The race to understand immunopathology in COVID-19: perspectives on the impact of quantitative approaches to understand within-host interactions

Gazeau, Sonia; Deng, Xiaoyan; Ooi, Hsu Kiang; Mostefai, Fatima; Hussin, Julie; Heffernan, Jane; Jenner, Adrianne L.; Craig, Morgan

Publisher's version / Version de l'éditeur:
https://doi.org/10.1016/j.immuno.2023.100021
ImmunoInformatics, 9, C, pp. 1-12, 2023-01-08

ImmunoInformatics 9 (2023) 100021

Contents lists available at ScienceDirect

ImmunoInformatics

journal homepage: www.journals.elsevier.com/immunoinformatics

The race to understand immunopathology in COVID-19: Perspectives on the impact of quantitative approaches to understand within-host interactions

Sonia Gazeau [a,b,1], Xiaoyan Deng [a,b,1], Hsu Kiang Ooi [c], Fatima Mostefai [d,e], Julie Hussin [d,e], Jane Heffernan [f,g], Adrianne L. Jenner [h], Morgan Craig [a,b,*]

[a] Department of Mathematics and Statistics, Université de Montréal, Montréal, Canada
[b] Sainte-Justine University Hospital Research Centre, Montréal, Canada
[c] Digital Technologies Research Centre, National Research Council Canada, Toronto, Canada
[d] Montréal Heart Institute Research Centre, Montréal, Canada
[e] Department of Medicine, Faculty of Medicine, Université de Montréal, Montréal, Canada
[f] Modelling Infection and Immunity Lab, Mathematics Statistics, York University, Toronto, Canada
[g] Centre for Disease Modelling (CDM), Mathematics Statistics, York University, Toronto, Canada
[h] School of Mathematical Sciences, Queensland University of Technology, Brisbane Australia

ARTICLE INFO

Keywords:
COVID-19
Mathematical modelling
Within-host dynamics
Computational modelling
Population genetics
Machine learning
Immunopathology
SARS-CoV-2

ABSTRACT

The COVID-19 pandemic has revealed the need for the increased integration of modelling and data analysis to public health, experimental, and clinical studies. Throughout the first two years of the pandemic, there has been a concerted effort to improve our understanding of the within-host immune response to the SARS-CoV-2 virus to provide better predictions of COVID-19 severity, treatment and vaccine development questions, and insights into viral evolution and the impacts of variants on immunopathology. Here we provide perspectives on what has been accomplished using quantitative methods, including predictive modelling, population genetics, machine learning, and dimensionality reduction techniques, in the first 26 months of the COVID-19 pandemic approaches, and where we go from here to improve our responses to this and future pandemics.

1. Introduction

Severe respiratory syndrome coronavirus 2 (SARS-CoV-2), which causes coronavirus disease 2019 (COVID-19), was first identified in Wuhan, China in December 2019 and set off a pandemic that we are still grappling with in mid-2022. In response to this global threat, the scientific community rapidly mobilized to study and better understand SARS-CoV-2 genomics, its spread between individuals, its effects within hosts, and prevention and treatment strategies. Within this scope, mathematical and computational modelling has been heavily leveraged to assist public health and clinical decision making. The COVID-19 pandemic is one of the best examples of the real-time implementation of applied mathematical modelling (especially computational modelling) to answer crucial questions about the within-host response to a virus from its emergence in the population. In this vein, here we *evaluate* the impact that within-host modelling has had on our ability to understand the multiple challenges presented by the COVID-19 pandemic so we can *learn* from the strengths and weakness of the modelling community's response. This evaluation is critical to *planning* our continued response to this and future emerging infectious diseases.

The manifestation of COVID-19 in individuals is highly variable, ranging from asymptomatic to life-threatening. The inflammatory response is particularly important for controlling SARS-CoV-2 infections, and explains the wide-ranging symptoms observed in COVID-19. As with previous beta-coronaviruses (SARS-CoV-1 and Middle East respiratory syndrome or MERS), patients with severe disease typically exhibit high degrees of uncontrolled inflammation that is absent in individuals with mild infections. A concerted and intensive research effort was thus deployed to better understand the immunological factors determinant of and contributing to disease severity. In combination with clinical and experimental efforts, researchers using mathematical and computational immunology have been relied upon to help untangle

* Corresponding author at: Department of Mathematics and Statistics, Université de Montréal and Sainte-Justine University Hospital Research Centre, Montréal, Canada.
E-mail address: morgan.craig@umontreal.ca (M. Craig).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.immuno.2023.100021
Received 17 June 2022; Received in revised form 16 November 2022; Accepted 3 January 2023
Available online 8 January 2023

complicated longitudinal immunological data, and to use model predictions to generate new hypotheses about factors influencing disease severity and dynamics. As highlighted below, these models build upon the well-established basic viral dynamics and target cell limited models that have been used extensively to characterize other viral infections, including influenza, HIV, HPV, and oncolytic viruses. In parallel, computational approaches characterizing genetic evolution have been critical for improving our understanding of emerging variants and within-host viral evolution. Advanced data visualization and artificial intelligence techniques have also been deployed to characterize clinical features of COVID-19 and design more effective treatments and vaccines (Fig. 1).

In this perspective, we review the state of within-host modelling, computational population genetics, and data science and machine learning approaches developed and applied to SARS-CoV-2 to date. We also summarize the types of findings obtained thanks to these tools when applied to analyse diverse features related to COVID-19. However, given the need for fast dissemination of these methods, we note that the results remain preliminary in many cases, such that our focus is to give critical insight into the methodological advancements rather than on the biological discoveries. Future modelling directions for the COVID-19 and future pandemics are then discussed as a guide for our continued response to this and future emerging infectious diseases. Note that this review does not cover epidemiological models. For further reading on contributions in this field, please see, for example, Iranzo and Pérez-González [1] or Saldaño and Velasco-Hernández [2]. Our review is divided into six sections. First, we provide a brief introduction to the state of the field at the beginning and throughout the pandemic. Next, we describe modelling approaches to study within-host dynamics. The third section covers the description of mathematical methods to study the genetic origins of SARS-CoV-2 and emerging mutations, before we describe the various methods of dimensionality reduction to study and visualize immunological data from COVID-19 infected individuals in a fourth section. The fifth section is dedicated to predictive machine learning approaches to study the immunopathology of SARS-CoV-2 and vaccine and drug development. We conclude with future perspectives to help navigate this and the next pandemic.

2. Within host-modelling in the two years of the COVID-19 pandemic

2.1. Within-host immunological mathematical models based on ordinary differential equations

Various mathematical models have been successfully applied to characterize viral load kinetics of infectious viruses including influenza [3–12], SARS-CoV-1 [13,14], and MERS [15,16]. Since the emergence of COVID-19, studies have used a series of viral dynamics models to capture critical features of SARS-COV-2 infection processes. In this section, we focus on the deterministic within-host models that describe viral expansion and the corresponding immune responses after infection. Mathematical models have also been used to capture and predict the pharmacodynamic effects of various therapies to study the efficacy of proposed or existing treatments [17–20], helping the search for effective therapeutic strategies.

The target cell limited model is the simplest model to capture the viral dynamics of SARS-CoV-2 and has been used in many studies [21–23]. According to this model, the basic reproduction number (R_0) that measures the infection persistency (i.e., the number of cells infected by a single virion) is given by $R_0 = \frac{p\beta T_0}{c\delta}$, where p, β, c, δ and T_0 are the virion production rate, the infectivity rate, the rate of viral elimination, the death rate of infected cells, and the initial amount of target cells, respectively [23,24]. Asymptotically, if $R_0 < 1$, the infection will be eradicated, which is the goal of anti-viral treatment, and if $R_0 > 1$, the infection will grow. Thus, the interpretation of the within-host infection persistency is the same as the R_0 in epidemiological models.

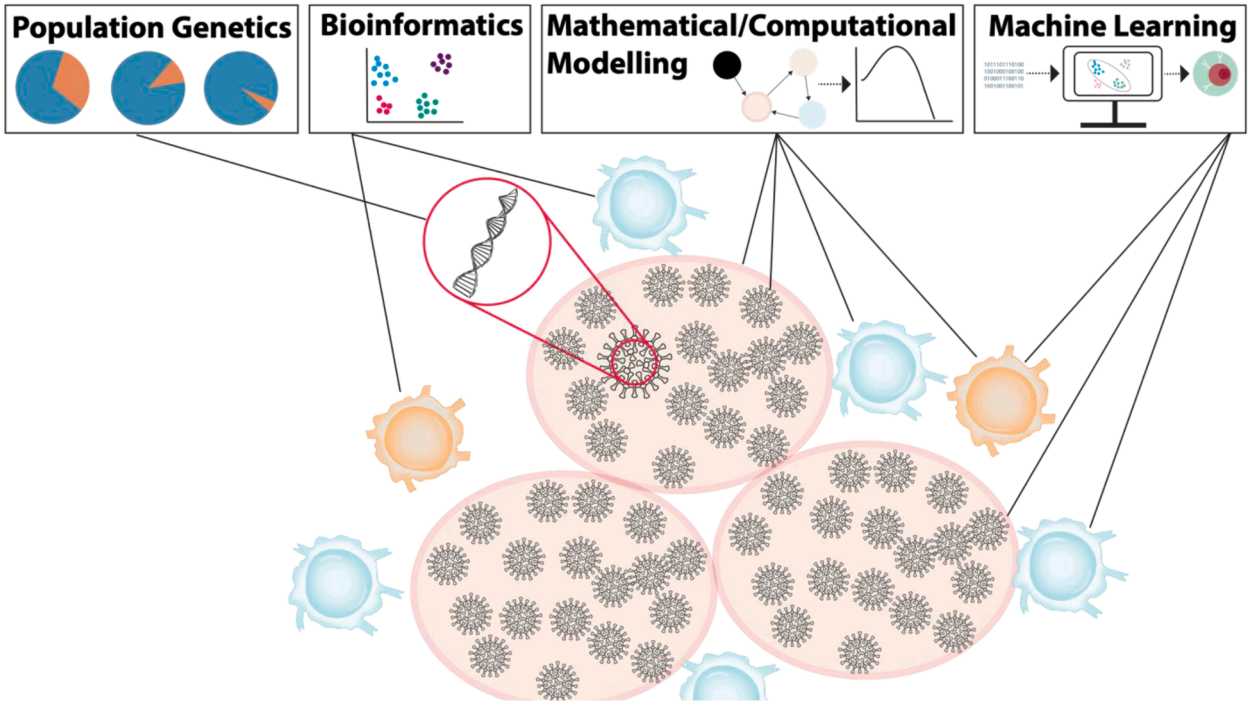

Fig. 1. Computational approaches to understanding the immune response and immunopathology in COVID-19 across scales. Beginning at the level of genes, the application of population genetics techniques enables the quantification of SARS-CoV-2 mutational patterns and dynamics (Section 3). Bioinformatics integrates computational and analytical methods to describe and interpret biological data through a variety of approaches, including dimensionality reduction (Section 4). Mathematical and computational modelling are means to quantitatively study and predict the immune response and immunopathology in COVID-19 (Section 2). Machine learning algorithms are able to effectively process multidimensional data and provide insights into complex systems that contribute to vaccine development and drug repurposing for COVID-19 (Section 5).

Unfortunately, this value becomes difficult to calculate as models become more complex.

Since infected cells usually take several hours to days before they start to produce infectious viral particles, a common extension of the target cell limited model is to consider an eclipse phase for infected cells which was first proposed by Baccam et al. [9] (see also, for example, Mittler et al. [25], Li and Shu [26], and reviews by Beauchemin and Handel [3] and Koelle et al. [27]). Based on the target cell limited model with an eclipse phase (during which cells are infected but not yet producing infectious virus), Néant et. al. [28] assumed that only a fraction μ of viral particles remained infectious while $1-\mu$ were noninfectious. They studied the relationship between viral kinetics and mortality on a cohort of French COVID+ hospitalized patients and explored which viral dynamics are associated with COVID-19 outcomes. For instance, they found that high viral loads in individuals were associated with mortality. Further, by integrating an antiviral drug model, their model predicted that a drug with 90% efficacy could accelerate viral clearance and decrease mortality.

SARS-CoV-2 causes infection in both the upper respiratory tract (URT) and the lower respiratory tract (LRT), with distinct infection dynamics in each tissue [29]. Ke et. al. [30] combined two extended target cell limited models and allowed for virus to move between the URT and LRT to capture the viral shedding dynamics in each compartment. By fitting the models to viral load data of nine SARS-CoV-2 infected patients reported in Wölfel et al. [31], their results indicated that the viral load dynamics in the URT provide an approximation for a person's infectiousness. They also determined that the long-term dynamics of SARS-CoV-2 are seeded by the continuous infection of new target cells in the LRT. Similarly, Wang et al. [32] constructed a model that included pneumocytes and lymphocytes as two groups of target cells. They concluded that their model significantly improved the fit of clinical data from both the URT and the LRT, through a comparison of model fits using the target cell limited model and its extended version with an eclipse phase. Moreover, they found that their extended model could exhibit a plateau after the initial viral load in the viral load curve which provides a better reflection of the underlying biology.

As viral loads have been found to be positively correlated to inflammation, several modelling studies have combined inflammation kinetics and viral kinetics. Starting from the basic SIV model, Fadai et al. [33] constructed a model with five components to capture inflammation kinetics, viral infection, and novel mechanisms of SARS-CoV-2, including recruited immune system cells, free SARS-CoV-2 virus, cells susceptible epithelial lung cells, infected cells, and pro-inflammatory mediators. Their model was able to capture the main clinical features observed in COVID-19 patients and their results indicated that early therapeutic intervention may effectively prevent the emergence of hyperinflammation, therefore decreasing the risk of severe disease. However, the finding of an unstable healthy steady state when the infection equilibrium exists suggests potential issues with this approach, given that individuals with moderate and severe disease can nonetheless successfully clear the virus.

As the first line of pre-existing defense in the host, innate immunity responds quickly to pathogens without requiring prior exposure. During the innate immune response, type-1 interferon (IFN) an important cytokine produced by infected cells can downregulate viral replication in infected and neighbouring cells [34], and activate immune cells during infection, including macrophages and natural killer cells, which can destroy infected cells [35–37]. Adaptive immunity develops over time after exposure to viruses and is mediated by lymphocytes including T cells and B cells. Therefore, researchers may use systems of delay differential equations to quantify the immune responses to viral dynamics. To study the immune responses to SARS-CoV-2, most studies are conducted considering both innate and adaptive responses rather than discussing their impacts separately.

In Goyal et al. [37], the authors extended the basic susceptible-infected-virus in-host model by considering the generation and function of effector T cells and characterizing both innate and adaptive immune responses to SARS-CoV-2 infection for individual patients in their cohort. For this, they included the stepwise production of effectors. Their model was shown to accurately characterize viral shedding kinetics, including viral expansion, a rapid decrease after an early peak, a slow decline period, and a final accelerating clearance phase for all patients. Subsequently, they studied the effects of drug timing on SARS-CoV-2 kinetics using their model. Their results improved our understanding of the immunopathology of SARS-CoV-2 infection and helped determine the optimal timing for anti-SARS-CoV-2 therapies. Overall, this work contributes to the development and optimization of therapeutic treatments in viral infections.

Jenner et al. [38] developed a mechanistic mathematical model to describe the within-host immune response to SARS-CoV-2 that modelled the interactions between epithelial cells, innate and adaptive immune cells (including $CD8^+$ T cells, neutrophils, macrophages, and monocytes), and cytokines. Furthermore, cytokine production dynamics and cytokine binding kinetics were explicitly considered by modelling both bound and free cytokine concentrations. After validation of the model against clinical and experimental data, Jenner et al. studied heterogenous COVID-19 severity in a virtual patient cohort. Their results identified key regulation processes of the immune response to SARS-CoV-2 infection in these virtual patients and suggested viable therapeutic targets, underlining the importance of a rational approach to studying novel pathogens using intra-host models.

Padmanabhan et al. [39] used a mathematical model of SARS-CoV-2 entry and dynamics to study the efficacy of repurposing drugs that block the activation of spike protein by two host proteases - TMPRSS2 and cysteine proteases Cathepsin B/Ls. Their results uncovered that treating both pathways independently provided successful prevention of virus entry. In their *in silico* study, Voutouri et al. [40] investigated the impact of risk factors such as age and existing comorbidities on disease progression to help establish optimal treatment courses. They developed a mathematical model predicting the expansion of infection that incorporated a patient's baseline health status. Their results indicated that the outcome of any therapy was strongly associated with the response rate of CD8+ T cells and balanced innate immune responses.

Immunological memory, including cellular and humoral memory produced by memory B cells, memory CD4+ and CD8+ T cells, and/or antibodies, is crucial for protection against re-exposure to infection and generating a long-term immune response to SARS-CoV-2. Numerous studies have successfully applied mathematical models to describe immune memory to influenza virus [4–6] and an increasing number of immunological studies [41–43] are revealing the different kinetics of B cell memory and T cell memory after SARS-CoV-2 infection. However, mathematical modelling studies quantifying immune memory to SARS-CoV-2 infections remain limited and warrant further study.

Memory responses are also particularly important to understand vaccine efficacy and establish optimal vaccination schedules. In that vein, Farhang-Sardoori et al. [44] constructed a mathematical model of the development of the memory immune response after adenovirus-based COVID-19 vaccines. Their model included antigen-presenting cells, CD8+ T cells, and cytokines including IL-6. The authors studied various vaccination strategies, including dose fractionation and extending the time between primers and boosting. For regimens with two standard doses or a standard dose followed by a low dose, they found that the minimum promoted antibody response was comparable with the neutralizing antibody level of 175 COVID-19 recovered patients. Their approach to investigating immune memory by introducing vaccine particles into within-host models provides a framework for vaccine selection and optimizing vaccination scheduling. This has been shown to be particularly salient for over the course of the pandemic, especially with respect to vaccine shortages and manufacturing delays. Similarly, Korosec et al. [45] used an extended version of the Farhang-Sardoori et al. model [44] to understand the

long-term humoral response to mRNA COVID-19 vaccines. Integrating a variety of data sources and using non-linear mixed-effects models to fit the data, they predicted an important decline in antibodies, notably a period longer than one month where an individual had less than 99% humoral immunity relative to peak immunity in the eight-month period following either Moderna or Pfizer mRNA vaccination.

The studies discussed above, and many others [46–50] (see also the recent perspective paper by Prague et al. [51]), highlight the use of mathematical modelling of the immune response to establish effective schedules or public-health vaccination strategies. However, a limitation of these deterministic models is their inability to capture the impact of stochasticity in COVID-19 severity. Thus, a large majority of the within-host modelling discussed above has concentrated on non-spatial effects and mean-field approximations of viral infectivity which may be necessary to understand the full extent of damage in severe COVID-19.

2.2. Computational, stochastic, and probabilistic models of SARS-CoV-2 dynamics

Computational stochastic modelling, including agent-based models (ABMs) [52], has become increasingly popular, particularly in oncology [52–54] and economics [55]. At the beginning of the COVID-19 pandemic, ABMs were rapidly deployed to model epidemiological spread at the population level [56–61]. For example, Warne et al. [62] used a stochastic epidemiological model combined with Bayesian methods to analyse the government response to COVID-19 in 158 countries and found that countries with the largest cumulative case tallies were characterized by a delayed response to the pandemic. Read et al. [60] developed an ABM of COVID-19 transmission to compare the impact of vaccination strategies while varying the temporal waning of vaccine efficacy following the first dose. They found no clear advantage of delaying the second dose with Pfizer-BioNTech. Garg et al. [63] constructed a stochastic model to predict antibody responses to different vaccine doses and timing. They found that reducing the first dose and increasing the time between doses results in improved responses, in agreement with clinical observations and the work of Farhang Sardroodi et al. [44] and Korosec et al. [45] discussed above.

While insightful for understanding heterogeneous population dynamics, such epidemiological models do not generally consider within-host dynamics or the immune response to infection. Given that immunopathology is characteristic of severe COVID-19, the ability to predict the kinetics of SARS-CoV-2 infection and the subsequent immune response at the tissue-level is essential for understanding COVID-19, its potential treatment, and the effects of vaccination. To that end, Sego et al. [64] developed an open-source platform for multiscale spatio-temporal simulation of epithelial tissue, viral infection, cellular immune responses, and tissue damage. This platform is specifically designed to be modular and extensible to support continuous updating and parallel development. By simulating the treatment of COVID-19, their results suggest that drugs that interfere with viral replication (e.g., remdesivir, an antiviral prodrug) yield substantially better infection outcomes when administered prophylactically, even at very low doses [65]. Similarly, using a community development approach, Getz et al. [66] constructed a cell-based model of SARS-CoV-2 infections and the subsequent systemic and tissue-level immune response based on the PhysiCell platform. Their framework was developed by concatenating models for receptor-mediated SARS-CoV-2 endocytosis, viral-induced pyroptosis, innate and adaptive immune responses and antigen presentation, type I interferon (IFN) dynamics, and the memory response in the lymph nodes. Using their combined model, Getz et al. simulated the effects of varying type I IFN dynamics, which have significant correlations with severe disease outcomes [67], and found that variable type I IFN dynamics induce large variations in immune cell numbers at the infection sites and determine the spatial distribution of these cells. These results provide us with an understanding of the spatial variation of local type I IFN dynamics and its impact on lung damage seen in human patients.

A large-scale community effort has also been put towards building an open-access, interoperable, and computable repository of SARS-CoV-2-virus-host interaction mechanisms called the COVID-19 disease map [68] that is a standardized knowledge repository guided by input from domain experts and based on published work. The map is a platform for visual exploration and computational analysis of molecular processes involved in SARS-CoV-2 entry, replication, and host-pathogen interactions, including immune responses, host cell recovery and repair mechanisms. The COVID-19 disease map is therefore a resource for graph-based analyses and disease modelling. For example, the map contains the pathways of the SARS-CoV-2 replication and its transcription including all relevant proteins and cellular mechanics. The goal of the map is to collate the fast-growing number of new SARS-CoV-2 publications in both human and machine-readable formats, support the research community in its understanding of this disease and to facilitate the development of efficient diagnostics and therapies.

In addition to multiscale and mechanistic modelling of acute infections within the host, computational methods can also be used to speed up the long and costly process of vaccine development [69]. For example, an early study searching for vaccine candidates used *in silico* methods to compare the sequence of N and S proteins of SARS-CoV-2 to B and T cell epitopes derived from SARS-CoV [70]. They identified epitopes for which no mutation had been observed in SARS-CoV-2 as of the 21st of February 2020 and proposed that immune targeted of these epitopes may potentially offer protection against this novel virus. Their findings provided a screened set of epitopes that can help guide experimental efforts toward the development of vaccines against SARS-CoV-2.

3. Population genetics of viral evolution

In early 2020, Wu et al. [71] reported the first genome sequence of SARS-CoV-2. The 30 Kilobase (Kb) genome consists of a single-stranded positive-sense RNA and codes for 16 non-structural proteins (nsp), 4 structural proteins, and 11 accessory proteins[72]. Since the release of the first genome sequence (NCBI Accession: NC_045512.2) [71], an unprecedented wealth of SARS-CoV-2 genomes have been sequenced internationally. Viral genomes accumulate mutations during the spread through human populations, with RNA viruses exhibiting the highest mutation rates of any group of organisms [73]. Since the introduction of SARS-CoV-2 RNA virus into human hosts, it has had a high mutational rate, despite its proofreading machinery involving the SARS-CoV-2-encoded 3′ exonuclease nsp14 [74]. Such high mutation rates increase the potential for fast viral adaptation and may hamper the development of vaccines and drugs.

Over the course of the pandemic, we have had to respond to new viral genetic variants with distinct characteristics, called variants of interest (VOI) or variants of concern (VOC), which have potential or confirmed impacts on transmission and human health. VOC are variants that are shown to be more transmissible or virulent or evade vaccines [75,76]. Understanding why these variants are concerning requires an investigation of the genomic epidemiology and evolution of SARS-CoV-2. In the last two years, evolutionary modelling techniques have been widely deployed to identify new emerging genetic variants raising widespread concerns, and to evaluate the impact of mutations on transmission, disease severity, immune response, and vaccine efficacy.

3.1. Phylogeny and population genetics to understand the origins and evolution of SARS-CoV-2

The large number of SARS-CoV-2 genomes generated in near real time led to myriad of data analysis approaches to understand the ongoing evolution of the virus. Phylogenetic approaches were first applied to multiple Sarbecovirus species genomes, and these analyses identified RaTG13, a CoV previously isolated from bat, as being the closest relative of SARS-CoV-2 [77,78]. Phylogenetic inference techniques have since been widely applied to large SARS-CoV-2 datasets,

predominantly based on maximum likelihood methods, such as TreeTime [79]. These approaches have brought informative inferences of evolutionary rates and time scale of the human outbreak. Using Bayesian phylogenetic analyses, Duchene et al. [80] argued that the phylodynamic threshold (the time at which the amount of observed molecular changes are sufficient for obtaining robust estimates from data) was met in March 2020 with hundreds of genomes, which allowed them to infer a time to the most recent common ancestor (TMRCA) between late October and mid-November 2019. However, phylogenetic tools have limitations in the context of millions of sequences, given the elevated computational demand and phylogenetic uncertainty due to highly similar sequences. Morel et al. [81] highlighted the difficulties of inferring reliable phylogenies given the high degree of sequence relatedness, calling for a cautious interpretation of the downstream inferred parameters. Furthermore, by only looking at the consensus sequence extracted from a sequenced sample, the classic phylogenetic approach also misses useful information to investigate the underlying mechanisms of viral evolution within hosts, which is of particular importance to understand viral-host dynamics and immune evasion.

On the other hand, as they are developed to study the evolution of populations using genetic sequences, population genetics models can accommodate varying levels of relatedness and divergence and can be applied at the level of the population or at the intra-individual scale to reveal the interplay between host-related mutational processes and transmission dynamics. Vasilarou et al. [82] modelled viral expansion in an approximate Bayesian computation (ABC) population genetics framework, an approach that uses stochastic simulations and summary statistics to bypass exact likelihood computation [83]. They investigated the evolution of early viral lineages and estimated the mutation rate of SARS-CoV-2 at 1.87×10^{-6} nucleotide substitutions per site per day as of April 2020. This means that each 30Kb genome will accumulate approximately 20 mutations per year, with the most recent estimate being up to 23.7 substitutions per year (https://nextstrain.org; accessed 1 February 2022). This increase in mutation rate within a two-year period reflects mutational bursts observed in sequences, leading to emerging lineages acquiring tens of mutations in a short amount of time. De Maio et al. [84] highlighted several genome sequence analysis pitfalls that can lead to inaccurate inference of mutation rates, such as assuming evolutionary equilibrium, not accounting for convergent mutations (recurring mutations, arising independently), and ignoring skewed mutational spectrum. Indeed, an excess of C-to-U mutations (40% of all single nucleotide variations) has been observed in SARS-CoV-2 and may be reminiscent of host-driven phenomenon, such as the action of human apolipoprotein B mRNA-editing enzyme, catalytic polypeptide (APOBEC) activity on single-strand RNA [85].

3.2. Genomic surveillance, natural selection, and variants of concern

The high prevalence of a newly arising mutation is determined by random drift and natural selection. When a limited number of viral particles establish a new large population during transmission, known as super-spreaders or founder events, the mutations present in their genome will increase in frequency regardless of their effects on viral fitness (its capacity to replicate and infect another host). For instance, Diez-Fuertes et al. showed that the earliest variants detected in Spain branched from a single viral clade, which they attributed to a founder effect [86,87]. Natural selection will also play a significant role in determining the fate of newly arising mutations, with those conferring a competitive advantage increasing in frequency (positive selection), and those reducing viral fitness being removed from the population of circulating viruses (negative selection).

Tracking new SARS-CoV-2 variants and distinguishing the ones that achieved high prevalence through positive selection from the ones that are random events is a key question for viral genomic surveillance. Extensive genomic surveillance data allow for the reconstruction of the dynamics of lineages locally, as done by Vöhringer et al. [88] in the UK between September 2020 and June 2021, leading to the identification of 71 different lineages across 315 English local authorities. The lineages were annotated using the Pangolin annotation system which is based on a computational approach that assigns to SARS-CoV-2 sequences the most likely lineage according to the Pango Lineage Nomenclature [89,90]. Using a Bayesian statistical model that estimates relative growth rates per lineage, this study tracked the fraction of genomes from different lineages in each local authority, accounting for differences in local epidemiological dynamics including in the rate of introduction of different lineages. Using classic population genetics statistical tools, Mostefai et al. [91] detected extensive population structure in viral genetic data from the first year of the pandemic, and characterized lineage expansion worldwide using changes in Tajima's D statistics [92] over time. Using birth-death processes, Scholer et al. [93] were able to quantify the impact of interventions on the extinction probability of deleterious SARS-CoV-2 variants, which is applicable in the initial outbreak of a new variant of concern. These studies showed how analysing SARS-CoV-2 genomic data using population genetics can be useful to predict the fate of VOC.

Population genetics modelling has also been used to detect mutations that give a competitive advantage with respect to viral replication, transmission, or escape from immunity. The first mutation inferred to be under positive selection was the D614G mutation in the spike glycoprotein, first detected in early March 2020 which then spread to become globally dominant in a few months. Volz et al. [94] used over 25,000 SARS-CoV-2 sequences and applied maximum likelihood phylogenetics reconstruction and an exponential growth coalescent model to contrast the growth rates of the 614 G and 614D sequences. Others have used the ratio of nonsynonymous to synonymous substitutions [95] and convergent evolution inference [96] to detect positive selection in SARS-CoV-2 genomes, but these methods did not all show conclusively evidence for positive selection at this mutation. This is in part because controlling for founder effects, population structure and sampling biases is a very difficult task, especially in the context of an emerging worldwide pandemic with low global viral genetic diversity. Nevertheless, *in vitro* data and animal models have confirmed the effects of D614G on receptor binding, indicating that 614 G viruses transmit more efficiently [97,98].

The D614G mutation is seen in all VOC, but Alpha, Delta and Omicron have all acquired a high number of lineage-characteristic mutations (22 [B.1.1.7], 20 [B.617.2] and 49–53 [BA.1, BA.2], respectively) [99]. These mutational bursts suggest significant increases in evolutionary rates for these variants, from evolutionary processes potentially occurring within chronically infected hosts [100] or via human-animal transmission [101]. All of these VOC outcompeted existing populations of circulating variants, strongly supporting that positive selection is the main driver of SARS-CoV-2 evolution at the population level.

3.3. Human-host genetic interactions

Host factors may play an important role in shaping SARS-CoV-2 genomic landscape. Indeed, for a set of mutations to become a variant segregating in a host population, they must survive intra-host selective pressures. The mechanisms for variant emergence can thus also be studied using intra-host genomic diversity. RNA viruses evolve rapidly by evading selective pressures from the host's immune response and adapting to the restrictive host environment. This leads to within-host selection for advantageous mutations, either generated from error-prone replications, or introduced by the host RNA-editing mechanisms [102,103]. These genetic variants within a host can be captured in next-generation sequencing reads as intra-host single nucleotide variants (iSNVs). Ramazotti et al. [104] introduced a methodological framework to characterize the intra-host genomic diversity of viral samples, revealing undetected infection chains and pinpointing mutations subjected to convergent evolution. Graudenzi et al. [105] identified specific distributions of nucleotide substitutions occurring within

hosts, which they call "non-overlapping mutational signatures", possibly impacted by purifying selection. Pathak et al. [106] found that many Delta (B.1.617.2) lineage-defining mutations appeared as iSNVs before getting fixed in the population. Finally, early bioinformatics analyses suggested that C-to-U mutations could be caused intra-host by APOBEC enzymes [107,108], and Kim et al. [85] added experimental support to these computational predictions, showing that APOBEC3A can target specific SARS-CoV-2 viral sequences for RNA editing, with the resulting mutations likely contributing to viral fitness.

Transmission bottleneck size, i.e., the size of the viral population transferred from the donor to the recipient, can also contribute to the intra-host viral diversity of the newly infected recipient individual. Popa et al. estimated the transmission bottleneck size of SARS-CoV-2 in the order of 1000 virion particles using viral genetic data [109]. However, re-examination of the same data set by Martin and Koelle demonstrated that SARS-CoV-2 exhibits a much narrower transmission bottleneck size (one to two virions) [110], a discrepancy they attributed to the low-frequency iSNVs, enriched for spurious mutations due to sequencing errors, called in the previous study. This illustrates the challenges in extracting meaningful information from sequencing data and the importance of stringent pre-processing of sequenced genomes to exclude artifacts.

Given the variation in disease symptoms and severity observed in the population, host genetic factors may explain differences in COVID-19 manifestations. Furthermore, susceptibility to infection may vary across individuals due to variability in genetically controlled pathogen clearance or persistence. Genetic association analysis in humans may thus allow for the identification of biological factors involved in the underlying progression and pathogenesis of the disease and in host susceptibility. A first Genome-Wide Association Study (GWAS) [111] was published in October 2020, detecting associations at two human genomic loci (3p21.31 and 9q34.2), both replicated in a meta-analysis of 46 cohorts [112] as well as in a trans-ancestry cohort of over one million research participants [113]. The 9q34 locus encompassed the ABO blood group locus and suggests that blood type O is protective against infection, unlike non-O blood types [114]. The chromosome 3 locus, which contains multiple candidate genes (including promising candidates such as *SLC6A20* (Solute Carrier Family 6 Member 20), *LZFTL1* (Leucine Zipper Transcription Factor Like 1), *CCR9* (CC Motif Chemokine Receptor 9), *CXCR6* (C-X-C Motif Chemokine Receptor 6), was strongly associated with severe respiratory outcomes. The strongest candidates at this locus are SLC6A20, a transporter protein potentially forming a complex with ACE2, as well as chemokine receptors CXCR6 and CCR9. Specifically, several studies have now proposed *CXCR6* as the causal gene, given its significant role, along with its ligand CXCL16, in the immunopathogenesis of severe COVID-19 [115–117], while epigenomic evidence also points to *CCR9* and *SLC6A20* as potential target genes [118]. Interestingly, Zeberg and Pääbo [119] identified a genomic segment within this locus that is inherited from Neanderthals, with each copy of this segment approximately doubling the risk of its carriers requiring intensive care when infected by SARS-CoV-2. Pairo-Casteneira et al. [120]. similarly found host genetic variants associated with critical illness in COVID-19 within *DPP9* (Dipeptidyl Peptidase 9) (19p13.3) and *IFNAR2* (Interferon Alpha And Beta Receptor Subunit 2) (21q22.1), as well as a gene cluster that encodes antiviral restriction enzyme activators (*OAS1, OAS2* and *OAS3* (2′−5′-Oligoadenylate Synthetase 1, 2, and 3)) on chromosome 12. A haplotype inherited from Neanderthals was also found at this latter locus, this time associated with reduced risk of becoming severely ill [121,122], suggesting that this haplotype may have been advantageous to modern humans throughout Eurasia in response to past RNA viruses. Association with *IFNAR2* polymorphisms, a gene which product mediates the cellular responses triggered by all type I IFN family members leading to the stimulation of antiviral genes [123], has been replicated in hospitalized patients with severe COVID-19 [124]. All the above-mentioned associated variants are commonly found in humans and do not show effect size heterogeneity between human populations, and therefore do not explain the differences in SARS-CoV-2 infection rates and hospitalization between Latino and African American compared to Americans from European ancestry [125,126], suggesting that the socioeconomical status of an individual might have a stronger effect on COVID19 outcomes. Finally, larger studies of sequencing datasets are starting to emerge to test the impact of rare genetic variants: for example, Horowitz et al. [127] identified a rare genetic variant close to ACE2, the cell surface receptor responsible for SARS-CoV-2 viral entry, that may confer protection against SARS-CoV-2 infection by modifying ACE2 expression levels. Further studies *in vivo* are warranted to investigate the causal impact of the identify associated loci on disease severity, and global efforts are now underway to analyse the genetics of individuals who are naturally resistant to SARS-CoV-2 infection [128].

4. Data visualization with dimensionality reduction techniques

As we can infer from the previous sections, vast amounts of viral, immunological, and sequencing data have been produced throughout the pandemic. Dealing with data, particularly complex immunological and genomics data, often requires developing and applying different visualization techniques to pre-process and understand them. Over the past few years, there have been significant improvements to such visualization approaches [129–134], with increasing application to questions involving biological data; COVID-19 is no exception.

To distinguish differences in immunological responses and search for connections between certain cell types and COVID-19 disease severity, Rébillard et al. [135] deployed flow cytometry, Phenograph [130] and FlowSom [136] to samples taken from Covid positive (Cov+) and Covid negative (Cov-) patients. In this study, cohorts of Cov+ and Cov- patients were matched according to pre-existing comorbidities. They were also compared to healthy controls (HCs). Rébillard et al. [135] used hierarchical clustering and uniform manifold approximation and projection (UMAP [132], see detailed explanation below) to further study the clinical features differentiating hospitalized SARS-COV-2 positive patients. These clinical characteristics included those typical of COVID-19 (e.g., fever and cough) in addition to the presence of chronic diseases (e.g., cancer, cardiovascular disease). To visualize the relationships between sampled immune cells, Rébillard et al. [135] performed FlowSom [136], using as an input, the number of clusters determined based on the modal value of clusters established by Phenograph [130], an algorithm similar to FlowSom that aims to detect communities (sets of highly connected nodes) that differ in density within the inferred interaction graph. Levine et al. [130] compared results using the Phenograph algorithm [130] to different techniques, including FLOCK [137], flowMeans [138], and SamSPECTRAL [139], and found that Phenograph gave better results in terms of the robustness and the overall quality of the final outcome. Using FlowSom and Phenograph, Rébillard et al. [135] discovered changes to the number of peripheral immune cell subpopulations (e.g. CD19+ B cells) in both Cov+ and Cov- severely ill patients as compared to health care workers, and a reduction of some specific immune cell subsets (e.g., CD27+ T cells) in Cov+ versus Cov- patients. To confirm and broaden these results, Rébillard et al. [135] then performed a hypothesis-driven analysis based on conventional manual gating and found a large increase in number of neutrophils connected to both disease severity (adverse outcomes) and age. However, this was not found to be characteristic of SARS-CoV-2 infections as it was observed in both Cov+ and Cov- patients, in contrast to the reduction of B cells and the increased percentage of some lymphocytes (e.g., CD38+ CD8+ killer T cells) which was typical of severe COVID-19. Furthermore, Rébillard et al. [135] noted a depletion of natural killer (NK) cells in severe Cov+ cases and Cov- patients compared to health care workers coupled with the reduction of CD4+ T cells expressing CD38 in hospitalized patients, regardless of the severity or age.

Manual gating using flow cytometry becomes rapidly unsuitable when dealing with larger and high-dimensional data [131]. Another

visualization and data-analysis technique that can be applied to help overcome these inconveniences is t-distributed Stochastic Neighbor Embedding (tSNE) [140]. Like FlowSom, tSNE [140] is a clustering technique that can be performed early within a data visualization pipeline (for example, as a basis for further analysis using FlowSom or Phenograph). However it has been shown that tSNE often fails to completely separate cell populations [131]. Other popular dimension reduction techniques include UMAP [132], which consists of searching for an optimal embedding by finding the fuzzy topological structure of the low dimensional data projection that is most similar to the original manifold. However, UMAP can only be performed under specific conditions, namely that data points should be uniformly distributed on the locally connected Riemannian manifold (i.e., there should be no isolated points), and the local Riemannian metric should be constant.

Principal component analysis (PCA) is another powerful technique to summarize genetic data and identify genetic structure and can be used to detect emerging viral sub-lineages from SARS-CoV-2 genetic data [58]. However, the dimensionality reduction required for final data visualization often depletes the quality of the resulting outcome. For example tSNE and PCA [141] suffer from sensitivity to noise, or do not preserve global (tSNE) or local (PCA) structures within the data [133]. To reduce these drawbacks, another technique called PHATE [133] was developed. Later PHATE was combined with improved diffusion condensation [142] to allow for large-scale visualization (i.e., Multiscale PHATE) [134]. In comparison to UMAP and tSNE, Multiscale PHATE gave significantly better results with regards to cell similarities (i.e., keeping proper distance between familiar and unfamiliar cell types). Further, the use of multiscale clusters in Multiscale PHATE distinguishes higher levels of data grouping than UMAP and tSNE. The most important advantages of using Multiscale PHATE are that the data can be visualized in all levels of granularity.

Multiscale PHATE was used by Kuchroo et al. [134] to evaluate 251 blood samples taken from the 168 Cov+ patients, resulting in the analysis of the 54 million cells. Patient similarities were analysed by creating patient manifolds based on multiresolution cluster estimation, a technique invented here by authors that combines work from [143,144]. This estimation was repeated for every sample to create a multiscale feature matrix, which was subsequently embedded using PHATE to obtain an improved visualization. The authors found that the number of pathogenic T, B and myeloid cells, in addition to granulocytes were increased in patients who died of COVID-19. T cells that expressed Granzyme B were found to be particularly strongly associated with the mortality. To uncover the connections between age/sex, disease severity, and outcomes, these clinical variables were mapped directly onto the patients' manifold. Using MELD [145], Kuchroo et al. [134] found that mortality was tightly linked to age and that male patients were more likely to experience severe COVID-19 with the need of oxygen support. Kuchroo et al. [134] then used DREMI [146] to find that female (and young) patients were more likely to have better outcomes, which was found to be related to their ability to mount strong T cells response as compared to men (and older) individuals. Moreover, their analysis showed that the increased expression of IL-2 and IL-6 cytokines was crucially associated with an adverse outcome of an infection.

5. Predictive machine learning approaches

In parallel to prospective modelling, machine learning (ML) has also played a prominent role throughout the pandemic, as it has been applied ubiquitously in many real-world applications that require the identification of trends and patterns. Multidimensional data are prevalent in many of these situations, and ML has proven to be proficient at providing insight into such complex data. ML capabilities lie in its unique ability to learn from training data, generalize patterns, and make inferences beyond the initial data. Continual improvement cycles based on the availability of new or real-time data make ML a suitable candidate to improve model prediction and adaptability. Within ML, the rapid advancement of deep learning (DL) allows for the inclusion of automatic feature extraction from the training data [147]. Hence, ML is deployed as an effective tool to address the rapidly changing nature of the COVID-19 pandemic at multiple scales. The use of machine learning in the context of the COVID-19 pandemic can be divided into a few broad categories:

(1) Vaccine development [148,149] and drug re-purposing [150–153].
(2) Patient triage for severity assessment and mortality prediction [154–156].
(3) Epidemiological forecasting [157,158].
(4) Detection of anomalies from medical images such as CT scans [159–161], chest X-rays [162], ultrasound images [163] and blood tests [164].

In this section, we focus on ML applied to understand immunopathology, specifically in the areas of vaccine development, and drug discovery and repurposing for COVID-19.

ML techniques are complementary to existing within-host ODE-based mathematical models like those discussed in previous sections and, in most cases, offer rapid prototyping of predictive models for immediate deployment for clinical use. Mathematical models can also complement machine learning as demonstrated by the work of Rosado et al. where the authors combined both techniques to provide an accurate and robust serological classification of individuals previously infected by SARS-CoV-2 [165]. ML classifiers were trained with multiplex data from these individuals using the random forest (RF) algorithm. Next, a Bayesian mathematical model was adopted to describe antibody kinetics informed by prior information from other coronaviruses. Together, the predictive capability of the Rosado et al. approach comes from a statistical estimator that gauges the seroprevalence of SARS-CoV-2 infections in very low-transmission settings. Farhang-Sardroodi et al. [166] used ML on within-host models calibrated to patients with either influenza or SARS-CoV-2 infections to distinguish features specific to either viral infection. They found that the ML classifiers were able to distinguish, with high accuracy, the kinetics of influenza and SARS-CoV-2 infections, suggesting that early viral dynamics differentiate these two viruses.

5.1. Machine learning approaches to COVID-19 vaccine development

To launch a safe and effective vaccine, the conventional vaccine research and development (R&D) pipeline requires significant financial investments over 5 to 10 years. Due to the urgent need for a COVID-19 vaccine, a paradigm shift in the regulatory process that espoused parallel clinical trials was required for the COVID-19 vaccine R&D. The genome-based vaccine design approach coined as reverse vaccinology (RV) was proposed by Rappuoli in 2000 [167]. Unlike conventional vaccines developed using pathogenic organisms, RV leverages expressed genetic sequence for vaccine discovery. Through the comparison of various classification techniques, including logistic regression (LR), support vector machine (SVM), k-nearest neighbour (KNN), RF, and extreme gradient boosting (XGB), various data resampling and performance metrics (AUPRC, AUROC) were established. The first viral sequence of SARS-CoV-2 was available in early 2020, and RV technology was already in place to take advantage of this information for rapid COVID-19 vaccine development. Vaxign-ML, which uses the XGB technique, was shown to be the best predictor of the original data [168]. Subsequently, a comprehensive RV webserver, Vaxign2 was developed to analyse SARS-CoV-2 vaccine candidates [148]. Vaxign2 based on Vaxign-ML was able to predict two critical candidates for vaccine development: spike (S) glycoprotein and non-structural protein 3 (nsp3). Putative target protein antigens can also be quickly identified using RV. To investigate target protein antigens, several predictive immunoinformatic tools were previously developed [168–173]. Supervised

ML classification techniques were mainly adopted for RV prediction of protective antigens. A machine learning workflow that combined the Markov model and propensity scale method was shown to analyse the proteome of SARS-CoV-2 and successfully identify putative T cell and B cell epitopes [174]. The identification of these epitopes spurred COVID-19 vaccine development. Similarly, an architecture combining a neuronal network architecture (SPAAN) and Hidden Markov Model (HMM) was developed as an RV technique for predicting COVID-19 vaccine candidates [175]. To screen for statistically significant epitope hotspot regions, other ML studies in this sphere include an AI approach to design a COVID-19 vaccine by generating a comprehensive epitope map from the NEC Immune Profiler tool and using the results as an input to Monte Carlo simulations [176]. Meanwhile, the screening of epitopes combined with the Deep Learning (DL) approach resulted in a framework called the DeepVacPred that showed that this DL approach can predict up to 26 potential vaccine subunits suitable for the design of a multi-epitope vaccine [177].

To develop a new vaccine, understanding peptide binding to major histocompatibility complex (MHC) is the single most selective biological process that determines a successful and optimal antigen processing and presentation. Hence, predicting peptide–MHC binding has become a focus in the field of immunoinformatics, where ML plays an important role. By leveraging the advancement in adversarial neural networks (ANN), the NNAlign framework has accurately characterized binding motifs [178]. In this framework, the optimal binding core of selected amino acids is searched for, and peptides matching a consensus motif or model bindings are predicted. NNAlign iteratively updates model parameters while minimizing the difference between the predicted and measured binding. This approach has become the basis for training NetMHC, NetMHCII and NetMHCIIpan. As a result, a study to identify peptides with epitope potentials for COVID-19 vaccines revealed 94 predicted peptides for 11 HLA alleles using NetMHC tools [179]. In another study, NetMHCpan was used to predict a global loss of SARS-CoV-2 T cell epitopes in individuals expressing HLA-B alleles of the B7 supertype family [180]. Similarly, two supervised neural network-driven tools (NetMHCpan4 and MARIA) were applied and were shown to screen potential T-cell epitopes for SARS-CoV- 2 close to the SARS-CoV-2 receptor-binding domain [149].

5.2. Machine learning approaches to drug repurposing for COVID-19

Throughout the first two years of the COVID-19 pandemic, ML was also leveraged to study existing drugs that have antiviral properties. Here the objective was to quickly predict drug-disease interactions and disease pathways by exploiting existing approved drugs that are proven to be safe. This approach was especially important to rapidly screen potential therapeutic drugs for COVID-19 and speed up new clinical trials. Drug repurposing applying graph Convolutional Network with Attentional mechanism (Att-GCN-DDI) allows us to better understand drug-disease interactions, and a few drug candidates that were eventually proven effective in clinical treatment were predicted by Att-CGN-DDI[150]. Similarly, Beck et al. [151] screened antiviral drugs against the SARS-CoV-2 virus and applied a pre-trained deep learning-based drug-target interaction model called Molecule Transformer-Drug Target Interaction (MT-DTI) to identify commercially available drugs. Using this framework, the group identified multiple antiviral drugs such as atazanavir and remdesivir as having high inhibitory potency against SARS-CoV-2.

5.3. Applying generative machine learning approaches for drug discovery

Machine learning has additionally been broadly adopted by the pharmaceutical industry to revolutionize drug discovery. In this vein, Variational AutoEncoders (VAE), a type of generative model that regularizes the encoding distribution during training to populate its latent space with desirable properties so that the final model can generate new data based on these properties [181], have been deployed. Several ML frameworks based on VAE can accurately generate novel molecular structures that capture chemical properties such as bond order and functional groups. The resulting novel molecules rank highly in metrics such as the quantitative estimate of drug-likeness (QED) score or synthetic availability score (SAS) [182,183]. A framework called Controlled Generation of Molecules (CogMol) applied the pre-training of a molecular Simplified Molecular Input Line Entry System (SMILES) VAE. CogMol targeted three SARS-CoV-2 target proteins and generated novel drug candidates with a high binding affinity to target proteins [184]. Alternatively, Tang et al. 2020 [185] proposed a fragment-based drug design methodology combined with a deep Q-learning network to speed up the generation of potential candidate compounds against SARS-CoV-2. This novel framework called the ADQN-FBDD was developed from a library of 284 known SARS-CoV-2 inhibitor molecules. It successfully generated a library of 4922 covalent lead compounds with unique valid structures, and 47 lead compounds were further identified for molecular docking evaluations [185].

From vaccine development to drug repurposing and drug discovery studies, machine learning has an important role to play in responding to an emerging infectious disease like SARS-CoV-2. Future work improving the integration of ML approaches, mechanistic mathematical and computational models, bioinformatics, and population genetics approaches will allow an even more rapid response to new pandemic scenarios, hopefully improving public health.

6. Reflections and future perspectives

As discussed above, quantitative approaches have been particularly prominent during the COVID-19 pandemic, owing to open-science endeavours, the availability of data, and the increased integration of quantitative scientists in biomedicine. Though epidemiological applications are at the forefront of mathematical and computational modelling for public health, genetic characterization of viruses and hosts, immunological applications related to discerning mechanisms of severe COVID-19, and treatment and vaccine development have also been developed and applied in tandem with experimental and clinical advances.

Despite the successes of the mathematical tools highlighted here, challenges remain for the current COVID-19 pandemic and future emerging infectious diseases (Table 1). Mathematical and computational modelling of the immunovirology of SARS-CoV-2 have elaborated the kinetics of infection [17,23,64,66,186], the actions of the immune response to infection [17,28,38], and the effects of vaccination [44,45], while data science approaches have elucidated mechanisms of disease in hospitalized patients [187]. Modelling, in particular, is heavily improved by densely sampled longitudinal data that can be difficult to

Table 1
Approaches needed to address challenges faced for the next pandemics.

Challenge	Future actions
Pace of emergence of longitudinal immunological data	Establish guidelines for data collection needed for modelling prior to epidemics/pandemics; establish translatable immunological models that can be rapidly adapted according to emerging data
Modelling networks and model sharing not operational until after beginning of outbreak	Set up and maintain working groups between different stakeholders (funding agencies, researchers, clinicians, and public health authorities) for rapid mobilization
Integrating immunological data with genetic information about variants and human hosts	Collect longitudinal viral sequencing data paired with clinical and immunological meta-data to leverage within-host genetic diversity to identify emerging variants

collect in the general population (e.g., individuals with mild infections who are not hospitalized); kinetic rates may be difficult or impossible to measure in humans, compounding data-related difficulties. This can be addressed through improved integration within quantitative fields (i.e., combining prospective and retrospective models) and between disciplines to prioritize decision making that is relevant to public health and clinical authorities. For an infectious disease like COVID-19 for which it may not be possible to achieve eradication, understanding the immunological features of the transition to endemicity is also a strength of predictive mathematical modelling [188]. Further, emphasis needs to be placed on timely model development to provide clinicians, and drug and vaccine developers with real-time predictions. Ultimately, these issues are not exclusive to modelling of novel infectious diseases like COVID-19 but become amplified during times of crisis.

The study of viral evolution has been a key component of our response to SARS-CoV-2. Population genetics modelling will need to be refined to predict the potential future of SARS-CoV-2 variants as we move to reopen societies and transition from a pandemic to an endemic context. These models will have to take into account the increasing evidence for recombination in SARS-CoV-2 resulting from co-infections [189], and consider the potential importance of animal-to-human transmission of SARS-CoV-2 [190]. The advent of machine learning has also allowed for accelerated approaches to vaccine development and drug discovery. For example, ML techniques such as deep learning, hidden Markov model and adversarial neural network identified important epitopes, antigen protein and peptide-MHC binding affinity to accelerate the development of a vaccine for COVID-19. At the height of the pandemic, drug repurposing was touted as a quick solution to use existing approved drugs to treat an infection. Several ML framework such as Att-GCN-DDI and MT-DTI were able to predict the efficacy of approved drugs and hence have an immediate effect on patients' disease outcome. For a long-term research strategy, ML can be used to generate novel drug molecules that are more effective than repurposing drugs. This requires more investment in time and resources to train the framework with emerging data on SARS-CoV-2. Whether it is vaccine development, drug re-purposing or generating new drug designs, ML has proven to substantially speed up the development time which plays an important role in mitigating the effect of the pandemic.

Overall, the outlook for the continued integration and use of predictive modelling to answer immunovirological questions is positive. Throughout the COVID-19 pandemic, the public has become more sensitized to modelling and quantitative methods. Concerted efforts to maintaining the scientific progress made over the past 26 months is critical to the success of these endeavours. Ultimately, our response to the next pandemic will depend on how well we can translate our current successes and address our failures and pitfalls to newly emerging infectious diseases, and we contend that a key component depends greatly on predictive modelling and analysis.

Funding

This work was supported by the Natural Sciences and Engineering Research Council of Canada (Discovery Grant RGPIN-2018–04546 (MC), Alliance COVID-19 Grant ALLRP 554923-20 (JH and MC)), the Coronavirus Variants Rapid Response Network (CoVaRR-Net) (JH) and the National Research Council of Canada (NRC) (JHKO and JMH).

CRediT authorship contribution statement

Sonia Gazeau: Conceptualization, Writing – review & editing. **Xiaoyan Deng:** Conceptualization, Writing – review & editing. **Hsu Kiang Ooi:** Conceptualization, Writing – review & editing, Funding acquisition. **Fatima Mostefai:** Conceptualization, Writing – review & editing. **Julie Hussin:** Conceptualization, Writing – review & editing, Funding acquisition. **Jane Heffernan:** Conceptualization, Writing – review & editing, Funding acquisition. **Adrianne L. Jenner:** Conceptualization, Writing – review & editing. **Morgan Craig:** Conceptualization, Writing – review & editing, Funding acquisition.

Declaration of Competing Interest

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

Data availability

No data was used for the research described in the article.

References

[1] Iranzo V, Pérez-González S. Epidemiological models and COVID-19: a comparative view. Hist Philos Life Sci 2021;43:104. https://doi.org/10.1007/s40656-021-00457-9.
[2] Saldaña F, Velasco-Hernández JX. Modeling the COVID-19 pandemic: a primer and overview of mathematical epidemiology. SeMA J 2022;79:225–51. https://doi.org/10.1007/s40324-021-00260-3.
[3] Beauchemin CAA, Handel A. A review of mathematical models of influenza A infections within a host or cell culture: lessons learned and challenges ahead. BMC Public Health 2011;11. https://doi.org/10.1186/1471-2458-11-s1-s7.
[4] Zarnitsyna VI, et al. Mathematical model reveals the role of memory CD8 T cell populations in recall responses to influenza. Front Immunol 2016;7. https://doi.org/10.3389/fimmu.2016.00165.
[5] Myers MA, et al. Dynamically linking influenza virus infection kinetics, lung injury, inflammation, and disease severity. Elife 2021;10. https://doi.org/10.7554/eLife.68864.
[6] Hancioglu B, Swigon D, Clermont G. A dynamical model of human immune response to influenza A virus infection. J Theor Biol 2007;246:70–86. https://doi.org/10.1016/j.jtbi.2006.12.015.
[7] Smith AM, Perelson AS. Influenza A virus infection kinetics: quantitative data and models. Wiley Interdiscip Rev Syst Biol Med 2011;3:429–45. https://doi.org/10.1002/wsbm.129.
[8] Boianelli A, et al. Modeling influenza virus infection: a roadmap for influenza research. Viruses 2015;7(10):5274–304. https://doi.org/10.3390/v7102875.
[9] Baccam P, Beauchemin C, Macken CA, Hayden FG, Perelson AS. Kinetics of influenza A virus infection in humans. J Virol 2006;80:7590–9.
[10] Smith AP, Moquin DJ, Bernhauerova V, Smith AM. Influenza virus infection model with density dependence supports biphasic viral decay. Front Microbiol 2018;9:1554. -1554.
[11] Boianelli A, et al. Modeling influenza virus infection: a roadmap for influenza research. Viruses 2015;7:5274–304. https://doi.org/10.3390/v7102875.
[12] Antia R, et al. Modeling within-host dynamics of influenza virus infection including immune responses. PLoS Comput Biol 2012;8. https://doi.org/10.1371/journal.pcbi.1002588.
[13] Zhou Y, Ma Z, Brauer F. A discrete epidemic model for SARS transmission and control in China. Math Comput Model 2004;40:1491–506. https://doi.org/10.1016/j.mcm.2005.01.007.
[14] Sugden B, et al. A quantitative model used to compare within-host SARS-CoV-2, MERS-CoV, and SARS-CoV dynamics provides insights into the pathogenesis and treatment of SARS-CoV-2. PLoS Biol 2021;19. https://doi.org/10.1371/journal.pbio.3001128.
[15] Yong B, Owen L. Dynamical transmission model of MERS-CoV in two areas. AIP Conf Proc 2016;1716:020010. https://doi.org/10.1063/1.4942993.
[16] Chang HJ. Estimation of basic reproduction number of the Middle East respiratory syndrome coronavirus (MERS-CoV) during the outbreak in South Korea, 2015. Biomed Eng Online 2017;16. https://doi.org/10.1186/s12938-017-0370-7.
[17] Goyal A, Cardozo-Ojeda EF, Schiffer JT. Potency and timing of antiviral therapy as determinants of duration of SARS-CoV-2 shedding and intensity of inflammatory response. Sci Adv 2020;6:eabc7112. https://doi.org/10.1126/sciadv.abc7112. -eabc7112.
[18] Tarek M, Savarino A. Pharmacokinetic basis of the hydroxychloroquine response in COVID-19: implications for therapy and prevention. Eur J Drug Metab Pharmacokinet 2020;45:715–23. https://doi.org/10.1007/s13318-020-00640-6.
[19] Conway JM, Abel Zur Wiesch P. Mathematical modeling of remdesivir to treat COVID-19: can dosing be optimized? Pharmaceutics 2021;13. https://doi.org/10.3390/pharmaceutics13081181.
[20] Hernandez-Vargas EA, Velasco-Hernandez JX. In-host mathematical modelling of COVID-19 in humans. Annu Rev Control 2020;50:448–56. https://doi.org/10.1016/j.arcontrol.2020.09.006.
[21] Kim KS, et al. A quantitative model used to compare within-host SARS-CoV-2, MERS-CoV, and SARS-CoV dynamics provides insights into the pathogenesis and treatment of SARS-CoV-2. PLoS Biol 2021;19. https://doi.org/10.1371/journal.pbio.3001128.

[22] Abuin P, Anderson A, Ferramosca A, Hernandez-Vargas EA, Gonzalez AH. Characterization of SARS-CoV-2 dynamics in the host. Annu Rev Control 2020;50:457–68. https://doi.org/10.1016/j.arcontrol.2020.09.008.

[23] Kim, K.S. et al. A quantitative model used to compare within-host SARS-CoV-2, MERS-CoV, and SARS-CoV dynamics provides insights into the pathogenesis and treatment of SARS-CoV-2. PLOS Biology. 2021 19(3): e3001128. https://doi.org/10.1371/journal.pbio.3001128.

[24] Hill AL, Rosenbloom DIS, Nowak MA, Siliciano RF. Insight into treatment of HIV infection from viral dynamics models. Immunol Rev 2018;285:9–25. https://doi.org/10.1111/imr.12698.

[25] Mittler JE, Sulzer B, Neumann AU, Perelson AS. Influence of delayed viral production on viral dynamics in HIV-1 infected patients. Math Biosci 1998;152:143–63. https://doi.org/10.1016/S0025-5564(98)10027-5.

[26] Li MY, Shu H. Impact of intracellular delays and target-cell dynamics on *in vivo* viral infections. SIAM J Appl Math 2010;70:2434–48. https://doi.org/10.1137/090779322.

[27] Koelle K, Farrell AP, Brooke CB, Ke R. Within-host infectious disease models accommodating cellular coinfection, with an application to influenza†. Virus Evol 2019;5. https://doi.org/10.1093/ve/vez018.

[28] Néant N, et al. Modeling SARS-CoV-2 viral kinetics and association with mortality in hospitalized patients from the French COVID cohort. Proc Natl Acad Sci 2021;118:e2017962118. https://doi.org/10.1073/pnas.2017962118.

[29] Chen PZ, et al. SARS-CoV-2 shedding dynamics across the respiratory tract, sex, and disease severity for adult and pediatric COVID-19. Elife 2021;10. https://doi.org/10.7554/eLife.70458.

[30] Ke R, Zitzmann C, Ho DD, Ribeiro RM, Perelson AS. *In vivo* kinetics of SARS-CoV-2 infection and its relationship with a person's infectiousness. Proc Natl Acad Sci 2021;118. https://doi.org/10.1073/pnas.2111477118.

[31] Wölfel R, et al. Virological assessment of hospitalized patients with COVID-2019. Nature 2020;581:465–9. https://doi.org/10.1038/s41586-020-2196-x.

[32] Wang S, et al. Modeling the viral dynamics of SARS-CoV-2 infection. Math Biosci 2020;328:108438. https://doi.org/10.1016/j.mbs.2020.108438.

[33] Fadai NT, et al. Infection, inflammation and intervention: mechanistic modelling of epithelial cells in COVID-19. J R Soc Interface 2021;18:20200950. https://doi.org/10.1098/rsif.2020.0950.

[34] Park A, Iwasaki A. Type I and type III interferons – induction, signaling, evasion, and application to combat COVID-19. Cell Host Microbe 2020;27:870–8. https://doi.org/10.1016/j.chom.2020.05.008.

[35] García-Sastre A, Biron CA. Type 1 interferons and the virus-host relationship: a lesson in détente. Science 2006;312:879–82. https://doi.org/10.1126/science.1125676.

[36] Mandelboim O, et al. Recognition of haemagglutinins on virus-infected cells by NKp46 activates lysis by human NK cells. Nature 2001;409:1055–60. https://doi.org/10.1038/35059110.

[37] Goyal A, Duke ER, Cardozo-Ojeda EF, Schiffer JT. Mathematical modeling explains differential SARS CoV-2 kinetics in lung and nasal passages in remdesivir treated rhesus macaques. bioRxiv 2020.

[38] Jenner AL, et al. COVID-19 virtual patient cohort suggests immune mechanisms driving disease outcomes. PLoS Pathog 2021;17:e1009753. https://doi.org/10.1371/journal.ppat.1009753. -e1009753.

[39] Padmanabhan P, Desikan R, Dixit NM. Modeling how antibody responses may determine the efficacy of COVID-19 vaccines. Nat Comput Sci 2022;2:123–31. https://doi.org/10.1038/s43588-022-00198-0.

[40] Voutouri C, et al. *In silico* dynamics of COVID-19 phenotypes for optimizing clinical management. Proc Natl Acad Sci 2021;118:e2021642118. https://doi.org/10.1073/pnas.2021642118.

[41] Dan JM, et al. Immunological memory to SARS-CoV-2 assessed for up to 8 months after infection. Science 2021;371. https://doi.org/10.1126/science.abf4063.

[42] Cohen KW, et al. Longitudinal analysis shows durable and broad immune memory after SARS-CoV-2 infection with persisting antibody responses and memory B and T cells. Cell Rep Med 2021;2. https://doi.org/10.1016/j.xcrm.2021.100354.

[43] Hartley GE, et al. Rapid generation of durable B cell memory to SARS-CoV-2 spike and nucleocapsid proteins in COVID-19 and convalescence. Sci Immunol 2020;5. https://doi.org/10.1126/sciimmunol.abf8891.

[44] Farhang-Sardroodi S, et al. Analysis of host immunological response of adenovirus-based COVID-19 vaccines. Vaccines 2021;9:861. https://doi.org/10.3390/vaccines9080861 (Basel)-861.

[45] Korosec CS, et al. Long-term durability of immune responses to the BNT162b2 and mRNA-1273 vaccines based on dosage, age and sex. Sci Rep 2022;12:21232. https://doi.org/10.1038/s41598-022-25134-0.

[46] Sadria M, Layton AT. Modeling within-host SARS-CoV-2 infection dynamics and potential treatments. Viruses 2021;13. https://doi.org/10.3390/v13061141.

[47] Nath BJ, Dehingia K, Mishra VN, Chu YM, Sarmah HK. Mathematical analysis of a within-host model of SARS-CoV-2. Adv Differ Equ 2021;2021. https://doi.org/10.1186/s13662-021-03276-1.

[48] Ghosh I. Within host dynamics of SARS-CoV-2 in humans: modeling immune responses and antiviral treatments. SN Comput Sci 2021;2. https://doi.org/10.1007/s42979-021-00919-8.

[49] Regoes RR, et al. SARS-CoV-2 viral dynamics in non-human primates. PLoS Comput Biol 2021;17. https://doi.org/10.1371/journal.pcbi.1008785.

[50] Pinky L, Dobrovolny HM. SARS-CoV-2 coinfections: could influenza and the common cold be beneficial? J Med Virol 2020;92:2623–30. https://doi.org/10.1002/jmv.26098.

[51] Prague M, Alexandre M, Thiébaut R, Guedj J. Within-host models of SARS-CoV-2: what can it teach us on the biological factors driving virus pathogenesis and transmission? Anaesth Crit Care Pain Med 2022;41. https://doi.org/10.1016/j.accpm.2022.101055.

[52] Metzcar J, Wang Y, Heiland R, Macklin P. A review of cell-based computational modeling in cancer biology. JCO Clin Cancer Inform 2019;2:1–13. https://doi.org/10.1200/cci.18.00069.

[53] Miller-Jensen K, Cess CG, Finley SD. Multi-scale modeling of macrophage—T cell interactions within the tumor microenvironment. PLoS Comput Biol 2020;16. https://doi.org/10.1371/journal.pcbi.1008519.

[54] Jenner AL, et al. Agent-based computational modeling of glioblastoma predicts that stromal density is central to oncolytic virus efficacy. iScience 2022;25. https://doi.org/10.1016/j.isci.2022.104395.

[55] Haldane AG, Turrell AE. Drawing on different disciplines: macroeconomic agent-based models. J Evol Econ 2018;29:39–66. https://doi.org/10.1007/s00191-018-0557-5.

[56] Hoertel, N. et al. Facing the COVID-19 epidemic in NYC: a stochastic agent-based model of various intervention strategies. medRxiv: the preprint server for health sciences, 2020.2004.2023.20076885 (2020). 10.1101/2020.04.23.20076885.

[57] Rockett RJ, et al. Revealing COVID-19 transmission in Australia by SARS-CoV-2 genome sequencing and agent-based modeling. Nat Med 2020;26:1398–404. https://doi.org/10.1038/s41591-020-1000-7.

[58] Maziarz M, Zach M. Agent-based modelling for SARS-CoV-2 epidemic prediction and intervention assessment: a methodological appraisal. J Eval Clin Pract 2020;26:1352–60. https://doi.org/10.1111/jep.13459.

[59] Estrada E. COVID-19 and SARS-CoV-2. Modeling the present, looking at the future. Phys Rep 2020;869:1–51. https://doi.org/10.1016/j.physrep.2020.07.005.

[60] Read AF, et al. Evaluation of COVID-19 vaccination strategies with a delayed second dose. PLoS Biol 2021;19. https://doi.org/10.1371/journal.pbio.3001211.

[61] Ogden NH, et al. Modelling scenarios of the epidemic of COVID-19 in Canada. Can Commun Dis Rep 2020;198-204. https://doi.org/10.14745/ccdr.v46i06a08.

[62] Warne DJ, et al. Hindsight is 2020 vision: a characterisation of the global response to the COVID-19 pandemic. BMC Public Health 2020;20. https://doi.org/10.1186/s12889-020-09972-z.

[63] Garg AK, Mittal S, Padmanabhan P, Desikan R, Dixit NM. Increased B cell selection stringency in germinal centers can explain improved COVID-19 vaccine efficacies with low dose prime or delayed boost. Front Immunol 2021;12. https://doi.org/10.3389/fimmu.2021.776933.

[64] Sego TJ, et al. A modular framework for multiscale, multicellular, spatiotemporal modeling of acute primary viral infection and immune response in epithelial tissues and its application to drug therapy timing and effectiveness. PLoS Comput Biol 2020. https://doi.org/10.1101/2020.04.27.064139.

[65] Ferrari Gianlupi J, et al. Multiscale model of antiviral timing, potency, and heterogeneity effects on an epithelial tissue patch infected by SARS-CoV-2. Viruses 2022;14. https://doi.org/10.3390/v14030605.

[66] Getz, M. et al. Rapid community-driven development of a SARS-CoV-2 tissue simulator. Biorxiv, 2020.2004.2002.019075-012020.019004.019002.019075 (2020). 10.1101/2020.04.02.019075.

[67] Trouillet-Assant S, et al. Type I IFN immunoprofiling in COVID-19 patients. J Allergy Clin Immunol 2020;4-8. https://doi.org/10.1016/j.jaci.2020.04.029.

[68] Ostaszewski M, et al. COVID-19 Disease Map, a computational knowledge repository of SARS-CoV-2 virus-host interaction mechanisms. Mol Syst Biol 2021;17:e10387.

[69] Hwang W, et al. Current and prospective computational approaches and challenges for developing COVID-19 vaccines. Adv Drug Deliv Rev 2021;172:249–74. https://doi.org/10.1016/j.addr.2021.02.004.

[70] Ahmed SF, Quadeer AA, McKay MR. Preliminary identification of potential vaccine targets for the COVID-19 coronavirus (SARS-CoV-2) based on SARS-CoV immunological studies. Viruses 2020;12. https://doi.org/10.3390/v12030254.

[71] Wu F, et al. A new coronavirus associated with human respiratory disease in China. Nature 2020;579:265–9. https://doi.org/10.1038/s41586-020-2008-3.

[72] Redondo N, Zaldívar-López S, Garrido JJ, Montoya M. SARS-CoV-2 accessory proteins in viral pathogenesis: knowns and unknowns. Front Immunol 2021;12. https://doi.org/10.3389/fimmu.2021.708264.

[73] Moya A, Holmes EC, González-Candelas F. The population genetics and evolutionary epidemiology of RNA viruses. Nat Rev Microbiol 2004;2:279–88. https://doi.org/10.1038/nrmicro863.

[74] Kockler ZW, Gordenin DA. From RNA world to SARS-CoV-2: the edited story of RNA viral evolution. Cells 2021;10. https://doi.org/10.3390/cells10061557.

[75] Willett BJ, et al. SARS-CoV-2 Omicron is an immune escape variant with an altered cell entry pathway. Nat Microbiol 2022;7:1161–79. https://doi.org/10.1038/s41564-022-01143-7.

[76] Wang R, Chen J, Hozumi Y, Yin C, Wei GW. Emerging vaccine-breakthrough SARS-CoV-2 variants. ACS Infect Dis 2022;8:546–56. https://doi.org/10.1021/acsinfecdis.1c00557.

[77] Li T, et al. Phylogenetic supertree reveals detailed evolution of SARS-CoV-2. Sci Rep 2020;10. https://doi.org/10.1038/s41598-020-79484-8.

[78] Zhou P, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 2020;579:270–3. https://doi.org/10.1038/s41586-020-2012-7.

[79] Sagulenko P, Puller V, Neher RA. TreeTime: maximum-likelihood phylodynamic analysis. Virus Evol 2018;4. https://doi.org/10.1093/ve/vex042.

[80] Duchene S, et al. Temporal signal and the phylodynamic threshold of SARS-CoV-2. Virus Evol 2020;6. https://doi.org/10.1093/ve/veaa061.

[81] Morel B, et al. Phylogenetic analysis of SARS-CoV-2 data is difficult. Mol Biol Evol 2021;38:1777–91. https://doi.org/10.1093/molbev/msaa314.

[82] Vasilarou M, Alachiotis N, Garefalaki J, Beloukas A, Pavlidis P. Population genomics insights into the first wave of COVID-19. Life 2021;11. https://doi.org/10.3390/life11020129.

[83] Beaumont MA, Zhang W, Balding DJ. Approximate bayesian computation in population genetics. Genetics 2002;162:2025–35. https://doi.org/10.1093/genetics/162.4.2025.

[84] De Maio N, et al. Mutation rates and selection on synonymous mutations in SARS-CoV-2. Genome Biol Evol 2021;13. https://doi.org/10.1093/gbe/evab087.

[85] Kim, K. et al. APOBEC-mediated editing of SARS-CoV-2 genomic RNA impacts viral replication and fitness. Biorxiv (2022). 10.1101/2021.12.18.473309.

[86] Díez-Fuertes F, et al. A founder effect led early SARS-CoV-2 transmission in Spain. J Virol 2021;95. https://doi.org/10.1128/jvi.01583-20.

[87] Zhang L, et al. SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. Nat Commun 2020;11. https://doi.org/10.1038/s41467-020-19808-4.

[88] Vöhringer HS, et al. Genomic reconstruction of the SARS-CoV-2 epidemic in England. Nature 2021;600:506–11. https://doi.org/10.1038/s41586-021-04069-y.

[89] O'Toole Á, et al. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. Virus Evol 2021. https://doi.org/10.1093/ve/veab064.

[90] OliverPybus. Pango Lineage Nomenclature: provisional rules for naming recombinant lineages, <https://virological.org/t/pango-lineage-nomenclature-provisional-rules-for-naming-recombinant-lineages/657>(2021).

[91] Mostefai F, et al. Population genomics approaches for genetic characterization of SARS-CoV-2 lineages. Front Med 2022;9. https://doi.org/10.3389/fmed.2022.826746 (Lausanne).

[92] Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 1989;123:585–95. https://doi.org/10.1093/genetics/123.3.585.

[93] Schiøler H, Knudsen T, Brøndum RF, Stoustrup J, Bøgsted M. Mathematical modelling of SARS-CoV-2 variant outbreaks reveals their probability of extinction. Sci Rep 2021;11. https://doi.org/10.1038/s41598-021-04108-8.

[94] Volz E, et al. Evaluating the effects of SARS-CoV-2 spike mutation D614G on transmissibility and pathogenicity. Cell 2021;184:64–75. https://doi.org/10.1016/j.cell.2020.11.020. e11.

[95] Zhan, X.Y. et al. Molecular evolution of SARS-CoV-2 structural genes: evidence of positive selection in spike glycoprotein. Biorxiv (2020). 10.1101/2020.06.25.170688.

[96] van Dorp L, et al. No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. Nat Commun 2020;11. https://doi.org/10.1038/s41467-020-19818-2.

[97] Hou YJ, et al. SARS-CoV-2 D614G variant exhibits efficient replication *ex vivo* and transmission *in vivo*. Science 2020;370:1464–8. https://doi.org/10.1126/science.abe8499.

[98] Plante JA, et al. Spike mutation D614G alters SARS-CoV-2 fitness. Nature 2020;592:116–21. https://doi.org/10.1038/s41586-020-2895-3.

[99] Mullen, J.L. et al. *outbreak.info*, <https://outbreak.info/>(2020).

[100] Wilkinson SAJ, et al. Recurrent SARS-CoV-2 mutations in immunodeficient patients. Virus Evolution 2022;8(2):veac050. https://doi.org/10.1093/ve/veac050.

[101] Oude Munnink BB, et al. Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans. Science 2021;371:172–7. https://doi.org/10.1126/science.abe5901.

[102] Di Giorgio S, Martignano F, Torcia MG, Mattiuz G, Conticello SG. Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. Sci Adv 2020;6. https://doi.org/10.1126/sciadv.abb5813.

[103] Desimmie BA, et al. Multiple APOBEC3 restriction factors for HIV-1 and one Vif to rule them all. J Mol Biol 2014;426:1220–45. https://doi.org/10.1016/j.jmb.2013.10.033.

[104] Ramazzotti D, et al. VERSO: a comprehensive framework for the inference of robust phylogenies and the quantification of intra-host genomic diversity of viral samples. Patterns 2021;2. https://doi.org/10.1016/j.patter.2021.100212.

[105] Graudenzi A, Maspero D, Angaroni F, Piazza R, Ramazzotti D. Mutational signatures and heterogeneous host response revealed via large-scale characterization of SARS-CoV-2 genomic diversity. iScience 2021;24. https://doi.org/10.1016/j.isci.2021.102116.

[106] Pathak AK, et al. Spatio-temporal dynamics of intra-host variability in SARS-CoV-2 genomes. Nucleic Acids Res 2022;50:1551–61. https://doi.org/10.1093/nar/gkab1297.

[107] Yi K, et al. Mutational spectrum of SARS-CoV-2 during the global pandemic. Exp Mol Med 2021;53:1229–37. https://doi.org/10.1038/s12276-021-00658-z.

[108] Simmonds P, Schwemmle M. Rampant C→U hypermutation in the genomes of SARS-CoV-2 and other coronaviruses: causes and consequences for their short- and long-term evolutionary trajectories. mSphere 2020;5. https://doi.org/10.1128/mSphere.00408-20.

[109] Popa A, et al. Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2. Sci Transl Med 2020;12. https://doi.org/10.1126/scitranslmed.abe2555.

[110] Martin MA, Koelle K. Comment on "Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2". Sci Transl Med 2021;13. https://doi.org/10.1126/scitranslmed.abh1803.

[111] The Severe Covid-19 GWAS Group. Genomewide association study of severe COVID-19 with respiratory failure. N Engl J Med 2020;383:1522–34. https://doi.org/10.1056/NEJMoa2020283.

[112] Niemi MEK, et al. Mapping the human genetic architecture of COVID-19. Nature 2021;600:472–7. https://doi.org/10.1038/s41586-021-03767-x.

[113] Shelton JF, et al. Trans-ancestry analysis reveals genetic and nongenetic associations with COVID-19 susceptibility and severity. Nat Genet 2021;53:801–8. https://doi.org/10.1038/s41588-021-00854-7.

[114] Zietz M, Zucker J, Tatonetti NP. Associations between blood type and COVID-19 infection, intubation, and death. Nat Commun 2020;11. https://doi.org/10.1038/s41467-020-19623-x.

[115] Kasela S, et al. Integrative approach identifies SLC6A20 and CXCR6 as putative causal genes for the COVID-19 GWAS signal in the 3p21.31 locus. Genome Biol 2021;22. https://doi.org/10.1186/s13059-021-02454-4.

[116] Dai Y, et al. Association of CXCR6 with COVID-19 severity: delineating the host genetic factors in transcriptomic regulation. Hum Genet 2021;140:1313–28. https://doi.org/10.1007/s00439-021-02305-z.

[117] Smieszek SP, et al. Elevated plasma levels of CXCL16 in severe COVID-19 patients. Cytokine 2022;152. https://doi.org/10.1016/j.cyto.2022.155810.

[118] Yao Y, et al. Genome and epigenome editing identify CCR9 and SLC6A20 as target genes at the 3p21.31 locus associated with severe COVID-19. Signal Transduct Target Ther 2021;6. https://doi.org/10.1038/s41392-021-00519-1.

[119] Zeberg H, Pääbo S. The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. Nature 2020;587:610–2. https://doi.org/10.1038/s41586-020-2818-3.

[120] Pairo-Castineira E, et al. Genetic mechanisms of critical illness in COVID-19. Nature 2020;591:92–8. https://doi.org/10.1038/s41586-020-03065-y.

[121] Zeberg H, Pääbo S. A genomic region associated with protection against severe COVID-19 is inherited from Neandertals. Proc Natl Acad Sci 2021;118. https://doi.org/10.1073/pnas.2026309118.

[122] Huffman JE, et al. Multi-ancestry fine mapping implicates OAS1 splicing in risk of severe COVID-19. Nat Genet 2022;54:125–7. https://doi.org/10.1038/s41588-021-00996-8.

[123] Ivashkiv LB, Donlin LT. Regulation of type I interferon responses. Nat Rev Immunol 2013;14:36–49. https://doi.org/10.1038/nri3581.

[124] Smieszek SP, Polymeropoulos VM, Xiao C, Polymeropoulos CM, Polymeropoulos MH. Loss-of-function mutations in IFNAR2 in COVID-19 severe infection susceptibility. J Glob Antimicrob Resist 2021;26:239–40. https://doi.org/10.1016/j.jgar.2021.06.005.

[125] Millett GA, et al. Assessing differential impacts of COVID-19 on black communities. Ann Epidemiol 2020;47:37–44. https://doi.org/10.1016/j.annepidem.2020.05.003.

[126] Rodriguez-Diaz CE, et al. Risk for COVID-19 infection and death among Latinos in the United States: examining heterogeneity in transmission dynamics. Ann Epidemiol 2020;52:46–53. https://doi.org/10.1016/j.annepidem.2020.07.007. e42.

[127] Horowitz JE, et al. Genome-wide analysis provides genetic evidence that ACE2 influences COVID-19 risk and yields risk scores associated with severe disease. Nat Genet 2022. https://doi.org/10.1038/s41588-021-01006-7.

[128] Andreakos E, et al. A global effort to dissect the human genetic basis of resistance to SARS-CoV-2 infection. Nat Immunol 2021;23:159–64. https://doi.org/10.1038/s41590-021-01030-z.

[129] Van Gassen S, et al. FlowSOM: using self-organizing maps for visualization and interpretation of cytometry data. Cytometry Part A 2015;87:636–45. https://doi.org/10.1002/cyto.a.22625.

[130] Levine JH, et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. Cell 2015;162:184–97. https://doi.org/10.1016/j.cell.2015.05.047.

[131] Toghi Eshghi S, et al. Quantitative comparison of conventional and t-SNE-guided gating analyses. Front Immunol 2019;10. https://doi.org/10.3389/fimmu.2019.01194.

[132] Becht E, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol 2018;37:38–44. https://doi.org/10.1038/nbt.4314.

[133] Moon KR, et al. Visualizing structure and transitions in high-dimensional biological data. Nat Biotechnol 2019;37:1482–92. https://doi.org/10.1038/s41587-019-0336-3.

[134] Kuchroo M, et al. Multiscale PHATE identifies multimodal signatures of COVID-19. Nat Biotechnol 2022. https://doi.org/10.1038/s41587-021-01186-x.

[135] Rébillard RM, et al. Identification of SARS-CoV-2-specific immune alterations in acutely ill patients. J Clin Invest 2021. https://doi.org/10.1172/JCI145853.

[136] Van Gassen S, et al. FlowSOM: using self-organizing maps for visualization and interpretation of cytometry data. Cytom A 2015;87:636–45. https://doi.org/10.1002/cyto.a.22625.

[137] Qian Y, et al. Elucidation of seventeen human peripheral blood B-cell subsets and quantification of the tetanus response using a density-based method for the automated identification of cell populations in multidimensional flow cytometry data. Cytom B: Clin Cytom 2010;78B:S69–82. https://doi.org/10.1002/cyto.b.20554.

[138] Aghaeepour N, Nikolic R, Hoos HH, Brinkman RR. Rapid cell population identification in flow cytometry data. Cytom Part A 2011;79A:6–13. https://doi.org/10.1002/cyto.a.21007.

[139] Zare H, Shooshtari P, Gupta A, Brinkman RR. Data reduction for spectral clustering to analyse high throughput flow cytometry data. BMC Bioinform 2010;11. https://doi.org/10.1186/1471-2105-11-403.

[140] van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9:2579–605.

[141] Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philos Trans R Soc A 2016;374. https://doi.org/10.1098/rsta.2015.0202.

[142] Brugnone N, et al. in *2019*. In: Proceedings of the IEEE international conference on big data (big data); 2019. p. 2624–33.
[143] Leeb W, Coifman R. Hölder–lipschitz norms and their duals on spaces with semigroups, with applications to earth mover's distance. J Fourier Anal Appl 2015;22:910–53. https://doi.org/10.1007/s00041-015-9439-5.
[144] Le T, Yamada M, Fukumizu K, Cuturi M. Tree-sliced variants of Wasserstein distances. Adv Neur In 2019:32. https://doi.org/10.48550/arXiv.1902.00342.
[145] Burkhardt DB, et al. Quantifying the effect of experimental perturbations at single-cell resolution. Nat Biotechnol 2021;39:619–29. https://doi.org/10.1038/s41587-020-00803-5.
[146] Krishnaswamy S, et al. Conditional density-based analysis of T cell signaling in single-cell data. Science 2014;346. https://doi.org/10.1126/science.1250689.
[147] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44. https://doi.org/10.1038/nature14539.
[148] Ong E, et al. Vaxign2: the second generation of the first web-based vaccine design program using reverse vaccinology and machine learning. Nucleic Acids Res 2021;49:W671–8. https://doi.org/10.1093/nar/gkab279.
[149] Fast, E., Altman, R.B. & Chen, B. Potential T-cell and B-cell epitopes of 2019-nCoV. Biorxiv, 1–9 (2020). 10.1101/2020.02.19.955484.
[150] Che M, Yao K, Che C, Cao Z, Kong F. Knowledge-graph-based drug repositioning against COVID-19 by graph convolutional network with attention mechanism. Future Internet 2021;13. https://doi.org/10.3390/fi13010013.
[151] Beck BR, Shin B, Choi Y, Park S, Kang K. Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model. Comput Struct Biotechnol J 2020;18:784–90. https://doi.org/10.1016/j.csbj.2020.03.025.
[152] Richardson P, et al. Baricitinib as potential treatment for 2019-nCoV acute respiratory disease. Lancet N Am Ed 2020;395:e30–1. https://doi.org/10.1016/s0140-6736(20)30304-4.
[153] Zhang H, et al. Deep learning based drug screening for novel coronavirus 2019-nCov. Interdiscip Sci: Comput Life Sci 2020;12:368–76. https://doi.org/10.1007/s12539-020-00376-6.
[154] Subudhi S, et al. Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19. npj Digit Med 2021;4. https://doi.org/10.1038/s41746-021-00456-x.
[155] Kar S, et al. Multivariable mortality risk prediction using machine learning for COVID-19 patients at admission (AICOVID). Sci Rep 2021;11. https://doi.org/10.1038/s41598-021-92146-7.
[156] Lalmuanawma S, Hussain J, Chhakchhuak L. Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: a review. Chaos Solit Fractals 2020;139. https://doi.org/10.1016/j.chaos.2020.110059.
[157] Pitzer VE, et al. Pandemic velocity: forecasting COVID-19 in the US with a machine learning & Bayesian time series compartmental model. PLoS Comput Biol 2021;17. https://doi.org/10.1371/journal.pcbi.1008837.
[158] Wang P, Zheng X, Li J, Zhu B. Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics. Chaos Solit Fractals 2020;139. https://doi.org/10.1016/j.chaos.2020.110058.
[159] Li L, et al. Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: evaluation of the diagnostic accuracy. Radiology 2020;296:E65–71. https://doi.org/10.1148/radiol.2020200905.
[160] Wang S, et al. A fully automatic deep learning system for COVID-19 diagnostic and prognostic analysis. Eur Respir J 2020;56. https://doi.org/10.1183/13993003.00775-2020.
[161] Qin L, et al. A predictive model and scoring system combining clinical and CT characteristics for the diagnosis of COVID-19. Eur Radiol 2020;30:6797–807. https://doi.org/10.1007/s00330-020-07022-1.
[162] Zargari Khuzani A, Heidari M, Shariati SA. COVID-Classifier: an automated machine learning model to assist in the diagnosis of COVID-19 infection in chest X-ray images. Sci Rep 2021;11. https://doi.org/10.1038/s41598-021-88807-2.
[163] Wang L, Lin ZQ, Wong A. COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. Sci Rep 2020;10. https://doi.org/10.1038/s41598-020-76550-z.
[164] Kukar M, et al. COVID-19 diagnosis by routine blood tests using machine learning. Sci Rep 2021;11. https://doi.org/10.1038/s41598-021-90265-9.
[165] Rosado J, et al. Multiplex assays for the identification of serological signatures of SARS-CoV-2 infection: an antibody-based diagnostic and machine learning study. Lancet Microbe 2021;2:e60–9. https://doi.org/10.1016/s2666-5247(20)30197-x.
[166] Farhang-Sardroodi S, Ghaemi MS, Craig M, Ooi HK, Heffernan JM. A machine learning approach to differentiate between COVID-19 and influenza infection using synthetic infection and immune response data. Math Biosci Eng 2022;19:5813–31. https://doi.org/10.3934/mbe.2022272.
[167] Rappuoli R. Reverse vaccinology. Curr Opin Microbiol 2000;3:445–50. https://doi.org/10.1016/s1369-5274(00)00119-3.
[168] He Y, et al. Vaxign-ML: supervised machine learning reverse vaccinology model for improved prediction of bacterial protective antigens. Bioinformatics 2020;36:3185–91. https://doi.org/10.1093/bioinformatics/btaa119.
[169] Pritam M, Singh G, Swaroop S, Singh AK, Singh SP. Exploitation of reverse vaccinology and immunoinformatics as promising platform for genome-wide screening of new effective vaccine candidates against Plasmodium falciparum. BMC Bioinform 2019;19. https://doi.org/10.1186/s12859-018-2482-x.
[170] Heinson A, et al. Enhancing the biological relevance of machine learning classifiers for reverse vaccinology. Int J Mol Sci 2017;18. https://doi.org/10.3390/ijms18020312.
[171] He Y, Xiang Z, Mobley HLT. Vaxign: the first web-based vaccine design program for reverse vaccinology and applications for vaccine development. J Biomed Biotechnol 2010;2010:1–15. https://doi.org/10.1155/2010/297505.
[172] Doytchinova IA, Flower DR. VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. BMC Bioinform 2007;8. https://doi.org/10.1186/1471-2105-8-4.
[173] Vivona S, et al. Computer-aided biotechnology: from immuno-informatics to reverse vaccinology. Trends Biotechnol 2008;26:190–200. https://doi.org/10.1016/j.tibtech.2007.12.006.
[174] Crooke SN, Ovsyannikova IG, Kennedy RB, Poland GA. Immunoinformatic identification of B cell and T cell epitopes in the SARS-CoV-2 proteome. Sci Rep 2020;10. https://doi.org/10.1038/s41598-020-70864-8.
[175] Ong E, Wong MU, Huffman A, He Y. COVID-19 coronavirus vaccine design using reverse vaccinology and machine learning. Front Immunol 2020;11. https://doi.org/10.3389/fimmu.2020.01581.
[176] Malone B, et al. Artificial intelligence predicts the immunogenic landscape of SARS-CoV-2 leading to universal blueprints for vaccine designs. Sci Rep 2020;10. https://doi.org/10.1038/s41598-020-78758-5.
[177] Yang Z, Bogdan P, Nazarian S. An in silico deep learning approach to multi-epitope vaccine design: a SARS-CoV-2 case study. Sci Rep 2021;11. https://doi.org/10.1038/s41598-021-81749-9.
[178] Nielsen M, Lund O. NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. BMC Bioinform 2009;10. https://doi.org/10.1186/1471-2105-10-296.
[179] Prachar M, et al. Identification and validation of 174 COVID-19 vaccine candidate epitopes reveals low performance of common epitope prediction tools. Sci Rep 2020;10. https://doi.org/10.1038/s41598-020-77466-4.
[180] Hamelin DJ, et al. The mutational landscape of SARS-CoV-2 variants diversifies T cell targets in an HLA-supertype-dependent manner. Cell Syst 2022;13:143–57. https://doi.org/10.1016/j.cels.2021.09.013. e143.
[181] Kingma, D.P., Welling, M. Auto-encoding variational Bayes. arXiv, 2014; 1–14. doi: 10.48550/arXiv.1312.6114.
[182] Bjerrum E, Sattarov B. Improving Chemical Autoencoder Latent Space and Molecular De Novo Generation Diversity with Heteroencoders. Biomolecules 2018;8. https://doi.org/10.3390/biom8040131.
[183] Grantham K, et al. Deep evolutionary learning for molecular design. IEEE Comput Intell Mag 2022;17:14–28. https://doi.org/10.1109/mci.2022.3155308.
[184] Chenthamarakshan V, et al. In: Proceedings of the 34th international conference on neural information processing systems. Curran Associates Inc.; 2020. Article 363.
[185] Tang B, et al. AI-aided design of novel targeted covalent inhibitors against SARS-CoV-2. Biomolecules 2022;12(6):746. https://doi.org/10.3390/biom12060746.
[186] Goyal S, Kim S, Chen ISY, Chou T. Mechanisms of blood homeostasis: lineage tracking and a neutral model of cell populations in rhesus macaques. BMC Bioinform 2015;13:85. https://doi.org/10.1186/s12915-015-0191-8. -85.
[187] Brunet-Ratnasingham E, et al. Integrated immunovirological profiling validates plasma SARS-CoV-2 RNA as an early predictor of COVID-19 mortality. Sci Adv 2021;7(48):eabj5629. https://doi.org/10.1126/sciadv.abj5629.
[188] Lavine JS, Bjornstad ON, Antia R. Immunological characteristics govern the transition of COVID-19 to endemicity. Science 2021;371:741–5. https://doi.org/10.1126/science.abe6522.
[189] Bolze A, et al. Evidence for SARS-CoV-2 Delta and Omicron co-infections and recombination. Med 2022;3(12):848–859.e4.
[190] Hobbs EC, Reid TJ. Animals and SARS-CoV-2: species susceptibility and viral transmission in experimental and natural conditions, and the potential implications for community transmission. Transbound Emerg Dis 2020;68:1850–67. https://doi.org/10.1111/tbed.13885.