All Influenza data (GISAID) [31]. These platforms have enabled the collection and sharing of global genomic sequences spanning multiple viruses, and have been significant catalyzers for worldwide research pertaining to viral evolution. The power of genomic surveillance was showcased throughout the recent COVID-19 pandemic, which has resulted in the rapid adaptation and drastic expansion of such data-sharing platforms. Specifically, GISAID became the largest amassment of genomic data belonging to a single virus, characterized by over 15 million genomic sequences collected globally over 4 years. This quantity of data constitutes a detailed geo-temporal account of the genomic evolution of SARS-COV-2, facilitating the development of highly data-rich fitness models. In addition to leveraging existing databases, the COVID-19 pandemic also resulted in the inception of novel data-sharing platforms. These included Pathoplexus, VirusSeq, and the discontinued COVID-19 GENOMICS UK CONSORTIUM (COG-UK) [32,33]. Notably, these initiatives shed light on the importance of data accessibility as well as stringent data quality controls. The Canadian initiative VirusSeq is an example of a data-sharing platform that sought to address both challenges [33]. Overall, these have demonstrated the ability of the scientific community to rapidly adapt to novel viral threats and generate highly specific data platforms, enabling the in-depth investigation of viral evolution in the context of epidemics and global pandemics. Importantly, the expansion in viral data collection and sharing observed during the COVID-19 pandemic has facilitated the development of computational frameworks aimed at predicting viral evolution.

3.2. Deep mutational scans

The development of high-throughput functional assays has enabled the interrogation of a large number of viral mutations, generating insight regarding viral fitness as well as protein functions. One such class of methods, Deep Mutational Scans (DMS), have played an essential role in deepening our understanding of viral evolutionary trajectories. They have effectively facilitated the interrogation of the mutational landscape of pathogens in the context of viral fitness, transmission, and immune evasion [26–30]. Early applications include the use of DMS to investigate the Influenza A viral protein Hemagglutinin (HA) by interrogating the impact of virtually all possible amino acid substitutions on viral fitness and immune recognition [26,28,29]. Briefly, deep mutational scans employ extensive barcoded mutant libraries that can be concomitantly assayed for multiple phenotypes. The tracking of specific mutant-phenotype can be subsequently conducted through deep sequencing [27]. Due to the ability of DMS to simultaneously probe the impact of large numbers of mutations on various viral functionalities, this approach proved to be a promising avenue for predicting viral evolution [34]. An early study by Lee et al. yielded comparative analyses of the mutational tolerance across HA proteins of various closely related Influenza strains [34]. This work suggested preferred mutational sites, thus providing valuable insight into the evolutionary trajectories of viral lineages.

The high-throughput component of DMS proved to be instrumental during the COVID-19 pandemic, by rapidly generating extensive muta-

Table 1
List of tools and studies aiming to predict viral evolution, with an indication of the viruses they were applied to. *All methods return per-mutation estimates.

Category		Name	General Approach	Predicted Feature(s)*	Virus Studied
Phylogenetic-based methods		ENCoM [39]	Molecular dynamic simulations	Structural stability changes	SARS-CoV-2
		PyR_0 [40]	Hierarchical Bayesian multinomial logistic regression	Mutation-driven lineage expansion	SARS-CoV-2
		Maher et al. 2022 [41]	Various methodologies	Driver mutations	SARS-CoV-2
		Bloom et al. 2024 [42]	UShER-based phylogenetic analysis	Mutational fitness effects	SARS-CoV-2
Energy-Based Models	**Non-LM**	Rodriguez-Rivas et al. 2022 [43]	Direct coupling analysis (DCA)	Epistatic fitness effects	SARS-CoV-2
		EVEscape (EVE) [44]	VAE; Gaussian mixture model	Mutation fitness and escape potential	SARS-CoV-2, influenza, HIV, Lassa, Nipah
		VPRE [45]	VAE; Gaussian Process	Chronological trajectories of protein evolution	SARS-CoV-2
		Hie et al. 2021 [46]	biLSTM	Mutational fitness landscapes	SARS-CoV-2, Influenza, HIV
	LM	EVEscape (TranceptEVE) [47]	Autoregressive transformer; VAE	Mutation fitness and escape potential	Multiple
		MLAEP [48]	Multi-task learning	Antigenic evolutionary trajectories	SARS-CoV-2
		CoVFit [49]	Multi-task learning	Viral fitness and escape	SARS-CoV-2
		TEMPO [50]	Transformer	Mutational impact scores	SARS-CoV-2, Influenza
		PRIEST [51]	Transformer	Mutational impact scores predictions	SARS-CoV-2
		Beguir et al. [52]	Transformer, structural modeling	Structural stability and binding affinity	SARS-CoV-2

tional datasets [35,36,6,37]. Studies applying this strategy investigated the impact of mutational events on a variety of biological features, including the SARS-CoV-2 Spike Glycoprotein (S protein) expression [6,38]; viral entry into cell hosts via the interaction between the S protein Receptor Binding Domain (RBD) and the human angiotensin converting enzyme 2 (ACE-2) [6,38]; and viral recognition with neutralizing antibodies [35,36,12]. Together, these investigations further established our understanding of prominent SARS-CoV-2 selective pressures, including the S-ACE2 binding interface as well as antibody evasion (non-exhaustive). However, the contributions of DMS experiments extend far beyond their direct identification of evolutionary selective pressures. The extensive array of DMS conducted on SARS-CoV-2 as well as other viruses has resulted in a highly comprehensive and ever-growing collection of mutation-phenotype interactions spanning numerous proteins and viruses. These datasets have been instrumental in both training and validating computational models aimed at forecasting pathogen evolutionary outcomes.

4. Computational advancements in forecasting viral evolution

In addition to experimental methods such as DMS, *in-silico* methods have shifted the paradigm of evolutionary and mutagenic studies. In particular, Machine Learning (ML) methods can leverage an unprecedented quantity of data, from collection, curation, and sharing efforts, towards forecasting viral evolution. In general, these methods estimate viral evolution by interrogating viral fitness, although they may implicitly or explicitly assume different definitions of fitness. For example, the epidemiological fitness of viral lineages may be inferred from mutation frequencies derived from a phylogenetic tree. Phylogenetic trees, where the length of branches reflect genetic divergence over time, enable the interrogation of evolutionary relationships. These effectively allow the identification of mutations responsible for successful viral lineages.

Other methods estimate fitness using energy-based representation of viral sequence landscape, where the fitness of a mutation can be estimated using the change in energy resulting from that mutation. Higher fitness represents higher change in energy. Probabilistic methods can estimate fitness *indirectly* using their probability assignments $P(Reference)$ and $P(Variant)$, which reflect the likelihood of observing the reference and mutated sequences, respectively, according to the model.

Typically, the fitness of a variant is estimated the following way:

$$Fitness = Energy(Reference) - Energy(Variant) \quad (1)$$

where $Energy \propto P^{-1}$, meaning sequences with higher probabilities correspond to lower-energy states. The probabilistic methods discussed largely operate by interrogating the sequence space either intra-species [50,51,45], across viral families [46,44,43], or by leveraging insight learned across domains of life [48]. Such models have been empirically shown to encode biological structure [43,46,53], suggesting that they can also encode replication and transmission fitness information. Since these affect the processes guiding viral evolution, such probabilistic methods may be well suited for predicting evolutionary forces such as immune evasion or therapeutic escape. Although certain approaches effectively inferred immune escape directly from the sequence space [46], others interrogated such phenomena using additional strategies. These include the incorporation of DMS data directly into learning architectures [48] as well as the identification of antibody evasion through surface accessibility metrics [44]. Furthermore, methodologies leveraging intra-species data likely capture epidemiological fitness alongside other biological factors [50,51,45], as estimates reflect mutation distributions in the training data, mirroring variant prevalence and success within a population.

One crucial aspect of predicting viral evolution lies in distinguishing between forecasting single mutations (gradual evolution) and forecasting combinations of mutations (saltation events). Although viral evolution can occur gradually, saltation events have been observed. Such events include the inception of variants in immunosuppressed patients as well as animal reservoirs. While the methods presented in this review generally aim to estimate the evolutionary advantage of single mutations, probabilistic energy-based models are particularly well suited to handling combinations of mutations [44,46,54,48]. These methods achieve this by allocating probability assignments to sequences, whether unmutated or mutated at multiple positions. In certain cases, such methods can be utilized to explore feasible mutational trajectories of existing variants [48]. Ultimately, the probabilistic methods discussed are best suited for assessing the selective advantage of viral mutations, which can act as a proxy for predicting the epidemiological success of variants within a specific host population. This was shown by several studies discussed below, wherein predictions were able to forecast with vary-

ing success the appearance of dominant variants. Nevertheless, directly interrogating the epidemiological success of variants is more effectively addressed through phylogenetic analyses.

In this section, we discuss both statistical methods leveraging phylogenetic-based frameworks as well as machine learning models designed to interrogate and forecast viral evolution (summarized in Table 1). In the way of machine learning models, we emphasize on methods such as Variational Autoencoders (VAEs) [44,45] and Language Models (LMs) [46,54] that can fit larger datasets to more flexible models, capturing more aspects of biology and leading to better estimates of viral fitness and evolution. Notably, all methods discussed here differ in whether they rely on "pandemic" or "pre-pandemic" data. Pandemic data corresponds to data collected over the course of an epidemic or pandemic, and is specific to the relevant virus (ex. SARS-CoV-2). In contrast, pre-pandemic data belongs to virus of the same family or genus and can provide evolutionary information about a virus of interest. In the case of the SARS-CoV-2 virus, it is data of coronoviridae family viruses such as the middle east respiratory syndrome (MERS) or sarbecoviruses [44].

4.1. Phylogenetic and other statistical models for viral fitness estimation

Several approaches utilize phylogenetic methodologies to investigate evolutionary trajectories and predict the impact of mutational events on viral fitness. These methods mainly seek to interrogate temporally-distributed viral sequences to genomic regions or positions under positive selection. In this regard, multinomial logistic regressions have proven successful in temporally assessing the fitness advantage of competing strains [55–57]. Such an approach was leveraged in PyR_0 to estimate the fitness of a mutation directly by fitting logistic growth curves from mutation frequencies. This can enable the estimation of fitness as it pertains to novel genomic mutational events [58,40,41]. Such methods successfully follow the genomic evolution of SARS-CoV-2 while identifying genomic positions associated with lineage expansion [40,41]. While highly varied in methodologies, these methods share their use of observed evolutionary trajectories to assess the epidemiological fitness of SARS-CoV-2 mutations. Importantly, these methods enable the capture of recurrent mutations and, as a result, convergent evolution. Convergent evolution is an important evolutionary concept, and refers to an evolutionary adaptation occurring independently in separate instances as a result of strong selective pressure [59]. Convergent evolution was shown to play an important role in SARS-CoV-2 evolution, as exampli-fied by the recurrence of the Spike N501Y mutation in multiple variants [60]. An additional advantage to this approach lies in their ability to interrogate factors of epidemiological fitness that go beyond viral biology, such as the impact of travel on variant frequencies [58]. Applied to SARS-CoV-2, these methods were able to recapitulate the increase in the viral fitness of variants observed throughout the COVID-19 pandemic.

Nevertheless, these studies primarily leverage observed evolutionary history and identify mutations of interest based on their presence or absence at the base of a viral lineage. As such, they are better suited to identify mutations beneficial to viral epidemiological viral fitness. However, their capacity to clearly differentiate between beneficial, neutral, and detrimental mutations is more limited. This was addressed in a study by Bloom and colleagues [42]. This genome-wide fitness-estimation method sought to compare the number of expected mutational events based on neutral evolutionary rates to observed mutational events along the phylogeny of SARS-CoV-2. By relying on observed mutations along the phylogeny, this method provides additional insight pertaining to the deleterious and neutral effects of mutations. One significant advantage of phylogenetic approaches over experimental strategies such as DMS lies in their ability to generate predictions for the full scope of viral proteins. This advantage was showcased by this method through the generation of mutation effect predictions extending beyond the relatively small selection of proteins interrogated by DMS experiments [42]. One advantage of this method lies in its ability to leverage large amounts of genomic data. Having been trained on millions of SARS-CoV-2 sequences, this model recapitulated DMS measurements more closely than other models [44,43,41], thus demonstrating its superior accuracy in the context of data-rich virus such as SARS-CoV-2. Nevertheless, the success of this model relies heavily on access to large quantities of virus-specific sequencing data. This effectively makes it better suited to later stages of pandemics rather than to emerging viruses.

Other non-phylogenetics statistical methods have been proposed. For instance, Rodriguez-Rivas and colleagues [43] uses a Markov Random Field fit to SARS-CoV-2 sequence data to estimate fitness from estimated probabilities, and is able to model coupling between pairs of residues in a sequence. Such an approach introduces an epistatic model that predicts mutable sites in SARS-CoV-2 proteins by leveraging Direct Coupling Analysis (DCA) on multiple sequence alignments (MSAs) of coronaviruses. The proposed method relies on genomic context across viral relatives, applicable when limited information specific to a virus of interest is available. The resulting DCA-based model for SARS-CoV-2, combined with MSA of genomes from Corovanidae family members and pandemic-specific SARS-CoV-2 evolutionary data incorporates insight pertaining to evolutionary constraints while identifying genomic regions under positive selection. This method confers the advantage of requiring a single viral sequence, provided genomic data from other members of its viral family are accessible. As such, it is well suited for use with emerging viruses, particularly when population sequencing data is scarce.

4.2. Variational autoencoders for viral evolution forecasting

While phylogenetic and statistical models have proven highly effective in estimating the fitness impact of mutations, advances in deep learning architectures have opened novel avenues to interrogate viral evolution. Indeed, deep learning has played a pivotal role in the analysis of biological data [61,62].

In recent years, the investigation of mutation effects on protein structure and function has been a central focus for the scientific community, closely tied to the concept of variant effect prediction. This task shares similarities with viral forecasting by evaluating the functional consequences of genetic variants. Variant Effect Predictors (VEP) consist of a class of computational tools that leverage biological data combined with a variety of model architectures to interrogate the functional impact of genetic variants [63]. Such methods also use probabilistic models to quantify mutation effects using equation (1). Several deep learning-based VEP have been developed in the context of human genetics, with the objective to identify pathogenic mutations relevant to human-centric diseases [64–70].

One such example is EVE (evolutionary model of variant effect). EVE is a Variational Autoencoder (VAE), an unsupervised generative model, trained on protein family-specific MSAs [70]. The model is interpreted through a Gaussian Mixture Model, used to estimate the pathogenicity of human genetic variants. While this method was originally intended for human genetic variants, a subsequent study adapted the framework for viral evolution [44]. The resulting tool, EVEscape, combined EVE with additional scores reflective of antibody evasion as well as changes in surface amino acid chemistry. Learning from MSAs specific to viral families of interest, EVEscape was shown to present state-of-the-art viral evolution forecasting, with the tool recapitulating SARS-CoV-2 pandemic evolution with significant accuracy. Beyond estimating single mutants, EVEScape was shown to estimate the evolutionary advantage of multi-mutation variants, recapitulating the selective advantage of observed Variants-Of-Concerns. Further analysis of forecasted SARS-CoV-2 variants by neutralizing assays demonstrated the ability of predicted mutations to evade mRNA booster vaccine-induced antibodies [71]. Importantly, EVEScape was shown to be adaptable across viral species, with predictions shown to recapitulate observed evolutionary trajectories for Influenza, Human Immunodeficiency Virus (HIV), Lassa and Nipah viruses [44] In a separate SARS-CoV-2-specific study [45], King

and colleagues use a Gaussian Process to predict and generate future viral genomic sequences based on the lower-dimensional representations learned by the VAE. They utilize a VAE to encode temporally distributed SARS-COV-2 genomic sequences into three-digit matrices. Manipulating and processing low-dimensional matrices is much less computationally costly than full viral genomes and encoded sequences, thus streamlining downstream analyses.

4.3. Protein language models to learn viral evolution

Recently, a class of machine learning models called language models (LM) have been applied to protein sequences for VEP and viral fitness estimation. Formally, a LM is a probability distribution over sequences of m words from lexicon $\mathcal{W}$:

$$P(X_1, \ldots, X_n), \quad X_i \in \mathcal{W} \tag{2}$$

While the earliest LM can be traced to early developments in probabilistic modeling, recent advances in neural network-based language models originated in foundational work on deep learning [72].

Such models use a neural network f_θ where θ are the learnable parameters to model the probability distribution of any word given it's context. Technically, the input is a distributed vector representation per input word, called a "word embedding". Word embeddings were traditionally computed independently of sequential context, using methods such as Word2Vec [73]. Methods that used contextual information to learn embeddings, such as ELMo [74], were subsequently found to have much better performance on downstream Natural Language Processing (NLP) tasks.

Intuitively, training a LM is akin to taking many Cloze tests. Cloze tests involve filling in missing words in a sentence. For example, given the sentence "The puppy played with the __" the LM predicts the missing word based on context. These models are then evaluated by "fine-tuning" them to do standard NLP tasks and benchmarking their performance [75]. In viral genomics, they are trained at predicting genomic sequences. This can be done for example by determining the likelihood of a specific nucleotide subsequence based on sequence context "(e.g. 'AAGTCG__TAG')." In recent years, LMs with billions of parameters have been trained on huge corpus's of natural text, demonstrating emergent capabilities [76,77]. Recent "pre-trained" LMs have been adapted to perform complex tasks including generating realistic news articles [78] and solving undergraduate-level homework problems [79]. This concept of pre-training originally applied to natural language, has been extended to genomic sequences, where models learn from large-scale sequence data to perform tasks such as predicting sequence function. For a more comprehensive review of LMs in general, we direct the reader to other published reviews [72].

Given the performance gain of LMs across a wide variety of NLP tasks, it is not surprising that there exists several attempts to use LMs to solve problems within biology [80–82]. Often these models operate on the protein level (where $|\mathcal{W}| = 20$ amino acids). LMs applied to protein sequences, known as Protein Language Models (PLM), aim to unveil the evolutionary and biological rules governing protein function, domain organization and selective pressures. Studies exploring this strategy were distinctly facilitated by the establishment and expansion of massive protein sequence databases such as UniREF [83]. Early implementations PLMs involved the use of Long Short-Term Memory (LSTM) networks applied to unlabeled protein sequences [84,85]. Such methods aim to distill vast arrays of inter-protein evolutionary and functional insight into embeddings that could in turn be interrogated with a variety of downstream machine learning models. The LSTM-based approach to language models was employed to explore the fitness landscape of a variety of viruses, including SARS-CoV-2 [86]. In this study, Hie and colleague train a bi-directional LSTM (BiLSTM) which was trained on viral sequences to predict viral evolution. Specifically, the authors highlight that grammar and semantic properties of natural language, corresponding to rules and meaning of text, are useful analogs of viral fitness and antigenic change respectively. By interrogating the output of the LM, this approach allows for the assessment of the viral sequence space to quantify fitness and immune evasion.

While previous LSTM models were trained on massive and diverse protein sequence databases, this model can be trained on virus-specific datasets, and three distinct models were trained for the Influenza A hemagglutinin (HA), HIV envelope glycoprotein (Env) and the SARS-CoV-2 spike glycoprotein (Spike). In all three cases, the respective model could recapitulate regions enriched in fitness and antigenic escape potential. LSTM models paved the way to numerous protein language investigations. Several of these took advantage of the development of the transformer architecture and its success in language models [87,88]. Initial implementations of such an approach trained transformers on >200 million protein sequences to unveil valuable information regarding protein biological properties, homology, secondary and tertiary structures, as well as Protein-Protein Interactions (PPI) [89,90,53]. Studies aimed at utilizing transformers to investigate protein evolutionary relationships can be categorized into two strategies: those learning from unaligned sequences [91,92] and those learning from MSAs [93]. A transformer model trained on 250 million unaligned sequences (UniRef100 [83]) was implemented by Notin and colleagues to shed light on protein fitness [47]. This model, named Tranception, sought to address the short-comings of MSA-based models, namely limitations pertaining to the depth and diversity of MSA alignments. Indeed, MSA-based models were found to have lesser performance with protein families characterized by shallow alignments. Tranception addresses this by employing an autoregressive transformer, which predicts the identity of each amino acid in a sequence based on the context provided by all preceding positions. This step-by-step prediction process allows the model to estimate positional fitness even when MSA data is sparse. It reinforces its predictions using MSA data when available, in a process called retrieval inference.

A subsequent study sought to leverage the advantages of both MSA-based and alignment-free models [47]. This was achieved with TranceptEVE, a model that maintains the overall autoregressive learning in the Tranception architecture while replacing the original MSA retrieval feature with EVE (see section 4.2 for more details). TranceptEVE was in fact applied to viral evolution by implementing it within the EVEscape evolution forecasting framework. Unlike the EVE fitness model, this approach additionally enabled the prediction of insertions and deletions. In addition, certain groups sought to incorporate real-world evolutionary trends into transformer-based models to improve predictions. This was done by combining transformer attention mechanisms with either time series [51] or phylogenetic tree sample strategies [50]. Such methods have the advantage of enhancing PLM-generated fitness predictions with observed evolution, thus providing a means of leveraging the massive amounts of virus-specific genomic data generated throughout a pandemic.

Given the importance of immune evasion in both viral fitness and long-term vaccine efficacy, significant interest lies in forecasting mutations promoting antibody evasion. While most approaches discussed above indirectly inferred antibody-evasion hot-spots from genomic sequences alone [86], others aimed to integrate sequence information with experimental and/or modeled antibody evasion data. Machine Learning-Guided Antigenic Evolution Prediction (MLAEP) represents such a strategy. This tool investigates the antigenic evolution of a virus by prioritizing putative variants with greater fitness and immune evasion potential [48]. This is achieved by searching the viral mutational space through a process of *in silico* directed evolution, followed by a query of the biophysical interactions between the resulting variants and the ACE2 protein as well as 8 antibodies. Briefly, the biophysical features are interrogated by taking advantage of Evolutionary Scale Modelling (ESM)-1b, a large language model trained on 250 million protein sequences, and fine-tuned on SARS-CoV-2 sequences as well as 3D structures of the Spike RBD in complex with ACE2 and 8 antibodies. Importantly, the directed evolution analyses enabled by generative modeling described here allows the

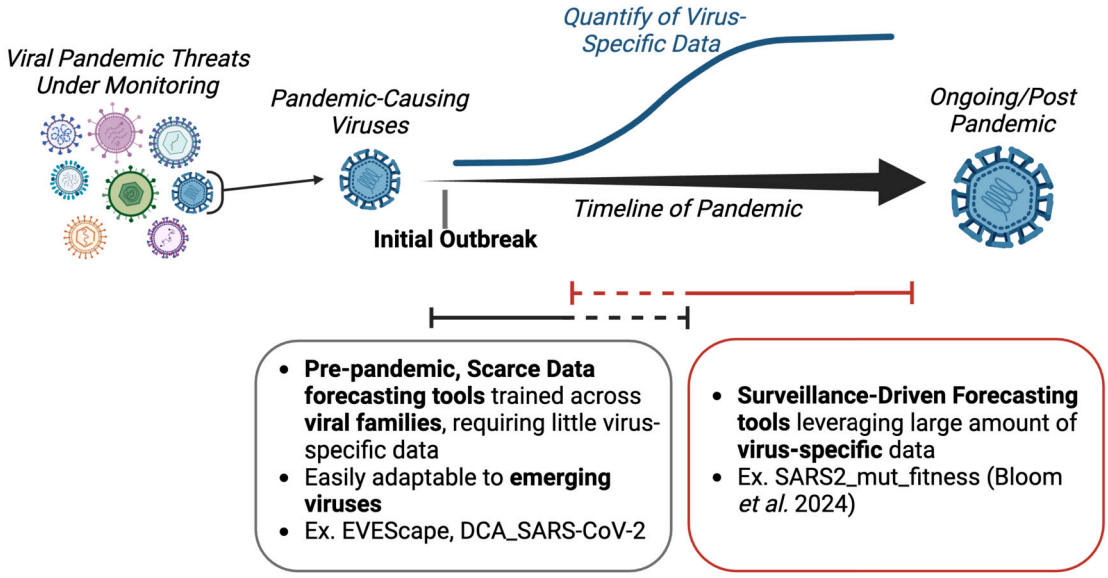

Fig. 3. Timeline of Variant Prediction Tools Usage Based on Data Availability During a Pandemic. Generalist tools easily adaptable to emerging variants are the ideal choice following the initial viral outbreak, when little viral data is available. Highly-specialized tools leveraging vast amounts of data are optimal during the later stages of a pandemic, when extensive data has been collected.

exploration of putative multi-mutation variants. This approach was able to capture naturally occurring variants. CoVFit, another notable example, achieved accurate SARS-CoV-2 antibody-evading as well as fitness predictions [49]. This was achieved by fine-tuning ESM-2 on coronaviridae genomes, SARS-CoV-2 DMS antibody-evasion data, and genotype-specific fitness information obtained form surveillance efforts.

5. The role of fitness measures and evolution prediction in pandemic preparedness

As demonstrated during the COVID-19 pandemic, the diversification of a virus during an ongoing pandemic represents a significant threat to effective public health interventions such as vaccination [94,95,12,13]. Beyond the COVID-19 pandemic and in preemption to future outbreaks, it is vital to develop frameworks that enable the rapid adaptation to emerging pathogens. The successful forecasting of viral evolution provides an opportunity to shape the actions to be taken in order to mitigate the impact of an evolving pathogen. In fact, the use of evolutionary predictions to better inform vaccine development has already been investigated in the context of SARS-CoV-2 [71].

In this section, we will discuss advancements made in the use of viral evolutionary prediction frameworks to mitigate viral pandemics. Particular attention will be given to *i)* leveraging evolutionary predictions to inform vaccine development as well as public health policy, *ii)* adapting existing frameworks to other viruses that may emerge or become of major concern, and *iii)* incorporating human population genetics into viral evolutionary forecasting schemes in order to interrogate host-pathogen co-evolution.

5.1. Anticipate variants to guide vaccine development and public health policies

The primary objective of vaccination interventions consists of providing long-lasting protection against infectious agents. However, the emergence of new viral variants poses a difficulty to the development of vaccines with sustained effectiveness. Challenges to vaccine efficacy arise when mutations associated with novel variants lead to the diversification of vaccine-targeted antigenic regions, [48,96], reducing their structural similarity to the original vaccine design. The corresponding reduction in vaccine-induced protection was strongly featured during the COVID-19 pandemic. As a well-documented example, the emergence of Omicron variants was shown to severely decrease the effectiveness of first-generation vaccines and booster vaccines as well as monoclonal antibody therapies by means of antibody evasion [94,95,12,13]. A promising avenue consists of developing vaccines in anticipation of potentially harmful viral variants [71].

DMS experiments have shown potential in identifying vaccine-evading variants [97], but their results are limited to a relatively small number pre-identified neutralizing antibodies. Additionally, interrogating the synergistic impact of co-occurring mutations remains difficult, particularly when thousands of strains are circulating. ML-based computational strategies can provide a powerful complimentary approach. Notably, Youssef and colleagues propose a computational framework forecasts antibody evading viral variants using EVEscape (see section 4.2 for details) [71]. The authors then leverage their predictions to propose, generate and assay vaccine constructs. Importantly, they claim their framework to be relevant in early stages of the pandemic due to its reliance on pre-pandemic data [71,44]. Frameworks such as this one can be used to anticipate future harmful variants to develop more vaccines in anticipation of the emergence of a variant. To this end, the development of booster vaccines modified to anticipate future variants was shown to result in superior neutralization over vaccines targeting circulating variants or ancestral vaccines [98,96]. Even without variant anticipation, ancestral-based booster vaccines have been shown to provide adequate immunization, reducing the likelihood of infection. One notable challenge with respect to the development of vaccines adapted to future variants lies in the potential gap between predicted and observed variants, which can hinder the integration of such frameworks in public health interventions. Nevertheless, the methods discussed in this review lay the groundwork for the incorporation of predictive strategies within public health interventions.

5.2. Adaptability to emerging viruses

The phenomenon of pathogen transmission from animal to human is known as spillover, and it has been the most common cause of pandemics in the last century [99]. It has been a major cause of concern for public health, especially when considering that the COVID-19 pandemic is of zoonotic origins [100,99]. To address this challenge, substantial efforts have been aimed at investigating the risk of spillover events as well as the identification of viruses that may pose putative pandemic threats [101–104]. Ultimately, such efforts aim to identify

Pandemic-Causing Viruses, which consist of viruses with high potential for leading to global pandemics [105,106]. Such endeavours aim to facilitate pathogen discovery, pathogen-host compatibility, as well as disease-risk mapping, with certain approaches leveraging panels of experts [101,107]. A notable example is SpillOver, an open-source public risk assessment web-based tool, also acting as a publicly accessible database, designed to prioritize zoonotic viruses that pose the greatest risk of causing a major spillover event in public health policies [101]. SpillOver was built on a framework that attributes a risk score to known or emerging viruses based on a set of risk factors determined through extensive literature review and specialist opinions. The higher the risk score estimated, the higher the urgency to come up with policies to mitigate the danger of the virus. The scores evaluated through the framework are then added to SpillOVer's ranking with a ranking score from 0 to 100. As of December 2nd 2024, SpillOver showcases 889 viruses where SARS-CoV-2 is the rank 1 virus with a risk score of 97, followed by the lassa virus with a score of 91. Beyond risk-assessment, this open-source platform provides a dynamic database encouraging the continued inclusion of under-studied viruses, promoting the monitoring of spillover events. Importantly, there exists a complementarity between risk assessment frameworks and evolutionary forecasting strategies. While the former identifies high-risk pathogens, the latter identifies regions at high risk of immune evasion and diversification, thus providing early blueprints for vaccine development. Studies have investigated the effectiveness of leveraging pre-pandemic strategies, by leading preemptive research that could be employed in cases of outbreaks [99]. One such method is EVEscape, which was capable of assessing and predicting specific SARS-CoV-2 variants that emerged during the pandemic, by leveraging pre-pandemic coronoviridae data. It also predicted emerging lassa virus and nipah virus variants, which coincidentally are among the top viruses on SpillOver's ranking. Thus, by leveraging risk assessment tools in identifying pathogens of greatest risk, variant prediction tools could be used to predict their most likely evolutionary pathway. This approach could greatly reduce the time needed to develop appropriate vaccines, which is crucial to lower the mortality rates in the early stages of a pandemic.

5.3. Host genetics and forecasting viral evolution

Whether built using phylogenetic or deep learning models, virtually all the evolutionary predictive methods discussed in this review rely heavily on viral genetic diversity to forecast evolution. While host genetic diversity is known to play a key role in the evolution of the virus, such information was generally not incorporated within these evolutionary predictive models. Nevertheless, interactions between viral pathogens and host genetics have been extensively reviewed, with host genetic diversity shown to act as a driver of viral evolution across various viral species [108–110]. For example, variations within the Major Histocompatibility Complex (MHC) region were shown to be associated with susceptibilities to a variety of viral pathogens including HIV, Tuberculosis, Hepatitis C Virus (HCV) and SARS-CoV-2 (non-exhaustive) [111–113]. The MHC region is responsible for the T cell arm of the adaptive immune system and has been identified as one of the most polymorphic regions of the human genome [114]. MHC polymorphism is characterized by both inter-individual and inter-population diversity, resulting in notable variations in MHC-disease associations and T cell responses [115,116]. In fact, the relationship between the MHC region and viral evolution constitutes a well-documented example of host-pathogen co-evolution that considers both host and viral diversity. MHC-driven viral polymorphisms have been extensively studied in both HIV and Influenza-A [117–120]. Interestingly, access to Influenza-A sequencing data spanning nearly a century has demonstrated the long-term antigenic diversification of T cell epitopes [120]. SARS-CoV-2 variants were shown to evade T cell immunity in an MHC allele-specific fashion [121,122,19]. As such, MHC-SARS-CoV-2 disease associations may suggest a putative role of MHC diversity in viral evolution. T cells were shown to be crucial in controlling viral infections, providing long-term immunity, and contributing to vaccine-induced immunity. Therefore identifying putative T cell escape events is essential to ensure the success of public health interventions. The development of frameworks enabling the anticipation of such mutations is particularly relevant to the effectiveness of T cell vaccines. These aim to specifically stimulate T cell immunity and constitute a promising vaccination avenue for individuals with B cell deficiencies.

Beyond the MHC, many genes have been associated with disease susceptibility in viral diseases. Notable examples include genes involved in interferon and Toll-like receptor signaling pathways, such as IFNAR2, which mediates cellular responses to type I interferons and has been associated with severe COVID-19 outcomes [123]. Other SARS-CoV-2-specific examples of host-pathogen genetic interactions include genomic variations within the Furin protease as well as TMPRSS2 [124–126]. These were both shown to affect viral entry into host cells, thereby modulating disease susceptibility. The genetic diversity within the ACE2 gene across human populations was also suggested as a putative source of selective pressure driving SARS-CoV-2 adaptation [127]. Additionally, RNA-editing mechanisms such as those mediated by APOBEC3A have been shown to contribute to intra-host genomic diversity, driving viral evolution and facilitating the emergence of advantageous mutations [128]. The investigation of intra-host evolution, which involves examining viral genetic diversity within individual hosts to explore within-host evolutionary dynamics [129–131], has also revealed several key host-pathogen interactions [132,133]. The incorporation of host genetics factors within viral evolution forecasting frameworks could possibly enhance the predictive power and accuracy of such frameworks, leading to the identification of mutations omitted by virus-centric models. The absence of such information across tools aimed at forecasting viral evolution likely lies in the scarcity of large datasets linking viral genetic diversity, host genetic diversity, and disease outcome. However, such datasets do exist. Several studies have attempted to interrogate cross-talks between host and pathogen genetics by conducting dual RNA-sequencing of both host and pathogen genomes across cohorts [134–137]. These investigations demonstrated the relationship between host genetic factors and pathogen polymorphism, recapitulating the role of HLA as well as interferon signaling pathways as drivers of pathogen evolution. Notably, Palmer and colleagues developed a Bayesian-based computational method to model the most likely viral evolutionary trajectories resulting from host factors [134]. While these approaches provide valuable insights, limitations in dataset availability and species coverage have prevented their integration within strategies aimed at forecasting pathogen evolution. Nevertheless, given the important influence of host genetic factors on viral evolution, leveraging such information could further improve the accuracy of viral evolution forecasting.

6. Perspectives: pandemic preparedness, a multi-strategic solution

A wide range of methods have been explored to anticipate viral evolution. Nevertheless, it is likely that successful pandemic mitigation will be best supported by a combination of strategies instead of a single approach (Fig. 3). For example, certain strategies were shown to be more suitable for early flagging of problematic mutations in emerging pathogens when limited information is available [44,86,43], denoted as "Scarce Data Forecasting tools" in Fig. 3. Leveraging evolutionary insights specific to viral families, these methods were conceived to be generalizable across species, making them particularly relevant to novel disease outbreaks. Importantly, they provide information regarding regions of the viral genome that are prone to frequent mutations and that may circumvent host immunity. Such knowledge may provide early guidance for experiments as well as vaccine development. As such, these methods will be particularly relevant during both pre-pandemic and early-stage pandemic timelines. These approaches may however be outperformed

by species-specific models during later stages of the pandemic, as more data becomes available [42,50,51,45], denoted as "Surveillance-Driven Forecasting tools" in Fig. 3. Such methods possess limited power when faced with data scarcity, but improve significantly as more data becomes available. While pre-pandemic tools rely on learning across viral families, pandemic tools aim to capture evolutionary relationships and biological rules within a single viral species. Effectively forecasting viral evolution prior to and during an ongoing pandemic will require appropriately leveraging early-pandemic and late-pandemic strategies to generate conducive to actionable insight. Beyond tool architecture, the incorporation of host genetic diversity within evolutionary forecasting frameworks could directly address evolutionary drivers currently not considered. Host factors, including genetic variability in immune response genes such as MHC and interferon-related pathways, play a crucial role in shaping viral evolution by exerting selective pressures on viral populations. However, the integration of host-pathogen interactions into predictive models remains challenging due to the scarcity of large-scale, well-characterized datasets that link viral mutations to host genetic backgrounds. While viral evolution forecasting offers a powerful means of pandemic preparedness, it also raises dual-use concerns, particularly regarding the potential misuse of predictive insights in viral pathogen engineering. Recent advances in generative AI for biodesign have raised concerns about its potential to inadvertently or intentionally create harmful viral agents with enhanced transmissibility and virulence [138–141]. As forecasting tools become increasingly sophisticated, responsible data-sharing frameworks and regulatory policies will be essential to ensuring that these technologies serve public health rather than pose unintended risks.

7. Conclusion

The recent COVID-19 pandemic has showcased the capacity of viral pathogens to rapidly adapt to host defence mechanisms while overcoming vaccination strategies. Advances in big data collection and analysis have unlocked new opportunities for forecasting viral evolution, providing a promising avenue to complement surveillance efforts and enable the development of preemptive public health interventions. If successfully implemented, this would facilitate the design of early warning systems while anticipating responses to future harmful viral variants. Recent advancements in computational methods, such as deep learning and language models, have the power to revolutionize the forecasting of viral evolution. There remains significant challenges to be addressed, which include data scarcity, handling of data collection biases, as well as the prediction of complex, saltation-like evolutionary events. Despite these logistical and implementation challenges, these methods are now poised for integration into public health and pandemic management strategies. These innovations, combined with the drastic growth of viral data collection and sharing strategies, have created a research environment uniquely suited to this task. Expanding global access to high-quality viral datasets will not only accelerate research but also enhance the development of preemptive public health strategies against future pandemics.

CRediT authorship contribution statement

D.J. Hamelin: Writing – review & editing, Writing – original draft, Funding acquisition, Conceptualization. **M. Scicluna:** Writing – original draft, Conceptualization. **I. Saadie:** Writing – original draft. **F. Mostefai:** Writing – review & editing. **J.C. Grenier:** Writing – review & editing. **C. Baron:** Writing – review & editing. **E. Caron:** Writing – review & editing. **J.G. Hussin:** Writing – review & editing, Writing – original draft, Supervision, Funding acquisition, Conceptualization.

Funding

The authors are supported by funding from the Canadian Institutes of Health Research (CIHR) Project Grant [174924]; CIHR operating grant to the Coronavirus Variants Rapid Response Network [CoVaRR-Net, ARR-175622]; DJH doctoral studies are supported by the Hydro Quebec Scholarship and Fonds de Recherche du Québec - Santé (FRQS) [350861]. MS doctoral studies are supported by a Canada Graduate Scholarship from Natural Sciences and Engineering Research Council of Canada (NSERC) [CGS D - 579121 - 2023]. FM doctoral studies are supported by the Hydro Quebec Scholarship. JGH holds a Tier 2 Canadian Research Chair from the NSERC [CRC-2024-00027].

Declaration of competing interest

The authors declare no competing interests.

Acknowledgements

We thank members of Dr. Hussin's group for helpful discussions and advice throughout this project, specifically Raphael Poujol and Pamela Mehanna.

References

[1] Chen Z, Azman AS, Chen X, Zou J, Tian Y, Sun R, et al. Global landscape of sars-cov-2 genomic surveillance and data sharing. Nat Genet 2022;54:499–507. https://doi.org/10.1038/s41588-022-01033-y.
[2] Stern A, Andino R. Viral evolution: it is all about mutations. Elsevier Inc.; 2016. p. 233–40.
[3] Sanjuán R. Quasispecies, vol. 1-5. Elsevier; 2008. pp. V4–359–V4–365.
[4] Markov PV, Ghafari M, Beer M, Lythgoe K, Simmonds P, Stilianakis NI, et al. The evolution of sars-cov-2. Nat Rev Microbiol 2023;21:361–79. https://doi.org/10.1038/s41579-023-00878-2.
[5] Wargo AR, Kurath G. Viral fitness: definitions, measurement, and current insights. Curr Opin Virol 2012;2:538–45.
[6] Starr TN, Greaney AJ, Hilton SK, Ellis D, Crawford KHD, Dingens AS, et al. Deep mutational scanning of sars-cov-2 receptor binding domain reveals constraints on folding and ace2 binding. Cell 2020;182:1295–131020.e. https://doi.org/10.1016/j.cell.2020.08.012.
[7] Lan J, Ge J, Yu J, Shan S, Zhou H, Fan S, et al. Structure of the sars-cov-2 spike receptor-binding domain bound to the ace2 receptor. Nature 2020;581:215–20. https://doi.org/10.1038/s41586-020-2180-5.
[8] Hayashi T, Abiko K, Mandai M, Yaegashi N, Konishi I. Highly conserved binding region of ace2 as a receptor for sars-cov-2 between humans and mammals. Vet Q 2020;40(1):243–9. https://doi.org/10.1080/01652176.2020.1823522.
[9] Liu Y, Liu J, Plante KS, Plante JA, Xie X, Zhang X, et al. The n501y spike substitution enhances sars-cov-2 infection and transmission. Nature 2022;602:294–9. https://doi.org/10.1038/s41586-021-04245-0.
[10] Zhang L, Mann M, Syed ZA, Reynolds HM, Tian E, Samara NL, et al. Furin cleavage of the sars-cov-2 spike is modulated by o-glycosylation. Proc Natl Acad Sci 2021;118. https://doi.org/10.1073/pnas.2109905118. https://www.pnas.org/doi/10.1073/pnas.2109905118.
[11] Hastie KM, Li H, Bedinger D, Schendel SL, Dennison SM, Li K, et al. Defining variant-resistant epitopes targeted by sars-cov-2 antibodies: a global consortium study. Science 2021;374:472–8. https://doi.org/10.1126/science.abh2315.
[12] Cao Y, Wang J, Jian F, Xiao T, Song W, Yisimayi A, et al. Omicron escapes the majority of existing sars-cov-2 neutralizing antibodies. Nature 2022;602:657–63. https://doi.org/10.1038/s41586-021-04385-3.
[13] Cao Y, Yisimayi A, Jian F, Song W, Xiao T, Wang L, et al. Ba. 2.12.1, ba.4 and ba.5 escape antibodies elicited by omicron infection. Nature 2022;608:593–602. https://doi.org/10.1038/s41586-022-04980-y.
[14] Newman J, Thakur N, Peacock TP, Bialy D, Elrefaey AM, Bogaardt C, et al. Neutralizing antibody activity against 21 sars-cov-2 variants in older adults vaccinated with bnt162b2. Nat Microbiol 2022;7:1180–8. https://doi.org/10.1038/s41564-022-01163-3.
[15] Patrick C, Upadhyay V, Lucas A, Mallela KMG. Biophysical fitness landscape of the sars-cov-2 delta variant receptor binding domain. J Mol Biol 2022;434(13):167622. https://doi.org/10.1016/j.jmb.2022.167622.
[16] Supasa P, Zhou D, Dejnirattisai W, Liu C, Mentzer AJ, Ginn HM, et al. Reduced neutralization of sars-cov-2 b. 1.1.7 variant by convalescent and vaccine sera. Cell 2021;184:2201–22117.e. https://doi.org/10.1016/j.cell.2021.02.033.
[17] Reuschl AK, Thorne LG, Whelan MV, Ragazzini R, Furnon W, Cowton VM, et al. Evolution of enhanced innate immune suppression by sars-cov-2 omicron subvariants. Nat Microbiol 2024;9:451–63. https://doi.org/10.1038/s41564-023-01588-4.
[18] Guo K, Barrett BS, Morrison JH, Mickens KL, Vladar EK, Hasenkrug KJ, et al. Interferon resistance of emerging sars-cov-2 variants. Proc Natl Acad Sci 2022. https://doi.org/10.1073/pnas.2203760119.
[19] Naranbhai V, Nathan A, Kaseke C, Berrios C, Khatri A, Choi S, et al. T cell reactivity to the sars-cov-2 omicron variant is preserved in most but not all individuals. Cell 2022;185:1041–10516.e. https://doi.org/10.1016/j.cell.2022.01.029.

[20] Moriyama M, Lucas C, Monteiro VS, Iwasaki A. Enhanced inhibition of mhc-i expression by sars-cov-2 omicron subvariants. Proc Natl Acad Sci USA 2023;120. https://doi.org/10.1073/pnas.2221652120.
[21] Pancera M, Changela A, Kwong PD. How hiv-1 entry mechanism and broadly neutralizing antibodies guide structure-based vaccine design. https://doi.org/10.1097/COH.0000000000000360, 5 2017.
[22] Dou D, Revol R, Östbye H, Wang H, Daniels R. Influenza a virus cell entry, replication, virion assembly and movement. https://doi.org/10.3389/fimmu.2018.01581, 7 2018.
[23] Volz EM, Koelle K, Bedford T. Viral phylodynamics. PLoS Comput Biol 2013;9:e1002947. https://doi.org/10.1371/journal.pcbi.1002947.
[24] Sagulenko P, Puller V, Neher RA. Treetime: maximum-likelihood phylodynamic analysis. Virus Evol 2018;4:vex042. https://doi.org/10.1093/ve/vex042.
[25] Ferreira RC, Wong E, Gugan G, Wade K, Liu M, Baena LM, et al. Covizu: rapid analysis and visualization of the global diversity of sars-cov-2 genomes. Virus Evol 2021;7. https://doi.org/10.1093/ve/veab092.
[26] Doud MB, Bloom JD. Accurate measurement of the effects of all amino-acid mutations on influenza hemagglutinin. Viruses 2016;8:155. https://doi.org/10.3390/v8060155.
[27] Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. Nat Methods 2014;11:801–7. https://doi.org/10.1038/nmeth.3027.
[28] Thyagarajan B, Bloom JD. The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. eLife 2014;3:e03300. https://doi.org/10.7554/elife.03300.
[29] Wu NC, Young AP, Al-Mawsawi LQ, Olson CA, Feng J, Qi H, et al. High-throughput profiling of influenza a virus hemagglutinin gene at single-nucleotide resolution. Sci Rep 2014;4:4942. https://doi.org/10.1038/srep04942.
[30] Wu NC, Thompson AJ, Xie J, Lin C-W, Nycholat CM, Zhu X, et al. A complex epistatic network limits the mutational reversibility in the influenza hemagglutinin receptor-binding site. Nat Commun 2018;9:1264. https://doi.org/10.1038/s41467-018-03663-5.
[31] Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 2018;34:4121–3. https://doi.org/10.1093/bioinformatics/bty407.
[32] Vecchia ED. Pathoplexus: towards fair and transparent sequence sharing. Lancet Microbe 2024. 100995doi:10.1016/j.lanmic.2024.100995. https://linkinghub.elsevier.com/retrieve/pii/S2666524724002635.
[33] Gill EE, Jia B, Murall CL, Poujol R, Anwar MZ, John S, et al. The Canadian virusseq data portal & duotang: open resources for sars-cov-2 viral sequences and genomic epidemiology. ArXiv 2024;16:35. https://github.com/cidgoh/DataHarmonizer.
[34] Lee JM, Huddleston J, Doud MB, Hooper KA, Wu NC, Bedford T, et al. Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human h3n2 influenza variants. Proc Natl Acad Sci 2018;115:E8276–85. https://doi.org/10.1073/pnas.1806133115.
[35] Greaney AJ, Starr TN, Barnes CO, Weisblum Y, Schmidt F, Caskey M, et al. Mapping mutations to the sars-cov-2 rbd that escape binding by different classes of antibodies. Nat Commun 2021;12:4196. https://doi.org/10.1038/s41467-021-24435-8.
[36] Greaney AJ, Starr TN, Gilchuk P, Zost SJ, Binshtein E, Loes AN, et al. Complete mapping of mutations to the sars-cov-2 spike receptor-binding domain that escape antibody recognition. Cell Host Microbe 2021;29:44–579.e. https://doi.org/10.1016/j.chom.2020.11.007.
[37] Taft JM, Weber CR, Gao B, Ehling RA, Han J, Frei L, et al. Predictive profiling of sars-cov-2 variants by deep mutational learning. bioRxiv 2021. https://doi.org/10.1101/2021.12.07.471580.
[38] Chan KK, Tan TJC, Narayanan KK, Procko E. An engineered decoy receptor for sars-cov-2 broadly binds protein s sequence variants. Sci Adv 2021;7:eabf1738. https://doi.org/10.1126/sciadv.abf1738.
[39] Teruel N, Mailhot O, Najmanovich RJ. Modelling conformational state dynamics and its role on infection for sars-cov-2 spike protein variants. PLoS Comput Biol 2021;17:e1009286. https://doi.org/10.1371/journal.pcbi.1009286.
[40] Obermeyer F, Jankowiak M, Barkas N, Schaffner SF, Pyle JD, Yurkovetskiy L, et al. Analysis of 6.4 million sars-cov-2 genomes identifies mutations associated with fitness. Science (New York, N.Y.) 2022;376:abm1208. https://doi.org/10.1126/science.abm1208.
[41] Maher MC, Bartha I, Weaver S, di Iulio J, Ferri E, Soriaga L, et al. Predicting the mutational drivers of future sars-cov-2 variants of concern. Sci Transl Med 2022;14:eabk3445. https://doi.org/10.1126/scitranslmed.abk3445.
[42] Bloom JD, Neher RA. Fitness effects of mutations to sars-cov-2 proteins. Virus Evol 2024;9:vead055. https://doi.org/10.1093/ve/vead055.
[43] Rodriguez-Rivas J, Croce G, Muscat M, Weigt M. Epistatic models predict mutable sites in sars-cov-2 proteins and epitopes. Proc Natl Acad Sci 2022. https://doi.org/10.1073/pnas.2113118119.
[44] Thadani NN, Gurev S, Notin P, Youssef N, Rollins NJ, Ritter D, et al. Learning from prepandemic data to forecast viral escape. Nature 2023;622:818–25. https://doi.org/10.1038/s41586-023-06617-0.
[45] King S, Chen XE, Ng SW, Rostin K, Hahn SV, Roberts T, et al. Forecasting sars-cov-2 spike protein evolution from small data by deep learning and regression. Front Syst Biol 2024:4. https://doi.org/10.3389/fsysb.2024.1284668.
[46] Hie B, Zhong ED, Berger B, Bryson B. Learning the language of viral evolution and escape. Science 2021;371(6526):284–8. https://doi.org/10.1126/science.abd7331.
[47] Notin P, Dias M, Frazer J, Marchena-Hurtado J, Gomez A, Marks DS, et al. Tranception: Protein fitness prediction with autoregressive transformers and inference-time retrieval. In: Proceedings of the 39 th International Conference on Machine Learning.
[48] Han W, Chen N, Xu X, Sahil A, Zhou J, Li Z, et al. Predicting the antigenic evolution of sars-cov-2 with deep learning. Nat Commun 2023;14:3478. https://doi.org/10.1038/s41467-023-39199-6.
[49] Ito J, Strange A, Liu W, Joas G, Lytras S, Sato K. A protein language model for exploring viral fitness landscapes. https://doi.org/10.1101/2024.03.15.584819, 3 2024. http://biorxiv.org/lookup/doi/10.1101/2024.03.15.584819.
[50] Zhou B, Zhou H, Zhang X, Xu X, Chai Y, Zheng Z, et al. Tempo: a transformer-based mutation prediction framework for sars-cov-2 evolution. Comput Biol Med 2023;152:106264. https://doi.org/10.1016/j.compbiomed.2022.106264.
[51] Saha G, Sawmya S, Saha A, Akil MA, Tasnim S, Rahman MS, et al. Priest: predicting viral mutations with immune escape capability of sars-cov-2 using temporal evolutionary information. Brief Bioinform 2024;25. https://doi.org/10.1093/bib/bbae218.
[52] Beguir K, Skwark MJ, Fu Y, Pierrot T, Carranza NL, Laterre A, et al. Early computational detection of potential high-risk sars-cov-2 variants. Comput Biol Med 2023;155:106618. https://doi.org/10.1016/j.compbiomed.2023.106618.
[53] Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci 2021;118. https://doi.org/10.1073/pnas.2016239118. https://www.pnas.org/doi/10.1073/pnas.2016239118.
[54] Notin P, Niekerk LV, Kollasch AW, Ritter D, Gal Y, Marks DS. TranceptEVE: Combining family-specific and family-agnostic models of protein sequences for improved fitness prediction. bioRxiv 2022. https://doi.org/10.1101/2022.12.07.519495.
[55] Abousamra E, Figgins M, Bedford T. Fitness models provide accurate short-term forecasts of sars-cov-2 variant frequency. PLoS Comput Biol 2024;20. https://doi.org/10.1371/journal.pcbi.1012443.
[56] Susswein Z, Johnson KE, Kassa R, Parastaran M, Peng V, Wolansky L, et al. Leveraging global genomic sequencing data to estimate local variant dynamics. https://doi.org/10.1101/2023.01.02.23284123, 1 2023. http://medrxiv.org/lookup/doi/10.1101/2023.01.02.23284123.
[57] Annavajhala MK, Mohri H, Wang P, Nair M, Zucker JE, Sheng Z, et al. Emergence and expansion of sars-cov-2 b. 1.526 after identification in New York. Nature 2021;597:703–8. https://doi.org/10.1038/s41586-021-03908-2.
[58] Lee B, Quadeer AA, Sohail MS, Finney E, Ahmed SF, McKay MR, et al. Inferring effects of mutations on sars-cov-2 transmission from genomic surveillance data. MedRxiv 2024. https://doi.org/10.1101/2021.12.31.21268591.
[59] Stern DL. The genetic causes of convergent evolution. https://doi.org/10.1038/nrg3483, 11 2013.
[60] Martin DP, Weaver S, Tegally H, San JE, Shank SD, Wilkinson E, et al. The emergence and ongoing convergent evolution of the sars-cov-2 n501y lineages. Cell 2021;184:5189–52007.e. https://doi.org/10.1016/j.cell.2021.09.003.
[61] Eraslan G, Avsec Žiga, Gagneur J, Theis FJ. Deep learning: new computational modelling techniques for genomics. https://doi.org/10.1038/s41576-019-0122-6, 7 2019.
[62] Li Y, Huang C, Ding L, Li Z, Pan Y, Gao X. Deep learning in bioinformatics: introduction, application, and perspective in the big data era. Methods 2019;166:4–21. https://doi.org/10.1016/j.ymeth.2019.04.008.
[63] Riccio C, Jansen ML, Guo L, Ziegler A. Variant effect predictors: a systematic review and practical guide. https://doi.org/10.1007/s00439-024-02670-5, 5 2024.
[64] Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using polyphen-2. In: Current protocols in human genetics; 2013.
[65] Kim HY, Jeon W, Kim D. An enhanced variant effect predictor based on a deep generative model and the born-again networks. Sci Rep 2021;11. https://doi.org/10.1038/s41598-021-98693-3.
[66] Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. Cadd: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res 2019;47:D886–94. https://doi.org/10.1093/nar/gky1016.
[67] Laine E, Karami Y, Carbone A. Gemme: a simple and fast global epistatic model predicting mutational effects. https://doi.org/10.1093/molbev/msz179, 11 2019.
[68] Qi H, Zhang H, Zhao Y, Chen C, Long JJ, Chung WK, et al. Mvp predicts the pathogenicity of missense variants by deep learning. Nat Commun 2021;12. https://doi.org/10.1038/s41467-020-20847-0.
[69] Zhang H, Xu MS, Fan X, Chung WK, Shen Y. Predicting functional effect of missense variants using graph attention neural networks. Nat Mach Intell 2022;4:1017–28. https://doi.org/10.1038/s42256-022-00561-w.
[70] Frazer J, Notin P, Dias M, Gomez A, Min JK, Brock K, et al. Disease variant prediction with deep generative models of evolutionary data. Nature 2021;599:91–5. https://doi.org/10.1038/s41586-021-04043-8.
[71] Youssef N, Ghantous F, Gurev S, Brock K, Jaimes JA, Dauphin A, et al. Deep generative models predict sars-cov-2 spike infectivity and foreshadow neutralizing antibody escape. bioRxiv.
[72] Li H. Language models: past, present, and future. Commun ACM 2022;65(7):56–63. https://doi.org/10.1145/3490443.
[73] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. In: 1st international conference on learning representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013; 2013. http://arxiv.org/abs/1301.3781.

[74] Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep contextualized word representations. In: Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long papers). New Orleans, Louisiana: Association for Computational Linguistics; 2018. p. 2227–37. https://aclanthology.org/N18-1202.
[75] Howard J, Ruder S. Universal language model fine-tuning for text classification. In: ACL, association for computational linguistics; 2018. http://arxiv.org/abs/1801.06146.
[76] Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). Minneapolis, Minnesota: Association for Computational Linguistics; 2019. p. 4171–86. https://aclanthology.org/N19-1423.
[77] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. https://openai.com/blog/better-language-models/.
[78] Zellers R, Holtzman A, Rashkin H, Bisk Y, Farhadi A, Roesner F, et al. Defending against neural fake news. https://doi.org/10.48550/ARXIV.1905.12616, 2019. https://arxiv.org/abs/1905.12616.
[79] Lewkowycz A, Andreassen A, Dohan D, Dyer E, Michalewski H, Ramasesh V, et al. Solving quantitative reasoning problems with language models. https://doi.org/10.48550/ARXIV.2206.14858, 2022. https://arxiv.org/abs/2206.14858.
[80] Hie B, Zhong ED, Berger B, Bryson B. Learning the language of viral evolution and escape. Science 2021;371(6526):284–8. https://doi.org/10.1126/science.abd7331. https://www.science.org/doi/pdf/10.1126/science.abd7331. https://www.science.org/doi/abs/10.1126/science.abd7331.
[81] Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci 2021;118(15):e2016239118. https://doi.org/10.1073/pnas.2016239118. https://www.pnas.org/doi/pdf/10.1073/pnas.2016239118. https://www.pnas.org/doi/abs/10.1073/pnas.2016239118.
[82] Ferruz N, Schmidt S, Höcker B. Protgpt2 is a deep unsupervised language model for protein design. Nat Commun 2022;13(1). https://doi.org/10.1038/s41467-022-32007-7.
[83] Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH. Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 2015;31:926–32. https://doi.org/10.1093/bioinformatics/btu739.
[84] Alley EC, Khimulya G, Biswas S, AlQuraishi M, Church GM. Unified rational protein engineering with sequence-based deep representation learning. Nat Methods 2019;16:1315–22. https://doi.org/10.1038/s41592-019-0598-1.
[85] Heinzinger M, Elnaggar A, Wang Y, Dallago C, Nechaev D, Matthes F, et al. Modeling aspects of the language of life through transfer-learning protein sequences. BMC Bioinform 2019;20. https://doi.org/10.1186/s12859-019-3220-8.
[86] Hie B, Zhong E, Berger B, Bryson B. Learning the language of viral evolution and escape. Science 2020. 2020.07.08.193946doi:10.1101/2020.07.08.193946.
[87] Vaswani A, Brain G, Shazeer N, Parmar N, Uszkoreit J, Jones L, et al. Attention is all you need, NIPS.
[88] Devlin J, Chang M-W, Lee K, Google KT, Language AI. Bert: pre-training of deep bidirectional transformers for language understanding. In: Proceedings ofNAACL-HLT; 2019. p. 4171–86. https://github.com/tensorflow/tensor2tensor.
[89] Nambiar A, Heflin M, Liu S, Maslov S, Hopkins M, Ritz A. Transforming the language of life: transformer neural networks for protein prediction tasks. In: Proceedings of the 11th ACM international conference on bioinformatics, computational biology and health informatics, BCB 2020. Association for Computing Machinery, Inc; 2020.
[90] Madani A, Mccann B, Naik N, Keskar NS, Anand N, Chu A, et al. Progen: Language modeling for protein generation. NeurIPS.
[91] Meier J, Rao R, Verkuil R, Liu J, Sercu T, Rives A. Language models enable zero-shot prediction of the effects of mutations on protein function. NeurIPS. https://github.
[92] Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, et al. Prottrans: toward understanding the language of life through self-supervised learning. IEEE Trans Pattern Anal Mach Intell 2022;44:7112–27. https://doi.org/10.1109/TPAMI.2021.3095381.
[93] Rao R, Liu J, Verkuil R, Meier J, Canny JF, Abbeel P, et al. Msa transformer; 2021. https://github.com/facebookresearch/.
[94] Kip KE, McCreary EK, Collins K, Minnier TE, Snyder GM, Garrard W, et al. Evolving real-world effectiveness of monoclonal antibodies for treatment of covid-19 a cohort study. Ann Intern Med 2023;176:496–504. https://doi.org/10.7326/M22-1286.
[95] Scheaffer SM, Lee D, Whitener B, Ying B, Wu K, Liang CY, et al. Bivalent sars-cov-2 mrna vaccines increase breadth of neutralization and protect against the ba.5 omicron variant in mice. Nat Med 2023;29:247–57. https://doi.org/10.1038/s41591-022-02092-8.
[96] Cromer D, Reynaldi A, Mitchell A, Schlub TE, Juno JA, Wheatley AK, et al. Predicting covid-19 booster immunogenicity against future sars-cov-2 variants and the benefits of vaccine updates. Nat Commun 2024;15(1). https://doi.org/10.1038/s41467-024-52194-9.
[97] Dadonaite B, Brown J, McMahon TE, Farrell AG, Figgins MD, Asarnow D, et al. Spike deep mutational scanning helps predict success of sars-cov-2 clades. Nature 2024;631:617–26. https://doi.org/10.1038/s41586-024-07636-1.
[98] Khoury DS, Docken SS, Subbarao K, Kent SJ, Davenport MP, Cromer D. Predicting the efficacy of variant-modified covid-19 vaccine boosters. Nat Med 2023;29(3):574–8. https://doi.org/10.1038/s41591-023-02228-4.
[99] Williams BA, Jones CH, Welch V, True JM. Outlook of pandemic preparedness in a post-covid-19 world. npj Vaccin 2023;8(1). https://doi.org/10.1038/s41541-023-00773-0.
[100] Holmes EC, Goldstein SA, Rasmussen AL, Robertson DL, Crits-Christoph A, Wertheim JO, et al. The origins of sars-cov-2: a critical review. Cell 2021;184(19):4848–56. https://doi.org/10.1016/j.cell.2021.08.017.
[101] Grange ZL, Goldstein T, Johnson CK, Anthony S, Gilardi K, Daszak P, et al. Ranking the risk of animal-to-human spillover for newly discovered viruses. Proc Natl Acad Sci 2021;118(15). https://doi.org/10.1073/pnas.2002324118.
[102] Hart WS, Park H, Jeong YD, Kim KS, Yoshimura R, Thompson RN, et al. Analysis of the risk and pre-emptive control of viral outbreaks accounting for within-host dynamics: Sars-cov-2 as a case study. Proc Natl Acad Sci 2023;120(41). https://doi.org/10.1073/pnas.2305451120.
[103] Golchin M, Di Marco M, Horwood PF, Paini DR, Hoskins AJ, Hickson R. Prediction of viral spillover risk based on the mass action principle. One Health 2024;18:100737. https://doi.org/10.1016/j.onehlt.2024.100737.
[104] Zhang T, Rabhi F, Chen X, Paik H-y, MacIntyre CR. A machine learning-based universal outbreak risk prediction tool. Comput Biol Med 2024;169:107876. https://doi.org/10.1016/j.compbiomed.2023.107876.
[105] Harrington WN, Kackos CM, Webby RJ. The evolution and future of influenza pandemic preparedness. https://doi.org/10.1038/s12276-021-00603-0, 5 2021.
[106] Mishra C, Meena S, Meena JK, Tiwari S, Mathur P. Detection of three pandemic causing coronaviruses from non-respiratory samples: systematic review and meta-analysis. Sci Rep 2021;11. https://doi.org/10.1038/s41598-021-95329-4.
[107] Salmanton-García J, Wipfler P, Leckler J, Nauclér P, Mallon PW, Bruijning-Verhagen PC, et al. Predicting the next pandemic: vaccelerate ranking of the world health organization's blueprint for action to prevent epidemics. Trav Med Infect Dis 2024;57. https://doi.org/10.1016/j.tmaid.2023.102676.
[108] Mogensen TH. Genetic susceptibility to viral disease in humans. https://doi.org/10.1016/j.cmi.2022.02.023, 11 2022.
[109] Kenney AD, Dowdle JA, Bozzacco L, McMichael TM, Gelais CS, Panfil AR, et al. Human genetic determinants of viral diseases. https:/doi.org/10.1146/annurev-genet-120116-023425, 11 2017.
[110] Kwok AJ, Mentzer A, Knight JC. Host genetics and infectious disease: new tools, insights and translational opportunities. https:/doi.org/10.1038/s41576-020-00297-6, 3 2021.
[111] Kasww RA, Carrjngton M, Apple R, Parj L, Munoz A, Saah AJ, et al. Influence of combinations of human major histocompatibility complex genes on the course of hiv-1 infection. http://www.nature.com/naturemedicine, 1996.
[112] Duggal P, Thio CL, Wojcik GL, Goedert JJ, Mangia A, Latanich R, et al. Genome-wide association study of spontaneous resolution of hepatitis c virus infection: data from multiple cohorts. Ann Intern Med 2013;158:235–45. https://doi.org/10.7326/0003-4819-158-4-201302190-00003.
[113] Sveinbjornsson G, Gudbjartsson DF, Halldorsson BV, Kristinsson KG, Gottfredsson M, Barrett JC, et al. Hla class ii sequence variants influence tuberculosis risk in populations of European ancestry. Nat Genet 2016;48:318–22. https://doi.org/10.1038/ng.3498.
[114] Robinson J, Barker DJ, Georgiou X, Cooper MA, Flicek P, Marsh SG. Ipd-imgt/hla database. Nucleic Acids Res 2020;48:D948–55. https://doi.org/10.1093/nar/gkz950.
[115] Medhasi S, Chantratita N. Human leukocyte antigen (hla) system: genetics and association with bacterial and viral infections. https://doi.org/10.1155/2022/9710376, 2022.
[116] Dendrou CA, Petersen J, Rossjohn J, Fugger L. HLA variation and disease. Nat Rev Immunol 2018;18. https://doi.org/10.1038/nri.2017.143.
[117] Payne R, Muenchhoff M, Mann J, Roberts HE, Matthews P, Adland E, et al. Impact of hla-driven hiv adaptation on virulence in populations of high hiv seroprevalence. Proc Natl Acad Sci USA 2014;111:E5393–400. https://doi.org/10.1073/pnas.1413339111.
[118] Brumme ZL, John M, Carlson JM, Brumme CJ, Chan D, Brockman MA, et al. Hla-associated immune escape pathways in hiv-1 subtype b gag, pol and nef proteins. PLoS ONE 2009;4. https://doi.org/10.1371/journal.pone.0006687.
[119] Goulder PJ, Walker BD. Hiv and hla class i: an evolving relationship. https://doi.org/10.1016/j.immuni.2012.09.005, 9 2012.
[120] Woolthuis RG, Dorp CHV, Keşmir C, Boer RJD, Boven MV. Long-term adaptation of the influenza a virus by escaping cytotoxic t-cell recognition. Sci Rep 2016;6. https://doi.org/10.1038/srep33334.
[121] Hamelin DJ, Fournelle D, Grenier J-C, Schockaert J, Kovalchik KA, Kubiniok P, et al. The mutational landscape of sars-cov-2 variants diversifies t cell targets in an hla-supertype-dependent manner. Cell Syst 2022;13(2):143–1573.e. https://doi.org/10.1016/j.cels.2021.09.013.
[122] Kovalchik KA, Hamelin DJ, Kubiniok P, Bourdin B, Mostefai F, Poujol R, et al. Machine learning-enhanced immunopeptidomics applied to t-cell epitope discovery for covid-19 vaccines. Nat Commun 2024;15:10316. https://doi.org/10.1038/s41467-024-54734-9. https://www.nature.com/articles/s41467-024-54734-9.
[123] Smieszek SP, Polymeropoulos VM, Xiao C, Polymeropoulos CM, Polymeropoulos MH. Loss-of-function mutations in ifnar2 in covid-19 severe infection suscepti-

bility. J Global Antimicrob Resist 2021;26:239–40. https://doi.org/10.1016/j.jgar.2021.06.005.

[124] Irham LM, Chou WH, Calkins MJ, Adikusuma W, Hsieh SL, Chang WC. Genetic variants that influence sars-cov-2 receptor tmprss2 expression among population cohorts from multiple continents. Biochem Biophys Res Commun 2020;529:263–9. https://doi.org/10.1016/j.bbrc.2020.05.179.

[125] Dobrindt K, Hoagland DA, Seah C, Kassim B, O'Shea CP, Iskhakova M, et al. Common genetic variation in humans impacts in vitro susceptibility to sars-cov-2 infection. https://doi.org/10.1101/2020.09.20.300574, 9 2020. http://biorxiv.org/lookup/doi/10.1101/2020.09.20.300574.

[126] Hou Y, Zhao J, Martin W, Kallianpur A, Chung MK, Jehi L, et al. New insights into genetic susceptibility of covid-19: an ace2 and tmprss2 polymorphism analysis. BMC Med 2020;18. https://doi.org/10.1186/s12916-020-01673-z.

[127] Devaux CA, Fantini J. Possible contribution of rare alleles of human ace2 in the emergence of sars-cov-2 variants escaping the immune response. Front Immunol 2023;14. https://doi.org/10.3389/fimmu.2023.1252367.

[128] Kim K, Calabrese P, Wang S, Qin C, Rao Y, Feng P, et al. The roles of apobec-mediated rna editing in sars-cov-2 mutations, replication and fitness. Sci Rep 2022;12. https://doi.org/10.1038/s41598-022-19067-x.

[129] Mostefai F, Grenier JC, Poujol R, Hussin J. Refining sars-cov-2 intra-host variation by leveraging large-scale sequencing data. NAR Genomics Bioinform 2024;6. https://doi.org/10.1093/nargab/lqae145.

[130] Gazeau S, Deng X, Ooi HK, Mostefai F, Hussin J, Heffernan J, et al. The race to understand immunopathology in covid-19: perspectives on the impact of quantitative approaches to understand within-host interactions. ImmunoInformatics 2023;9:100021. https://doi.org/10.1016/j.immuno.2023.100021.

[131] Fuhrmann L, Jablonski KP, Beerenwinkel N. Quantitative measures of within-host viral genetic diversity. https://doi.org/10.1016/j.coviro.2021.06.002, 8 2021.

[132] Leigh DM, Peranić K, Prospero S, Cornejo C, Dsignurković-Perica MC, Kupper Q, et al. Long-read sequencing reveals the evolutionary drivers of intra-host diversity across natural rna mycovirus infections. Virus Evol 2021;7. https://doi.org/10.1093/ve/veab101.

[133] Gu H, Quadeer AA, Krishnan P, Ng DY, Chang LD, Liu GY, et al. Within-host genetic diversity of sars-cov-2 lineages in unvaccinated and vaccinated individuals. Nat Commun 2023;14. https://doi.org/10.1038/s41467-023-37468-y.

[134] Palmer DS, Turner I, Fidler S, Frater J, Goedhals D, Goulder P, et al. Mapping the drivers of within-host pathogen evolution using massive data sets. Nat Commun 2019;10. https://doi.org/10.1038/s41467-019-10724-w.

[135] Ansari MA, Pedergnana V, Ip CL, Magri A, Delft AV, Bonsall D, et al. Genome-to-genome analysis highlights the effect of the human innate and adaptive immune systems on the hepatitis c virus. Nat Genet 2017;49:666–73. https://doi.org/10.1038/ng.3835.

[136] Nuss AM, Beckstette M, Pimenova M, Schmühl C, Opitz W, Pisano F, et al. Tissue dual rna-seq allows fast discovery of infection-specific functions and riboregulators shaping host-pathogen transcriptomes. Proc Natl Acad Sci USA 2017;114:E791–800. https://doi.org/10.1073/pnas.1613405114.

[137] Westermann AJ, Förstner KU, Amman F, Barquist L, Chao Y, Schulte LN, et al. Dual rna-seq unveils noncoding rna functions in host-pathogen interactions. Nature 2016;529:496–501. https://doi.org/10.1038/nature16547.

[138] Gati NS, Altinok OA, Kumar S, Ferrando VA, Kurtz J, Quante M, et al. Integrating evolutionary aspects into dual-use discussion: the cases of influenza virus and enterohemorrhagic escherichia coli. https://doi.org/10.1093/emph/eoab034, 2021.

[139] Wheeler NE. Responsible ai in biotechnology: balancing discovery, innovation and biosecurity risks. https://doi.org/10.3389/fbioe.2025.1537471, 2025.

[140] Musunuri S, Sandbrink JB, Monrad JT, Palmer MJ, Koblentz GD. Rapid proliferation of pandemic research: implications for dual-use risks. mBio 2021;12. https://doi.org/10.1128/mBio.01864-21.

[141] Bloomfield D, Pannu J, Zhu AW, Ng MY, Lewis A, Bendavid E, et al. Ai and biosecurity: the need for governance. Science 2024. https://doi.org/10.1126/science.adq1977.