UNIVERSITA' DEGLI STUDI DI PAVIA

FACOLTA' DI INGEGNERIA
DIPARTIMENTO DI INGEGNERIA INDUSTRIALE E DELL'INFORMAZIONE

DOTTORATO DI RICERCA IN BIOINGEGNERIA, BIOINFORMATICA E TECNOLOGIE PER LA SALUTE
XXXVIII - 2025

Artificial Intelligence Approaches for Genomic Surveillance, Epidemiological Modeling, and Evolutionary Forecasting of SARS-CoV-2

PhD Thesis by

Simone Rancati

Advisors:
Prof. Riccardo Bellazzi
Prof. Giovanna Nicora

PhD Program Chair:
Prof. Riccardo Bellazzi

I wish to dedicate this PhD thesis:
To Prof. *Laura Insalata and Patrizia Angeletti*, with heartfelt thanks.
To Prof. Gina Giovannini who always believed in my scientific ability.
To the *secondary school teachers* who reignited my love of knowledge.

Acknowledgments

I am deeply grateful to my supervisors, Professor Riccardo Bellazzi and Professor. Giovanna Nicora, whose inspiration, insight, and generous feedback have been invaluable. Thank you for always believing in me,even when I did not believe in myself.

I warmly thank everyone at the Bio-Medical Informatics "Mario Stefanelli" Laboratory (BMI) for the stimulating environment and daily support. It has been a privilege to learn from and collaborate with you over these years.

Special thanks to Professor Marco Salemi, Professor Mattia Prosperi, and Professor Simone Marini for hosting me at the University of Florida and for your steadfast support throughout my PhD.

Finally, I am profoundly grateful to my family for their unwavering encouragement, including financial support, throughout this journey; to my best friend for always being by my side,especially in the moments when I wanted to give up; and to my friends, including those in the United States, for their constant encouragement.

Abstract (Italiano)

Nelle emergenze pandemiche, la sorveglianza genomica è cruciale per l'allerta precoce e l'anticipazione di varianti ad alto potenziale di diffusione; tuttavia la sua efficacia è spesso limitata da tempi di risposta lenti e dall'individuazione tardiva delle varianti, quando queste sono già ampiamente diffuse nella popolazione a rischio. L'intelligenza artificiale (AI) può trasformare la sorveglianza genomica da flusso di reporting retrospettivo a capacità proattiva di supporto alle decisioni. Con l'aumento del sequenziamento e l'evoluzione della diversità virale, l'imprevedibilità deriva da tre fattori principali: eterogeneità e squilibri dei dati, ritardi e lacune nelle conoscenze di riferimento, e assenza di standard condivisi per valutazioni quantitative in contesti operativi reali. Questa tesi affronta tali criticità (i) delimitando gli ambiti in cui i modelli "sequence-aware" sono affidabili, (ii) sviluppando strumenti non supervisionati e generativi per anticipare traiettorie evolutive, e (iii) proponendo rappresentazioni pan-virali riusabili a supporto della sanità pubblica.

Il Capitolo 1 inquadra lo stato dell'arte della sorveglianza genomica basata su AI, evidenziando risultati, problemi aperti e contesto operativo (governance dei dati, privacy, standard di valutazione). Il Capitolo 2 descrive le interazioni tra l'evoluzione della pandemia COVID-19 e la gestione da parte delle autorità sanitarie, mediante un modello autoregressivo su indicatori settimanali europei (inclusi casi COVID-19, ricoveri, decessi, vaccinazioni, varianti, indici di efficienza alla risposta pandemica, come l'indice di stringenza,) e utilizza modelli avanzati di AI, inclusi i foundation models per la previsione a breve termine dell'incidenza. Il Capitolo 3 introduce DeepAutoCoV, un framework non supervisionato per identificare varianti del SARS-CoV-2 che diventeranno Future Dominant Lineages (FDL). Il Capitolo 4 presenta SARITA, un modello autoregressivo domain-adapted per la generazione di sequenze genomiche di SARS-CoV-2: la qualità delle sequenze sintetiche è valutata in base al realismo evolutivo, per esempio la capacità del modello di predire varianti realmente comparse, nonostante non siano state presenti nel dataset di allenamento. Il Capitolo

5 sviluppa MistrVirus, un embedding pan-virale basato sul modello LLM open source dell'azienda Mistral in grado di rappresentare le sequenze genomiche appartenenti a diverse specie di virus in uno spazio multidimensionale; tale rappresentazione può essere usata per allenare modelli di classificazione, per esempio per distinguere i genomi virali dai genomi non virali. Il Capitolo 6 riporta conclusioni complessive, limitazioni e una roadmap verso AI interpretabile e "sequencing-aware" per una sanità pubblica anticipatoria.

Gli studi dei Capitoli 2–5 sono stati condotti presso l'Università di Pavia (Laboratorio di Informatica Biomedica BMI "Mario Stefanelli") e presso l'Università della Florida, sfruttando infrastrutture HPC (Hipergator) offerte dall'Università della Florida.

Abstract (English)

In pandemic emergencies, genomic surveillance is crucial for early warning and anticipation of variants with high transmission potential; however, its effectiveness is often limited by slow response times and the late detection of variants, when they are already widespread in at-risk populations. Artificial intelligence (AI) can shift genomic surveillance from a retrospective reporting pipeline to a proactive, decision-support capability. As sequencing scales up and viral diversity evolves, unpredictability stems from three main factors: data heterogeneity and imbalance, delays and gaps in reference knowledge, and the lack of shared standards for quantitative evaluation in real-world operational settings. This thesis addresses these challenges by (i) delineating the domains in which "sequence-aware" models are reliable, (ii) developing unsupervised and generative tools to anticipate evolutionary trajectories, and (iii) proposing reusable pan-viral representations to support public-health analytics. Chapter 1 surveys the state of the art in AI-enabled genomic surveillance, highlighting achievements, open problems, and the operational context (data governance, privacy, and evaluation standards). Chapter 2 examines the interplay between the evolution of the COVID-19 pandemic and public-health management through an autoregressive model on European weekly indicators (including COVID-19 cases, hospitalizations, deaths, vaccinations, variants, and indices of pandemic response efficiency, such as the stringency index), and leverages advanced AI,including foundation models,for short-term incidence forecasting. Chapter 3 introduces DeepAutoCoV, an unsupervised framework for identifying SARS-CoV-2 variants that will become Future Dominant Lineages (FDLs). Chapter 4 presents SARITA, a domain-adapted autoregressive model for generating SARS-CoV-2 genomic sequences; the quality of synthetic sequences is assessed via evolutionary realism,for example, the model's ability to predict variants that later emerged despite being absent from the training set. Chapter 5 develops MistrVirus, a pan-viral embedding based on the Mistral company's open source LLM model capable of representing genomic sequences belonging to different virus species in a multidimensional space; this representation can be used to train classifiers, for instance to distinguish viral from non-viral genomes.

Chapter 6 provides overall conclusions, limitations, and a roadmap toward interpretable, "sequence-aware" AI for anticipatory public health. The studies in Chapters 2–5 were conducted at the University of Pavia (Biomedical Informatics Laboratory "Mario Stefanelli") and the University of Florida, with computation on the HiPerGator HPC infrastructure provided by the University of Florida.

Contents

1 Background and Scientific Context 1

1.1. Major pandemics: from 1918 to the modern era 1

1918-1919 Spanish Influenza 3

1957-1958 Asian Influenza (H2N2) 4

1968-1970 Honk Kong Influenza 5

2009-2010 H1N1 "Swine Flu" Pandemic 6

2019-2023 COVID-19 (SARS-CoV-2) Pandemic 8

From SARS-CoV-1 to SARS-CoV-2: species jump and global spread 10

The Spike protein: molecular structure, functional domains, immunological target 15

Functional Domains of the Spike Protein 16

The Spike Protein as an Immunological Target 17

Mechanisms of Mutation, Selection, and Emergence of SARS-CoV-2 Variants of Concern 20

Genetic Mechanisms Driving SARS-CoV-2 Variation 21

Spike Protein Mutations and Viral Fitness 22

Key Variants of Concern (VOCs) 25

Public response and Severity metrics: from the Stringency Index to other measures 28

Severity Metrics and Public Response: From the Stringency Index to Other Measures 30

1.2. The Role of Artificial Intelligence in Surveillance and Health Policy Decision-Making 35

AI Applications in SARS-CoV-2 Genomic Surveillance 36

Early Detection of Emerging Variants with Machine Learning and Deep Learning 37

AI for Variant Tracking and Phylogenetic Analysis 44

AI for Transmissibility and Impact Analysis 46

AI for Real-Time Policy Support and Decision-Making 49

Epidemic Forecasting and Early Warning 49

Decision Support for Non-Pharmaceutical Interventions 50

Augmenting Contact Tracing and Outbreak Control 52

Resource Allocation and Healthcare Operations 52

1.3. Main Aim of Thesis 54

2 Epidemiological Modeling with AI: mlVAR and TimeGPT 56

2.1. Data collection and preprocessing 58

2.2. Dataset Structure 61

Stringency Index Dataset 61

Sequencing Rate Dataset 63

2.3. Analysis of the role of containment policies, vaccination strategies and virus variants in the COVID-19 pandemic 64

Multilevel Autoregression Model 64

Correlation analysis 64

Population Model Design 65

Results 69

Lag-length selection 69

Fixed effects 69

Temporal progression of the coefficients 74

2.4. Sequencing Intensity and Epidemiological Trends in Europe's COVID-19 Pandemic ... *76*
Correlation Analysis ... 76
Time series forecasting with TimeGPT ... 77
Results ... 78
Three Year Correlation ... 78
Years Wise Correlation ... 79
TimeGPT Analysis ... 81
2.5. General Conclusions: Linking Genomic Surveillance and Policy via TimeGPT and mlVAR (Europe, 2020–2022) ... *82*
2.6. Implications for public health and European policy making ... *86*

3 Anomaly Detection for Dominant Lineage Forecasting: DeepAutoCoV 90
3.1. Problem definition: forecasting Future Dominant Lineages (FDLs) ... *92*
3.2. Genomic data preprocessing and GISAID filtering ... *94*
3.3. DeepAutoCoV architecture and training configuration ... *96*
3.4. Weekly genomic surveillance simulation workflow ... *98*
3.5. Evaluation metrics: lead time, PPV, prioritization ... *100*
3.6. Comparison with alternative detection methods ... *102*
3.7. Results ... *103*
Anticipating the Future: Lead Time in FDL Detection ... 103
Precision Under Pressure: Positive Predictive Value (PPV) ... 104
Ranking What Matters: Prioritization of True Positives ... 106
Mutation Analysis ... 107
Comparison with competing methods ... 108
3.8. Discussion and Future Directions ... *109*
Strengths and contributions ... 110
Limitations ... 110
Future Directions ... 112

4 SARITA: Predictive Generation of SARS-CoV-2 Spike Using LLMs ... 113
4.1. Motivation: toward proactive viral surveillance through sequence generation ... *114*
4.2. SARITA architecture: decoder-only transformer and continual learning ... *116*
4.3. Training and Test Set Design: Filtering, Balancing, and Selection ... *119*
4.4. Sequence Generation and Evaluation ... *121*
4.5. Model Comparison ... *124*
Non-parametric random generators ... 124
Decoder-only protein language models ... 125
4.6. Results ... *126*
Sequence quality assessment: validity, length, biological realism ... 126
Evolutionary similarity to real-world data: Levenshtein, PAM30, FMR ... 129
Reproduction of VOC/VOI Signature Mutations and Structure-Based ΔΔG Analysis of ACE2 and Monoclonal Antibody Binding ... 132
Mutation density and positional analysis: SARITA vs empirical data ... 134
4.7. Discussion and Conclusion ... *137*

5 Pan-Viral Genome Embedding with Mistral ... 140
5.1. Motivation and Scope ... *141*
5.2. Dataset Creation and Curation ... *142*
5.3. Methods: Tokenization and Model Architecture ... *147*
5.4. Experimental Design and Evaluation Metrics ... *150*
5.5. Results ... *152*
5.6. Conclusion and Limitations ... *159*

6 Conclusion and Future Directions 163
6.1. Summary of key contributions 163
6.2. Impact on Public Health and Biomedical Engineering 165
6.3. Methodological limitations and ethical considerations 167
6.4. Future Directions: multi-omics, federated learning, explainability 169

7 Appendix 172

References 199

Chapter 1

1 Background and Scientific Context

During pandemics, the paramount objective is to reduce infections; conventional public-health measures,testing–tracing–isolation, masking, distancing, vaccination campaigns, and targeted closures,are effective but often blunt instruments with delayed feedback and high social cost when applied broadly. The rapid maturation of AI prompts the question of whether data-driven prediction, genomic analysis, and decision support can enable earlier, more precise, and less disruptive interventions toward the same goal. This chapter frames that question and delineates the evidence base used throughout the thesis: it first reviews the strengths and limitations of traditional control strategies and the epidemiological metrics used to evaluate them (e.g., effective reproduction number, hospitalization and Intensitive Care Unit (ICU) occupancy, infection-fatality ratios, vaccine uptake). It then introduces AI-enabled capabilities,including nowcasting/forecasting, anomaly detection in genomic surveillance, representation learning for variant assessment, and generative scenario modeling,using SARS-CoV-2 as the primary case study. The integration of these elements motivates a program that links sequence-level signals to population-level outcomes. The chapter concludes by stating the objectives that guide the remainder of the dissertation: quantifying variant impacts, developing early-warning pipelines from sequence embeddings, and exploring generative approaches for anticipatory sequence design and pan-viral inference.

1.1. Major pandemics: from 1918 to the modern era.

Throughout human history, pandemics have profoundly influenced societies, leading to substantial demographic changes, economic disruption, social upheaval, and major advancements in medical care and public health systems [1]. A pandemic is defined as the worldwide spread of a new infectious disease, typically caused by a pathogen to which the population has little or no pre-existing immunity, resulting in sustained human-to-human transmission across multiple countries or continents [1]. Each pandemic, shaped by the biological characteristics of its causative agent and the sociopolitical context in which it unfolds, represents a critical stress test for global health systems and societal resilience. Over the past century alone, several pandemics (Figure 1) have severely tested global health and social systems, ranging from the 1918 Spanish influenza, which struck at a time when medical science and public health infrastructure were still underdeveloped, to the recent SARS-CoV-2 pandemic, which challenged even the most advanced healthcare systems worldwide.

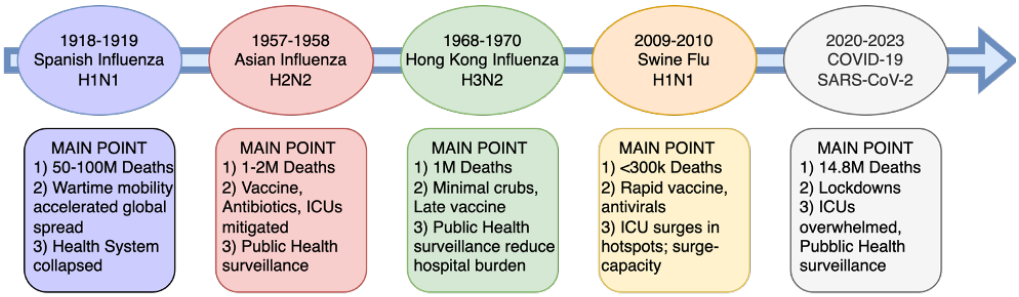


Figure 1: Timeline of the five major 20th- and 21st-century respiratory viral pandemics. From left to right, the schematic summarizes: (i) the 1918–1919 "Spanish" influenza (H1N1), responsible for an estimated 50–100 million deaths and characterized by rapid wartime spread and health-system collapse; (ii) the 1957–1958 "Asian" influenza (H2N2), whose impact (~1–2 million deaths) was mitigated by the first pandemic influenza vaccine, antibiotics and the emergence of intensive-care units; (iii) the 1968–1970 "Hong Kong" influenza (H3N2), causing ~1 million deaths but fewer societal disruptions thanks to improved surveillance and a timelier vaccine; (iv) the 2009–2010 swine-origin influenza (H1N1), with <300 000 deaths and notable for the rapid deployment of vaccines and antivirals alongside surge-capacity ICU strategies; and (v) the 2020–2023 COVID-19 pandemic (SARS-CoV-2), which has led to an estimated 14.8 million excess deaths worldwide, widespread lockdowns, overwhelmed critical-care services and unprecedented reliance on real-time public-health surveillance. Notably, COVID-19 accelerated the clinical maturation and global deployment of next-generation vaccine

platforms, especially mRNA and adenoviral-vector vaccines, enabling rapid redesign against emerging variants and marking a major milestone in modern vaccinology [2]

1918-1919 Spanish Influenza

The 1918 influenza pandemic, commonly known as the “Spanish flu”, remains one of the deadliest infectious crises ever recorded. The causative H1N1 virus infected an estimated 500 million people and was responsible for roughly 50 million deaths worldwide, with some estimates reaching as high as 100 million [1]. Unusually, the mortality curve peaked among otherwise healthy adults aged 20–40 years, and the resulting excess deaths caused a marked decline in life expectancy across multiple nations by the end of 1918. Rapid, successive waves of infection enveloped the globe within a single year (Figure 2). Governments attempted to curb transmission through school closures, bans on public gatherings, and quarantines, yet implementation was heterogeneous. World War I,characterized by troop mobilizations, overcrowded barracks, and wartime censorship,both accelerated viral spread and hampered coordinated public-health responses [1] Widespread absenteeism, labor shortages, and elevated mortality among working-age adults destabilized economies and social networks, leaving millions of widows, orphans, and psychologically traumatized survivors [3]. Health-care systems of the period were quickly overwhelmed. Hospitals and clinics overflowed, and in many urban centers every available bed was occupied. The crisis was exacerbated by a critical shortage of medical personnel: large contingents of physicians and nurses were deployed to military service, while many who remained contracted the disease themselves [3]. As capacity limits were reached, emergency wards were improvised in schools, churches, and other public buildings, with nurses,often volunteers or students,providing essential frontline care [3]. Overall, the impact on health-care institutions and workers was catastrophic, characterized by exhausted staff, acute shortages of critical supplies, and the establishment of makeshift morgues as cities exhausted their capacity to handle the deceased.

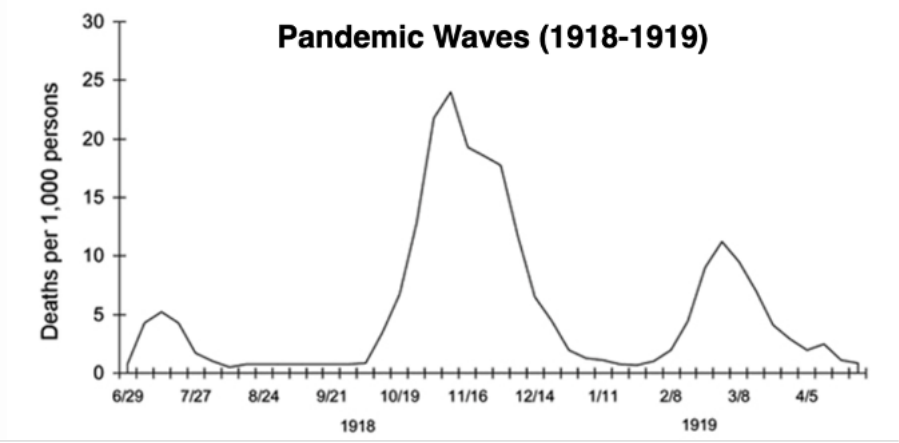


Figure 2: Weekly death rates (per 1,000 population), derived from United Kingdom influenza-and-pneumonia vital-registration data, illustrate the classic triphasic course of the "Spanish" influenza: a relatively mild 'herald' wave in early July 1918, a far more lethal second wave peaking in late October 1918, and a third, intermediate wave during February–March 1919 [4]. Estimates may be conservative, as wartime conditions could have affected diagnosis and death certification.

1957-1958 Asian Influenza (H2N2)

The 1957–1958 "Asian flu" pandemic originated in southern China with the emergence of a novel H2N2 influenza A virus and attained global distribution within months [5]. Although the virus was highly transmissible, its overall virulence proved moderate: contemporary estimates attribute roughly 1–2 million deaths worldwide, corresponding to a case-fatality ratio of about 0.1–0.2 %, dramatically lower than that of the 1918 H1N1 pandemic [5, 3]. Partial pre-existing immunity in older cohorts,likely derived from prior exposure to antigenically related strains,helped blunt disease severity, and while a second, smaller wave followed in 1958, the cumulative mortality burden ultimately resembled that of a severe seasonal epidemic rather than a catastrophic event. Public-health interventions were comparatively limited and short-lived: some jurisdictions instituted brief school closures and isolation of cases, but widespread lockdowns were neither deemed necessary nor enacted. Post-war improvements in hygiene and the rapid development of a monovalent vaccine, released late in 1957 [1], further contained societal disruption, and the economic impact was largely confined to transient spikes in absenteeism and minor productivity losses. Advances in biomedicine and clinical technology substantially mitigated pressure on healthcare services. Hospitals experienced short-term surges of

admissions, yet the availability of antibiotics allowed effective treatment of secondary bacterial pneumonias,one of the principal killers in 1918,thereby reducing overall lethality [1]. Newly established intensive-care units equipped with mechanical ventilators and oxygen therapy supported patients with acute respiratory failure, improving survival prospects, while the World Health Organization's influenza-surveillance network, inaugurated in 1952, enabled rapid identification of the H2N2 strain and timely selection of vaccine candidates for global distribution (Figure 3) [1]. Consequently, although the Asian flu imposed a significant burden of morbidity and mortality, most health systems coped without collapse, underscoring how antibiotics, critical-care infrastructure, and real-time surveillance can attenuate pandemic impact even in the absence of broad societal restrictions [1].

Figure 3: Monovalent inactivated influenza A vaccine ("Asian" strain, 1957). Five-millilitre multi-dose vial (200 CCA units mL^{-1}) and original carton produced by The National Drug Company, Philadelphia (United States of America (USA)). The antigen,virus harvested from chick-embryo allantoic fluid and inactivated with formalin,was deployed in the first large-scale pandemic vaccination campaign during the 1957–1958 H2N2 "Asian influenza" outbreak, marking a milestone in post-war influenza-control efforts [6].

1968-1970 Honk Kong Influenza

The 1968–1970 "Hong Kong flu" arose from a reassortant H3N2 influenza A virus that supplanted the prevailing H2N2 virus and spread worldwide within months . Although highly transmissible, its virulence

was moderate: global excess-mortality studies place the death toll at roughly one million (commonly cited range 0.5–2 million), a burden comparable to a severe seasonal influenza rather than a catastrophic pandemic [1]. Clinical severity was concentrated in older or otherwise vulnerable adults with a median age of fatal cases was in the 60s, while many younger individuals benefited from partial immunity conferred by the conserved N2 neuraminidase inherited from the 1957 strain [1]. Public recognition of a distinct pandemic was muted; no widespread lockdowns or extraordinary restrictions were imposed, schools and businesses largely remained open, and major public events. Governments promoted vaccination once a monovalent H3N2 vaccine became available, though distribution occurred after the first wave and uptake was limited. Economic repercussions were correspondingly minor, confined mainly to transient spikes in absenteeism and modest productivity losses [1]. Biomedical advances further moderated the impact on healthcare systems. Surges in hospital admissions were observed, yet case loads generally remained within existing capacity, thanks in part to widespread antibiotic availability for secondary bacterial pneumonias, the growing use of mechanical ventilators and oxygen therapy, and supportive-care protocols refined during the 1957 pandemic [1]. The World Health Organization's influenza-surveillance network, established in 1952, enabled rapid viral characterisation and timely vaccine-strain selection, facilitating a coordinated international response [1]. No large-scale quarantine or travel restrictions were deemed necessary, and healthcare-worker absenteeism, while elevated, did not exceed typical seasonal-flu levels. Overall, the Hong Kong flu imposed only modest and short-lived strain on global health infrastructures, underscoring the value of continual influenza surveillance, antibiotic stewardship, and critical-care capacity in mitigating pandemic effects even in the absence of stringent societal measures.

2009-2010 H1N1 "Swine Flu" Pandemic

In April 2009 a novel influenza A(H1N1) virus, colloquially termed "swine flu" because of its genomic links to swine influenza, was detected almost simultaneously in Mexico and the United States and achieved global dissemination within weeks [1]. Unlike typical seasonal influenza, the new strain disproportionately affected children and young adults; the mean age of fatal cases was approximately 37 years, likely because older cohorts retained partial immunity from exposure to earlier H1N1 lineages [1]. Although the World Health Organization (WHO) ultimately

confirmed about 18 600 laboratory-verified deaths, subsequent excess-mortality analyses place the true toll between 150 000 and 250 000,and up to 284 000 when respiratory and cardiovascular sequelae are included,yielding an overall case-fatality ratio of only 0.01–0.1%, markedly lower than the 1918, 1957, or 1968 pandemics [1]. Public alarm was initially high after WHO declared a Phase 6 pandemic, yet fear subsided once it became evident that most infections were clinically mild [1]. Governments promoted hand hygiene, respiratory etiquette, and voluntary home isolation; some Asian countries briefly recommended face masks, and a handful of cities enacted short-term school closures or event cancellations. Nevertheless, no wide-scale lockdowns or long-lasting travel restrictions were imposed, and global commerce continued with only modest sector-specific losses,the most notable involving tourism, air travel, and pork production. A monovalent H1N1pdm vaccine, developed and licensed within months, was deployed during the first year of the pandemic, ultimately covering 20–40 % of populations in North America and comparable proportions elsewhere, while antiviral stockpiles of oseltamivir were used for treatment and prophylaxis in high-risk groups [1]. For healthcare systems the 2009 pandemic represented a rigorous, albeit manageable, test of preparedness. Most countries experienced only moderate strain, yet regional hotspots reported sharp surges of severe cases, particularly among pregnant women and young adults with influenza-associated acute respiratory distress syndrome. During the austral winter, for example, Australia and New Zealand recorded peaks of roughly 7.4 H1N1 ICU admissions per million inhabitants, stretching critical-care capacity; a substantial proportion required prolonged mechanical ventilation, and extracorporeal membrane oxygenation (ECMO) was deployed in selected centres [7]. Similar but geographically confined ICU pressures were documented in Canada, the United Kingdom and the United States. Surge protocols were activated,elective surgeries deferred, high-dependency beds repurposed, and antiviral therapies prioritised for vulnerable patients,yet, unlike COVID-19 a decade later, most hospitals avoided systemic collapse. By late 2010 the H1N1pdm09 virus had stabilised as the new seasonal H1N1 strain [7]. Retrospective analyses underscore that even a clinically “moderate” pandemic can challenge intensive-care resources and highlight the importance of flexible ICU surge capacity, rapid vaccine rollout, and robust global surveillance,lessons that proved prescient for the SARS-CoV-2 response [8].

2019-2023 COVID-19 (SARS-CoV-2) Pandemic

The COVID-19 pandemic, driven by the novel coronavirus SARS-CoV-2, emerged in late 2019 and within months triggered a global health crisis without precedent in the modern era. By December 2021 more than 287 million laboratory-confirmed cases and approximately 5.4 million officially reported deaths had been recorded [9], yet subsequent WHO excess-mortality modelling indicates roughly 14.8 million pandemic-related deaths in 2020–2021,nearly three times the reported toll,and, by mid-2025, official deaths have surpassed 6.9 million, making COVID-19 one of the deadliest events of the past century [9](Figure 4).

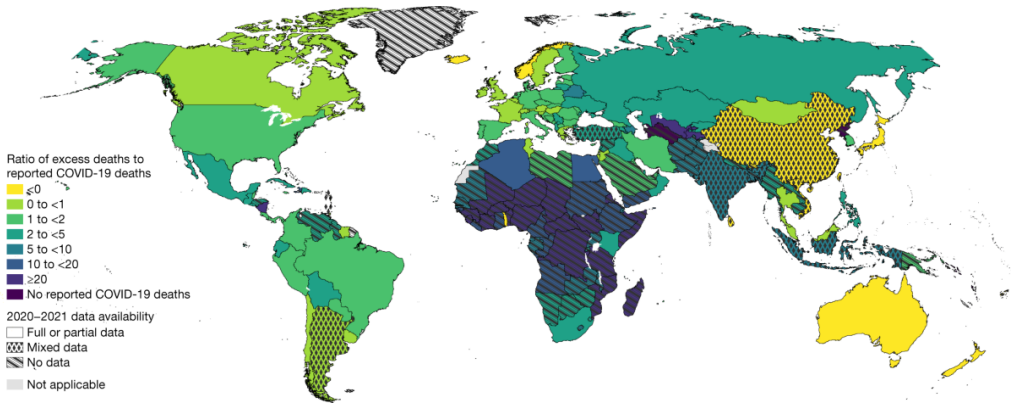


Figure 4: Country-level under-ascertainment of COVID-19 mortality, 2020-2021. The choropleth maps the ratio of model-estimated excess deaths to officially reported COVID-19 deaths for 194 countries and territories during 1 January 2020 – 31 December 2021. The figure highlights wide geographic heterogeneity in death registration and surveillance capacity, with the largest gaps concentrated in sub-Saharan Africa, South and Southeast Asia, and parts of Latin America [9].

Although the infection-fatality rate is markedly lower than that of the 1918 influenza, the enormous size of the susceptible global population resulted in an unparalleled absolute mortality burden, with the elderly and individuals with comorbidities most affected, but substantial losses also observed among younger adults before widespread vaccination. The societal consequences were equally profound. In early 2020 governments worldwide implemented unprecedented non-pharmaceutical interventions (Figure 5),nationwide lockdowns, school and business closures, bans on gatherings, and severe restrictions on international travel,measures that disrupted daily life on a scale not seen since World War II [10].

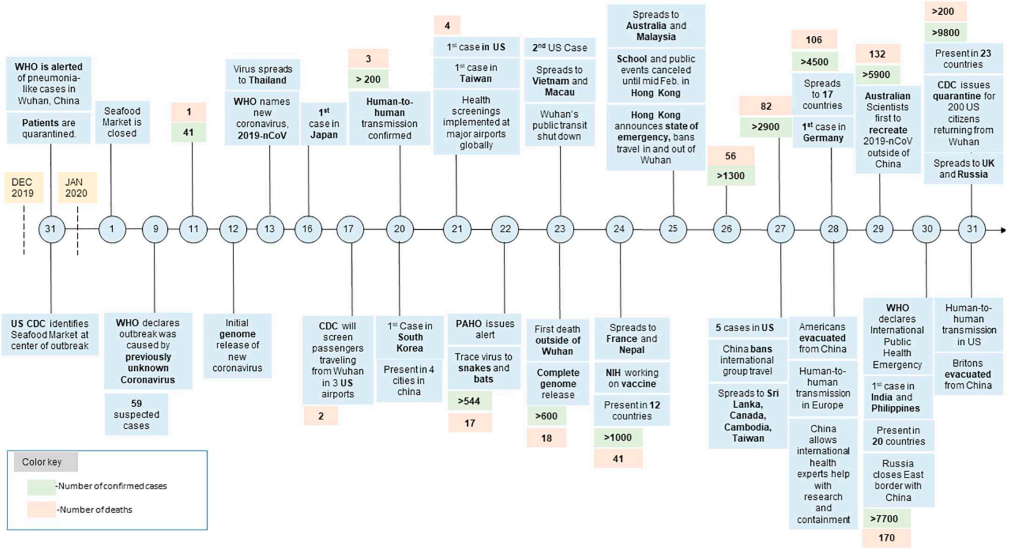


Figure 5: Early chronology of the COVID-19 outbreak (31 December 2019 – 31 January 2020). The horizontal timeline charts the first month of SARS-CoV-2: from the World Health Organization's 31 December 2019 alert on pneumonia in Wuhan, through market closure, pathogen identification and genome release, to the 30 January 2020 declaration of a Public Health Emergency of International Concern. Above the axis, pale-blue boxes mark key events,first international exportations (Thailand, Japan, United States, France, Germany, Australia), confirmation of human-to-human transmission, transport shutdowns and airport screening,while coloured insets (green for confirmed cases, salmon for deaths) show cumulative counts to each date. The figure highlights the virus's spread to more than 20 countries within one month and the rapid escalation of international response measures [11].

Mask-wearing, social distancing, and remote work and learning became commonplace, though adherence varied and pandemic fatigue grew over time. Economically, the "Great Lockdown" precipitated a 3 % contraction in global GDP in 2020, the steepest decline since the Great Depression, as supply chains fractured, unemployment surged, and entire sectors such as tourism and hospitality stalled, prompting governments to deploy massive fiscal-stimulus packages to avert systemic collapse [10]. Social isolation, school closures, and pervasive uncertainty generated widespread psychological stress, accelerating mental-health crises that are still being quantified. The arrival and global rollout of effective vaccines during 2021–2022 mitigated disease severity and enabled gradual reopening, yet recurrent waves driven by variants such as Delta and Omicron forced repeated recalibration of public-health measures. The

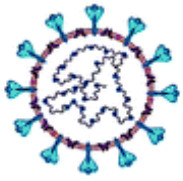
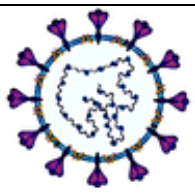
impact on healthcare systems was unprecedented. Successive surges of patients with severe viral pneumonia and respiratory failure overwhelmed hospital wards and intensive-care units in multiple epicentres,Wuhan, Lombardy, New York City,necessitating the conversion of operating theatres, recovery rooms, and even car parks into ad-hoc COVID wards [12]. Early shortages of personal protective equipment exposed frontline staff to high infection risk; WHO estimates that approximately 115 000 health and care workers died from COVID-19 in the first 18 months of the pandemic, while burnout and psychological distress were pervasive [12]. Elective procedures were cancelled, staff were redeployed to critical care, and severe resource scarcities forced crisis-standards-of-care triage in some regions. The rapid expansion of telemedicine accelerated adoption of evidence-based therapies (e.g., corticosteroids, prone ventilation, anticoagulation), and global genomic surveillance of emerging variants illustrated the capacity of biomedical science to adapt under duress, yet also exposed structural weaknesses,insufficient surge capacity, fragile supply chains, and inequitable vaccine distribution. By 2022, vaccination, natural immunity, and improved treatments had stabilised hospital pressures in many countries, but the pandemic's legacy of deferred care, workforce fatigue, and renewed emphasis on resilient health-system design continues to reshape global preparedness planning [12]. The COVID-19 has redefined the scale and speed with which an infectious threat can disrupt interconnected societies, economies, and healthcare infrastructures. The lessons extracted,from the critical value of early surveillance and rapid vaccine development to the necessity for flexible intensive-care capacity, robust supply chains, and mental-health support,now underpin strategic frameworks aimed at safeguarding population health against future pandemics.

From SARS-CoV-1 to SARS-CoV-2: species jump and global spread

The COVID-19 pandemic was caused by SARS-CoV-2. This section examines the biological features of this virus in comparison with related coronaviruses. Over the past quarter-century, two closely related zoonotic coronaviruses,SARS-CoV-1 (2002–2003) and SARS-CoV-2 (2019),emerged roughly 17 years apart (Table 1). Both belong to the subgenus Sarbecovirus (lineage B betacoronaviruses) and trace their ecological origins to horseshoe bats (Rhinolophus). SARS-CoV-1 spilled over to humans via civet cats in live-animal markets, whereas for SARS-CoV-2 an intermediate host has not been definitively established. Subtle

differences in viral adaptation and in the socio-ecological contexts of 2003 versus 2019 reshaped the consequences of the spillover [13].

Table 1: Comparative taxonomic and epidemiological features of the two human Sarbecoviruses. The tabulated summary contrasts the original severe acute respiratory syndrome coronavirus (SARS-CoV-1; first detected November 2002 in Guangdong, China) with the pandemic strain SARS-CoV-2 (identified December 2019 in Wuhan, China) [14].

DISEASE	COVID-19	SARS
Disease Causing Virus	 SARS-CoV-2	 SARS-CoV
Order	Nidovirus	Nidovirus
Family	Coronaviridae	Coronaviridae
Sub-Family	Orthocoronaviridae	Orthocoronaviridae
Genus	Betacoronavirus	Betacoronavirus
Sub-Genus	Serbecovirus	Serbecovirus
Date of outbreak	December 2019	November 2002
Origin	Wuhan, China	Guangdong, China

Basic Reproductive Number	2 – 2.5	3
Case Fatality Rate	3.4%	9.6 – 11%
Incubation Time	4 – 14 Days	2 – 7 Days
Region of Spread	Worldwide	29 Countries
Confirmed Cases	>100 million	2519
Deaths	>2 million	774

For SARS-CoV, genomic and serological investigations converge on masked palm civets traded alive in Guangdong's wet markets as the amplifying hosts: virus isolates from civets were >99 % identical to early human strains, and wildlife traders showed markedly elevated antibody titres, indicating repeated zoonotic exposure. The species jump required few molecular adjustments,two spike-protein (i.e; glycoprotein on the surface of certain enveloped viruses that binds host-cell receptors and triggers membrane fusion to let the virus enter the cell) substitutions and a 29-nt deletion in ORF8 ,so once the virus escaped market cages it was already competent for binding human ACE2 (Angiotensin-Converting Enzyme 2) and for limited person-to-person transmission [15, 16, 17] (Figure 6). In fact mutations in the Spike protein may enhance its binding affinity to the ACE2 receptor, resulting in a higher likelihood of successful host cell infection. SARS-CoV-2 likewise descends from bat sarbecoviruses, the best-sampled relative (RaTG13) sharing ≈96 % genome identity; however, every known bat virus sits several decades of evolution away from the pandemic strain. The missing branches imply cryptic circulation and adaptation in an unsampled intermediate host. Raccoon dogs, civets, mink and other small carnivores sold live in Wuhan in late 2019 fit the epidemiological trace-backs, and early human genomes reveal at least two independent animal-to-human introductions at the Huanan Seafood Market. Crucially, SARS-CoV-2 crossed into humans already armed with a high-affinity receptor-binding domain and a unique polybasic furin-cleavage site at the S1/S2 junction. This motif is recognized and cut by the host protease furin, splitting Spike into two subunits: S1, which contains the receptor-binding domain that engages ACE2, and S2, which drives membrane fusion. By "priming" the fusion

machinery through S1/S2 pre-cleavage, the furin site can make viral entry into airway epithelial cells more efficient than in related coronaviruses lacking this feature (including SARS-CoV-1), and it is widely regarded as a key contributor to the rapid spread of COVID-19 [15, 16, 17].

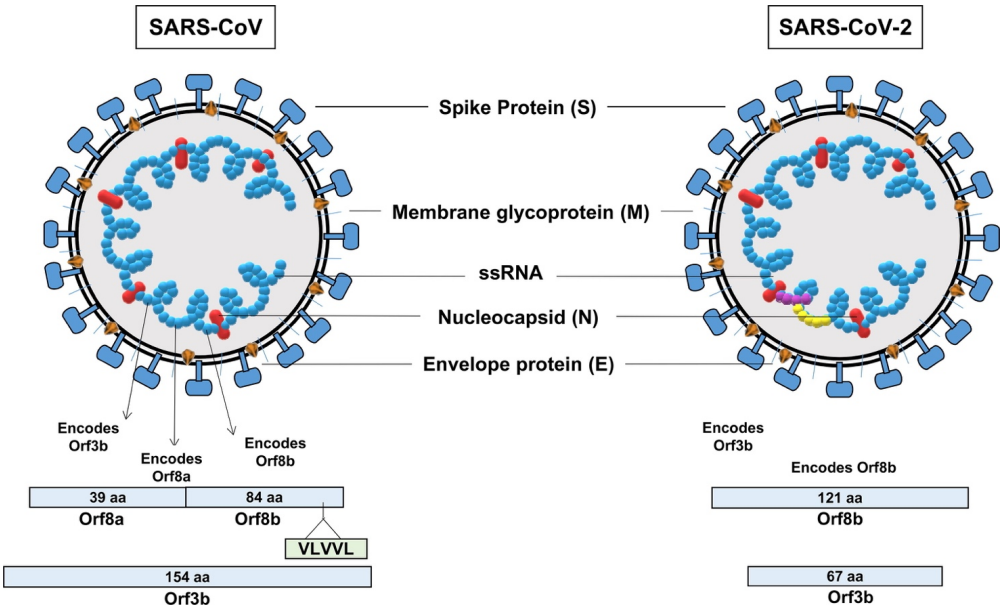


Figure 6: Structural layout and accessory-gene differences between SARS-CoV (2002/03) and SARS-CoV-2 (2019). Both Sarbecoviruses share the canonical coronavirus envelope,trimeric spike glycoproteins (S, blue), membrane proteins (M, gold), and small envelope proteins (E, brown),surrounding a positive-sense single-stranded RNA genome that is packaged by nucleocapsid protein (N, teal). The principal genomic divergence lies in the 3′ accessory region: SARS-CoV expresses a bicistronic ORF8a (39 aa) / ORF8b (84 aa) pair and a 154-aa ORF3b, whereas SARS-CoV-2 encodes a single 121-aa ORF8 and a markedly truncated 67-aa ORF3b. These changes, including the loss of the VLWVL motif in SARS-CoV-2 ORF8, are thought to modulate innate-immune antagonism and may partly underlie the distinct transmission and pathogenesis profiles of the two viruses [18].

Once established in people, the two viruses diverged sharply in behaviour. SARS-CoV's basic reproduction number hovered near 2–3; infectiousness peaked five to seven days after symptom onset, and asymptomatic carriage was vanishingly rare. These properties made conventional public-health countermeasures,temperature screening, prompt isolation, rigorous contact tracing,proportionally effective. Within four months, global case counts plateaued at about 8 000 and transmission chains were snuffed out. SARS-CoV-2, by contrast,

incubates a day or two longer, sheds maximally at or before the first cough, and generates a large iceberg of mild or silent infections. The initial R_0 overlapped SARS-CoV's but subsequently escalated as successive variants accrued transmissibility advantages. Infected yet unsuspecting air passengers and holidaymakers dispersed the virus worldwide during the 2019-2020 festive travel surge; airport fever scanners missed presymptomatic carriers, and local contact-tracing systems were swiftly overwhelmed. Within a single quarter, COVID-19 case counts eclipsed SARS ten-fold; within six months, they exceeded ten million. Superspreading illustrates the contrast vividly [15, 16, 17]. Both pathogens exhibit over-dispersion, but while SARS outbreaks centred on nosocomial clusters,inadequately protected healthcare workers, ward visitors, or a single apartment block contaminated by aerosolised sewage,COVID-19 surged through choirs, nightclubs, call centres, meat-packing lines and other everyday community settings where asymptomatic individuals exhaled virus-laden aerosols for hours. The same biological edge,upper-airway tropism enhanced by the furin site,also rendered ordinary speech and breathing sufficient for airborne dissemination, making ventilation and masking decisive battlegrounds. Virulence further modulated impact. SARS-CoV's case-fatality ratio averaged 9–10 % (>50 % in seniors), compelling hospitalisation of the majority of patients. SARS-CoV-2 kills a smaller fraction,roughly 1–2 % during the first wave,but its vastly larger denominator translated into a death toll order of magnitude higher and stretched critical-care capacity in every region. In both diseases, age and comorbidity stratified risk, yet COVID-19's milder clinical spectrum in many adults concealed spread until health systems faced sudden surges. Why, then, did two genetically related viruses follow such divergent trajectories? The answer lies at the intersection of molecular virology and globalization (Petersen E, 2020). SARS-CoV-2 inherited several adaptive features before the first reported pneumonia cluster,high ACE2 affinity, furin-site cleavage, and efficient aerosol formation,that SARS-CoV only partly possessed. At the same time, international air traffic volumes in 2019 were nearly triple those of 2003, and supply chains moved live wildlife from rural farms to metropolitan markets in hours. The result was a pathogen biologically primed for rapid, silent spread released into a hyper-connected world. Both outbreaks nevertheless deliver a common lesson: wildlife trade and dense live-animal markets constitute a high-risk interface where bat-borne coronaviruses can recombine, adapt, and breach the species barrier. Preventing the next spillover demands active surveillance of bat populations, strict regulation (or elimination) of live-

wildlife commerce, and rapid data-sharing once unusual clusters appear. Containment remains possible,SARS-CoV proves as much,yet only if detection precedes widespread cryptic transmission, a window that narrows as viral fitness and global mobility expand [19, 20].

The Spike protein: molecular structure, functional domains, immunological target

Both SARS-CoV and SARS-CoV-2 are characterized by the presence of a highly mutable component that enables viral entry into host cells, known as the Spike protein (Figure 6). In the case of SARS-CoV-2 virus, monitoring mutations in the Spike protein has been of critical importance for tracking viral evolution and guiding socio-health measures aimed at containing the spread of the virus [21]. The SARS-CoV-2 Spike (S) protein is a large (~180 kDa) trimeric class I fusion glycoprotein that protrudes from the viral envelope, giving coronaviruses their characteristic crown-like appearance [22, 23]. Each Spike trimer consists of three identical protomers (each 1273 amino acids in SARS-CoV-2) arranged in a prefusion conformation with a mushroom-like shape. The globular S1 subunits form the head of the spike, and the membrane-anchored S2 subunits form the stalk that attaches to the virus membrane [22, 23]. A single Spike protomer contains two main functional subunits: S1, which is responsible for receptor binding, and S2, which mediates membrane fusion [22, 23]. During virus assembly, host proteases (furin or furin-like proteases) cleave the Spike at the S1/S2 junction, producing the S1 and S2 fragments (which remain non-covalently associated in the mature virion [22, 23]. Upon encountering a target cell, the Spike's receptor-binding domain in S1 engages the host receptor (human ACE2 for SARS-CoV-2), and a second proteolytic cleavage (at the S2′ site, often by TMPRSS2 or cathepsin) activates S2, releasing the fusion peptide and triggering the conformational changes that drive viral-host membrane fusion [22, 23]. In the post-fusion state, the Spike rearranges into a elongated needle-like structure that facilitates entry of the viral RNA into the host cell.

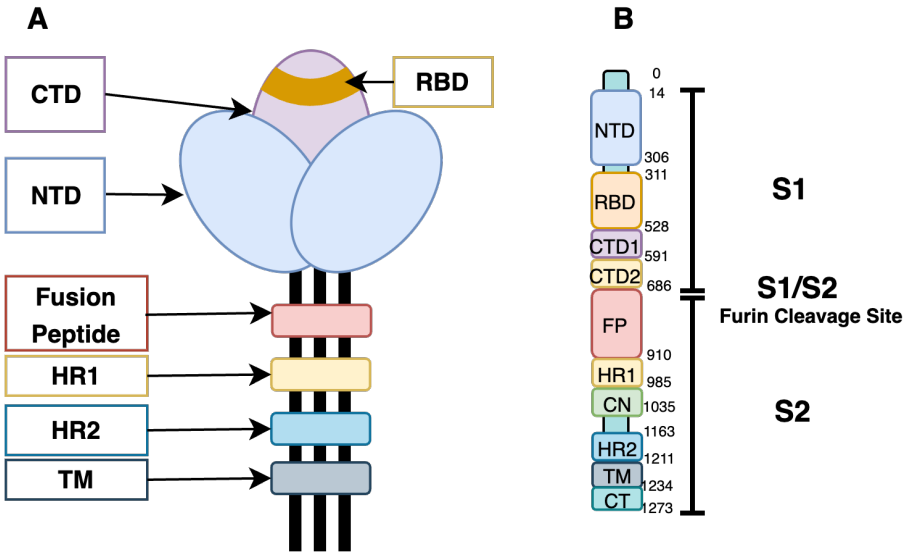


Figure 7: Domain organisation of the SARS-CoV-2 spike glycoprotein. (A) Side-view cartoon of the prefusion trimer embedded in the viral envelope. Each protomer is colour-coded by functional module: N-terminal domain (NTD) and C-terminal domain (CTD) form most of the S1 "cap", while the receptor-binding domain (RBD) hinges upward to engage ACE2. The stalk (S2 subunit) contains the fusion peptide (FP), heptad repeat 1 (HR1), heptad repeat 2 (HR2) and the trans-membrane helix (TM).
(B) Linear map of one spike polypeptide (residue numbers for the Wuhan-Hu-1 reference). The bracket delineates the receptor-binding S1 region (aa 14–685) from the fusion-machinery S2 region (aa 686–1273), separated by the furin-cleavage site (S1/S2). CTD is further divided into CTD1 and CTD2; CN indicates the connector segment, and CT the intraviral cytoplasmic tail. This modular layout physically segregates host-cell attachment (NTD/RBD) from membrane-fusion functions (FP, HR1/HR2, TM), enabling the large conformational rearrangements that drive viral entry [2].

Functional Domains of the Spike Protein

S1 Subunit (amino-acid (aa) residues 14–685): The S1 subunit (Figure 7) is the region of the Spike protein that mutates most rapidly and is responsible for binding to the host cell [24]. It contains an N-terminal domain and a receptor-binding domain, along with two smaller subdomains (SD1 and SD2). The N-Terminal Domain (NTD) can bind to sugar moieties on the host cell surface (e.g. sialic acids) and may play an auxiliary role in viral attachment [22, 23].The C-Terminal Domain (CTD) of S1 includes the Receptor-Binding Domain (RBD), which directly recognizes and binds to the host receptor. In SARS-CoV-2, the RBD

(within the CTD of S1) specifically binds to the ACE2 receptor on human cells, conferring the virus its tropism for ACE2-expressing tissues [22, 23]. The RBD is a dynamically moving domain: in the prefusion Spike, it can alternate between a "down" (receptor-inaccessible) and an "up" (receptor-accessible) position, the latter being required for ACE2 engagement. This conformational flexibility is a key aspect of Spike's function, as one RBD must swing up to attach ACE2, after which the S1 subunit can detach to allow the S2 subunit to refold and mediate fusion.

S2 Subunit (aa residues 686–1273): The S2 subunit (Figure 7) is responsible for fusion of the viral envelope with the host cell membrane. It contains several critical domains: the fusion peptide (FP) at the N-terminus of S2, followed by two heptad repeat regions (HR1 and HR2), a transmembrane domain, and a short cytoplasmic tail tissues [22, 23]. Upon activation, the hydrophobic fusion peptide (a ~15–20 amino acid segment) is thrust into the host cell membrane, where it anchors the virus and initiates the merging of the two membranes [25]. The HR1 and HR2 regions then fold back onto each other to form a six-helix bundle, a structural collapse that pulls the viral and cellular membranes together and drives the fusion process [25]. The transmembrane (TM) domain of S2 is a single-pass helix that anchors Spike firmly in the viral envelope attachment [22, 23]. Inside the virion, the S2 subunit ends with a short cytoplasmic tail (CT). This tail contains sorting signals and interacts with other viral components (such as the M protein) to ensure Spike is incorporated into new virions during assembly attachment [23]. Notably, the SARS-CoV-2 Spike has a unique polybasic furin cleavage site at the S1–S2 boundary (absent in the original SARS-CoV), which is cleaved during virus egress and is thought to enhance infectivity by priming the Spike for rapid entry into host cells [22, 23].

The Spike Protein as an Immunological Target

The Spike glycoprotein is the principal target of the immune system during infection and the cornerstone antigen for COVID-19 vaccines (Figure 8). It is highly immunogenic – patients who recover from SARS-CoV-2 infection develop robust antibody responses against [23]. In particular, the RBD of Spike is an immunodominant region: many potent neutralizing antibodies in convalescent plasma are directed against epitopes on the RBD [23, 25]. By binding to the RBD, these antibodies block the Spike from attaching to ACE2, thereby preventing the virus from entering cells (Figure 8). Other regions of Spike can also induce neutralizing antibodies – for example, certain antibodies target the NTD

"supersite," and others target more conserved S2 subdomains like the fusion peptide – but the RBD is considered the major target for neutralization [23, 25]. Neutralizing antibodies not only arise from natural infection but have also been isolated and developed as therapeutics; several monoclonal antibody treatments consist of antibody combinations that bind the Spike RBD and adjacent regions, neutralizing the virus by blocking receptor engagement [23]. In addition to humoral (antibody) responses, Spike-specific T cell responses ($CD4^+$ and $CD8^+$ T cells) have been detected in COVID-19 patients, underscoring Spike's strong immunogenicity across multiple arms of the adaptive immune system [23, 25]. Unsurprisingly, due to its critical functions and immunogenicity, the Spike protein has been the primary target for vaccine development and antiviral [23, 25]. Beyond vaccines and monoclonal antibodies, Spike has also been explored as a direct target for entry inhibitors. Structural studies revealed a conserved free-fatty-acid binding pocket in Spike that can bind linoleic acid and stabilize a more "locked" conformation, motivating the investigation of linoleic-acid derivatives as potential Spike-targeting antivirals [26]. In parallel, fusion-inhibitory peptides targeting the S2 heptad-repeat interface (e.g., EK1 and its derivative EK1C4) can block formation of the six-helix bundle required for membrane fusion [27, 28]. Finally, small molecules such as the repurposed laxative bisoxatin have been proposed (primarily from in-silico screening) as RBD-binding entry inhibitors, although these approaches remain largely preclinical [29, 30]. Virtually all current COVID-19 vaccines (Figure 8) (mRNA vaccines like Pfizer-BioNTech and Moderna, viral vectored vaccines like AstraZeneca's, protein subunit vaccines like Novavax, etc.) are designed to elicit an immune response against Spike. These vaccines present either the full-length Spike (often stabilized in the prefusion conformation) or key portions of it, with the aim of generating neutralizing antibodies that block virus entry [23, 25]. mRNA and viral-vectored vaccines deliver genetic instructions for Spike, leading host cells to transiently produce the antigen in situ and trigger immune priming, whereas protein-subunit vaccines directly administer purified Spike (or its domains) together with an adjuvant to elicit the same neutralizing response [23, 25]. The success of this approach is evident in the high efficacy of several Spike-based vaccines, which induce strong neutralizing antibody titers and protective immunity in humans. Likewise, most neutralizing monoclonal antibodies used for COVID-19 treatment or prophylaxis target the Spike – typically the RBD, given its exposed position and essential role in receptor binding [23, 25]. This intense immune focus on Spike, however, also means that viral evolution

tends to concentrate in the Spike gene: the ongoing emergence of SARS-CoV-2 variants has often involved mutations in the Spike protein (especially in the RBD and NTD regions) that enhance ACE2 affinity or confer resistance to some antibodies [23, 25]. These adaptations highlight the need to continuously monitor Spike mutations. More broadly, they emphasize why a comprehensive understanding of Spike's structure–function relationship is crucial – it helps us predict how changes in the Spike protein might impact viral infectivity, immune escape, and the effectiveness of vaccines or antibody therapies. To conclude the Spike protein's molecular architecture and immunological profile make it the central focus in both the pathogenesis of SARS-CoV-2 infection and the development of medical countermeasures to combat COVID-19.

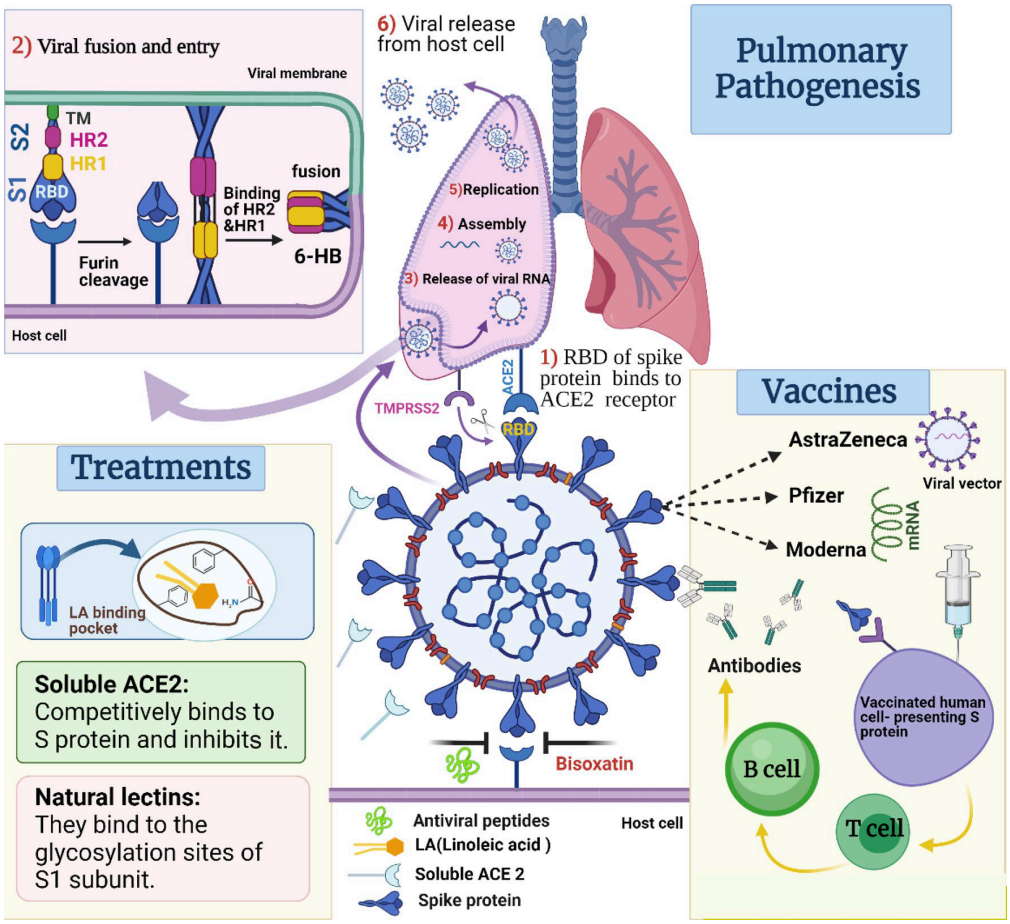


Figure 8: Schematic representation of the SARS-CoV-2 life-cycle in airway epithelial cells, its translation into pulmonary pathology, and the principal biomedical counter-measures. The spike's receptor-binding domain binds ACE2, TMPRSS2 primes the spike, and furin cleavage at S1/S2 allows S2 (HR1/HR2) to form a six-helix bundle that fuses viral and

cellular membranes, releasing the positive-sense RNA genome. The RNA is translated and replicated; virions assemble in the endoplasmic-reticulum–Golgi intermediate compartment and bud from the apical surface, spreading to the lower lung where alveolar infection drives COVID-19 lung damage. Entry-stage interventions include linoleic-acid derivatives that lock the spike closed, soluble ACE2 decoys, glycan-binding lectins, and fusion-inhibitory peptides (for example, bisoxatin). Current vaccines,adenoviral-vector (AstraZeneca) and lipid-nanoparticle mRNA (Pfizer-BioNTech, Moderna),deliver a prefusion-stabilized spike to elicit neutralizing antibody and T-cell responses. Together, these points connect entry mechanisms, pathology, and intervention targets central to biomedical-engineering strategies against SARS-CoV-2 [25].

Mechanisms of Mutation, Selection, and Emergence of SARS-CoV-2 Variants of Concern

SARS-CoV-2 (the virus causing COVID-19) has shown a remarkable capacity to generate genetic diversity through multiple mechanisms. Genetic mutations arise continually as the virus replicates, and natural selection then acts on these mutations. Changes in the virus's Spike protein – especially in the receptor-binding domain (RBD) – can significantly alter how easily the virus infects cells, how quickly it spreads, and its ability to evade immune defenses. Over the course of the pandemic, these processes have led to the emergence of several variants of concern (VOCs) – notably Alpha, Beta, Gamma, Delta, and Omicron – which had attributes (such as higher transmissibility or immune escape) that gave them a fitness advantage (Figure 9). After the unprecedented expansion of Omicron, no new SARS-CoV-2 variant has accumulated enough phenotypic change to justify a fresh WHO (World Health Organization) "Variant of Concern" label. Instead, in March 2023 the WHO revised its tracking system so that the focus shifted from broad VOC categories to a finer-grained, lineage-based framework that evaluates individual Omicron descendants on their own merits. In the Pango nomenclature [31] a lineage is a genetically coherent, monophyletic cluster of viral genomes that share a characteristic constellation of mutations and show clear evidence of sustained transmission; it functions as a stable genomic "address" that can be tracked across laboratories and datasets, irrespective of the public-health risk that cluster may (or may not) pose. This lineage-centred approach recognises that, in the post-Omicron era, viral evolution proceeds mainly through incremental diversification within the Omicron clade rather than through the emergence of wholly new variant backbones, and it allows

risk assessments to be updated lineage-by-lineage as fresh data accumulate.

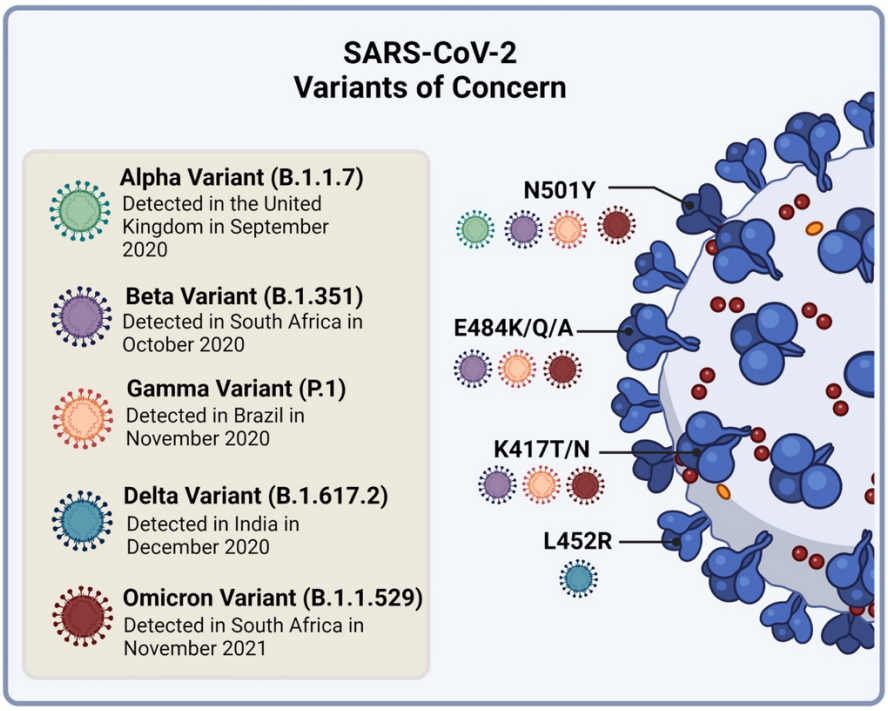


Figure 9: Global emergence of major SARS-CoV-2 "variants of concern" and their signature receptor-binding-domain (RBD) mutations. The first five variants designated by the World Health Organization,Alpha (B.1.1.7, United Kingdom, September 2020), Beta (B.1.351, South Africa, October 2020), Gamma (P.1, Brazil, November 2020), Delta (B.1.617.2, India, December 2020) and Omicron (B.1.1.529, South Africa, November 2021) ,are shown schematically on the left. On the right, a cut-away view of the trimeric spike illustrates four RBD substitutions repeatedly selected across these lineages: N501Y, which tightens ACE2 binding and increases intrinsic transmissibility; E484K/Q/A and K417T/N, which diminish the affinity of many class-1 and class-2 neutralising antibodies; and L452R, which simultaneously enhances receptor affinity and confers partial resistance to cytotoxic-T-cell recognition. The convergence of these mutations underlines the twin evolutionary pressures of host-receptor optimisation and immune escape that have shaped SARS-CoV-2 diversification [32].

Genetic Mechanisms Driving SARS-CoV-2 Variation

SARS-CoV-2 is an RNA virus that accumulates point mutations during replication, because its RNA-dependent RNA polymerase occasionally makes copying errors; although the virus possesses a proofreading

exonuclease (ExoN) that slows the mutation rate compared with many other RNA viruses, substitutions still accrue at a steady pace and give rise to genetic variants over time [31]. A well-known early case was the Spike D614G substitution, which emerged in early 2020 and quickly became globally dominant after conferring roughly a 20 % transmission advantage over the ancestral strain [31]. Coronaviruses also undergo recombination because their replication is "discontinuous," allowing the polymerase to switch templates when two distinct SARS-CoV-2 variants coinfect the same cell, so the resulting progeny can be mosaics of the two parental genomes [31]. Recombinant lineages are denoted with an "X" prefix in the PANGO nomenclature [31]. The first documented example, XA, combined material from the Alpha variant (B.1.1.7) and lineage B.1.177 in the United Kingdom, while another recombinant designated XC arose from an Alpha–Delta coinfection in Japan [31]. As different strains co-circulated, additional recombinants such as XD (a Delta × Omicron BA.1 hybrid) and XBB (formed by two Omicron BA.2 sublineages) appeared (Figure 10, Figure 11) [31]. Although most early recombinants had little epidemiological impact, the success of XBB demonstrated that template switching can occasionally generate highly fit progeny [31]. Beyond substitutions and recombination, SARS-CoV-2 genomes experience small deletions and occasional insertions that modulate viral properties. Recurrent deletions in the Spike gene are noteworthy: the Alpha variant lost amino acids 69–70 in the N-terminal domain (NTD), a region targeted by several neutralizing antibodies, and this H69/V70 deletion helped the virus partially evade immune recognition [31]. Alpha also harbored a Y144 deletion in the NTD, further altering an antigenic "supersite" and reducing neutralization by convalescent and vaccine sera [33]. Although rarer, insertions can be significant; Omicron, for instance, possesses an insertion in its Spike NTD (ins214EPE) that likely arose through template switching and may affect antigenicity [34]. Collectively, point mutations, recombination, and indels underscore the genetic flexibility that drives SARS-CoV-2 evolution.

Spike Protein Mutations and Viral Fitness

Many of the most consequential SARS-CoV-2 mutations occur in the Spike protein, which is responsible for binding to host cell receptors and is also the primary target of the immune response. Mutations in Spike – especially in the RBD – can affect two key aspects of viral fitness: infectivity/transmissibility and immune evasion. Impact on Infectivity and Transmissibility: The Spike protein's RBD mediates attachment to the

ACE2 receptor on human cells. Mutations in the RBD can increase the strength or efficiency of this interaction, thereby increasing infectivity. A prominent example is the N501Y substitution (found in Alpha, Beta, and Gamma VOCs), which enhances binding affinity to ACE2 and was shown to increase viral transmission rates [33]. Another example is L452R (present in Delta and some others), which stabilizes RBD interactions and has been associated with both improved ACE2 binding and partial immune escape. Besides the RBD, mutations near the furin cleavage site of Spike can also boost transmissibility. The furin cleavage site (at the S1–S2 junction) is critical for efficient cell entry, and both Alpha and Delta evolved mutations in this region (e.g. P681H in Alpha and P681R in Delta) that enhance Spike cleavage by furin [31]. These changes contributed to Alpha and Delta being approximately 50–60 % more transmissible than prior strains [31]. In short, mutations that allow the virus to enter cells more efficiently or replicate to higher levels in the respiratory tract will be favored and can lead to faster spread in the population. Impact on Immune Evasion: Equally important is the virus's ability to evade neutralizing antibodies produced by the host (through infection or vaccination). Spike is the main target of these antibodies, and thus mutations in Spike can reduce antibody binding. The RBD and NTD of Spike contain the dominant antigenic sites. Changes in these regions can prevent antibodies from recognizing the virus, leading to "immune escape." For example, the E484K mutation (first notable in Beta and Gamma) alters an RBD residue that is a key contact for many potent neutralizing antibodies. E484K has been linked to markedly reduced neutralization by convalescent plasma and vaccine-elicited sera [33]. This mutation alone was sufficient to reduce the effectiveness of certain monoclonal antibody therapies and to diminish vaccine protection, prompting concern and close monitoring [33]. Beta and Gamma both possessed E484K along with K417T/N and N501Y in the RBD – this combination not only increased ACE2 binding but also conferred significant resistance to neutralizing antibodies [33]. In the NTD, a cluster of mutations or deletions can likewise abrogate binding of NTD-specific neutralizing antibodies. Beta, for instance, had an array of NTD changes (e.g. L18F, D80A, Δ242-244, R246I) that, together with its RBD mutations, significantly impaired neutralization by both monoclonal antibodies and convalescent/vaccine sera [33]. These mutations gave Beta a pronounced ability to escape immunity – at the cost of some viral fitness, perhaps, since Beta did not spread as globally as the more transmissible variants. It is important to note that there can be trade-offs. A mutation that improves immune evasion might slightly reduce ACE2 affinity or

replication efficiency, and vice versa, so the virus's evolution finds a balance. Over time, as population immunity increased (through infection and vaccination), variants with greater immune escape had increasing advantages. The Omicron variant (B.1.1.529 lineage), which emerged in late 2021, represents the extreme end of this spectrum: it accumulated ~30 Spike mutations, including >15 in the RBD, resulting in a virus that could infect people with previous immunity (from vaccines or prior COVID-19) at far higher rates than earlier variants [31]. Omicron's mutations (e.g. at positions 346, 440, 446, 477, 478, 484, 493, 501, 505 in the RBD, and deletions at 69–70, 143–145 in NTD among others) caused a major antigenic shift, allowing it to evade the vast majority of neutralizing antibodies from pre-Omicron immunity [31]. Thus, Spike mutations have been central to SARS-CoV-2's adaptation: some increased contagiousness, others enabled escape from the immune system, and the most successful variants often combined both types.

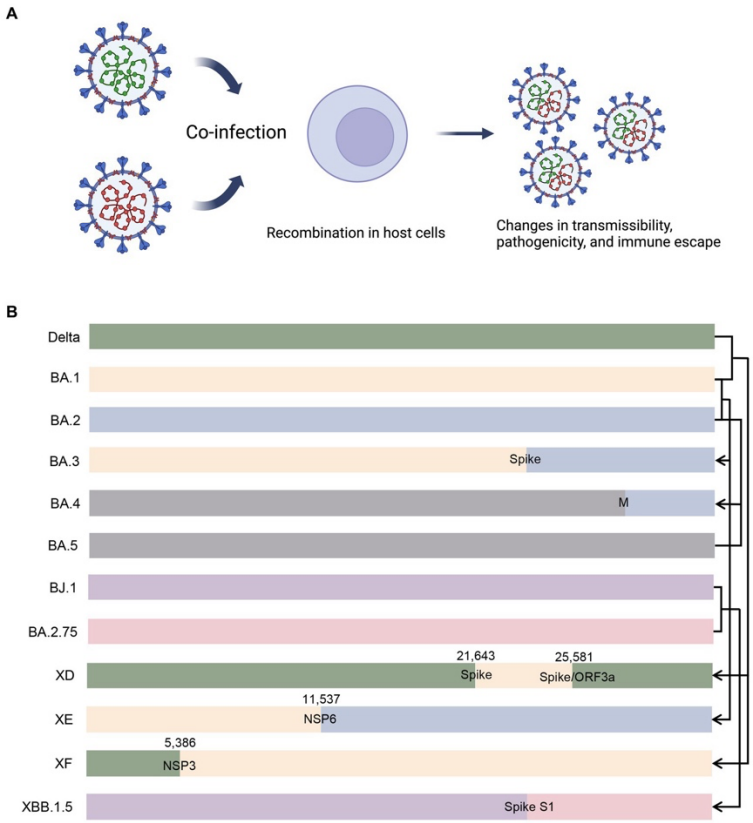


Figure 10: Mechanism and genomic signature of recombinant SARS-CoV-2 lineages ("X-series" variants). When two genetically distinct variants,illustrated in green and red,co-infect the same host cell, the viral RNA polymerase can switch templates, producing mosaic genomes that

blend sequence blocks from each parent; the resulting progeny may acquire new levels of transmissibility, pathogenicity or immune escape (panel A). Panel B aligns the full-length genomes of the principal parental groups,Delta and the Omicron sub-lineages BA.1, BA.2, BA.3, BA.4, BA.5, BJ.1 and BA.2.75,with the World Health Organization–tracked recombinants XD, XE, XF and XBB.1.5. Coloured bars denote genomic segments, while arrows mark the recombination break-points and their nucleotide positions (e.g., 21 643 and 25 581 within the spike gene and at the spike/ORF3a junction; 11 537 in NSP6; 5 386 in NSP3). The dendrogram on the right situates each recombinant within the Omicron-dominated phylogeny, underscoring that RNA-template switching, alongside stepwise mutation, is a powerful engine of SARS-CoV-2 diversification [35].

Key Variants of Concern (VOCs)

During the COVID-19 pandemic, several major variants,Alpha, Beta, Gamma, Delta, and finally Omicron,arose on different continents. Each carried distinctive mutations that boosted its fitness, enabling it to spread widely at the global or regional level. Once Omicron appeared, the WHO stopped assigning new Greek letters to its offshoots; instead, the entire family of Omicron descendants has been treated as a single variant class. From 2023 through 2025, every circulating strain belongs to this Omicron lineage (for example, the various XBB subvariants). Consequently, surveillance has shifted from tracking brand-new Greek-letter variants to monitoring the ongoing evolution taking place within Omicron itself.

Alpha (B.1.1.7 lineage)

First identified in the UK (late 2020), Alpha (Figure 9) was characterized by ~23 mutations, including Spike N501Y (increasing ACE2 affinity) and the 69–70 deletion in the NTD. Alpha's transmissibility was substantially higher than previous strains (roughly 50% or more greater) [31]. Its mutations (which also included P681H near the furin site and Y144del in NTD) were thought to improve viral entry and allow some immune escape, such as evading certain NTD-targeted antibodies [33]. Alpha caused a large surge in cases globally and became dominant in many countries in early 2021. Fortunately, studies indicated that vaccines remained effective against Alpha, as it did not have the most problematic RBD escape mutations.

Beta (B.1.351 lineage)

Emerged in South Africa (late 2020) and was notable for its immune evasiveness. Beta (Figure 9) had three critical RBD mutations – K417N, E484K, and N501Y – that both increased binding to ACE2 and conferred resistance to many neutralizing antibodies [33]. It also had multiple NTD deletions/mutations (e.g. Δ242–244) that helped it escape immune recognition [33]. Beta showed approximately 50% increased transmissibility over earlier strains in its region [33], but its defining feature was the reduced effectiveness of antibodies against it. Convalescent plasma and some vaccine-induced sera were markedly less potent at neutralizing Beta [33]. This variant was even able to cause reinfections in people who had recovered from other variants. While Beta did not become globally dominant (likely due to a fitness disadvantage in transmission compared to Alpha and later Delta), it raised alarms because it demonstrated the virus could evolve to significantly erode vaccine efficacy. The mutations in Beta foreshadowed the need for updated vaccines if such variants spread widely.

Gamma (P.1 lineage)

First detected in early 2021 among travelers from Brazil (and associated with the Manaus outbreak), Gamma (Figure 9) had a similar mutational profile to Beta. It carried N501Y, E484K, and K417T in the Spike RBD, along with several NTD changes [33]. Gamma was highly transmissible and struck Brazil at a time when an estimated 70% of the population had already been infected in earlier waves, suggesting extensive immune escape. In Manaus, a six-fold increase in hospitalizations was observed in Jan 2021 (Gamma wave) compared to Dec 2020, despite high seroprevalence [33]. This indicated many reinfections and potentially more severe disease. Gamma's RBD/NTD mutations mirrored Beta's in function, allowing it to partially evade neutralization. It became the dominant strain in Brazil and spread to many countries in South America. Though Gamma, like Beta, was somewhat less fit than Delta in absolute transmission, it reinforced concerns about variants causing reinfection and reduced antibody protection.

Delta (B.1.617.2 lineage)

Emerged in India (early 2021) and became the most dominant global variant by mid-2021. Delta (Figure 9) was a highly transmissible variant

– about 55–60% more transmissible than Alpha [31], making it several times more contagious than the original Wuhan strain. Key Spike mutations in Delta included L452R and T478K in the RBD, which enhanced ACE2 binding and helped the virus partially evade certain class of antibodies, and P681R at the furin cleavage site, which markedly improved Spike processing and viral entry [31, 33]. Delta also had changes in other genes that may have increased viral replication. Clinically, Delta caused rapid, high-volume outbreaks and was associated with higher viral loads and a potential increase in severity (some data suggested a higher risk of hospitalization with Delta infections). While Delta did show some ability to evade immunity (e.g., reduced neutralization by one dose vaccines or certain monoclonal antibodies [33]), vaccines still provided good protection against severe disease from Delta. Delta's success was primarily due to its formidable transmissibility – it outcompeted Alpha across the globe and remained the dominant lineage until the advent of Omicron.

Omicron (B.1.1.529 lineage and sublineages)

First reported in Southern Africa in November 2021, Omicron (Figure 9) represented a dramatic evolutionary leap. It harbored over 50 mutations, with around 30 in the Spike protein alone. Notably, Omicron's Spike had multiple RBD mutations (including some at the same sites as Beta/Gamma, like E484 and K417, but often different amino acid changes, e.g. E484A and K417N in BA.1) as well as many new ones (such as G446S, Q498R, N440K, and the combination of S371L/S373P/S375F, etc.). The result was a variant with unprecedented immune escape: Omicron could efficiently infect individuals who were fully vaccinated or previously infected by earlier variants [31]. Laboratory tests showed that neutralizing antibody titers against Omicron were dramatically lower; first-generation vaccines and even antibodies from those who recovered from Delta had trouble neutralizing Omicron sublineages [31]. This led to a wave of breakthrough infections and reinfections worldwide. Despite this immune evasion, Omicron (especially the initial BA.1 strain) was observed to cause somewhat less severe disease on average than Delta, possibly due to differences in cell tropism (Omicron prefers upper respiratory tract replication more than lung tissue) and the fact that many hosts had some immune priming. Omicron's intrinsic transmissibility was high, though some of its spread was due to immune evasion allowing it to find new hosts. It very rapidly displaced Delta globally. Multiple sublineages of Omicron (BA.1, BA.2, BA.4, BA.5, and others like XBB

which is a recombinant of Omicron strains) sequentially rose to dominance through 2022 and 2023, demonstrating the virus's continued evolution even within the Omicron family. Omicron's emergence marked a shift to a new antigenic variant so distinct that some experts argued it should be considered a new "serotype" of SARS-CoV-2 [31]. This variant forced updates in vaccines and treatments, as discussed next.

Public response and Severity metrics: from the Stringency Index to other measures

Public Response and Genomic Surveillance

SARS-CoV-2 has demonstrated a remarkable capacity to evolve (Figure 11) under the selective pressures exerted by host immune responses and public health interventions. Widespread immunity,whether from prior infection or vaccination,along with measures like antiviral treatments and lockdowns, creates an environment where the virus is pressured to adapt for continued spread [36]. This dynamic has produced VOCs with higher transmissibility or immune escape [36]. For example, the Beta and Omicron variants carried mutations (like E484K or multiple spike protein changes) that significantly reduced neutralization by antibodies, leading to immune evasion and reinfections even in previously immune individuals [36]. Such variants underscore how the virus can exploit evolutionary opportunities despite ongoing public health measures. These variant-driven adaptations reinforce the critical need for continuous genomic surveillance and real-time virological monitoring worldwide. Genomic surveillance involves sequencing virus samples to detect new mutations and variant lineages as early as possible. The importance of this was evident as new VOCs like Delta and Omicron were identified and tracked through global databases (e.g., GISAID [37]), enabling rapid scientific analysis. Robust surveillance systems serve as an early warning mechanism, allowing health authorities to spot concerning mutations and assess their impact on transmissibility, severity, or immune escape [36].

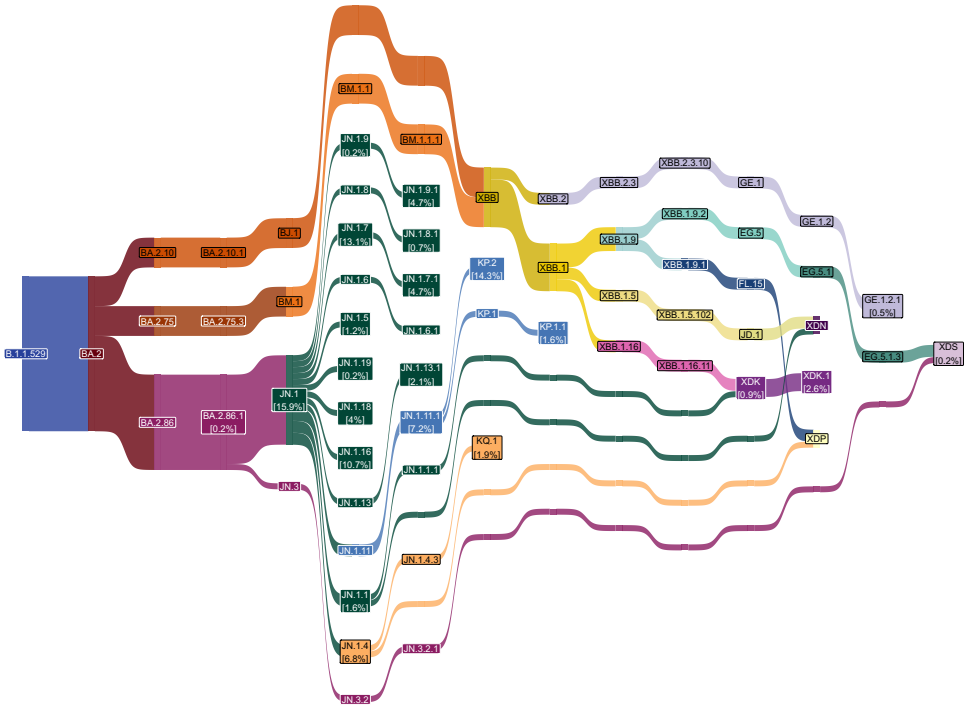


Figure 11: Sankey diagram illustrating the evolutionary trajectory and diversification of Omicron sublineages, beginning with the parental lineage B.1.1.529 and branching into multiple BA.2-derived variants. The diagram tracks the progressive emergence of key lineages such as BA.2.75, BA.2.86, JN.1, XBB, KP.2, and KQ.1, with percentages indicating their relative prevalence at the population level. Notably, XBB and its sublineages (e.g., XBB.1.5, XBB.1.16, XBB.2.3)represent major evolutionary nodes that subsequently generated further high-impact variants including EG.5, FL.15, JD.1, XDK, and GE lineages.From an epidemiological perspective, these branching patterns reflect successive waves of viral dominance: early BA.2 expansions provided the genetic substrate for the explosive rise of XBB derivatives, which were associated with widespread global surges in 2023. Later, EG.5 and JN.1 continued this trajectory, driving subsequent increases in transmission and immune escape [38]

In practice, once a variant with immune-escape properties is identified, interventions can be adjusted – for instance, tightening travel restrictions or advisories if a highly transmissible variant emerges, or reintroducing mask mandates if needed. Crucially, vaccine developers can update vaccine compositions to target new strains. Indeed, mRNA vaccine platforms offer the flexibility to rapidly tweak the antigen design in response to evolving variants [36]. During 2021–2022, booster doses and bivalent vaccines were rolled out as responses to variants (like Omicron)

that were able to partially evade first-generation vaccines [36]. By integrating genomic surveillance with real-time epidemiological data, public health officials can proactively implement or adjust interventions (such as targeted testing, localized lockdowns, or updated treatment protocols) before a variant becomes widespread. Notably, the interaction between viral evolution and public health measures is a two-way street. While measures like lockdowns and travel bans can slow the spread, they also impose selective pressures that may favor variants with certain advantages (such as higher infectivity). Conversely, the emergence of a more transmissible or virulent variant often forces governments to strengthen interventions. For example, the global spread of the Delta variant (with higher virulence) prompted many countries to reinstate strict distancing rules in mid-2021, and the highly transmissible Omicron variant led to re-imposition of travel restrictions and accelerated booster vaccination campaigns in late 2021. These examples illustrate how viral evolution and policy responses influence each other, necessitating an agile approach to pandemic management [36].

Severity Metrics and Public Response: From the Stringency Index to Other Measures

An evidence-based pandemic response hinges on quantifiable metrics that capture both the disease's impact and the strictness of interventions. Quantitative metrics have become crucial for assessing COVID-19 pandemic severity and guiding policy decisions in real time and also for genomic surveillance. They allow authorities to gauge how fast and severe an outbreak is, and to determine when to tighten or relax public health measures. One widely used composite indicator of government interventions is the Oxford COVID-19 Government Response Tracker's Stringency Index (Figure 12).

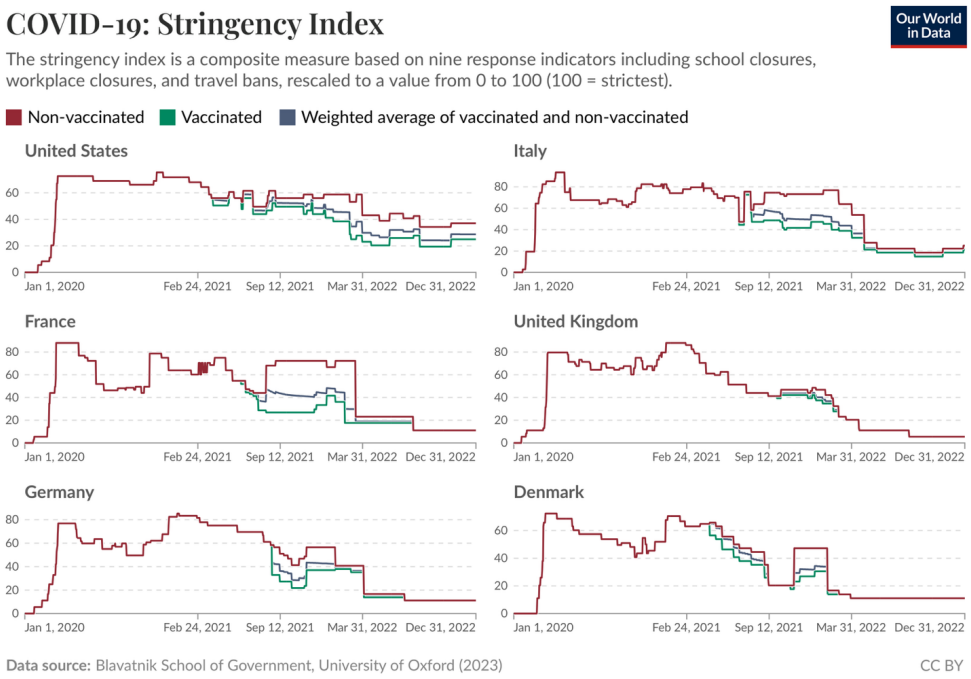


Figure 12: Evolution of government response stringency during the COVID-19 pandemic (January 2020 – December 2022). The Oxford–Blavatnik composite Stringency Index (0 = no restrictions, 100 = full lockdown) is plotted separately for non-vaccinated populations (red), vaccinated populations (green) and the weighted national average (blue) in six high-income countries: United States, Italy, France, United Kingdom, Germany and Denmark. All nations imposed near-maximal restrictions in March 2020, followed by gradual easing through 2021. After widespread vaccine rollout, differential policies,such as "green passes" and workplace mandates,produced a persistent gap in index values between vaccinated and non-vaccinated groups, most evident in continental Europe. By mid-2022 the index converged toward low levels across all groups as mass-vaccination, natural immunity and variant evolution permitted relaxation of school closures, mobility bans and workplace limits [39].

This Stringency Index distills various containment and closure policies into a single number on a scale from 0 to 100, reflecting the overall strictness of a country's response [39]. It incorporates measures such as school and workplace closures, cancellation of public events, restrictions on gatherings and travel, stay-at-home orders, public information campaigns, and other social-distancing rules [39]. A higher Stringency Index score indicates more stringent restrictions (with 100 being the maximum severity of lockdown policies) at a given time. This metric has enabled global comparisons of policy aggressiveness – for instance, to

compare how different nations responded to surges – and provides context for interpreting health outcomes. It is important to note that the Stringency Index records the strictness of government policies, not their effectiveness [39]. However, by tracking this index alongside epidemiological data, researchers and policy-makers can study correlations between intervention strictness and outcomes like case numbers or mortality. In addition to policy strictness indices, a range of clinical and epidemiological metrics are monitored to assess pandemic severity and the burden on health systems. These include measures of disease spread, health care capacity, and fatality risk, each offering essential insights into different dimensions of the crisis. Key metrics that have been widely tracked include:

- **Hospitalization Rates** – the number of COVID-19 hospital admissions (often scaled per 100,000 population or 1,000,000 population) indicates the burden of severe disease in the community. A rising hospitalization rate signals a growing wave of severe infections, which can forewarn of stress on the healthcare system. Health authorities have relied on hospital admission data to decide when to implement emergency measures (for example, pausing elective surgeries or opening field hospitals). Declining hospitalization rates, especially after vaccine rollouts, have signaled an easing of pandemic severity. In fact, as vaccination coverage increased in late 2021, many regions observed a significant drop in COVID hospitalization rates compared to earlier waves [40], reflecting the protective effect of vaccines against severe outcomes (Figure 13).
- **ICU Occupancy** – the percentage of intensive care unit beds occupied by COVID-19 patients is a critical indicator of health system capacity. High ICU occupancy rates can indicate that a healthcare system is near its limit in caring for the most critical patients. Public health agencies closely watch this metric, as sustained ICU occupancy above certain thresholds (e.g. >75% or >90% capacity) is associated with increased mortality risk for patients due to strained resources. For example, during the Delta variant surge, many hospitals saw ICU beds fill up, with some regions reporting ICU occupancy rates well above 80–90%, necessitating crisis standards of care. Ensuring that ICU occupancy remains below dangerous levels is key for saving lives; thus this metric often triggers interventions like inter-hospital patient transfers or regional lockdowns to curb transmission (Figure 13).

- **Infection–Fatality Ratio (IFR)** – the IFR represents the proportion of infected individuals who die from the disease. It quantifies the likelihood of an individual dying once infected with the pathogen [Variation in the COVID-19 infection–fatality ratio by age, time, and geography during the pre-vaccine era: a systematic analysis]. This metric is fundamental for understanding the true severity of COVID-19 in the population (as opposed to the case-fatality ratio, which uses only confirmed cases in the denominator). IFR can vary by factors such as age, risk profile, variant, and level of care available. During the pandemic, the overall IFR evolved over time – early estimates in 2020 often ranged around 0.5–1%, but by 2022 the IFR in some countries had fallen to around 0.1% [41]. This improvement was driven by better treatments, younger age distributions of cases, and especially widespread vaccination of high-risk groups. Tracking IFR over time provides insight into whether the virus is becoming more or less deadly (intrinsically or due to interventions). Moreover, reliable IFR estimates, together with infection hospitalization ratios (IHR), help modelers and health officials forecast the number and timing of severe outcomes, thereby allowing hospitals to prepare for surges [41]. For instance, if an emerging variant or waning immunity causes IFR to rise, that would warn of potentially higher death tolls and justify stricter control measures.
- **Vaccination Coverage** – the proportion of the population vaccinated (with initial series and boosters) is a pivotal metric for public health planning. High vaccination coverage, especially among vulnerable groups, typically correlates with lower rates of severe disease and death, thus bending down the hospitalization and IFR metrics. Public dashboards have tracked the percentage of people fully vaccinated and boosted; these figures inform projections of herd immunity levels and susceptibility of the population. Regions with higher vaccination coverage have not only seen milder impacts from new variants, but also faster recovery from surges [40]. For example, as noted, when vaccine coverage in a community rose, subsequent waves like the Omicron wave resulted in far fewer hospitalizations and ICU cases than earlier waves in the unvaccinated era. Conversely, gaps in vaccination coverage can signal where the next outbreaks or severe clusters may occur, guiding targeted vaccine campaigns or resource allocation (such as sending more support to areas with low coverage before a surge

hits). Thus, vaccination coverage is both a metric of population-level protection and a predictor of future pandemic severity.

Each of these metrics provides a different lens on the pandemic's status: policy strictness (Stringency Index), healthcare burden (hospital and ICU numbers), and disease severity (IFR), as well as population immunity (vaccination rates). Effective policy responses depend on the integration and interpretation of all these indicators in tandem. No single number can capture the full picture; for instance, case counts alone became less informative once mass vaccination began, so governments started weighing hospitalizations and ICU occupancy more heavily to guide reopening plans. Multidisciplinary teams – including epidemiologists, clinicians, data scientists, and public policy experts – are needed to analyze and contextualize these data for decision-making. Interpreting the metrics requires understanding of epidemiology (to infer trends and model projections), medical insight (to gauge healthcare capacity and outcomes), and social science (to assess adherence to interventions and impacts of policies). Indeed, generating actionable insights from pandemic data requires staff with a range of skills working together to check, analyze, and translate data for policy audiences [42].

Furthermore, robust data infrastructure is vital to support this metrics-driven approach. High-quality, timely data from diverse sources (hospitals, laboratories, surveillance systems, etc.) must feed into unified platforms to give decision-makers an up-to-date picture of the pandemic. Many countries have developed live dashboards and data pipelines that aggregate these metrics at national and regional levels. However, ready access to comprehensive, multi-dimensional data has proven challenging in many places [42]. The COVID-19 pandemic exposed gaps in data-sharing and health information systems; thus, strengthening data infrastructure and governance has become a priority for future preparedness. Investments in real-time data collection, integration (for example linking epidemiological data with genomic data), and analytics are essential. With improved data systems, authorities can rapidly identify when a combination of indicators – say, a sharp rise in test positivity alongside increasing hospital admissions – warrants an immediate public health response By basing decisions on a holistic dashboard of indicators (from stringency levels to ICU bed usage and fatality rates), policymakers can better balance public health measures with socio-economic considerations, ultimately leading to more effective and adaptive pandemic management [42, 36].

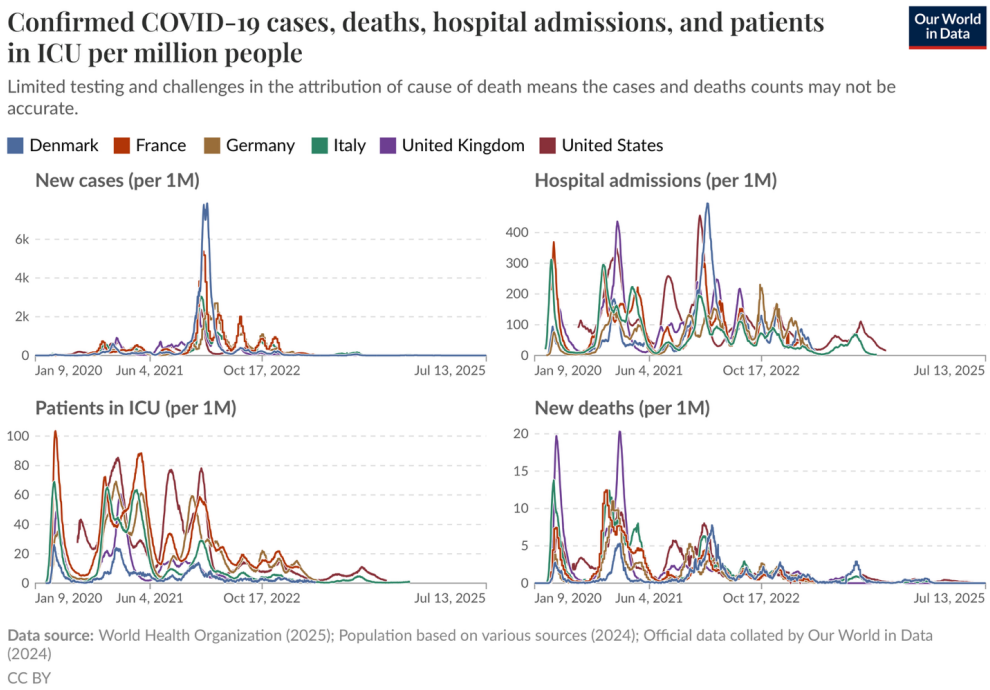


Figure 13: Trajectories of COVID-19 burden indicators in six high-income countries, January 2020 – July 2025. Time-series panels report per-million population averages for (top left) new laboratory-confirmed cases, (top right) hospital admissions, (bottom left) patients in intensive-care units (ICU), and (bottom right) new deaths in Denmark, France, Germany, Italy, the United Kingdom and the United States. All six nations experienced synchronous first-wave surges in early 2020, followed by successive variant-driven peaks that became progressively decoupled: while Omicron-era case incidence briefly exceeded 6 000 new infections / 1 M, corresponding ICU occupancy and mortality peaks were markedly blunted, reflecting vaccine- and infection-acquired immunity, improved therapeutics, and targeted shielding of high-risk groups. By mid-2025 all indicators approached baseline, yet inter-country heterogeneity persisted,most conspicuously in early hospital-pressure spikes (e.g., United Kingdom) and prolonged ICU burdens (e.g., United States) [43].

1.2. The Role of Artificial Intelligence in Surveillance and Health Policy Decision-Making

The COVID-19 pandemic has had an unprecedented global impact [44] with the rapid emergence of multiple SARS-CoV-2 VOCs carrying mutations that altered viral transmissibility, virulence, or immune escape properties [45]. Early detection of these variants was essential for updating vaccines and therapeutics and for guiding public health

responses [45]. In response, the global scientific community launched expansive genomic surveillance efforts. By May 2022, over 11 million SARS-CoV-2 genomes had been sequenced and shared via the GISAID platform [46, 37], enabling near real-time tracking of viral evolution on a scale never before seen. However, translating this deluge of genomic data into actionable insights in real time posed a major challenge. Artificial intelligence (AI) and machine learning (ML) rapidly emerged as critical tools to augment human analysis, capable of rapidly detecting patterns and making predictions that support pandemic management [44]. Indeed, during COVID-19, AI has shown potential to enhance viral surveillance, streamline epidemiological predictions, and support decision-making for public health policy (Figure 14) [44].

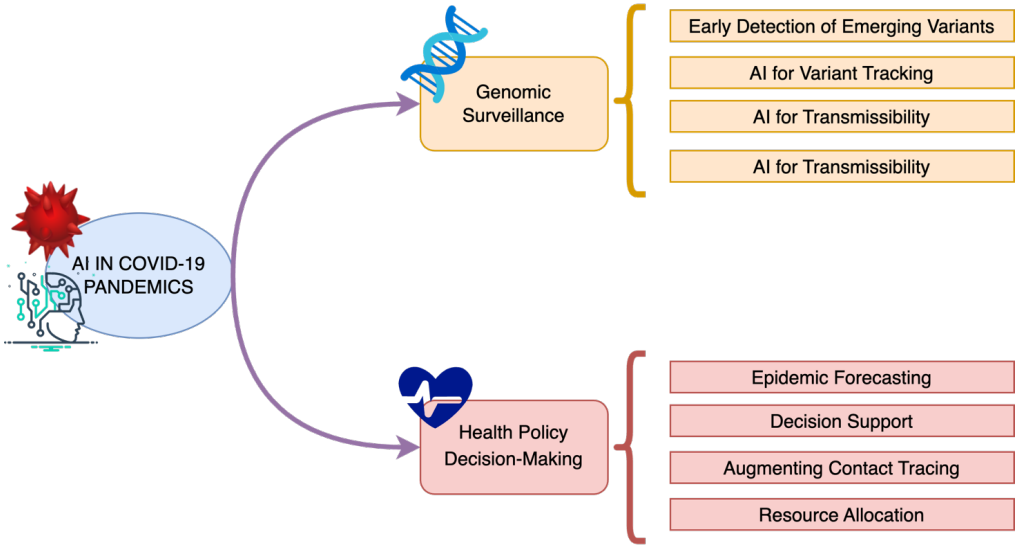


Figure 14: Role of Artificial Intelligence in Genomic Surveillance and Health Policy Decision-Making During the COVID-19 Pandemic. This schematic outlines the dual impact of AI applications during the COVID-19 pandemic across two key domains. In genomic surveillance, AI facilitated the early detection of emerging SARS-CoV-2 variants, enhanced variant tracking, and supported the assessment of viral transmissibility. In health policy decision-making, AI contributed to epidemic forecasting, strategic decision support, improved contact tracing, and optimal resource allocation. Together, these applications highlight the integrative role of AI in both scientific and operational pandemic response efforts [2].

AI Applications in SARS-CoV-2 Genomic Surveillance

The COVID-19 pandemic prompted an unprecedented genomic surveillance effort, yielding millions of SARS-CoV-2 genome sequences worldwide. Analyzing this deluge of data in real-time has been challenging for traditional epidemiological methods. AI and machine learning ML have therefore been leveraged to detect patterns in viral mutations, track emerging variants, and predict variant behavior. Between 2020 and 2024, 135 studies explored AI-driven techniques – from deep learning models to statistical learning and clustering algorithms – to strengthen genomic surveillance. These studies aim to identify new VOC earlier, classify and cluster viral lineages, and predict key outcomes such as a variant's spread, transmissibility, or immune escape potential.

Early Detection of Emerging Variants with Machine Learning and Deep Learning

A primary goal of genomic surveillance is early detection of new or high-risk variants before they spread widely. Traditional identification (via lab assays or phylogenetics) often lags behind variant emergence, spurring development of AI models for faster warnings. Unsupervised learning has shown particular promise in flagging novel variants from large sequence datasets without prior labels. For instance, Nicora et al. [47] introduced an automated genomic surveillance system using a One-Class Support Vector Machine (SVM) trained on early neutral SARS-CoV-2 spike protein sequences to detect anomalies suggestive of emerging variants of concern (VOCs) or interest (VOIs). The training process (Figure 15) begins by selecting a set of "neutral" variants,those circulating before the emergence of known variants of concern (VOCs),from the early phase of the pandemic (February–June 2020). These sequences are transformed into numerical vectors using 3-mer encoding, a standard bioinformatics technique that captures short amino acid patterns. The One-Class SVM is then trained exclusively on these neutral sequences, learning the statistical distribution of what constitutes "normal" variation in the spike protein. Importantly, the model is not trained to recognize known VOCs directly, but rather to identify outliers,sequences that deviate from this baseline distribution and may thus signal emerging evolutionary trajectories.Every week, the model is applied to new sequences deposited in GISAID. Any sequence predicted as an outlier is flagged as a potential non-neutral variant (i.e., a candidate VOC or VOI). If during that week health authorities officially classify a new VOC or

VOI, the model is retrained: the original neutral set is updated by incorporating all previously predicted inliers,that is, sequences the model had judged as normal,up to that week. This dynamic retraining approach enables the classifier to adapt to shifts in the viral population while preserving its ability to highlight truly novel deviations. Notably, false negatives (i.e., VOCs misclassified as inliers) are not excluded from the retraining set, which reflects a pragmatic compromise: the goal is not to catch all outliers but to detect at least some representative early signals that can be further investigated in the lab.

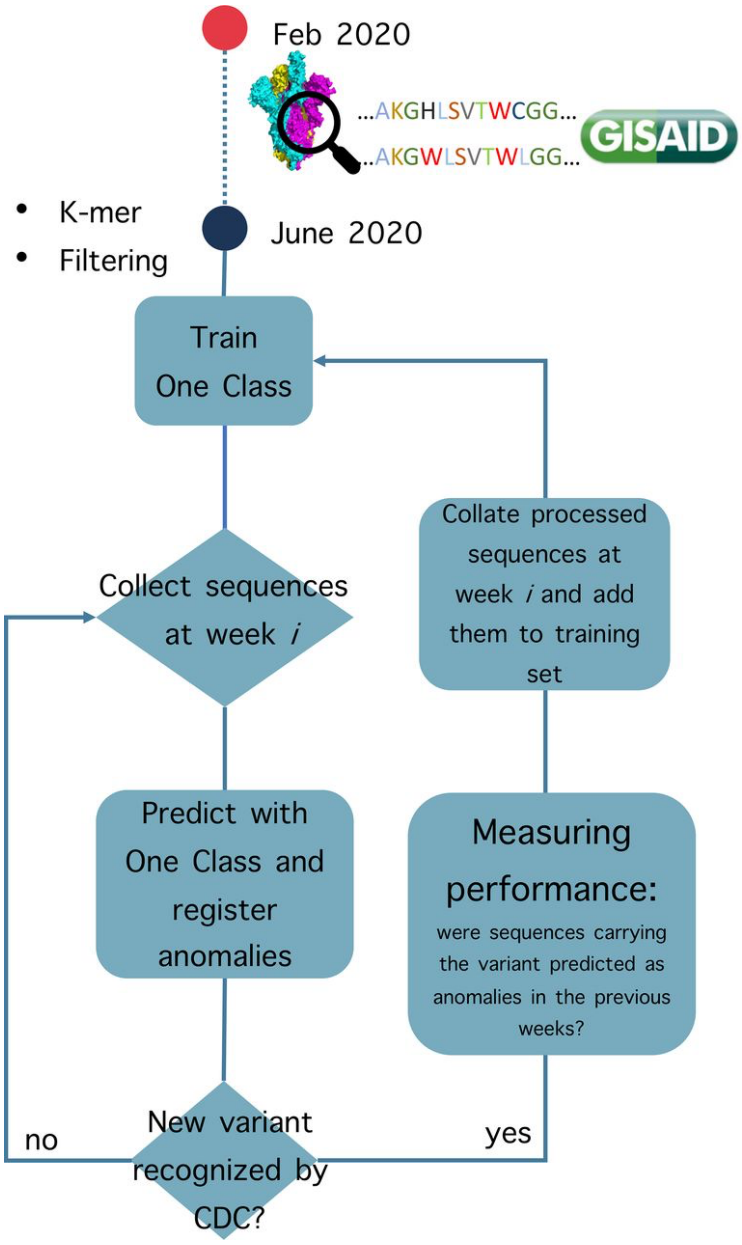


Figure 15: Schematic representation of automatic variant surveillance system. SARS-CoV-2 amino acid sequences deposited in GISAID from February 2020 to June 2020 are collected, transformed into k-mers and filtered. The anomaly detection system (One Class support vector machine (SVM)) is trained on this set of neutral variants. Then, at each following week, the newly uploaded sequences are predicted as either outlier or not. Predicted outliers are registered as anomalies. If authorities have recognised a new variant of concern (VOC)/variants of interest (VOI) in that week, the model is tested by evaluating whether the newly recognised VOC/VOI has already been [47].

The system identified VOCs such as Delta and Lambda up to 20 weeks in advance, and Mu even 21 weeks prior to designation by health authorities, with a median lead time of 10 weeks across all variants. Despite being unsupervised, the algorithm maintained a high precision (>98% after initial ramp-up), thus minimizing false positives and the associated laboratory burden [47]. This study demonstrates how lightweight anomaly detection methods can serve as early warning tools in pandemic settings, rapidly surfacing potentially dangerous variants from large-scale genomic data streams, well before their public health impact becomes evident [47].

Another unsupervised strategy is to cluster co-occurring mutation patterns. Zhuang et al. [48] developed a machine learning pipeline (Figure 16), termed ICA-Var, designed to detect emerging SARS-CoV-2 variants directly from wastewater sequencing data, independently from clinical testing. Their method relies on multivariate independent component analysis (ICA) to extract co-varying and time-evolving mutation patterns from high-dimensional viral genomic signals collected in environmental samples. The ICA-Var method transforms mutation frequency data into statistically independent components,each representing a potential variant-specific mutation pattern. These components are then projected back onto weekly wastewater data using dual-regression, allowing the detection of variants even when present at low frequencies across multiple samples.The results demonstrated that ICA-Var could detect major variants,such as Delta (B.1.617.2), Omicron sublineages (BA.1, BA.2, BA.4, BA.5), and recombinant XBB variants,one to several weeks earlier than Freyja [49]. For example, the EG.5 variant was identified by ICA-Var in early June 2023, several weeks before clinical confirmation [48].

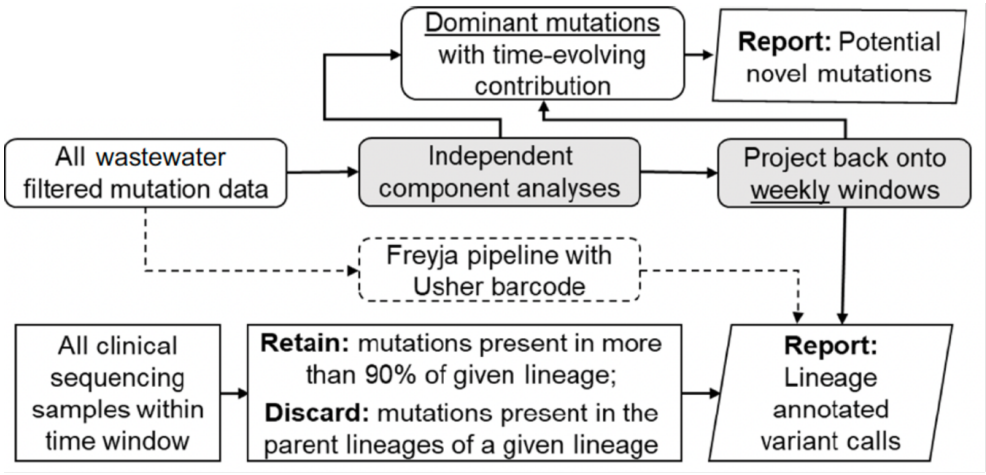


Figure 16: Bioinformatic workflow for real-time variant tracking through wastewater metagenomics and clinical genomes. SARS-CoV-2 single-nucleotide-variant counts obtained from weekly composite wastewater samples (upper left) are subjected to independent component analysis (ICA) to disentangle co-circulating mutation bundles. ICA returns the dominant mutation sets together with their time-varying weights, enabling (i) projection back onto successive 7-day windows to reconstruct variant prevalence curves and (ii) automatic flagging of previously unseen mutations for early-warning reports. In parallel, all patient-derived sequences collected in the same window (lower left) are filtered so that only mutations present in >90 % of a given Pango lineage and absent from its parental lineages are retained; shared or background changes are discarded. These lineage-defining mutation barcodes feed the Freyja [49] deconvolution pipeline (dashed arrows) and the UShER phylogenetic placement tool, which map ICA-derived bundles to specific lineages. The combined output is a weekly list of lineage-annotated variant calls, accompanied by alerts on putative novel mutations, providing a near-real-time picture of viral evolution at population scale [48].

Instead Sapoval et al. [50] introduced QuaID (Quasi-unique mutation-based Anomaly Identification in wastewater), a bioinformatics pipeline specifically designed to enhance the early detection of SARS-CoV-2 VOCs through wastewater genomic surveillance (Figure 17). QuaID operates by identifying and tracking “quasi-unique” mutations,genomic alterations that are highly specific to a single lineage but rare or absent in others,thus enabling more precise variant recognition even when sequencing coverage is sparse or degraded. Unlike traditional approaches that rely solely on single nucleotide variants (SNVs), QuaID incorporates both SNVs and indels (insertions and deletions), significantly improving sensitivity and robustness in complex environmental samples. QuaID was able to detect the Delta variant two weeks before the first clinical

sequence was reported in Texas, and similarly identified the Omicron variant in wastewater a full two weeks prior to its clinical confirmation. Additionally, it consistently detected the BA.2 and BA.5 sublineages of Omicron well in advance of standard methods. In simulated datasets replicating varying degrees of genome degradation, QuaID maintained high detection accuracy even when only 10–25% of the relevant SNV data were retained. Beyond real-time surveillance, QuaID achieved >95% precision across multiple VoCs in synthetic validation experiments involving over 32,000 samples, confirming its reliability and resilience to sequencing artifacts. While QuaID does not estimate variant abundances or perform full phylogenetic placement, its core strength lies in early detection under suboptimal conditions,an increasingly important capability as clinical testing declines. By leveraging the full spectrum of mutational signals, including deletions often ignored by other pipelines, QuaID exemplifies how AI-driven unsupervised analysis of mutational profiles can substantially accelerate and refine variant monitoring at the population level [50].

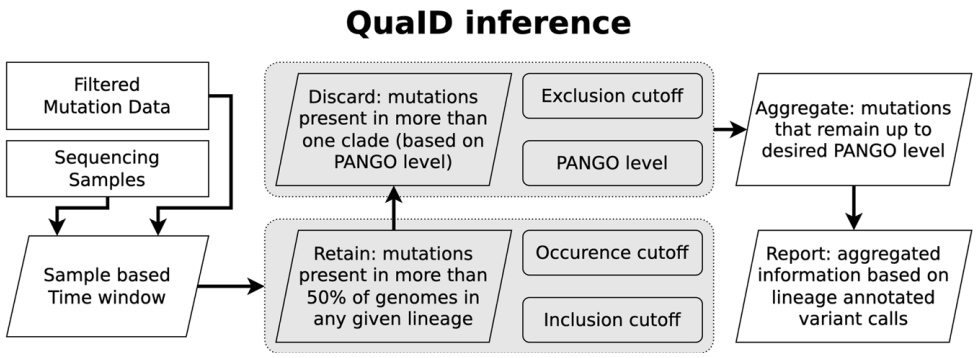


Figure 17: QualID inference workflow for lineage-resolved mutation aggregation in SARS-CoV-2 genomic surveillance. Filtered mutation lists and their corresponding sequencing reads are first grouped into user-defined time windows. For each window, QualID retains mutations found in > 50 % of genomes belonging to any single Pango lineage (occurrence/inclusion cutoff) and discards mutations that occur in multiple higher-order clades at the chosen Pango level (exclusion cutoff). The lineage-specific mutations that remain are then aggregated to the desired taxonomic rank (sub-lineage, lineage, or clade) to produce a concise report of lineage-annotated variant calls, enabling high-specificity, time-resolved tracking of circulating variants [50].

In parallel, supervised learning models have been devised to predict which new viral lineages are “high-risk” – likely to spread widely or evade immunity. HiRisk-Detector [51] exemplifies this approach. It uses a

LightGBM machine learning algorithm trained on haplotype network features of SARS-CoV-2 genomes to computationally rank variant risk [51]. This model, trained on 7.6 million viral genomes, retrospectively detected all 13 known high-risk variants (all WHO-designated VOCs/VOIs up to that time) on average 27 days earlier than the official WHO announcementsn [51]. Even with a quarter of the sequencing data, it only lost ~3.8 days of early warning, demonstrating robustness [51]. HiRisk-Detector also generalized well to sub-lineages of Omicron, achieving high ROC-AUC and PR-AUC in flagging the most transmissible Omicron strains [51]. The key insight was incorporating graph-based features (e.g. node degree, network centrality in a haplotype network) which capture evolutionary relationships and transmission patterns (Figure 18) [51]. These features improved the model's ability to foresee which mutational clusters were gaining a fitness advantage (a proxy for transmissibility) [51]. Beyond these new methods, earlier computational frameworks set the stage. Reference-free approaches like one-class Support Vector Machines on k-mer profiles were used to distinguish VOC/VOI sequences without needing a reference genome [51]. Likewise, convolutional neural networks on dinucleotide composition data attempted to predict viral adaptation and fitness [51]. Reference-based models also emerged: VarEPS combined genomic mutations and structural data in a random forest to assess variant risk, while a "spreading mutations" logistic regression pipeline predicted which amino acid changes would rise in frequency [51]. Another tool, PyR_0, applied a Bayesian model to detect lineages with increased reproductive fitness from genomic surveillance data [51]. These diverse AI approaches – spanning machine learning classifiers, deep learning, and Bayesian inference – all aimed to automate the early warning of concerning variants, reducing reliance on slower human expert designation. By 2022–2023, the convergence of huge global genomic datasets and advanced ML models enabled proactive variant detection systems that often beat conventional surveillance timelines by weeks.

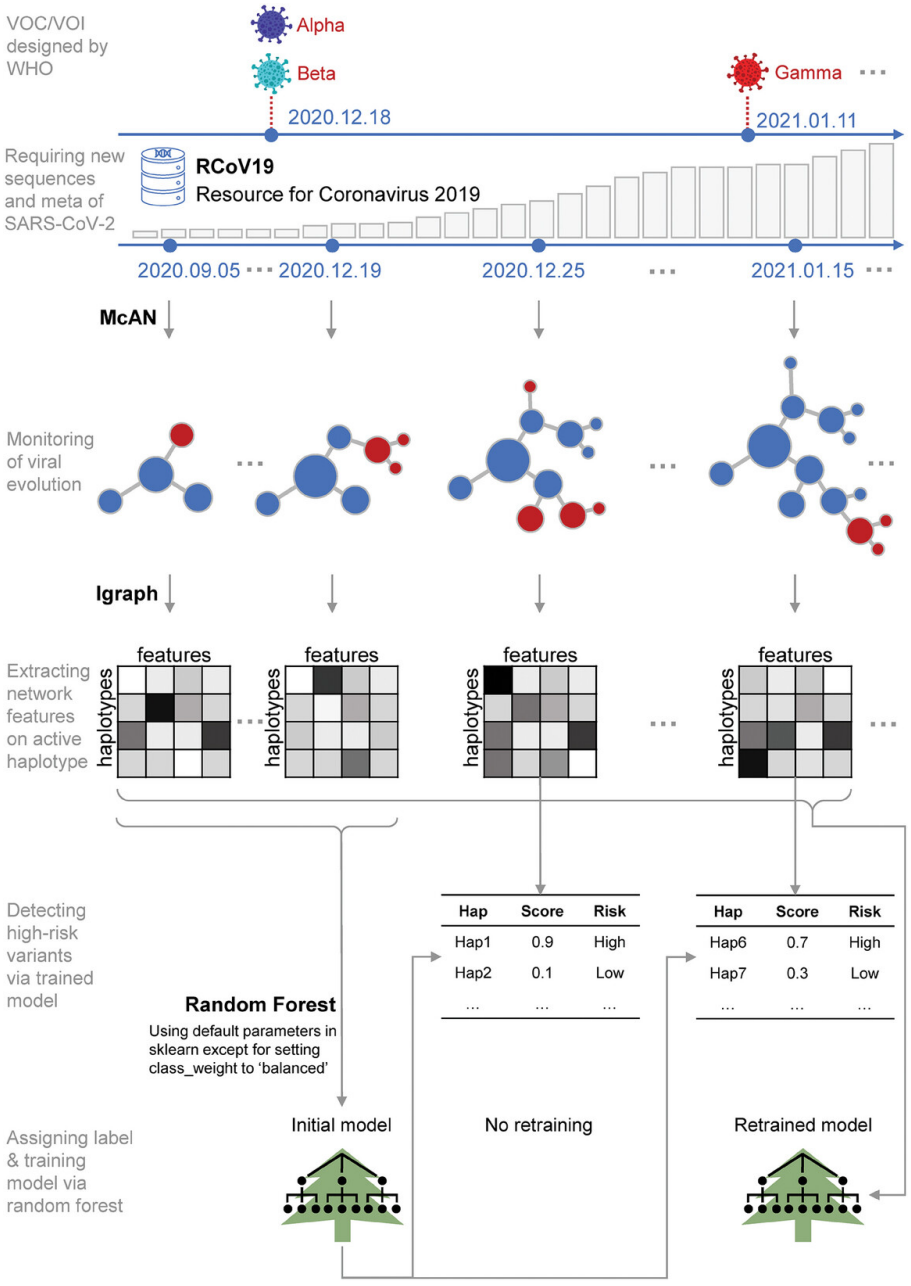


Figure 18: HiRisk-Detector framework for early detection of emerging high-risk SARS-CoV-2 variants using haplotype network features. The workflow integrates daily genomic data from the RCoV19 repository with network-based metrics to monitor viral evolution. Newly sequenced SARS-CoV-2 genomes are processed using McAN to generate haplotype networks, in which nodes represent distinct viral haplotypes (blue = circulating, red = emerging/VOI/VOC). Topological features (e.g., node degree, centrality, clustering) are extracted using igraph and compiled into feature matrices for each time point. These features feed a random forest classifier, initially trained on retrospectively labelled data (e.g., Alpha, Beta, Gamma), which assigns a risk score to each haplotype. High-risk scores trigger early warnings for potentially dangerous variants. The

model can be periodically retrained as new variant outcomes are confirmed, enabling dynamic refinement of predictions over time. This framework supports proactive genomic surveillance by identifying concerning evolutionary patterns before formal designation by public health authorities [51].

AI for Variant Tracking and Phylogenetic Analysis

Once variants are identified, AI and computational models assist in tracking their spread and assessing transmissibility. Phylogenetic methods combined with AI have been crucial to map how variants disseminate geographically and through populations. For instance, large-scale Bayesian phylodynamic models were used to infer the transmission dynamics (basic reproduction number, R_0) of emerging variants directly from genomic data [52]. Geidelberg et al. (2021) and others applied Bayesian inference on sequence phylogenies to estimate that certain variants had higher R_0, quantifying their transmissibility advantages early [52]. These analyses guided public health by indicating which new lineages were likely driving surges.

AI has also been used to integrate genomic data with time and location data for outbreak forecasting. An example is Strainflow (Nagpal S, 2022), which employed language model embeddings of viral genomes coupled with predictive modeling to anticipate case surges caused by variant waves (Figure 19). Strainflow treated genomes like sentences (using codon-level "word" embeddings) and identified latent genomic features that correlate with increased spread. Remarkably, the model prospectively signaled COVID-19 case increases about 2 months ahead of the Delta and Omicron variant surges in multiple countries [53]. For instance, it predicted a coming spike in India as early as the beginning of November 2021, well before Omicron's impact was evident [53]. This demonstrates how combining deep learning feature extraction (unsupervised sequence embeddings capturing viral genetic drift) with supervised learning can yield a powerful early-warning tool for variant-driven outbreaks. The Strainflow study also introduced interpretability by analyzing entropy in the latent feature space, suggesting an "explore–exploit" cycle in viral evolution that precedes large [53]. Such insights help explain model predictions in biological terms.

Phylogenetic clustering and network analysis techniques, sometimes augmented with AI, have further enriched variant tracking. Contact tracing can be augmented by ML-based clustering of genomes to infer likely transmission chains. For example, AI-based outbreak

reconstruction models have been proposed to "finely-detect transmission networks" by linking cases with similar genomic fingerprints [52]. During 2020–2022, health agencies increasingly utilized genomic clustering tools (like Nextstrain's phylogeny) to visualize variant spread; AI now builds on these by automating anomaly detection in phylogenetic trees or predicting future cluster growth. Graph neural networks and network science metrics (as seen in haplotype network models) began to be used to spot superspreading clusters or international spread routes. AI methods – from deep learning predictors like Strainflow to Bayesian phylogenetics and graph-based algorithms – significantly enhanced our ability to track how SARS-CoV-2 variants emerge and move through populations. These tools provided quantifiable measures of variant transmissibility and helped target interventions to regions likely to see the next surge.

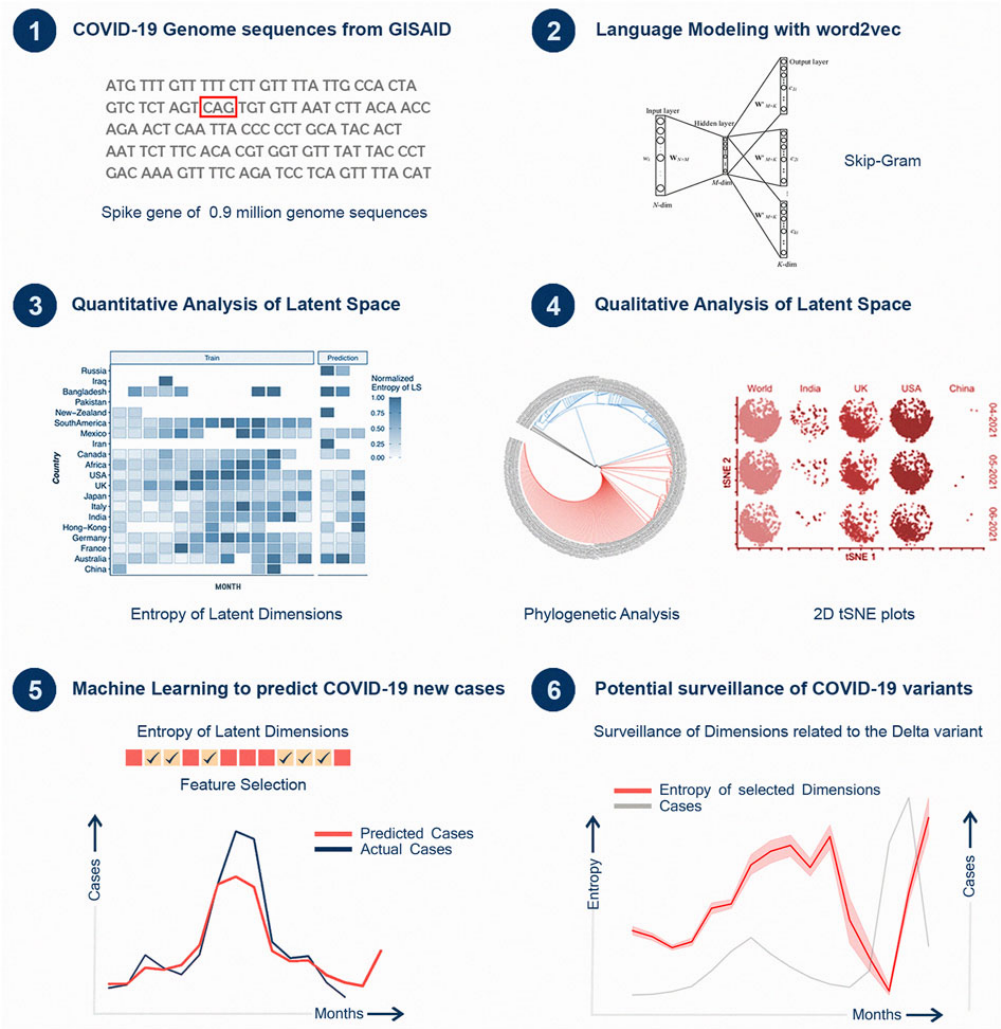


Figure 19: Strainflow framework for tracking and predicting COVID-19 dynamics using latent representations of SARS-CoV-2 genome sequences. (1) Over 900,000 spike gene sequences from the GISAID database are

tokenized into k-mers and (2) embedded into a latent space using a word2vec Skip-Gram language model trained to capture contextual relationships between nucleotide motifs. (3) The quantitative analysis of latent space entropy across dimensions enables spatiotemporal profiling of genomic diversity by country and month, while (4) qualitative projections (e.g., 2D t-SNE plots and phylogenetic trees) help visualize clustering of variants over time. (5) Latent dimensions with high entropy are selected as predictive features in a machine learning model (e.g., random forest or linear regression) to forecast new COVID-19 case counts, aligning well with observed trends. (6) Specific latent dimensions can be retrospectively associated with major variants (e.g., Delta) and monitored in real time, enabling unsupervised genomic surveillance that anticipates outbreaks by tracking entropy changes in relevant regions of sequence-embedding space [53].

AI for Transmissibility and Impact Analysis

Beyond detection and spread, researchers applied AI to understand variant impact, including changes in transmissibility, virulence, and immune escape. One important application has been predicting how mutations affect the virus's ability to evade immunity (from prior infection or vaccines). A landmark study by Hie et al. (2021) used an NLP-based protein language model (Figure 20) to learn the "grammar" of viral evolution and predict which amino acid substitutions would enable immune escape while preserving viral fitness [54]. Trained on sequence data from SARS-CoV-2 and other rapidly evolving viruses (HIV, influenza), this model identified likely escape mutations that traditional methods might miss. Notably, when new variants like Alpha (UK) and Beta (South Africa) arose, the researchers applied their model and flagged specific spike protein mutations in those variants that merited concern for vaccine escape [54]. Their analysis (in early 2021) pinpointed viral genetic changes that should be experimentally monitored for reducing vaccine efficacy [54]. This demonstrated how AI can proactively assess the antigenic risk of emerging variants. Similarly, subsequent models (e.g. EVEscape [55]) have combined evolutionary modeling and machine learning to forecast viral escape potential from historical data [54]. These approaches are really usefull vaccine updates by indicating which new strain might dodge antibody responses.

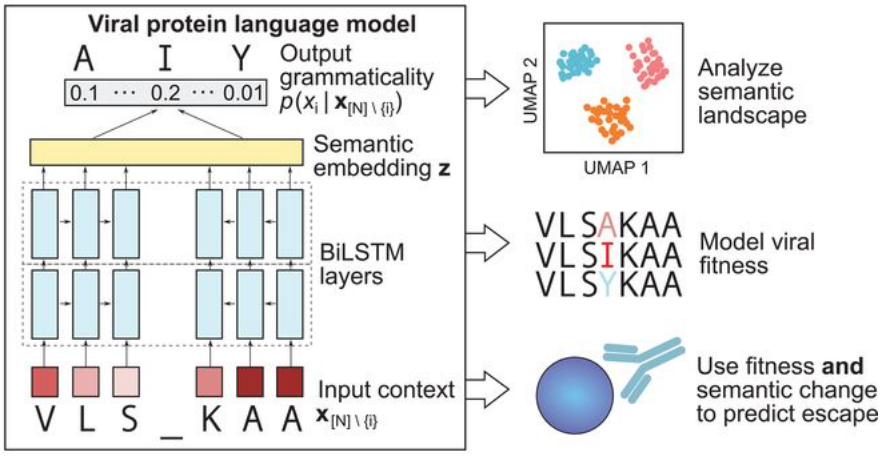


Figure 20: HIV Env model for predicting protein-level mutation effects on fitness and immune escape. The model architecture encodes amino acid sequences using a bidirectional LSTM (BiLSTM) trained to assign grammaticality scores to each residue in context,i.e., the probability of a given amino acid given the surrounding sequence. From this, the model outputs both a semantic embedding vector (z) capturing functional similarity, and a probabilistic score indicating how likely a mutation is to preserve protein structure and function. UMAP projection of semantic embeddings allows visualization of the latent landscape of viral evolution, while systematic substitution of aa residues (e.g., S→A, S→I, S→Y) reveals how sequence variation affects viral fitness. Combining semantic drift with predicted fitness loss or gain enables forecasting of immune escape potential, helping identify mutations that may allow the virus to evade neutralising antibodies while maintaining viability [54].

AI has also been leveraged to study variant virulence and clinical severity. By late 2021, certain mutations (like Spike D614G or N501Y) were suspected to increase transmissibility or severity. ML methods helped confirm and discover such associations. Huang et al. (2022) applied a feature-selection based machine learning approach (Figure 21) to link genomic mutations with COVID-19 severity in patients [56]. They analyzed genomes from patients with varying clinical outcomes, using algorithms (Boruta, LASSO, LightGBM) to identify mutations most correlated with severe disease [56]. Their model rediscovered well-known mutations (e.g. D614G in Spike, which was associated with higher infectivity [56]) and revealed new candidates (mutations in nsp14 like A320V and I164ILV) that had not been previously reported [56]. By building classifiers with these features, they could predict patient severity from the viral genome to some extent [56]. They even generated simple decision rules illustrating how certain mutation combinations distinguish mild vs. severe cases [56]. While host factors usually dominate clinical outcomes, such studies show that viral genomic data can contribute to understanding pathogenicity. In a related vein, supervised ML models have integrated viral lineage data with patient

clinical variables to predict outcomes. For example, a 2022 study developed models to predict hospitalization risk using a combination of patient demographics, comorbidities, and viral lineage-specific genomic signatures [56]. These models found that although clinical features were more predictive, certain viral genomic patterns (notably those in strains just preceding major VOCs) had measurable associations with severe disease [56]. This implies that the genetic makeup of a variant can subtly influence disease severity or transmissibility, and AI can help disentangle those effects by sifting through high-dimensional data.

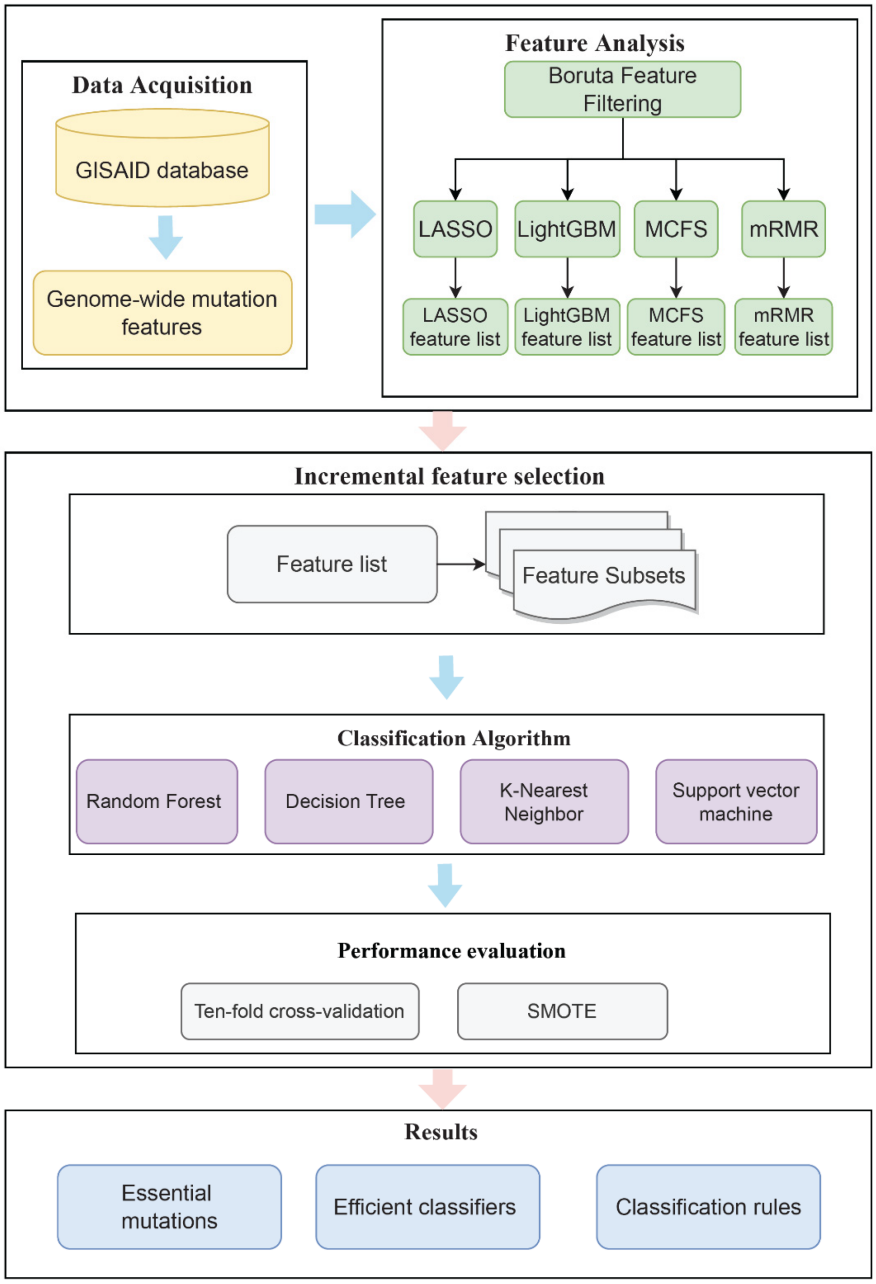


Figure 21: Workflow of the analytical pipeline. Genome-wide mutation features were collected from GISAID and the Coronavirus Antiviral Research Database. Patients were labeled as *mild* or *severe* based on clinical status. Feature selection was performed with Boruta, yielding four feature sets, and features were also prioritized using four independent ranking algorithms. Using Incremental Feature Selection (IFS), we identified the optimal classifier–feature combinations. Finally, decision rules were extracted from the best decision-tree (DT) models to provide interpretable criteria for distinguishing patient clinical status [56].

Another facet of impact is vaccine and therapeutic effectiveness against new variants. AI tools have been used to predict changes in spike protein structure and binding. Some deep learning models, originally developed for protein folding or interaction prediction, were repurposed to assess if a variant's mutations might reduce neutralizing antibody binding or alter ACE2 receptor affinity. For example, deep mutational learning approaches trained neural networks on large-scale experimental screens of RBD mutant libraries to predict how *combinatorial* mutations affect ACE2 binding and escape from monoclonal antibodies, enabling rapid in silico scoring of both circulating and hypothetical variants [57]. Similarly, EVEscape integrates deep learning–based evolutionary constraints with structural/biophysical terms to forecast mutations with high antibody-escape potential while maintaining receptor-binding competence, supporting early-warning prioritization of antigenically risky changes [58]. Complementary deep mutational scanning maps provide experimental "ground truth" on how specific substitutions shift ACE2-binding affinity and antibody escape in VOC backgrounds and are often used to train/validate these predictive models [59]. For instance, random forest models combining genomic and structural features (VarEPS [60]) attempted to rank variants by potential immune escape or transmissibility [51].

AI for Real-Time Policy Support and Decision-Making

While AI aided in tracking the virus's genome, it simultaneously played a pivotal role in translating data into policy action [61, 62, 63] During the pandemic, governments and health agencies worldwide faced critical decisions about interventions (lockdowns, travel restrictions, social distancing mandates), healthcare resource allocation, and vaccine distribution under conditions of deep uncertainty. AI-driven tools – from predictive models to decision-support platforms – were used to analyze trends, forecast future scenarios, and optimize responses, thereby supporting policymakers in making informed, timely decisions.

Epidemic Forecasting and Early Warning

One of the most direct applications of AI in policy support was forecasting COVID-19 trajectories. From national agencies to research consortia, many groups built machine learning models to predict cases, hospitalizations, and deaths weeks into the future, aiming to guide intervention timing and medical preparedness. For instance, the U.S. Centers for Disease Control and Prevention (CDC) leveraged AI to integrate diverse data streams (clinical data, mobility patterns, social media signals, etc.) and track the pandemic's spread in real time [61]. By combining data from electronic health records, web and news trends, and other sources, AI systems at the CDC could detect upticks or shifts in transmission faster than manual analyses [61]. This provided decision-makers with early warnings of regional surges, allowing more proactive resource deployment and public advisories. More generally, experience from COVID-19 highlighted that for predictive models to effectively support public health decisions, they must be both accurate and trusted by officials. Models that demonstrated reliable performance were used to inform decisions like when to tighten or relax social restrictions. For example, many governments convened expert task forces to synthesize outputs from multiple AI-driven models (an ensemble approach) in order to increase confidence in forecasts guiding policy. The pandemic thus underscored both the opportunity and the challenge of AI forecasting: it can greatly enhance situational awareness and planning, but its utility depends on rigorous validation and clear communication of uncertainty [62].

Decision Support for Non-Pharmaceutical Interventions

Beyond passive forecasting, AI was used in a what-if analysis capacity to help policymakers evaluate the potential impact of different intervention strategies. An example comes from a machine-learning powered policy simulator developed for Xiamen, China [63]. Wu et al. built a compartmental epidemic model augmented with ML (Figure 22) that could simulate COVID-19 transmission under various user-defined scenarios (e.g. changing combinations of interventions, varying their timing and intensity, adjusting vaccination rates, or introducing new viral variants) [63].

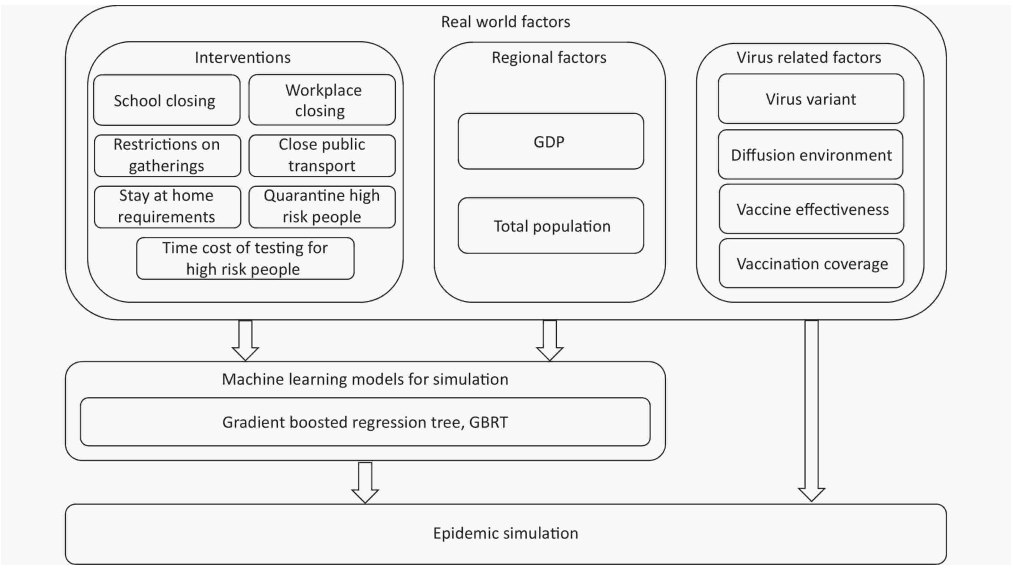


Figure 22: Structure of a machine learning–driven epidemic simulation framework incorporating interventions and real-world factors. The model integrates three main input domains: (1) public health interventions (e.g., school and workplace closures, gathering restrictions, quarantine policies, and testing logistics for high-risk individuals); (2) regional characteristics(such as gross domestic product and total population); and (3) virus-specific features (including circulating variants, diffusion environment, vaccine effectiveness, and vaccination coverage). These heterogeneous inputs are fed into a machine learning model based on gradient boosted regression trees (GBRT) [64], which is trained to simulate and predict epidemic dynamics. The system allows scenario testing by simulating how various combinations of interventions and contextual variables influence infection trajectories, thereby supporting evidence-based policy planning and outbreak response [63].

The system featured a web-based interface where local decision-makers could toggle parameters – such as implementing school closures, travel bans, or mask mandates – and immediately see the projected outcomes in terms of case [63]. This AI-driven tool enabled quantitative comparisons of policy options. In the case of Xiamen's Delta-variant outbreak, the model estimated that without any interventions the city would have seen over 4.6 million infections, whereas under the strictest combined measures the total cases would be reduced to just 78 [63]. When configured to mirror the actual interventions Xiamen implemented, the simulator predicted an outbreak duration of 52 days with a final caseload within a small margin of the real observed values [63]. Such results gave officials confidence in the model's fidelity and in the benefits of aggressive control measures [63]. More broadly, this example shows how AI can act as a "virtual laboratory" for policymakers: by running counterfactual scenarios, leaders can anticipate the epidemiological and societal outcomes of decisions before enacting them.

This approach supports evidence-based policy by highlighting, for instance, how much a partial lockdown might delay the epidemic peak or how combining moderate restrictions could achieve similar benefits as one drastic measure. During COVID-19, countries like New Zealand, Israel, and various U.S. states similarly used AI-enabled scenario modeling to plan phased reopening strategies and to prepare for potential resurgence waves, illustrating a new paradigm of AI-assisted policy simulation in public health planning [63].

Augmenting Contact Tracing and Outbreak Control

AI also contributed to on-the-ground containment efforts which, in turn, informed policy adjustments. A prominent case study is China's early deployment of AI in outbreak hotspots, which demonstrated how technology could amplify traditional public health measures. In Guangdong province in early 2020, authorities applied AI algorithms to vast datasets (including mobile phone location data and demographic information) to trace infection pathways and predict high-risk transmission networks [44]. This AI-enhanced contact tracing reportedly helped break chains of transmission more quickly, contributing to significant reductions in local case growth within the first month of the outbreak [44]. In major Chinese cities, AI-based systems were also integrated into transportation hubs – for example, thermal imaging cameras combined with facial recognition software automatically identified travelers with fevers or those violating quarantine orders [44]. These real-time surveillance tools enabled authorities to isolate suspected cases in crowded places with greater efficiency and speed than manual screening. By curbing undetected spread in train stations and airports, such AI applications supported policy measures to restrict mobility only where necessary, minimizing broader disruption. Furthermore, AI-assisted analytics informed localized policy tightening or relaxation: if the algorithms signaled rising transmission in a particular district, city officials could swiftly impose targeted restrictions there, while avoiding citywide lockdowns. Conversely, when AI monitoring indicated sustained low risk, authorities gained confidence to safely ease certain measures. In essence, the infusion of AI into contact tracing and enforcement allowed for more granular and responsive policy decisions, tailoring interventions to real-time risk levels in different communities.

Resource Allocation and Healthcare Operations

Another vital policy domain during COVID-19 was the allocation of scarce healthcare resources – from ICU beds and ventilators to testing kits and vaccines. AI tools were applied to optimize these logistics under pressure. For instance, artificial intelligence helped in vaccine distribution planning, where the goal was to maximize public health impact under

constraints. An AI model analyzed demographic and health data along with geospatial information to determine the most strategic locations for vaccination sites [61]. By identifying areas with high population density, vulnerable groups, or low existing healthcare access, the model proposed site locations that would ensure equitable and efficient vaccine coverage. This informed policy decisions on where to set up mass vaccination centers and how to allocate doses across regions.

Likewise, hospital systems employed AI-based predictive models to forecast patient loads and supply needs. An example is the ALFABETO project in Italy, which developed a Bayesian Network (BN) model to support triage decisions during COVID-19 patient admissions (Figure 23). The model integrated clinical and radiological features, including comorbidities and deep learning-derived chest X-ray features, to predict whether a patient should be hospitalized or treated at home. Unlike black-box models, the BN framework allowed explicit incorporation of medical knowledge and provided interpretable cause-effect relationships among [65]. This transparency helped align model outputs with clinical reasoning, making it more acceptable for real-world healthcare operations. During both the first and third pandemic waves, the BN showed comparable performance to state-of-the-art ensemble methods like Random Forests (AUC ~0.80), while maintaining interpretability and generalizability across changing patient populations [65].

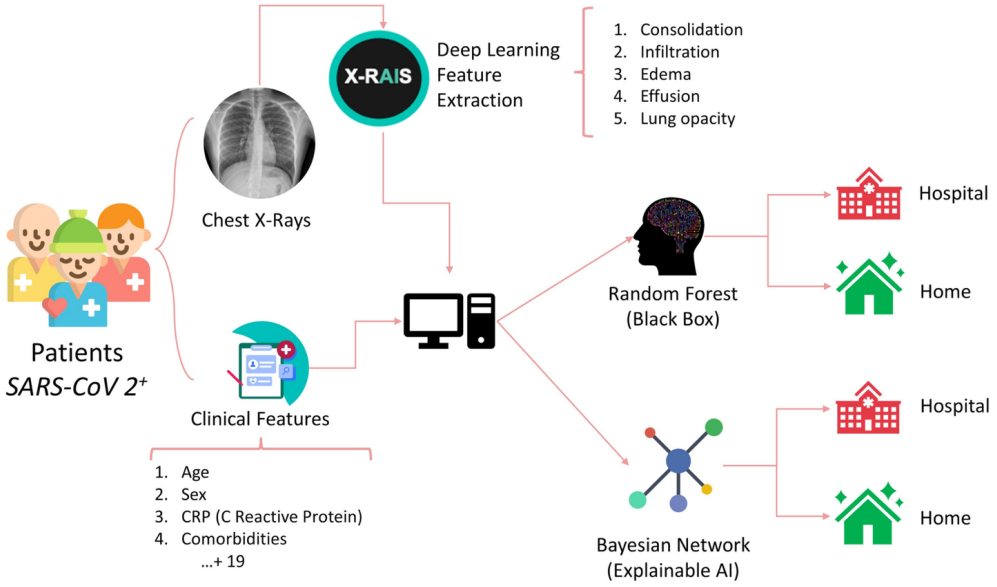


Figure 23: Integrated AI framework for hospitalization prediction in SARS-CoV-2–positive patients. This schematic illustrates a decision-support system combining chest X-ray imaging and clinical features to predict hospitalization needs in COVID-19 patients. Chest radiographs are processed through X-RAIS, a deep learning model extracting radiological features such as consolidation, infiltration, edema, effusion, and lung opacity. In parallel, structured clinical data (e.g., age, sex, CRP levels,

comorbidities) are collected. Both data sources feed into a computational system leveraging two predictive models: a Random Forest classifier (black-box) and an Explainable AI model based on Bayesian networks. The models classify patients as requiring hospitalization or suitable for home care, with the Bayesian approach offering enhanced interpretability to support clinical decision-making [65].

Furthermore, AI models were also used to support dynamic triage and capacity management: real-time data on hospital occupancy and regional case trajectories informed decisions to activate emergency field hospitals or reroute patients to facilities with available capacity. By supporting such data-driven allocation decisions, AI helped policymakers mitigate the pandemic's strain on health systems and supported equitable, timely responses in rapidly evolving scenarios.

1.3. Main Aim of Thesis

The unprecedented global emergency caused by the COVID-19 pandemic has exposed critical gaps in our ability to monitor, understand and anticipate viral evolution in real time. As described in the previous chapters, the rapid succession of SARS-CoV-2 variants,Alpha through Omicron,challenged public-health authorities across Europe, forcing governments to implement non-pharmaceutical interventions under considerable uncertainty. This thesis sets out to develop methodological and computational contributions that respond to these shortcomings.

The central aim of the thesis is four-fold:

- **Quantify the impact of successive SARS-CoV-2 variants in Europe.** Leveraging a complementary suite of analytical approaches,ranging from classical statistical models to state-of-the-art foundation time-series architectures,this thesis assesses how variant-specific prevalence and the depth of genomic surveillance predict key epidemiological indicators (daily confirmed cases, hospital admissions, fatalities) as well as policy responses such as the Oxford Stringency Index across EU Member States. By integrating insights from both methodological families, the studies clarifie the direct epidemiological impact of viral evolution and quantifies the added forecasting value conferred by timely sequencing data.
- **Design an early-warning system for lineage dominance.** Moving beyond retrospective description, we propose an anomaly-detection pipeline that flags lineages exhibiting growth patterns indicative of future dominance. In contrast with earlier efforts that focused

exclusively on VoCs [Predicting emerging SARS- CoV- 2 variants of concern through a One Class dynamic anomaly detection algorithm, Machine Learning Early Detection of SARS-CoV-2 High-Risk Variants, Frontiers | Genomic Surveillance of COVID-19 Variants With Language Models and Machine Learning], the proposed approach models emerging Omicron sub-lineages as statistical anomalies within the high-dimensional space of deep sequence embeddings, thereby enabling proactive intervention before these lineages become dominant in the population.

- **Explore generative models for anticipatory genomics.** Motivated by the recent surge of large generative models, we investigate whether autoregressive and diffusion-based architectures can synthesize biologically plausible spike-protein sequences that anticipate the mutational landscape likely to emerge in the medium term. Generating credible synthetic sequences serves two purposes: it supports scenario planning for future epidemic waves and provides in-silico antigens that can inform adaptive vaccine design,an area still sparsely explored compared with text generation.
- **Develop a cross-species viral embedding.** Finally, we extend the scope beyond SARS-CoV-2 by training a unified embedding model capable of representing the full breadth of viral diversity. This multi-species embedding allows us to address heterogeneous tasks,including viral versus non-viral genome discrimination, DNA/RNA typing, fine-grained SARS-CoV-2 variant classification, influenza segment identification, and HIV-1/2 speciation,within a single, flexible framework.

These objectives converge toward a common goal: to demonstrate how deep-learning techniques, and more in general AI model duly integrated with public-health data streams, can transform genomic surveillance from a reactive tool into a predictive and decision-support engine.

Chapter 2

2 Epidemiological Modeling with AI: mlVAR and TimeGPT

Chapter 2 proposes two pipelines (Figure 24) to examine how European countries managed the balance between protecting public health and maintaining socio-economic stability during the COVID-19 pandemic. The unprecedented pressure on health, economic, and social systems, together with highly heterogeneous policy responses,from generalized lockdowns to mass vaccination campaigns,provides the rationale for the study. The central question concerns which factors determined the adoption and subsequent lifting of containment measures at different stages of the pandemic and the extent to which vaccines and viral variants modulated their impact on clinical outcomes. Specifically, the analysis aims to identify which epidemiological indicators (cases, hospitalizations, deaths) most influenced the tightening of restrictions over time, whether increased vaccination coverage reduced the need for non-pharmaceutical interventions, and how the emergence of new SARS-CoV-2 variants reshaped these relationships. To address these questions, correlational and predictive time-series techniques are applied to weekly data from European countries (2020–2022 and 2020–2023, respectively). First, a multilevel vector autoregression (mlVAR, [66]) model integrates policy-stringency indices (OxCGRT [67]), epidemiological variables (ECDC, [68]) and genomic frequencies (GISAID, [69]) within a single analytical framework, simultaneously estimating population-level effects and country-specific deviations to capture both common patterns and national particularities. Second, mlVAR is complemented with TimeGPT [70], a transformer-based forecasting model used to predict monthly new

cases with hospital admissions, ICU admissions, deaths, and weekly sequencing rates included as exogenous features; this configuration assesses the contribution of sequencing rates to pandemic dynamics relative to other epidemiological variables.

This dual approach enables the joint analysis of temporal associations (mlVAR) and practical predictability (TimeGPT) within the same empirical setting, while evaluating the early-warning value of genomic surveillance. Building on earlier analyses confined to 2020 [71] where the variable consiserated was only epidemiological variabiles (i.e.; New Cases, Deaths, Admission ICU, Admission Hospitalization), the studies [72, 73] incorporate the impact of first, second, and booster vaccination doses, the spread of the Alpha, Delta, and Omicron variants, and sequencing rates across European countries, thereby providing a comprehensive view of the pandemic's evolution. Taken together, this body of work establishes a robust methodological framework for analysing future pandemics and supports public health authorities in identifying the key indicators needed to anticipate and interpret outbreak trajectories.

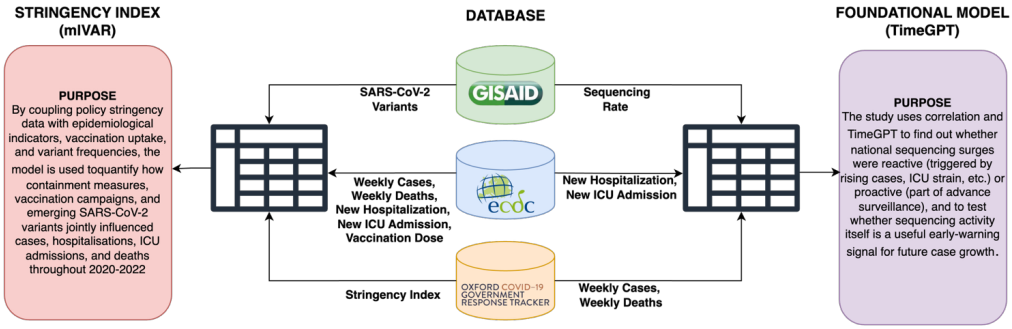
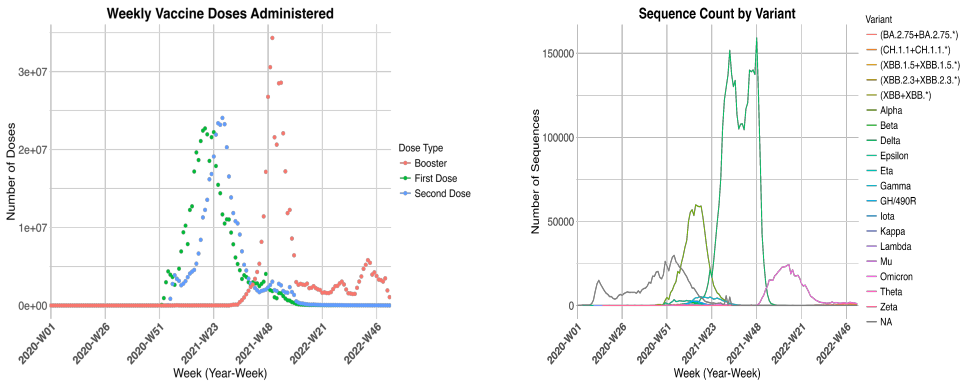


Figure 24: Integrative modeling framework to assess the impact of SARS-CoV-2 variants, containment policies, and sequencing activity during the COVID-19 pandemic. The figure illustrates two complementary study purposes and data workflows. On the left, a modeling pipeline combines data from GISAID (variant frequencies), ECDC (epidemiological indicators), and the Oxford COVID-19 Government Response Tracker (policy stringency, epidemiological indicator) to quantify the joint influence of containment measures, vaccination campaigns, and variant emergence on cases, hospitalizations, ICU admissions, and deaths from 2020 to 2022. On the right, the second study uses correlation analyses and the time-series model TimeGPT to determine whether increases in sequencing activity were reactive (responding to rising clinical pressure) or proactive (part of early surveillance), and whether sequencing itself can serve as an early-warning indicator of future case growth [72, 73, 2].

2.1. Data collection and preprocessing

These studies draw on three primary data sources,the Oxford COVID-19 Government Response Tracker (OxCGRT; Oxford University, 2025), the European Centre for Disease Prevention and Control (ECDC, 2025), and the Global Initiative on Sharing All Influenza Data (GISAID; CCDC, 2025),covering the period from January 2020 to November 2023. Data were compiled for 25 European countries: Austria, Belgium, Bulgaria, Cyprus, Czechia, Estonia, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Liechtenstein, Lithuania, Luxembourg, Malta, the Netherlands, Portugal, Romania, Slovakia, Slovenia, Spain, and Sweden. For the analysis of the impact of variants and vaccination on the Stringency Index (SI), the sample was restricted to Spain, France, the Netherlands, Latvia, Slovenia, Greece, Ireland, Cyprus, and Estonia due to higher data completeness. Preprocessing was performed in RStudio and PyCharm, beginning with filtering all datasets to the study window and partitioning them by country to enable consistent cross-source alignment.



(A) (B)

Figure 25: A) Weekly distribution of COVID-19 vaccine doses administered over time, categorized in first dose, second dose, and booster doses (third and fourth doses). Each colored dot corresponds to a different dose type, highlighting the trends in vaccine uptake and the progression of the vaccination campaign over time from 2020 to 2023. B) Number of sequences uploaded to the GISAID database for each COVID-19 variant from 2020 to 2022. Each colored line represents a different variant, showing the number of protein sequences submitted over time

The Oxford COVID-19 Government Response Tracker (OxCGRT) dataset (Figure 24) documents daily variations in governmental responses to COVID-19 [67], including public health measures (COVID-19 Cases, COVID-19 Deaths), economic policies, and mobility restrictions. A central metric is the Stringency Index (SI) computed as the unweighted average of nine policy indicators,school closures, workplace closures, cancellation of public events, restrictions on public gatherings, public transport closures, stay-at-home requirements, public information campaigns, restrictions on internal movements, and international travel controls,each ranging from 0 to 100, with higher values indicating stricter measures [39]. When policies differed at subnational levels, the SI reflected the strictest regional measure. The SI does not assess policy effectiveness, only their degree of restrictiveness. While other measures such as the Policy Intensity Scores (PISs) offer more granular categorisations [74], the SI was chosen for this study due to its continuous availability through the first months of 2023 and its concise interpretability (more detail on 1.1.4 subsection). Daily SI values were aggregated into calendar weeks (53 weeks for 2020, 52 for subsequent years) to align with the weekly resolution of the other datasets and to reduce noise and irregularities present in daily reporting. From this dataset, in particular for Foundational Model, has been used also the Daily number of COVID-19 Cases and COVID-19 Deaths. Each metric has been aggregated weekly and normalised per 100,000 inhabitants to ensure comparability across countries (Table 2).

The European Centre for Disease Prevention and Control (ECDC) [68] provides weekly epidemiological data (Figure 24), including COVID-19 case incidence, hospital admissions, Intensity Care Unit (ICU) admissions, deaths, and vaccination coverage (first, second, and booster doses (Booster Dose) (Figure 25 A)). In this study, booster doses,covering third and subsequent doses,were combined into a single variable (Table 2). Each metric was normalised per 100,000 inhabitants to ensure comparability across countries. Missing data were addressed through last observation carried forward (LOCF) imputation [75]: missing values were replaced by the most recent non-zero valid entry in the same variable, or with zero when no previous value was available or the last value was zero. This approach was justified by the low percentage of missing data (median 5.7% across countries), mostly concentrated in early 2020 when either no data were recorded or SARS-CoV-2 circulation was absent. Weekly aggregation was applied where necessary to harmonise temporal resolution across datasets, further reducing short-term volatility.

The Global Initiative on Sharing All Influenza Data (GISAID) database [69] was used to incorporate genomic surveillance information (Figure 24), focusing on the Spike protein due to its central role in viral entry and high mutation rate, which drives the emergence of variants with altered transmissibility or immune escape potential. A total of 16,187,950 Spike protein sequences, dated from December 2019 to November 2023, were initially retrieved along with their lineage classifications. Data were filtered to retain only sequences from the selected countries, ex and retained only those whose Spike protein length was between 1241 and 1301 amino acids,i.e., the expected Spike length ±30 amino acids [21]. This filtering produced a curated dataset of 4,506,983 high-quality sequences. For each country and variant, the relative weekly frequency (Figure 25 B) was calculated as the proportion of that variant among all sequences uploaded from that country in the same week. Variants considered included Alpha, Beta, Delta, Gamma, Omicron, and various sublineages such as BA.2.75, CH.1.1, XBB, and XBB.1.5. Missing variant frequencies for a given week and country were imputed as zero, assuming no epidemiologically relevant circulation, an assumption partly supported by the fact that European countries accounted for approximately 48% of all SARS-CoV-2 sequences uploaded globally to GISAID during the study period. On the filtered dataset, we also computed the sequencing rate defined as the number of genome sequences uploaded per week for each country (Table 2).

Table 2: Summary of datasets, variables, and preprocessing steps used in the analysis. The table outlines the key data sources,OxCGRT, ECDC, and GISAID,together with the extracted variables and the corresponding preprocessing pipelines, including temporal filtering, country-level aggregation, normalization per 100 000 population, imputation strategies, and dataset merging procedures.

DATASET	VARIABLES	PREPROCESS
OxCGRT	Stringency Index(2020-2022); Weekly COVID-19 Cases (2020-2023); Weekly COVID-19 Deaths (2020-2023).	Period filter; split by country; daily → weekly; normalization per 100k population; aggregation; merge with other datasets on week-ID

ECDC	Weekly COVID-19 Cases (2020-2022); Weekly COVID-19 Deaths (2020-2022); New Hospitalization (2020-2022); New ICU Admission (2020-2022); Vaccination Dose (2020-2022).	Period filter; split by country; normalization per 100k population; LOCF + zero imputation; weekly aggregation if needed
GISAID	SARS-CoV-2 Variants Frequency (2020-2022); Sequencing rate (2020-2023).	Period filter; remove sequences with missing AAs or out-of-range length; select target countries; compute weekly variant frequencies; normalize sequencing volume per 100k population; zero imputation for absent variants

2.2. Dataset Structure

Upon completion of preprocessing, a dataset comprising all features listed in Table 2 was obtained. For each analysis (Stringency Index and Sequencing Rate), a dedicated subset (Stringency Index Dataset and Sequencing Rate Dataset) was derived, each aligned to its respective study period.

Stringency Index Dataset

The dataset employed to study the effect of Variants and vaccines on SI integrates heterogeneous data sources encompassing governmental policy measures (OxCGRT), epidemiological indicators and vaccination coverage (European Centre for Disease Prevention and Control (ECDC)), and SARS-CoV-2 genomic variant frequencies (GISAID). All data cover nine European countries,Spain, France, the Netherlands, Latvia, Slovenia, Greece, Ireland, Cyprus, and Estonia,over the period from January 2020 to December 2022. For each year of the analysis we created a data subset containing the following features: weekly Stringency Index (SI weekly epidemiological indicators, including new COVID-19 cases (New Cases, NC), hospital admissions (admHosp, aH), intensive care unit (ICU) admissions (admICU, aI), deaths, and vaccination coverage (first dose (First Dose, FD), second dose (Second Dose), and booster doses (Booster Dose)), weekly frequencies of SARS-CoV-2 variants, with the set of tracked variants evolving over time: in 2020, Alpha (Al), Beta (Be), Delta (De), and Zeta (Ze) were monitored. In 2021, the scope expanded to 13 variants, including Omicron (Om); and in 2022, the focus shifted to Omicron sublineages such as BA.2.75, CH.1.1, and XBB (Table 3, Table 4).

Table 3: Example a datset used to the evaluate the effect of SARS-CoV-2 Variants and the Vaccine Dose on SI during 2020

SI	NC	aH	aI	deaths	FD	SD	BD	Alpha	Beta	Delta	Zeta	Country	Week
11.1	3	3.40	0.08	0.05	120.6	0	0	90	10	0	0	France	45

Table 4: SARS-CoV-2 variants included in the dataset for each simulation year. The table presents the variants considered in the models for 2020, 2021, and 2022, highlighting the increase in variant diversity in 2021 and the subsequent reduction in 2022, with the emergence of new subvariants such as BA.2.75, CH.1.1, XBB, and XBB.1.5

Year	Variants Included

2020	Alpha (Al), Beta (Be), Delta (De), Zeta (Ze)
2021	Alpha (Al), Beta (Be), Delta (De), Zeta (Ze), Theta (Th), Gamma (Ga), Epsilon (Ep), Eta (Et), Kappa (Ka), Iota (Io), Mu (Mu), Lambda (La), GH490R (GH), Omicron (Om)
2022	Delta (De), GH490R (GH), Omicron (Om), Alpha (Al), BA.2.75, CH.1.1, XBB, XBB.1.5

Sequencing Rate Dataset

The dataset used to study how sequencing intensity relates to pandemic dynamics integrates epidemiological (OxCGRT, ECDC) and genomic (GISAID) sources at a weekly resolution, spanning January 2020–November 2023 across 25 European countries: Austria, Belgium, Bulgaria, Cyprus, Czechia, Estonia, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Liechtenstein, Lithuania, Luxembourg, Malta, the Netherlands, Portugal, Romania, Slovakia, Slovenia, Spain, and Sweden. For each year of analisys and for each nation we have a relative subdataset (Table 5) with the follow features: From OxCGRT, we extracted daily records of confirmed cases and deaths and aggregated them to weeks; from the ECDC, we obtained weekly indicators,new cases (Weekly Cases), hospital admissions (Weekly Hospitalization), Deaths (Weekly Deaths), and ICU admissions (Weekly ICU Admission); and from GISAID, we collected the number of SARS-CoV-2 genomes uploaded each week per country to compute weekly sequencing counts (and the corresponding sequencing rate (Sequencing Rate)). All series were harmonised to a common weekly calendar and merged by country.

Table 5: Example of dataset used to evaluate the effect of Sequencing Rate on pandemic dynamics in generic country

Sequencing Rate	Weekly Cases	Weekly Hospitalization	Weekly ICU Admission	Weekly Deaths
20	3	3.40	0.08	0.05

2.3. Analysis of the role of containment policies, vaccination strategies and virus variants in the COVID-19 pandemic

As define above the aim of this study is to provide a global overview of how the public health policy makers reacted to the spreading of the virus, both at the initial stage of the pandemic and in the subsequent phases, characterized by the introduction of large vaccination campaigns and the birth of new, often more infectious, variants in nine European Countries. After the creation of dataset, the analysis uses a multilevel vector auto-regressive mixed-effects model (mlVAR), considering each country as a member of a population and including them as random effects in the model, in order to take into account both the general relationships among the different variables in the entire population and the local effects for each individual country. This approach allows modeling how the complex interactions among different aspects of the pandemic evolved in time, using epidemiological and socio-political data that we collected from a variety of diverse sources.

Multilevel Autoregression Model

Correlation analysis

To assess potential relationships between variables (Table 3, Table 4, Subchapter 2.2), the correlation between the SI and each epidemiological and genetic variable has been assessed for each year analyzed (i.e.; from 2020 to 2022) using Pearson's linear correlation coefficient:

$$\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \quad (1)$$

where X and Y are two numerical variables, $Cov(X,Y)$ indicates their covariance and σ indicates the variance. The score can range from -1 to

1. A positive value indicates that an increase in one variable is associated with an increase in the other variable, while a negative score indicates that the two variables grow in opposite directions. 0 indicates an absence of correlation. The results for the 2020 anayisis are presented in Table 6.

Table 6: The contemporaneous bivariate correlations between each index and the weekly epidemiological and genetic variables for 2020 were analyzed across all countries. Significant results ($p < 0.05$) are highlighted in bold. Correlations were calculated by combining data from all available countries into a single vector for each variable

	NC	aH	aI	Deaths	FD	SD	BD	Alpha	Beta	Delta	Zeta
SI	**0.33**	**0.41**	**0.42**	**0.48**	**0.11**	7e-5	0.08	**0.27**	0.07	0.08	0.03

Table 6 shows that in 2020 SI was significantly and positively correlated with all epidemiological variables, especially with admissions to hospitals and ICU, and deaths, which have the highest correlations scores. Concerning the other variables, among all variants, only the Alpha has a significant, although weak, correlation, while the others, which appeared at the end of the year, appear not to be correlated to policy intensity. First vaccine doses, which began towards the end of the year, also have a significant correlation, although the low score indicates a substantial absence of association. The correlations between each index and the weekly epidemiological and genetic variables for 2021 and 2022 are reported in Table A1 and Table A2, respectively.

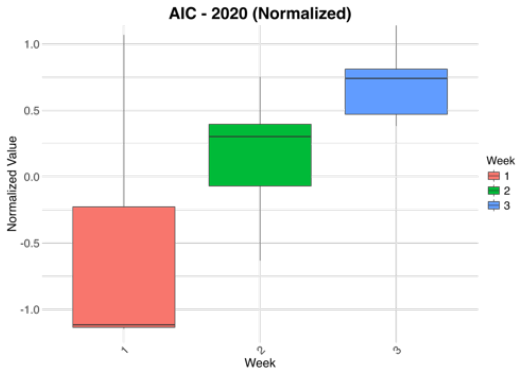
Population Model Design

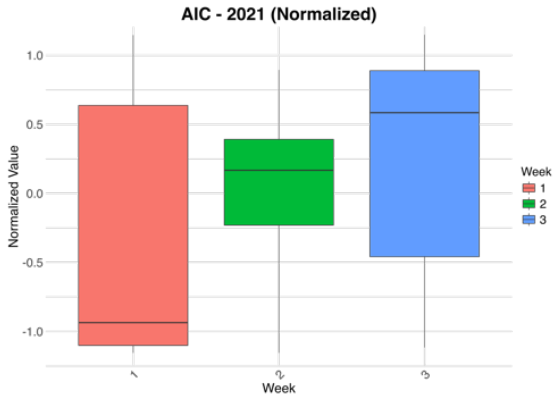
Building on the correlation analysis that highlighted systematic relationships between the SI and epidemiological and genomic covariates, a multi-country autoregressive framework was developed to quantify how lagged and contemporaneous conditions relate to policy stringency. The modelling strategy adopts a one-model-per-year design (2020, 2021, 2022) to accommodate regime shifts in data availability and viral dynamics (e.g., the emergence of new variants and vaccine rollout) while preserving cross-country comparability [71, 76]. Autoregressive models inherently incorporate time lags to capture temporal dependencies within the data. This allows us to assess how past values influence current trends. The optimal lag length was determined by evaluating candidate lag orders and

selecting the specification that minimized the Akaike Information Criterion (AIC) (Figure 26); the AIC is defined in Equation 2:

$$AIC = 2k - 2ln(L) \quad (2)$$

where k represents the number of parameters estimated by the model, and L is the maximum likelihood of the model.





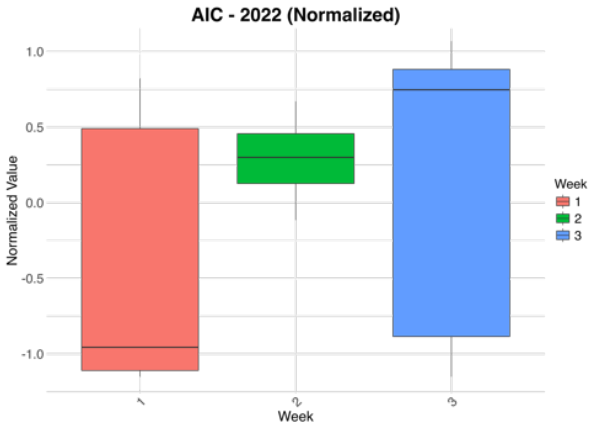


Figure 26: Normalized AIC for the model applied to all three years with a lag of one, two and three weeks. The model performs better with a lag of one week in 2020 (Kruskall-Wallis P-Value = 0.008). No significant differences are visible for 2021 and 2022, although the AIC is generally still lower with a one-week delay [72]. The larger AIC dispersion observed in 2021–2022 likely reflects increased cross-country heterogeneity and measurement noise driven by vaccination rollout and variant turnover, which partially decoupled incidence from severe outcomes and led to more differentiated policy responses and reporting practices across countries.

For each analysis year, an mlVAR model was estimated to integrate policy, epidemiological, and genomic time series across countries. The mlVAR framework is suited to populations of related units, capturing shared structure while accommodating unit-specific heterogeneity [77, 78]. Countries are treated as units drawn from a common population. Population-level (fixed) effects capture relationships that are stable across Europe, whereas country-specific (random) effects model deviations reflecting national idiosyncrasies,such as differences in policy timing, healthcare capacity, or compliance. This hierarchical (partial-pooling) structure increases statistical efficiency and out-of-sample stability by borrowing strength across countries without imposing a single uniform dynamic. By jointly modeling contemporaneous interactions and lagged feedback among the Stringency Index (SI), epidemiological indicators, and variant frequencies, mlVAR provides a flexible, consistently estimated representation of pandemic dynamics that balances generalizability with country-level nuance. In implementation, each of the nine selected European countries was treated as an individual unit within the broader population framework. Population-wide effects are represented as fixed effects (FE), i.e., parameters held constant across countries. Equation 3 illustrates the fixed-effects specification using the Stringency Index as the dependent variable for the 2020 analysis.

$$\mathrm{I}(t) = \beta_{01} + \beta_{11} \cdot I(t-1) + \beta_{21} \cdot NewCases(t-1) + \beta_{31} \cdot admHosp(t-1) + \beta_{41} \cdot admICU\,(t-1) + \beta_{51} \cdot deaths(t-1) + \beta_{61} \cdot FirstDose(t-1) + \beta_{71} \cdot SecondDose(t-1) + \beta_{81} \cdot BoosterDose(t-1) + \beta_{91} \cdot Alpha(t-1) + \beta_{101} \cdot Beta(t-1) + \beta_{111} \cdot Delta(t-1) + \beta_{121} \cdot Zeta(t-1) \quad (3)$$

On the other hand, the Random Effects (RE) capture country-specific deviations from the population model. For each coefficient β_j, the country-level parameter β_{ij} is modeled as a Gaussian perturbation around

the population average, with variance quantifying cross-country heterogeneity (Equation 4):

$$\beta_{i,j} \sim N\left(\beta_{j,pop}, \sigma_{\beta_{j,pop}}\right) \quad (4)$$

In statistical terms, an autoregressive model is a linear stochastic difference equation whose output depends on past values augmented by a random innovation, a standard formulation for time-dependent phenomena. The multilevel vector autoregression was estimated in R (RStudio) using the mlVAR [79] package for the VAR structure and lme4 [80] to encode the hierarchical (random-effects) component. The estimation dataset (Tables 3–4) comprises the policy Stringency Index (SI); four epidemiological indicators,new cases, hospital admissions (admHosp), ICU admissions (admICU), and deaths; three vaccination covariates,first-, second-, and booster-dose coverage; year-specific variant-frequency columns tracking circulating SARS-CoV-2 lineages (Figure 25, Table 6); and a country identifier linking observations to national panels. For each simulation year, nine country-specific panels were assembled to ensure that country-level patterns and trends were explicitly captured. The primary outputs are the estimated regression coefficients of the autoregressive system. As an illustrative case, Equation 5 presents the SI equation for 2020, in which the conditional term denotes the random effects, indicating that the contribution varies by country ID. By adopting this hierarchical specification, population-level relationships common across countries and country-specific deviations shaping epidemiological and policy dynamics are jointly recovered, while preserving the temporal dependencies fundamental to autoregressive modelling.

$$\begin{aligned} I(t) \sim I(t-1) + NewCases(t-1) + admHosp(t-1) + admICU\ (t-1) \\ + death(t-1) + NewCases(t) + admHosp(t) + admICU\ (t) + \\ death(t) + FirstDose(t-1) + SecondDose(t-1) + BoosterDose(t- \\ 1) + Alpha(t-1) + Beta(t-1) + Delta(t-1) + Zeta(t-1) + \\ NewCases(t) + admHosp(t) + admICU\ (t) + death(t) + FirstDose(t) + \\ SecondDose(t) + BoosterDose(t) + Alpha(t) + Beta(t) + Delta(t) + \\ Zeta(t) + ((1|ID) + (I(t-1)|ID) + (NewCases(t-1)|ID) + (admHosp(t \\ - 1)|ID) + (admICU\ (t-1)|ID) + (death(t-1)|ID) + (FirstDose(t- \\ 1)|ID) + (SecondDose(t-1)|ID) + (BoosterDose(t-1)|ID) + (Alpha(t \\ - 1)|ID) + (Beta(t-1)|ID) + (Delta(t-1)|ID) + (Zeta(t-1)|ID)) \end{aligned}$$

(5)

Results

Lag-length selection

A one-week lag was adopted for the mlVAR specification following targeted analyses and methodological considerations. This choice is supported by epidemiological theory and empirical model selection. Although the incubation period of SARS-CoV-2 spans approximately 2–14 days [81, 82] the effects of non-pharmaceutical interventions (NPIs) on observable indicators (e.g., case and hospitalization rates) typically emerge within about one week, reflecting symptom onset, testing/reporting cycles, and weekly surveillance rhythms. Empirically, measurable shifts within this horizon have been documented: case growth declined within 1–5 days after stringent distancing orders in U.S. counties [83]; statewide mask mandates reduced daily case growth during the first 1–20 days post-implementation [84]; large-scale anti-contagion policies produced significant reductions in infection growth within the "first week" across six countries [85]; and early community-mitigation packages yielded inflections in case trends within days [86]. Hospitalization growth rates fell by ~3 percentage points in the first two weeks following mask mandates [87].

One-, two-, and three-week lags were compared using AIC (Figure 26). While the improvement from one to two weeks reached statistical significance only in 2020, the one-week lag achieved the best or near-best AIC across all three years, indicating adequacy for detecting policy effects on the analyzed outcomes. Adopting a uniform one-week lag also ensures comparability across years and continuity with previous work [71]. Taken together, these theoretical and empirical considerations justify the selection of a one-week lag in the mlVAR analyses.

Fixed effects

Fixed effects quantify average (population-level) relations among variables across countries. Year-specific fixed effects are estimated to reflect shifting conditions across 2020–2022, including vaccination rollout and the emergence of new variants.

Results of the 2020 model

In 2020, prior to vaccine availability, the ancestral lineage and the Alpha variant accounted for most infections in Europe; additional variants emerged toward the end of the year, when initial vaccinations also began. From the significant entries (Figure 27) a strong persistence of the Stringency Index (SI) is observed, with SI at week tt largely determined by its value at t–1, and positively associated with hospital admissions at t–1. Epidemiological indicators exhibit temporal coupling: increases in new cases are followed one week later by higher hospital admissions, ICU admissions, and deaths; higher hospital admissions are followed by increased ICU admissions; and deaths tend to rise after increases in hospital admissions. Conversely, increases in deaths are associated with a subsequent decrease in hospital occupancy after one week. Vaccination variables show some significant coefficients; however, these should be interpreted cautiously because vaccination began only at year's end, limiting the temporal coverage required to establish robust trends.

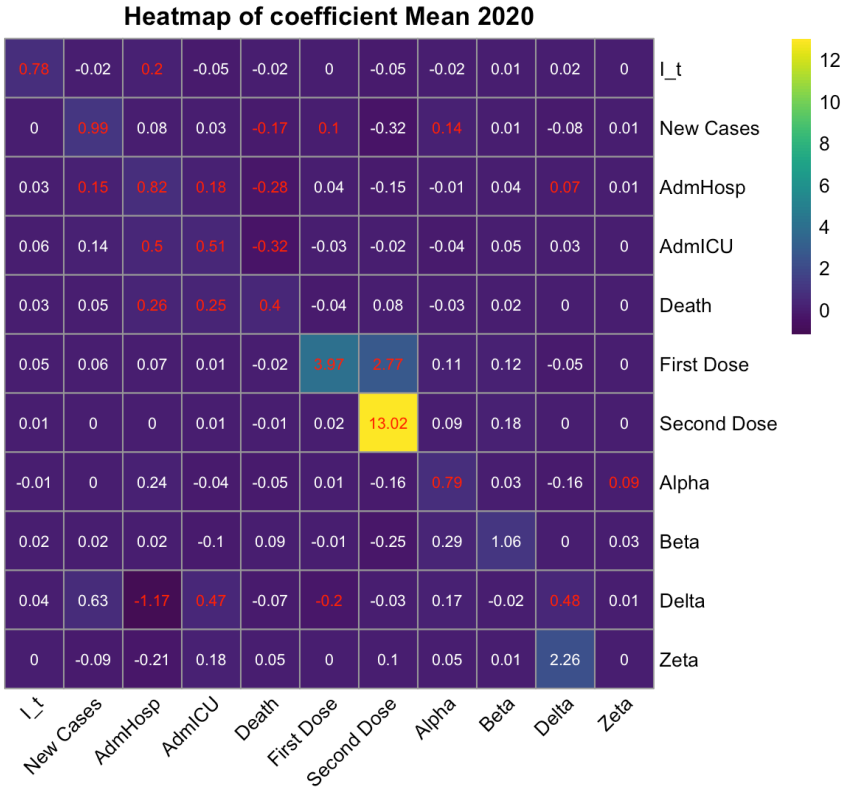


Figure 27: Heatmap of the fixed effects' coefficients for the year 2020. The significant ones (P-Value < 0.05) are highlighted in red. The rows indicate the variables at time t, i.e. the outcomes, whereas the columns indicate the variables at time t – 1, i.e. the covariates. For instance, in the first row, I_t is the outcome variable, while each column (I_t, New Cases, AdmHosp, AdmICU, Death, First Dose, Second Dose, Alpha, Beta, Delta, Zeta) represents a covariate used by mlVAR to predict this outcome [72].

Results of the 2021 model

The central year of the pandemic was highly eventful, with multiple infection waves and emerging variants coinciding with progressive vaccination uptake (Figure 25). From the significant coefficients (highlighted in red (Figure 28, Figure A1)), several patterns emerge. SI tends to increase following rises in new cases and hospital admissions,the latter remaining the strongest epidemiological driver by magnitude,whereas ICU admissions do not display the same positive association. Vaccination coverage is associated with reductions across epidemiological outcomes: first doses are linked to decreases in new cases and hospital admissions; booster doses to decreases in hospital and ICU admissions; and second doses to decreases in all three variables (cases, hospital, ICU). The SI–ICU coefficient contrasts with the broader pattern, suggesting that higher ICU occupancy is associated with lower subsequent stringency; this finding may reflect the clinical lag from infection to ICU admission, with policy responses being more sensitive to leading indicators such as new cases and general admissions.Variant dynamics are consistent with competitive replacement: the Alpha signal diminishes as Delta increases, reflecting Delta's rapid rise to prevalence in 2021. Moreover, Omicron is associated with a marked increase in new cases, as indicated by a positive, statistically significant, and comparatively large coefficient (Figure 28).

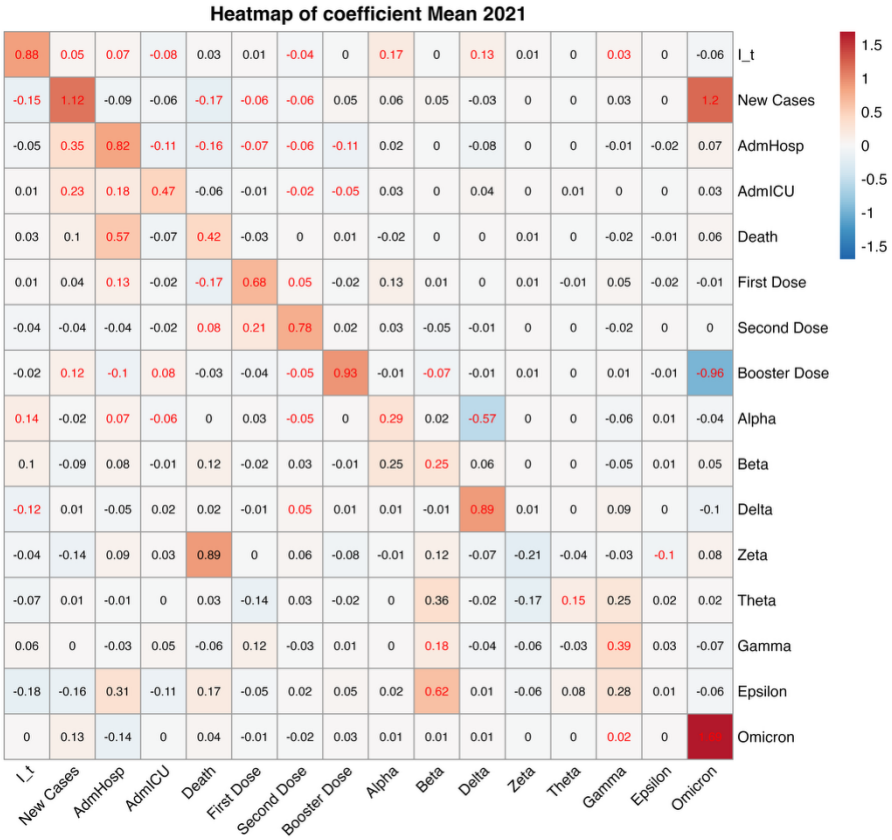


Figura 28: Heatmap of the relevant fixed effects' coefficients for the year 2021. The significant ones (P-Value < 0.05) are highlighted in red. The rows indicate the variables at time t, i.e. the outcomes, whereas the columns indicate the variables at time t – 1, i.e. the covariates. For instance, in the second row, New Cases is the outcome variable, while each column (I_t, New Cases, AdmHosp, AdmICU, Death, First Dose, Second Dose, Alpha, Beta, Delta, Zeta, Theta, Gamma, Epsilon,Omicron) represents a covariate used by mlVAR to predict this outcome [72].

Results of the 2022 model

During 2022, restrictions were progressively relaxed worldwide. With rising population immunity,driven by vaccination uptake and prior infections,and the spread of the Omicron lineage, which exhibited lower severity despite high transmissibility [88], hospitalizations and deaths displayed a downward trend. In this context, the Stringency Index shows no robust association with the epidemiological indicators. Residual temporal dependence is still detectable among the epidemiological variables themselves, and several coefficients involving variant covariates attain statistical significance; however, because these covariates are constructed from relative sequencing frequencies (and thus are sensitive

to sequencing coverage), such associations are not interpreted as substantively meaningful (Figure 29).

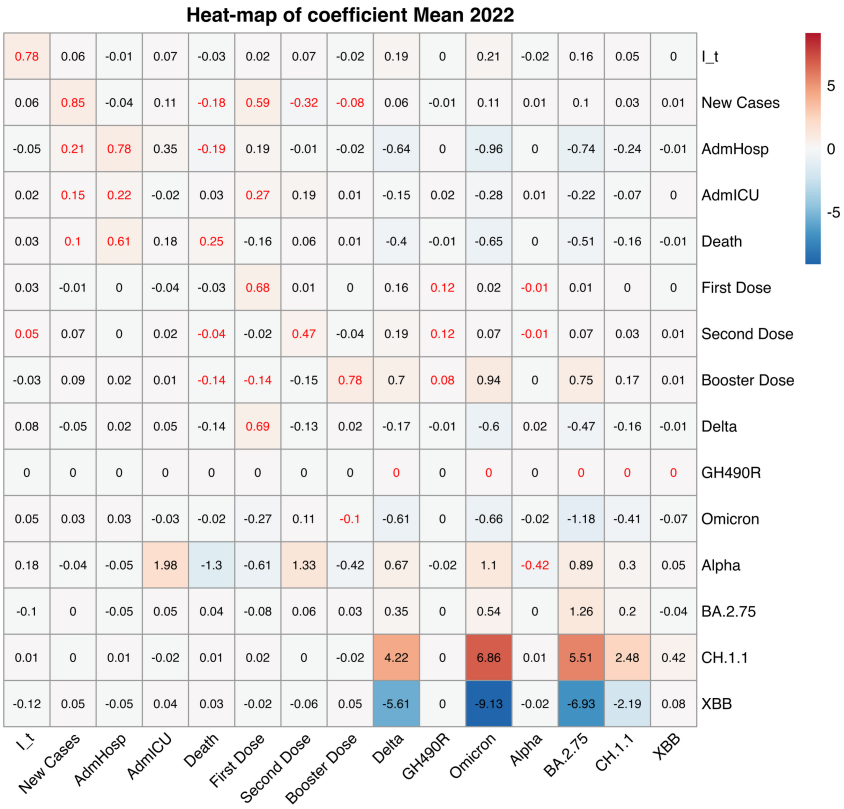


Figure 29: Heatmap of the fixed effects' coefficients for the year 2022. The significant ones (P-Value < 0.05) are highlighted in red. The rows indicate the variables at time t, i.e. the outcomes, whereas the columns indicate the variables at time t – 1, i.e. the covariates. As an example, in the Third row, I_t is the outcome variable, while each column (I_t, New Cases, AdmHosp, AdmICU, Death, First Dose, Second Dose, Booster Dose, Delta, GH490R, Omicron, Alpha, BA.2.75, CH.1.1, XBB) represents a covariate used by mlVAR to predict this outcome [72].

Random Effects

While fixed effects capture population-level relations among variables, random effects quantify country-specific deviations in the model coefficients, thereby accommodating heterogeneity attributable to socio-demographic and institutional contexts. Considered alongside the fixed effects, several patterns emerge. In 2020, SI is significantly associated only with its own lag and with hospital admissions at t–1 (Figure 30); the random effects indicate that the hospitalization influence is stronger in Ireland and attenuated in Spain, Slovenia, Latvia, and Cyprus. In 2021

(Figure A2), the hospitalization effect appears largely homogeneous across countries, whereas other covariates display greater dispersion; for example, new cases exert a comparatively larger impact on SI in the Netherlands than in Spain or Ireland. In 2022 (Figure A3), country-specific differences persist, although the only significant fixed effect concerns the lag of SI; this autoregressive effect is highest in Latvia, Slovenia, and Cyprus, and lowest in Spain.

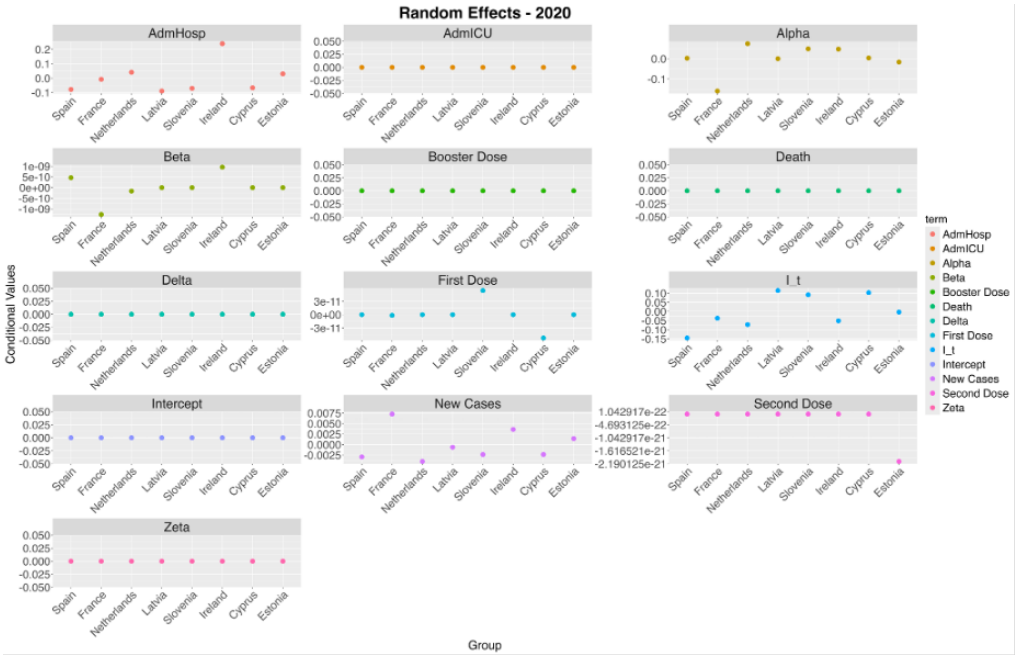


Figure 30: Random effects (country-wise variation on the model intercept) for all predictors with stringency index as outcome in 2020 [72].

Temporal progression of the coefficients

A useful representation of how the general mutual impacts between the epidemiological variables and the policies varied in all the considered countries can be obtained by plotting the value progressions of the fixed effects coefficients in the three years of the pandemic. Figure 31 depicts these trajectories with the stringency index as the dependent variable. Hospital admissions (Figure 31) exhibits the strongest association with the stringency index in 2020 and 2021, particularly in the early phase of 2020, when more severe variants predominated and vaccines were unavailable. In 2021, as vaccination coverage increased and variants with higher transmissibility but comparatively lower hospitalization rates circulated, the influence of hospital admissions on policy stringency diminished, while new cases gained importance. The Second Dose coefficient is

negative in 2020 and 2021, indicating that increases in second-dose administrations are followed by policy relaxation. The coefficient rises in 2022; however, the heatmap indicates that this estimate is not significant, which is consistent with the high and stable second-dose coverage during that year. Considering the stringency index as a covariate rather than an outcome reveals additional interdependencies. Figure A4 reports the time-varying coefficients with new cases as the dependent variable. An increase in stringency does not appear to reduce new cases in 2020,consistent with the heatmap results,but becomes more effective in 2021, whereas the 2022 estimate is not significant. Vaccination effects display a threshold-like pattern: reductions in new cases emerge when coverage is sufficiently high, with negative and significant coefficients for first and second doses in 2021 and for the booster in 2022. The booster nonetheless significantly reduces hospital admissions already in 2021 (Figure 28). The relatively large 2022 coefficients for first and second doses are not considered reliable, given that coverage for these doses was already high and nearly constant.

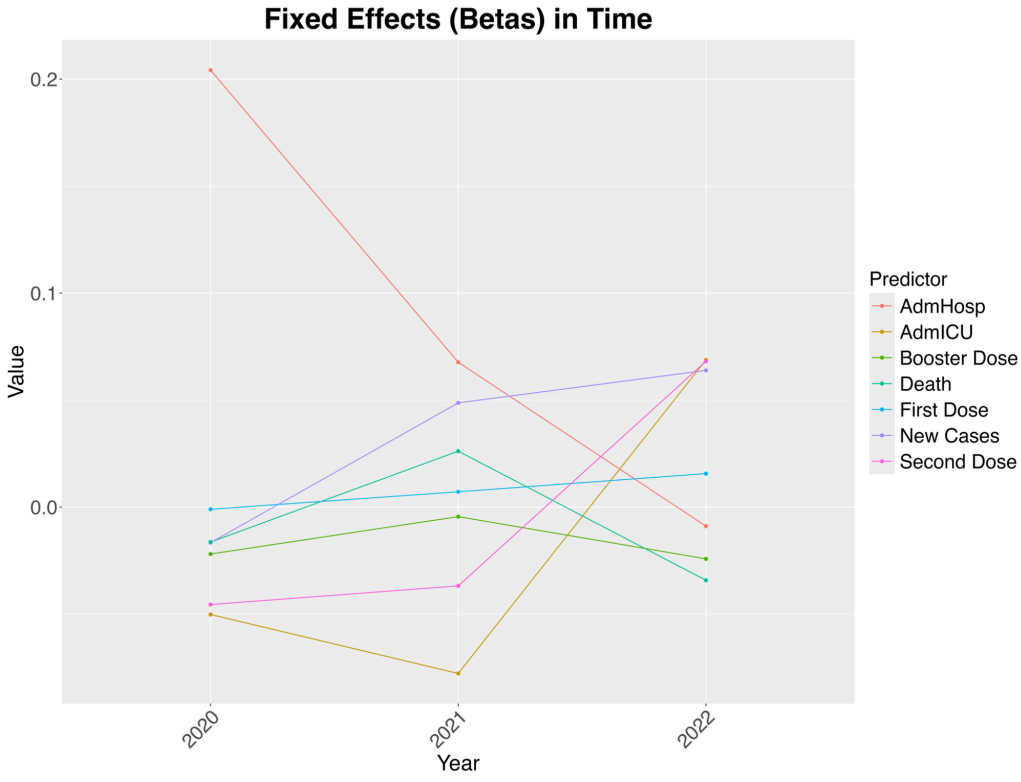


Figure 31: Variations of the fixed effects coefficients indicating the effects of the epidemiological variables on the stringency index in the three considered years [72].

2.4. Sequencing Intensity and Epidemiological Trends in Europe's COVID-19 Pandemic

This analysis examines the relationship between SARS-CoV-2 sequencing rates across European nations, as reported in public databases (e.g., GISAID), and key epidemiological parameters, including cases, deaths, and ICU admissions. The objective is to assess whether sequencing efforts were primarily reactive,triggered by surges in cases or ICU strain,or proactive, driven by broader public-health strategies. Statistical correction methods (Figure 32), hypothesis testing, and transformer-based time-series models such as TimeGPT [89] are employed to characterize the dynamics between sequencing capacity and epidemiological trends in Europe. These insights clarify the decision-making processes underlying sequencing efforts and delineate their role in managing the pandemic's evolving landscape.

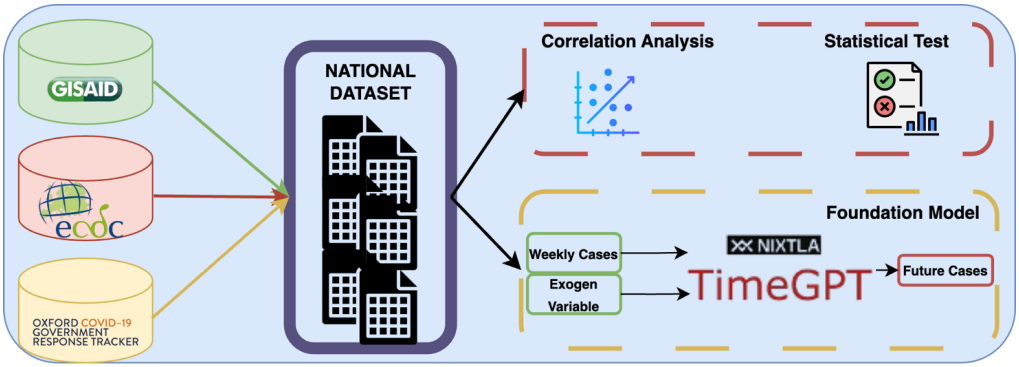


Figure 32: Analytical workflow integrating correlation analysis, statistical testing, and foundation time-series modeling to investigate COVID-19 dynamics. Data from GISAID (variant frequencies), ECDC (epidemiological indicators), and the Oxford COVID-19 Government Response Tracker (policy stringency) are harmonized into a unified national dataset. This dataset is first analyzed using correlation methods to identify relationships between epidemiological trends and exogenous variables, followed by statistical tests to assess their significance. In parallel, weekly case counts and exogenous variables are provided as inputs to the Nixtla TimeGPT to forecast future COVID-19 case trajectories [72, 73, 2].

Correlation Analysis

To investigate the relationships between weekly sequencing data and various epidemiological indicators, a correlation analysis (Figure 32) was

performed using Pearson correlation coefficients, as expressed in Equation 1. Associations between weekly sequencing rates and four key metrics,Weekly Hospitalizations, Weekly ICU Admissions, Weekly Deaths, and Weekly Cases,were analyzed. The analysis was carried out both over the entire three-year study period (January 2020 to November 2023) and on a yearly basis. The three-year analysis assessed the overall correlation for each country across the full period, while the year-level analysis examined temporal variations in correlation strength, allowing the identification of changes in these relationships for each year of the pandemic. This approach enabled a comprehensive assessment of how correlations evolved over time, providing insights into the dynamics between sequencing efforts and epidemiological trends.

Time series forecasting with TimeGPT

To investigate relationships between weekly sequencing data and epidemiological trends, TimeGPT [89],a transformer-based time-series forecasting model that leverages self-attention to weight the importance of time points,was employed. The self-attention mechanism enables the capture of long-range dependencies and complex temporal patterns, enhancing forecast accuracy. The model uses an encoder–decoder architecture: the encoder transforms historical series into latent representations that summarize trends and seasonality, while the decoder combines these representations with positional information for future time steps to generate forecasts; a final linear layer produces the output [8]. Positional encoding within the encoder preserves the sequential order of observations and thus their temporal relationships.

TimeGPT is well suited to epidemic time series because COVID-19 incidence is typically non-stationary, characterized by multi-wave dynamics, abrupt regime shifts (e.g., interventions, variant turnover), and heteroskedasticity driven by changes in testing and reporting. Unlike classical statistical forecasters [90, 91], which often assume more stable seasonal structure and require dataset-specific model identification, TimeGPT leverages transformer attention to model nonlinear dependencies **and** long-range temporal effects without hand-crafted seasonality terms. In addition, compared with training task-specific deep models from scratch on a single country series, TimeGPT is foundation-model–like: it is pre-trained on large collections of time series and can be applied in a zero-/few-shot manner, which is advantageous when country-level histories are short or disrupted. Finally, TimeGPT differs

from mechanistic compartmental models (e.g., SEIR [92]) by not imposing epidemiological assumptions; instead, it provides a data-driven benchmark that can directly test the incremental predictive value of exogenous signals such as sequencing rates. For these reasons, TimeGPT was applied to predict monthly new COVID-19 cases for each country. For each forecast, the upcoming month was predicted from the preceding ten weeks (input window = 10). To quantify the incremental contribution of covariates, four exogenous variables were considered individually: weekly hospitalizations, weekly ICU admissions, weekly deaths, and weekly sequencing rates. Countries lacking any of these variables due to missingness were excluded from the corresponding analysis. Model performance was evaluated using the root mean squared error (RMSE) [93], as defined in Equation 6, where n is the sample size, x_i the observed values, and $\widehat{x_i}$ the predicted values. Comparing RMSE across scenarios identified the exogenous variable yielding the lowest error, thereby indicating the most influential epidemiological indicator for forecasting monthly case counts.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{x_i})^2} \quad (6)$$

Results

Three Year Correlation

Sequencing trends appeared to be influenced by different factors across countries. In 14 countries, sequencing was closely related to ICU admissions (Austria, Belgium, Bulgaria, Cyprus, Estonia, Germany, Ireland, Lithuania, Luxembourg, Netherlands, Romania, Slovakia, Slovenia, Spain). In 10 countries, significant correlations were observed with hospital admissions (Bulgaria, Cyprus, Estonia, France, Liechtenstein, Luxembourg, Netherlands, Romania, Slovakia, Slovenia). Additional associations emerged with case numbers in five countries (Belgium, Bulgaria, Luxembourg, Portugal, Romania) and with deaths in 11 countries (Bulgaria, Cyprus, Estonia, Germany, Greece, Italy, Luxembourg, Netherlands, Romania, Slovakia, Slovenia). Several countries exhibited multiple simultaneous relationships. Belgium (Figure 33A) showed significant, albeit weak, correlations with ICU admissions (ρ = 0.30, p = 7.0×10^{-5}) and cases (ρ = 0.18, p = 0.009). Bulgaria (Figure

33B) displayed moderate correlations with hospitalizations (ρ = 0.52, p = 2.12×10^{-10}), ICU admissions (ρ = 0.53, p = 1.0×10^{-11}), and deaths (ρ = 0.55, p = 2.2×10^{-16}), and a weak correlation with cases (ρ = 0.35, p = 2.27×10^{-5}). A similar pattern was observed in Luxembourg, where sequencing correlated with all parameters, particularly ICU admissions (ρ = 0.45, p = 2.21×10^{-9}) and deaths (ρ = 0.48, p = 7.71×10^{-9}). Countries with higher population density, such as France, showed significant correlations with hospital admissions (ρ = –0.20, p = 0.006). Germany exhibited significant correlations with ICU admissions (ρ = 0.48, p = 3.32×10^{-11}) and deaths (ρ = 0.19, p = 0.01).

A B

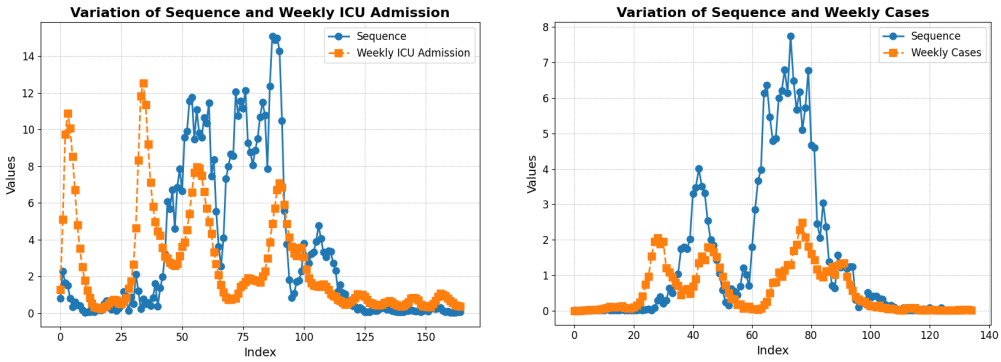


Figure 33: A) Relationship between Weekly Sequencing and Weekly Admission in ICU in Belgium. B) Relationship between Weekly Sequencing and Weekly cases in Bulgaria [73].

Years Wise Correlation

National responses varied over time and according to epidemiological conditions. In some countries,France being a representative example,Weekly Sequencing Rates were significantly correlated with most analyzed variables, including Weekly ICU Admissions, Weekly Cases, and Weekly Deaths. This was particularly evident in 2020 and 2022, when Weekly Sequencing Rates showed significant associations with cases and deaths in 48% and 40% of countries, respectively. Across the study period, sequencing consistently demonstrated significant correlations with cases and deaths in many countries, supporting its utility as an indicator. Germany, for instance, showed a moderate correlation between Weekly Sequencing Rates and Weekly Cases in 2020

(ρ = 0.45, p = 0.001) and 2021 (ρ = 0.54, p = 2.5×10^{-5}), strengthening to a strong correlation in 2022 (ρ = 0.74, p = 1.27×10^{-10}) and 2023 (ρ = 0.92, p = 1.8×10^{-9}). Conversely, Ireland showed no significant correlation between these rates and the analyzed variables in several years (Figure 34). A plausible explanation is that Ireland's sequencing activity may have been driven more by operational policy (e.g., fixed throughput, targeted outbreak/travel investigations, or batching of samples) than by proportional sampling of incident cases. Such targeted sampling and variable reporting delays can decouple weekly sequencing counts from contemporaneous case, ICU, and death dynamics, attenuating rank correlations even when sequencing remains epidemiologically valuable.
A tendency toward smaller p-values was observed in 2021 and 2023 across multiple countries, although smaller nations such as Belgium and Luxembourg continued to display significant correlations for all variables even in 2023.

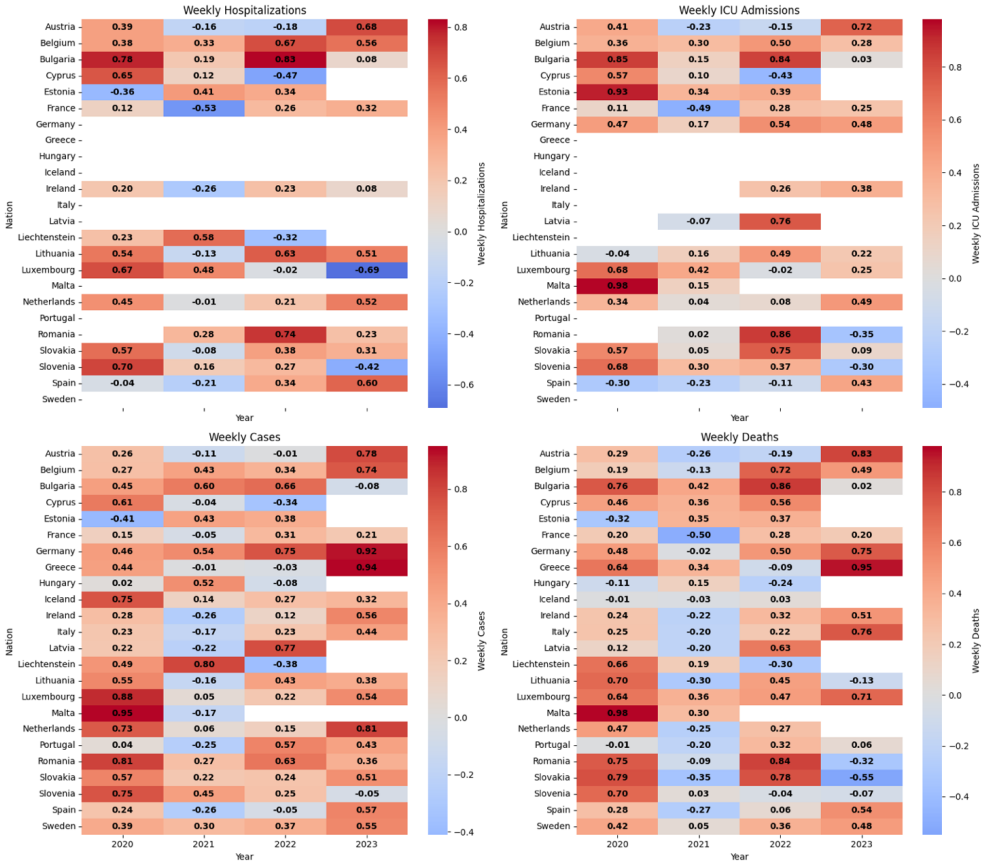


Figure 34: HeatMaps showing the correlations between weekly sequencing rates and four key metrics: Weekly Hospitalizations (Hospital_Corr), Weekly ICU Admissions (ICU_Corr), Weekly Cases

(Case_Corr) and Weekly Deaths (Deaths_Corr). Blank spaces indicate that data was not available to compute the correlation [73].

TimeGPT Analysis

Across countries, the most informative exogenous regressor for improving TimeGPT forecasts of next-month cases was the Weekly Sequencing Rate, identified in 54% of nations (notably Austria, Belgium, Bulgaria, Cyprus, Germany, Ireland, Luxembourg, and Slovakia). By comparison, Weekly Deaths was most informative in 6% of countries, Weekly ICU Admissions in 20%, and Weekly Hospitalizations in the remaining 20%. Predictive accuracy varied markedly across settings, with a median RMSE of 78.84 cases per 100,000 inhabitants (IQR: 78.84) across all nations (Figure 35). The best performance was observed in Ireland and Bulgaria, with RMSEs of 17 and 26 cases per 100,000 inhabitants, respectively (Figure 36–37). Additional correlation results are reported in Figure 34.

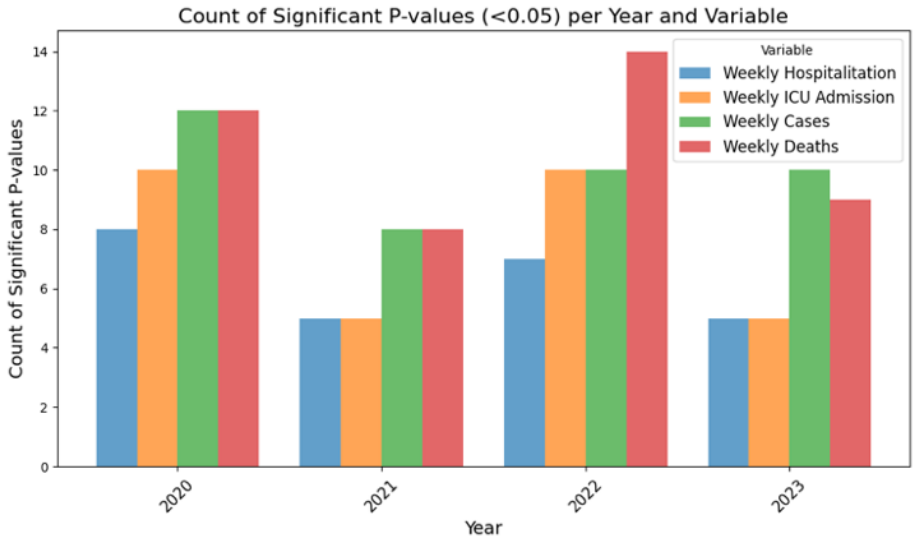


Figure 35: Counting of significant P-values in different years of Analysis (2020-2023) [73]

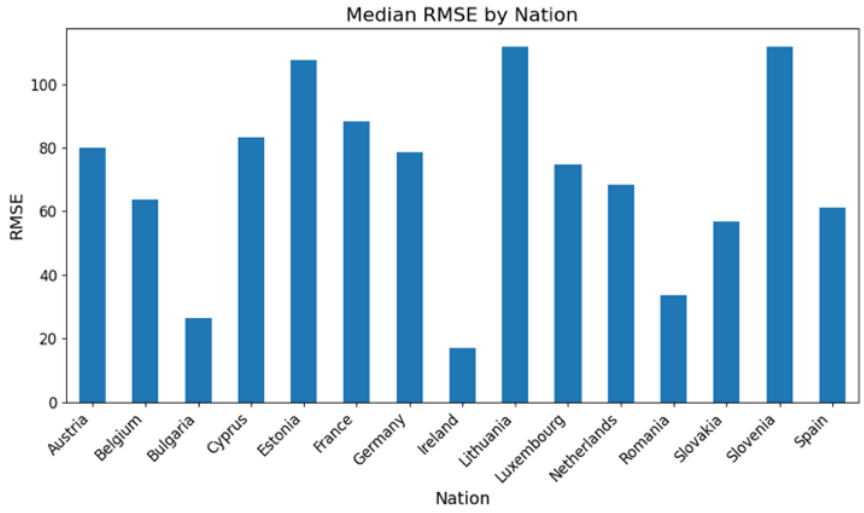


Figure 36: RMSE obtained by TimeGPT in predicting Weekly cases every month [73]

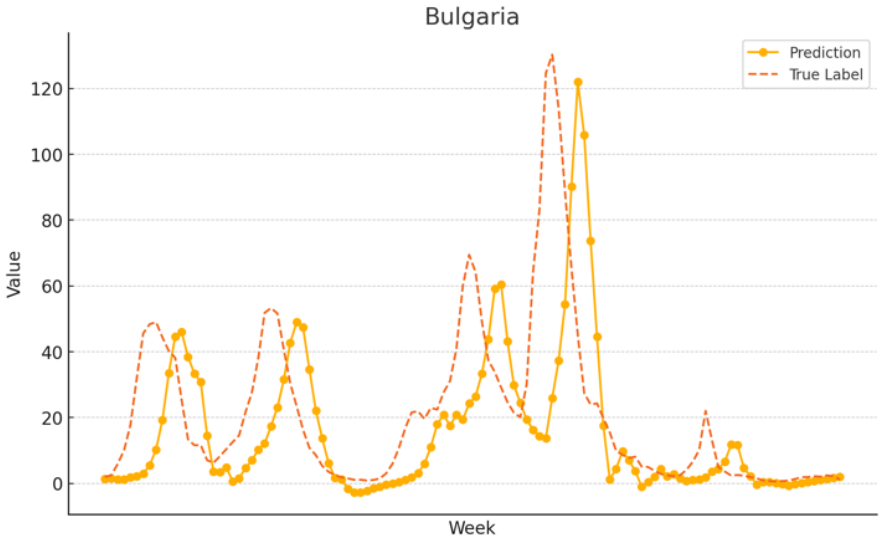


Figure 37: Weekly trends of predictions vs true values for Bulgaria (2020-2023) [73]

2.5. General Conclusions: Linking Genomic Surveillance and Policy via TimeGPT and mlVAR (Europe, 2020–2022)

The COVID-19 pandemic, driven by a highly contagious and initially unknown virus, precipitated a three-year global crisis that compelled governments to implement containment policies to curb social

interactions, travel, and economic activities. In the early stages, policy responses were largely reactive to rapidly changing public health conditions rather than grounded in data-driven evaluation, and they entailed substantial socioeconomic and mental-health costs. Identifying policy portfolios that achieve maximal epidemiological impact with minimal collateral damage is therefore essential for future preparedness. While prior studies [94, 71], focused mainly on the initial phase and often ignored the evolving variant landscape and the progressive roll-out of vaccination, the analysis proposes a comprehensive, multi-year, multi-country framework that jointly analyzes epidemiological indicators, vaccination coverage, viral variants, and policy stringency. The framework explicitly models population-level effects alongside country-specific deviations, acknowledging that epidemiological and sociopolitical processes are mutually influential and geographically heterogeneous. Two complementary analyses are integrated.

First, a vector autoregressive mixed-effects model (mlVAR) was fit to weekly data from nine European countries (Spain, France, the Netherlands, Latvia, Slovenia, Greece, Ireland, Cyprus, and Estonia) spanning 2020–2022. Second, an orthogonal analysis evaluated genomic sequencing effort as both a correlational indicator and a predictive exogenous feature,via TimeGPT,across a broader European panel, with attention to cross-country and year-to-year variability. The mlVAR results reveal marked temporal shifts yet considerable cross-country consonance in the relationships among variables. In 2020, when severity and uncertainty were highest, hospitalizations were the primary epidemiological correlate of policy stringency, consistent with decision-makers reacting to acute health-system pressure. This association was strongest in Ireland and relatively attenuated in Slovenia, Latvia, Spain, and Cyprus. In the short one-week horizon examined, higher stringency did not translate into a statistically significant reduction in hospitalizations, though longer lags might capture delayed policy effects. In 2021, as testing availability expanded and more infectious variants circulated, new cases emerged as an additional driver of stringency; this relationship was weakest in Ireland and Spain and comparatively stronger in countries such as the Netherlands. Unlike 2020, a significant reverse association appeared in 2021: higher stringency was followed by a measurable reduction in new cases. Vaccination coverage was consistently associated with decreases in new cases and hospitalizations, with dose-specific dynamics: first doses were most consequential in 2021, while second and booster doses showed stronger effects in 2022, suggesting that benefits become more apparent once meaningful

population coverage is achieved. Notably, increases in second-dose coverage in 2021 were the only vaccination changes robustly associated with reductions in stringency across all countries. Regarding variants, Delta (2021) stood out as the only lineage significantly associated with more stringent policy responses, whereas Omicron's emergence (2022) produced sharp increases in cases without commensurate rises in hospitalizations or stringency in the countries studied.

The sequencing-focused analysis complements these findings by treating viral genomic surveillance both as a window onto molecular evolution and as a population-level indicator that can track,and in some settings anticipate,pandemic dynamics. Over a three-year horizon, sequencing trends correlated with different epidemiological endpoints depending on country context: in 14 countries they aligned most closely with ICU admissions; in 10 others with hospital admissions; and elsewhere with cases or deaths. These patterns likely reflect differences in health-system capacity, surveillance strategies, and population density. Belgium and Bulgaria (Figure 33) exhibited multi-endpoint relationships (ICU, hospitalizations, deaths), while France (Figure 34) showed a negative correlation with hospital admissions suggestive of proactive surveillance coincident with efforts to mitigate hospital burden; Luxembourg maintained moderate correlations across all endpoints, and Germany displayed particularly strong alignment with ICU admissions and deaths. Year-specific analyses further clarified the dynamics: correlations were widely significant in 2020,when many countries scaled sequencing aggressively,and again in 2022 amid Omicron's rapid expansion and concerns about immune escape [67]. In contrast, significance generally weakened in 2021 and 2023, plausibly reflecting the onset and maturation of vaccination coverage and a corresponding reduction or refocusing of routine sequencing in many settings. Smaller countries such as Belgium and Luxembourg sustained significant, multi-endpoint correlations even in 2023 (Figure 34), consistent with steady surveillance intensity.

A predictive perspective reinforces sequencing's operational value. Using TimeGPT with exogenous drivers, Weekly Sequencing Rates emerged as the single most informative external feature for forecasting next-month cases in 54% of the nations analyzed. Countries including Austria, Belgium, Bulgaria, Cyprus, Germany, Ireland, Luxembourg, and Slovakia benefited most clearly from incorporating sequencing as a predictive covariate (Figure 36). Although deaths, ICU admissions, and hospitalizations also contributed in some settings, they were less frequently the top predictors. The strongest predictive performance was

observed in Ireland and Bulgaria (Figure 37), suggesting relatively stable, policy-relevant linkages between sequencing intensity and near-term case trajectories in those contexts.

These results must be interpreted in light of limitations inherent to observational surveillance data. Epidemiological indicators depend on testing availability, ascertainment practices, and reporting protocols, while sequencing volumes reflect laboratory capacity and sampling strategies that may vary within and across countries and over time. The mlVAR granularity is national, potentially averaging over sub-national heterogeneity in both policies and epidemic dynamics. Correlational and autoregressive associations, while informative, do not establish causality. Incidental or policy-induced correlations can arise (e.g., between first doses and new cases in 2022), and short lags may understate the delayed effects of interventions. Despite these caveats, the integrated framework demonstrates value on two fronts: it recovers known relationships (e.g., the dominant role of hospital pressure in 2020) and quantifies less appreciated ones (e.g., second-dose coverage associated with reduced stringency in 2021), while showing that sequencing rates can serve as a pragmatic, early-moving indicator and a useful exogenous predictor in real-time forecasting.

Missing data represent an additional limitation. Coverage of key covariates (especially ICU admissions, hospitalization series, vaccination stratifications, and sequencing volumes) was heterogeneous across countries and time, leading to variable sample sizes across analyses and, in some cases, exclusion of countries/periods with incomplete reporting. Because missingness is unlikely to be random, often reflecting changes in surveillance capacity, reporting policy, or epidemic pressure, it can bias correlations and autoregressive estimates (e.g., attenuating associations when under-reporting coincides with peaks) and reduce power to detect lagged effects. To mitigate this, analyses were run on complete-case subsets for the variables required in each model; nevertheless, results should be interpreted as conditional on the available reporting patterns, and future work could incorporate principled imputation and sensitivity analyses to assess robustness.

From a policy and preparedness standpoint, the combined approach, hierarchical time-series modeling of epidemiology, vaccination, variants, and stringency, augmented by sequencing-aware forecasting, offers a scalable template for future epidemics. It accommodates evolving pathogen characteristics, changing surveillance intensity, and heterogeneous national responses, yet preserves comparability across settings. Practically, timely investment in high-quality health monitoring

(testing, hospitalization/ICU tracking, robust data pipelines) and sustained, strategically targeted genomic surveillance can accelerate learning about transmission and severity, support adaptive policy calibration, and improve near-term forecasting of demand on health services. In future outbreaks, deploying a flexible, multilevel analytic workflow of this kind could shorten the time to actionable insight, help identify the most responsive levers (and appropriate lags) for intervention, and tailor strategies to country-specific constraints and goals.

The analysis present an integrated, generalizable framework that spans the full pandemic arc and bridges population-level regularities with country-specific nuances. Across 2020–2022, policy stringency aligned first with hospitalizations and later with cases, in parallel with the rise of vaccination coverage and the shifting variant landscape; vaccination reduced cases and hospitalizations overall, with second-dose coverage uniquely associated with lower stringency in 2021; Delta coincided with tighter policies, while Omicron primarily drove cases without comparable increases in hospitalization or stringency. Across a broader European set, sequencing trends tracked severe outcomes in many countries and, when used as an exogenous feature in TimeGPT, frequently improved short-horizon case forecasts, with particularly strong performance in Ireland and Bulgaria. Although constrained by the limits of surveillance data and national-level aggregation, the evidence supports sequencing-informed, context-aware analytics as a core component of epidemic policy design, linking early molecular signals to actionable epidemiological intelligence and enabling faster, more calibrated responses in the face of future threats.

2.6. Implications for public health and European policy making

Taken together, the mlVAR analysis across nine European countries and the sequencing-informed forecasting study using TimeGPT point toward a pragmatic, evidence-linked way of governing epidemic risk in Europe. The findings indicate that the policy signals most informative for decision making shift as the epidemic context evolves: during the earliest phase, pressure on hospital services aligned most closely with tightening of measures, while in later phases,amid widespread testing, expanding vaccine coverage, and more transmissible but generally less severe variants,new case trends became more salient and, crucially, were

followed by measurable reductions when stringency increased. Vaccination coverage consistently tracked declines in cases and hospitalizations, with dose-specific dynamics that became more visible as coverage rose; notably, increases in second-dose uptake aligned with subsequent relaxation of measures across countries. Variant emergence had heterogeneous policy consequences: Delta coincided with higher stringency in 2021, whereas Omicron produced sharp incidence spikes without commensurate increases in hospital burden or policy tightening in the settings examined. These results argue for decision rules that adapt over time, explicitly linking trigger sets to vaccine milestones, circulating variant profiles, and health-system capacity, rather than relying on static indicators such as cases alone or hospitalizations alone.

The sequencing analysis complements this picture by showing that genomic surveillance is not only a tool for variant discovery but also an operational signal with predictive value. Weekly sequencing rates correlated with severe outcomes in many countries and, when treated as an exogenous feature in TimeGPT, frequently improved short-horizon case forecasts. This suggests that public health returns accrue not only from peak sequencing capacity during surges but also from the *consistency* of sequencing over time, which stabilizes early warning and enhances planning for staffing, oxygen supply, and test distribution. Countries such as Austria, Belgium, Bulgaria, Cyprus, Germany, Ireland, Luxembourg, and Slovakia particularly benefited from incorporating sequencing as a predictive covariate, while others exhibited different correlation structures that likely reflect differences in surveillance strategy, population density, and health-system organization. Rather than treating this heterogeneity as a flaw, European governance can leverage it by promoting comparable modeling standards and common data definitions at the EU level while allowing Member States to calibrate thresholds to their own fitted relationships.

Methodologically, the multilevel structure of the mlVAR framework, combining population-wide fixed effects with country-specific random effects, maps naturally onto the balance Europe must strike between coordination and subsidiarity. This architecture enables comparative learning across Member States without erasing local nuance in how indicators map to policy action. In operational terms, it favors a routine, transparent modeling cadence aligned with epidemiological delay structures (e.g., weekly updates on standardized feeds) and encourages documentation of lags and thresholds so that the rationale for policy changes is traceable. A durable European “epi-operations” stack would therefore invest in harmonized weekly data pipelines for epidemiology,

vaccination, health-system utilization, policy measures, and genomic surveillance; publish versioned model artifacts and nowcast/forecast packages that blend sequencing-informed signals with traditional indicators; and provide open dashboards where policymakers and the public can see how shifts in indicators translate into recommended actions.

At the same time, the limits of observational surveillance data require guardrails. Measurement error from variable testing, reporting practices, and sampling strategies can introduce incidental associations; national-level aggregation may obscure sub-national dynamics and inequities; and autoregressive associations do not establish causality. European guidance built on these analytics should therefore include explicit sensitivity analyses, minimum data-quality criteria for activating or relaxing measures, independent monitoring for model drift, and procedures for revising thresholds as vaccination, variant severity, or clinical management changes alter the mapping between indicators and outcomes. Communicating these caveats clearly can help maintain trust while still extracting decision value from imperfect data.

Translating these insights into policy, Europe can move from reactive crisis management to anticipatory, context-aware governance. A practical implementation would maintain a weekly rhythm of country-specific multilevel models, integrate sequencing-based leading indicators into short-term forecasts, and pair the outputs with transparent action bands that consider hospitalization growth, ICU occupancy, second-dose and booster milestones, and sustained sequencing-informed signals. When severity decouples incidence from clinical burden, as with Omicron in well-vaccinated populations, hospital and ICU trends should be weighted more heavily than raw case counts, and sequencing should be used to detect shifts in severity or immune escape early enough to recalibrate thresholds. Where surveillance is noisier, targeted EU support to improve the steadiness of sequencing and data completeness can yield outsized gains by sharpening both situational awareness and forecast skill.

To conclude, the combined evidence supports a European public health strategy that treats genomic surveillance as an operational early-warning system, institutionalizes multilevel, country-calibrated modeling for routine decision support, and adapts policy triggers to evolving vaccination coverage and variant characteristics. By standardizing data streams, clarifying decision cadences, and aligning analytics with action, European institutions and Member States can shorten the time from signal to response, tailor interventions to local realities, and reduce the societal

cost of future waves or novel threats while preserving a coherent, transparent policy framework across the Union.

Chapter 3

3 Anomaly Detection for Dominant Lineage Forecasting: DeepAutoCoV

The previous chapter provided regarding how epidemiological indicators describe pandemic dynamics, with a particular focus on the European context. The scope of this thesis extends beyond the analysis of the impact of epidemiological and genomic factors, aiming also to propose a practical methodology capable of contributing to the prevention of the establishment and spread of dominant lineages. In this context, the present chapter introduces an approach for the early identification of emerging SARS-CoV-2 lineages in the population, prior to their widespread diffusion and before associated increases in the Stringency Index are observed. The proposed framework is based on automated detection through anomaly-detection techniques.

Previous studies [47, 51] have employed either machine-learning–based anomaly detection or traditional statistical surveillance to identify Variants of Interest (VOIs) and Variants of Concern (VOCs). Following the emergence of Omicron, however, no macro-variant with comparable disruptive properties has been observed, leading to a shift in surveillance priorities from broad variant classes to specific lineages of concern and their sublineages [31]. In response to this shift, the methodology presented in this chapter focuses explicitly on Omicron lineages and sublineages, and, in contrast to prior work, is grounded in deep learning.

The approach integrates an autoencoder-based anomaly detection model to flag newly arising lineages whose genomic and epidemiological profiles suggest a risk of future dominance. Candidate lineages are

prioritized through a composite ranking that combines biologically motivated features with purely statistical indicators to produce an interpretable ordering. This ranked output is intended to support biologists and public-health authorities in decisions related to vaccine strain updates and in the timely adjustment of containment policies to mitigate viral spread.

At a broader level, the chapter promotes the use of AI-enabled genomic surveillance to reduce the time interval between sequence detection and actionable decision-making by governmental bodies and pharmaceutical developers, thereby accelerating the design and implementation of proportionate interventions (Figure 38). In this way, the methodological contribution of the thesis aligns with the overarching objective of transitioning from retrospective assessment to proactive prevention in pandemic response.

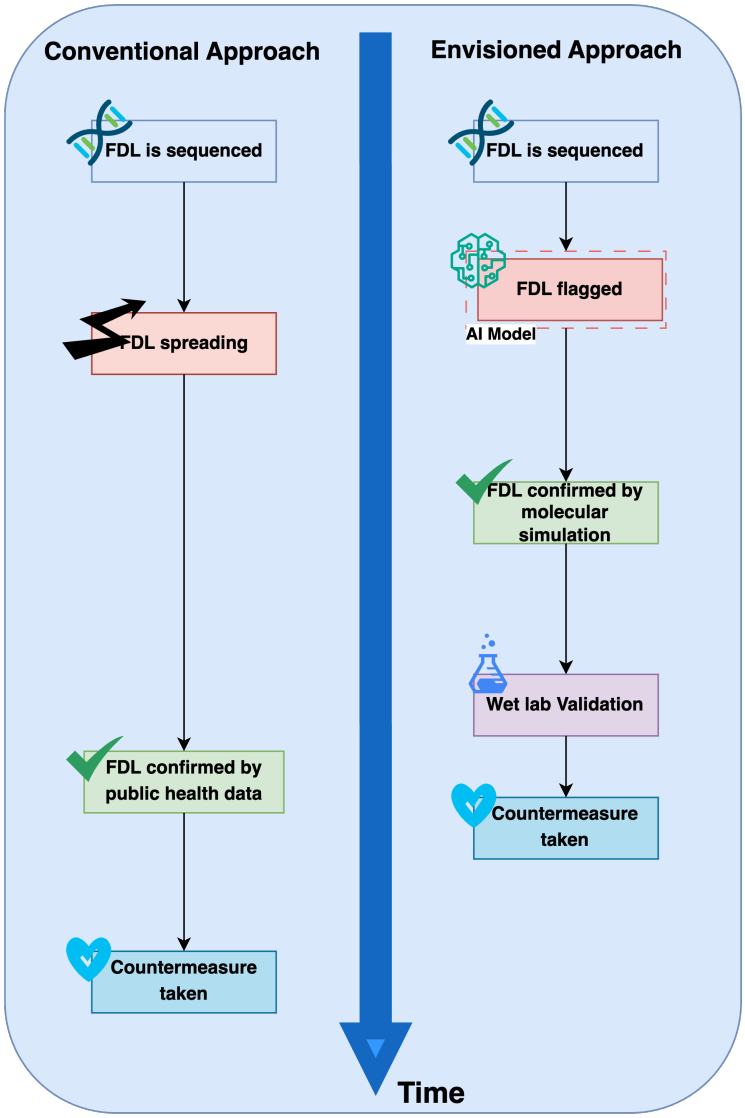


Figure 38: Comparison between the conventional and the envisioned AI-assisted approaches for genomic surveillance. In the conventional approach, FDLs are identified only after widespread transmission and confirmation through public health data, delaying countermeasure implementation. The envisioned approach integrates AI-based detection immediately after sequencing, followed by molecular simulation and wet-lab validation, enabling earlier confirmation and faster adoption of countermeasures [21, 2].

3.1. Problem definition: forecasting Future Dominant Lineages (FDLs)

The COVID-19 pandemic has been marked by the sequential emergence of SARS-CoV-2 variants, lineages, and sublineages that have progressively replaced previously circulating forms as a consequence of increased transmissibility, immune evasion, or other fitness advantages [95]. Public health agencies, including the CDC [4] and WHO, classify certain variants as VBM (Variant Being Monitored) or VOC when epidemiological evidence indicates a significant public health impact, such as increased hospitalizations, mortality, or reduced vaccine and therapeutic efficacy [96]. In practice, however, the designation of a lineage as a public health concern is often reactive, occurring only after its spread is already evident in epidemiological data (Figure 38). This delay reduces the effectiveness of early countermeasures, such as targeted containment strategies or timely vaccine updates.

The concept of Future Dominant Lineages (FDLs) has been introduced in this thesis to address this limitation. A viral (sub)lineage is defined as an FDL if it is projected to reach or exceed a predefined prevalence threshold,typically 10% of all sequenced viruses in a specific geographical region,within a specified time frame [21]. Early identification of such lineages, ideally when prevalence remains extremely low could enable a transition in genomic surveillance from a reactive to a proactive approach.

From a methodological perspective, forecasting FDLs represents a complex challenge. Conventional AI and bioinformatics approaches for SARS-CoV-2 surveillance are primarily focused on: (i) classifying sequences into known lineages [97, 98, 99, 100], (ii) predicting mutations likely to occur at specific protein sites [101], or (iii) estimating immune escape or binding affinity of Spike protein variants. These approaches rely on prior knowledge of target mutations or established lineages and are therefore unable to detect truly novel emerging variants with pandemic potential. Furthermore, supervised learning methods depend on labeled

examples of past VOCs, which are inherently unavailable for lineages that have not yet emerged.

Within this context, the problem can be reformulated as an unsupervised anomaly detection task [47]: given a constantly evolving viral population represented by its genomic sequences, the objective is to identify anomalous Spike protein sequences that are likely to belong to future high-prevalence lineages. These anomalies are hypothesized to exhibit mutation patterns that confer a selective advantage, enabling them to outcompete contemporaneous circulating strains.

Addressing this problem involves overcoming several challenges:

- The high dimensionality of genomic data, even when restricted to the Spike protein, necessitates effective dimensionality reduction and feature encoding strategies.
- The rapid evolution of the virus and the resulting data drift require periodic retraining of models to preserve predictive performance.
- Low signal-to-noise ratio during early emergence, when FDL sequences are rare and can be confounded with non-advantageous mutations.
- Operational constraints in genomic surveillance, whereby only a limited number of flagged sequences can undergo in vitro validation, make accurate detection and effective prioritization of candidates essential.

The DeepAutoCoV [21]framework has been designed to address these challenges by implementing an autoencoder-based anomaly detection system trained exclusively on currently circulating sequences, without assumptions regarding mutation sites. Autoencoders (AEs) are deep neural networks trained to reconstruct as output the input provided by the user [21]. For example, AEs are widely exploited in image processing to reduce image noise. When the input of an AE deviates from the training set used to train the network, the AE ability to precisely reconstruct the input decreases. In our framework, the input of the AE is represented by Sars-CoV-2 genomic sequences. Sequences with AE reconstruction errors significantly deviating from the expected range are identified as potential anomalies, which are subsequently filtered using biological plausibility criteria (e.g., PAM scoring). This approach enables the prioritization of FDL candidates weeks to months before they reach dominance. By providing such early warnings, the framework has the potential to shorten the response time available to public health systems, thereby

facilitating timely interventions to mitigate the spread of high-risk variants.

3.2. Genomic data preprocessing and GISAID filtering

The genomic data preprocessing pipeline (Figure 39) was designed to extract high-quality Spike protein sequences from the GISAID repository, ensuring both accuracy in lineage assignment and suitability for downstream anomaly detection analysis. The dataset encompassed all available SARS-CoV-2 Spike protein sequences collected between December 24, 2019, and November 8, 2023, for a total of 16,187,950 proteins across 198 epidemiological weeks, each annotated with its corresponding Pango lineage classification .

The decision to focus exclusively on the Spike glycoprotein was motivated by its central role in antibody-mediated immunity, its designation as the primary antigen in COVID-19 vaccines, and the concentration of mutations affecting viral transmissibility and immune escape within its sequence. Restricting the analysis to the amino acid representation of the Spike protein, rather than the full viral genome, significantly reduced the feature space while preserving the most biologically relevant information for the prediction task.

A series of stringent quality control filters was applied to ensure the reliability of the data. Sequences containing missing or unrecognized amino acids were excluded, as were sequences whose lengths deviated by more than ±30 amino acids from the median length of their corresponding lineage. This filtering step retained sequences within a length range of 1241 to 1301 amino acids, effectively removing incomplete or misassembled proteins [21]. To resolve inconsistencies in lineage annotation for identical sequences, the most recent label in temporal order was retained. Additionally, all Omicron variant sequences with reported collection dates preceding November 2021 were manually removed to eliminate artifacts from retrospective lineage assignment.

Following quality control, the dataset was reduced to 4,506,983 high-quality Spike proteins. These were organized into five datasets: four representing individual countries (USA, UK, France, and Denmark) and one global (“World”) dataset comprising sequences from all available countries worldwide. This subdivision enabled evaluation of model performance under varying levels of sequencing coverage, ranging from high-volume surveillance systems (USA: 1,294,936 sequences; UK:

962,496) to more limited sequencing efforts (France: 140,757; Denmark: 312,612) (Figure 39).

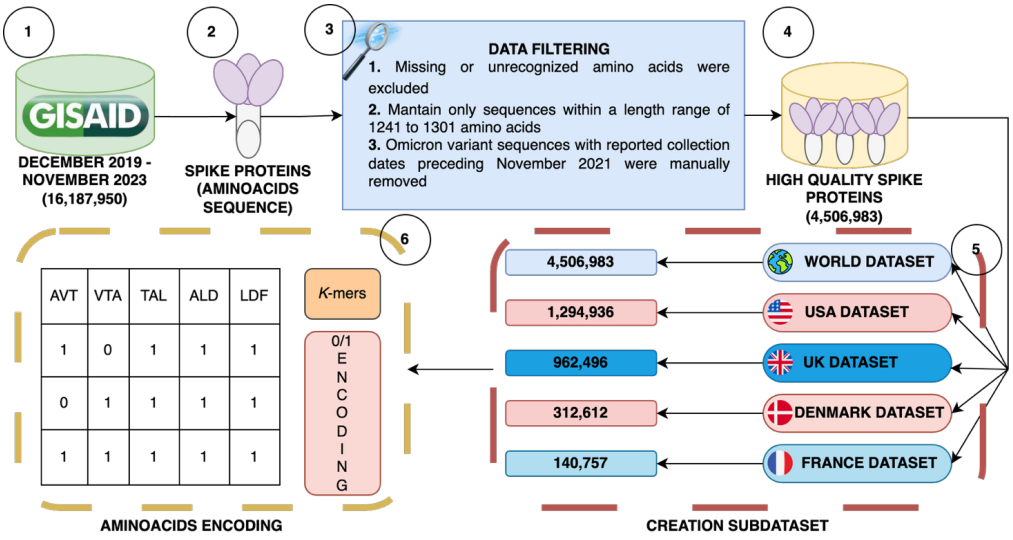


Figure 39: Overview of the data processing workflow for SARS-CoV-2 spike protein sequences retrieved from GISAID. The pipeline includes filtering steps (3) to exclude sequences with missing or unrecognized amino acids, retain only sequences within a length range of 1241–1301 amino acids, and remove Omicron variant sequences collected before November 2021. Following filtering, sequences are transformed into *k*-mer representations (6) and subsequently encoded in binary (0/1) format. The resulting datasets comprise the global dataset and country-specific subsets for the USA, UK, Denmark, and France (5), with sequence counts reported for each [2].

For model input preparation, amino acid sequences were transformed into fixed-length binary feature vectors using a 3-mer (k=3) encoding strategy, consistent with prior studies on viral sequence analysis [47]. In bioinformatics, a k-mer is defined as a subsequence of length k obtained by sliding a fixed-size window across a biological sequence [102, 103],in this case, the Spike protein's amino acid chain. For example, given the peptide "NSH", this would constitute a 3-mer when k=3. The complete set of k-mers extracted from a sequence represents all possible contiguous amino acid triplets it contains. Each *k*-mer was assigned to a unique position in the feature vector, with binary values indicating its presence (1) or absence (0) in the given sequence (Figure 39). This representation captures local sequence patterns that may be associated with structural or functional changes in the Spike protein, providing a compact yet

informative input for anomaly detection models. The preprocessing pipeline thus ensured that only high-fidelity, biologically relevant, and consistently annotated genomic data were passed to the DeepAutoCoV anomaly detection framework. By filtering out low-quality sequences, standardizing feature encoding, and stratifying datasets by geographical origin, this stage established a robust foundation for the reliable identification of potential Future Dominant Lineages in real-world genomic surveillance settings.

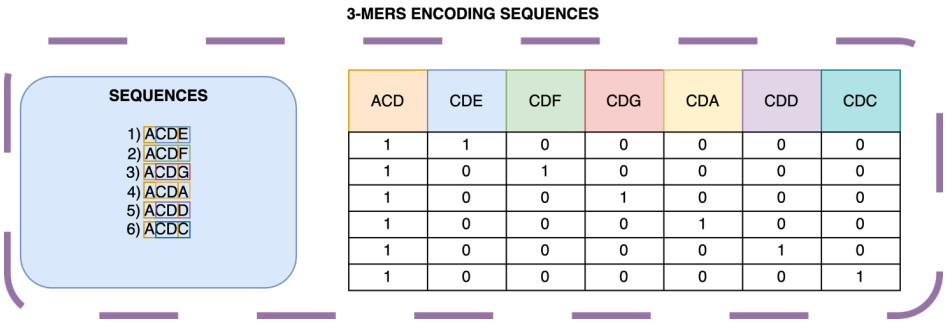


Figure 40: Illustration of the *k-mer* encoding process applied to amino acid sequences. The example shows six sequences, each decomposed into overlapping triplets *(k-mers)* of amino acids. The unique k-mers identified across all sequences are listed on the right, forming the vocabulary used for binary (0/1) encoding in subsequent computational analyses [2].

3.3. DeepAutoCoV architecture and training configuration

The DeepAutoCoV (Figure 41) system was implemented as a deep learning AutoEncoder specifically designed to process SARS-CoV-2 Spike protein sequences represented through k-mer (Figure 40) encoding. Following established approaches for viral genome analysis [103, 102, 47], as explained above, each amino acid sequence was transformed into a fixed-length binary vector, in which each element indicated the presence (1) or absence (0) of a specific contiguous 3-amino-acid k-mer. This choice of k=3 allowed a balance between capturing sufficient sequence context and maintaining computational tractability, while significantly reducing the feature space compared to whole genome or nucleotide-level representations [47]. The binary input vector was fed into the AutoEncoder architecture (Figure 41A), which consisted of multiple fully connected encoding layers with progressively decreasing neuron counts (from 1024 to 64), converging towards a central bottleneck layer of 32

neurons. This bottleneck layer incorporated Gaussian noise, a regularization strategy intended to improve the model's generalization ability by preventing overfitting to specific training patterns. The decoding path mirrored the encoding structure, with layers expanding symmetrically from 64 back to 1024 neurons, culminating in an output layer tasked with reconstructing the original k-mer input representation. A dropout layer was integrated into the architecture to further reduce overfitting risks by randomly deactivating a fraction of neurons during training, thus encouraging the network to learn more robust latent features [104, 105].

The training procedure aimed to minimize the reconstruction error between the input and the decoded output, adopting the mean squared error (MSE) loss function and the Adam optimizer to ensure stable and efficient convergence (Figure 41B). The model was trained for 50 epochs with a batch size of 256, using the TensorFlow and Keras libraries [106, 107]. The anomaly detection mechanism was based on the principle that FDLs would present patterns of k-mer occurrence that deviated from the learned steady-state viral landscape. After training on sequences representing the initial epidemiological baseline, the AutoEncoder produced a reconstruction error for each new sequence; those whose MSE exceeded 1.5 standard deviations above the median were classified as anomalies, following established practices in unsupervised outlier detection [108, 109].

To refine the anomaly set and reduce false positives, a secondary filtering step based on the Point Accepted Mutation-30 (PAM30, Figure 42) scoring method [110] was applied. The PAM30 is a substitution scoring matrix specifically optimized for comparing sequences that are closely related, typically with less than 30% evolutionary divergence. It assigns scores to each possible amino acid substitution based on the probability of that substitution occurring over an evolutionary distance corresponding to 30 PAM units, where 1 PAM represents a 1% change in amino acids. Higher scores indicate substitutions that are more likely to occur in evolutionarily related proteins, while lower or negative scores correspond to rare or deleterious substitutions. The PAM30 is particularly suitable for detecting fine-scale differences in highly similar protein sequences, such as those of emerging SARS-CoV-2 lineages [110]. For each dataset, the average PAM30 score between all training set FDL proteins and the original Wuhan reference sequence was computed. Sequences whose PAM30 score exceeded this average were excluded from the anomaly list, as they were considered too divergent to represent plausible emerging lineages within the observed evolutionary context. In

addition, proteins identical to any in the training set were automatically removed from consideration, as their known circulation excluded them from being novel FDLs.

The training and evaluation of DeepAutoCoV followed a rolling simulation protocol (Figure 41B) designed to emulate real-world genomic surveillance (Figure 41C). The model was initially trained on sequences collected during the first available week for each geographical dataset, which captured the steady-state viral population of that region at the start of observation. To control feature dimensionality, only k-mers present in at least 1% of sequences in the initial training set were retained for model input. Subsequently, each week's newly uploaded sequences were processed by the trained model, and those flagged as anomalies were subjected to PAM30-based filtering. When a lineage reached the FDL definition threshold,set at a weekly prevalence of at least 10% ,the model was retrained to include all sequences belonging to that FDL observed thus far, and the PAM threshold was recalculated. This threshold is user-defined and can be changed to ref lect epidemiology and/or public health considerations. Retraining ensured that the model adapted to the evolving viral landscape, maintaining sensitivity to new anomalies without repeatedly flagging sequences from already dominant lineages. More details about the simulation pipeline are in the next subchapter.

3.4. Weekly genomic surveillance simulation workflow

The weekly genomic surveillance simulation was designed to reproduce a realistic operational framework in which genomic data are collected, processed, and analyzed in near real time to identify FDLs before they reach widespread prevalence (Figure 41B). The simulation begins with the assembly of an initial training set consisting of Spike protein sequences collected during the first available week for each dataset. This initial set represents the steady-state viral population for the given geographical area and serves as the baseline reference for anomaly detection. To reduce feature dimensionality and computational overhead, only k-mers occurring in at least 1% of the sequences in this first training set are retained as model input features. The AutoEncoder is trained on this baseline set to learn a compressed representation of the dominant sequence patterns present in the early-stage viral landscape.

Subsequent weeks are processed in chronological order, with newly uploaded Spike sequences undergoing k-mer encoding and being passed

through the trained AutoEncoder. For each sequence, the reconstruction error, calculated as the mean squared error between the input and output, is evaluated, and sequences exceeding a threshold of 1.5 standard deviations above the median reconstruction error are classified as anomalies. These anomalies are further refined through a PAM30 scoring step, which compares each flagged sequence to the original Wuhan reference sequence. Sequences with a PAM score higher than the mean score observed among known FDLs in the training set are excluded from the anomaly list, as they are deemed excessively divergent to represent realistic emerging lineages within the simulated epidemiological context. The sequences flagged as anomalies are ranked based on the RMSE and PAM30. Identical sequences to those already present in the training set are also discarded to avoid redundant detections.

If, during the simulation, a lineage reaches or exceeds the predefined dominance threshold of 10% weekly prevalence, it is designated as an FDL. At this point, the model undergoes a retraining phase that incorporates all sequences belonging to the newly emerged FDL observed up to that week. This retraining serves to update the AutoEncoder's internal representation of the circulating viral population, ensuring that the model remains tuned to detect novel anomalies without repeatedly flagging sequences from already dominant lineages. The PAM threshold and the *k*-mers are recalculated during each retraining cycle to maintain consistency between anomaly detection and sequence divergence evaluation.

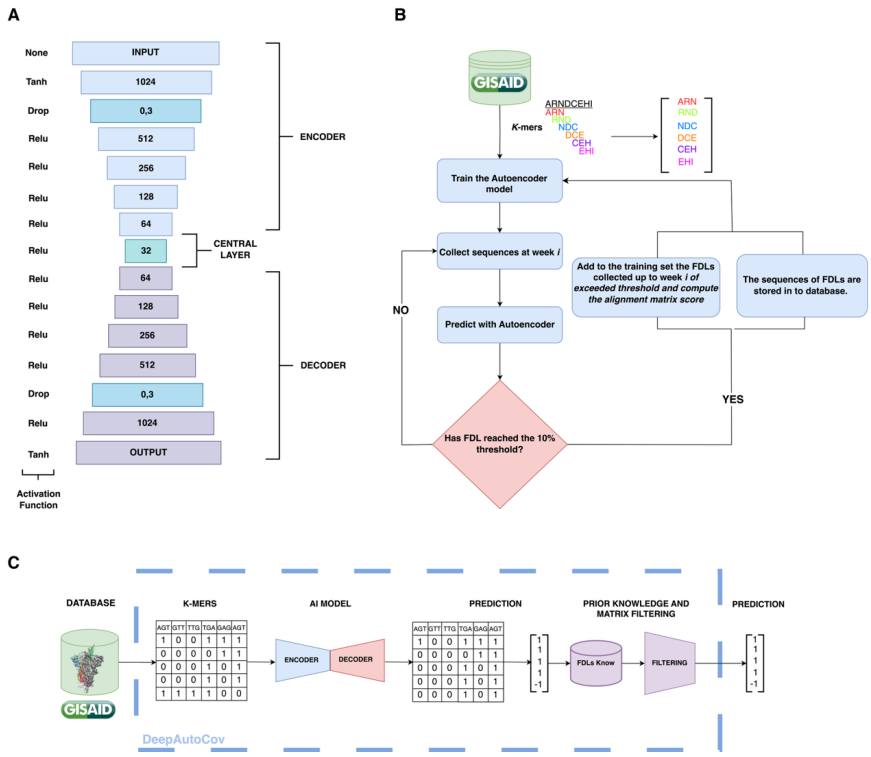


Figure 41: DeepAutoCoV architecture. A) Layered architecture with decreasing neuron counts in the encoding layers (from 1024 to 64) and increasing counts in the decoding layers (from 64 to 1024). A central layer with 32 neurons introduces Gaussian noise to enhance generalization. The decoding layers mirror the encoding structure. The output layer reconstructs the original input data. B) Schematic of the genomic surveillance simulation. SARS-CoV-2 Spike protein sequences are gathered, converted into k-mers, and filtered. An AutoEncoder outlier detection system is trained. Each week, newly uploaded sequences are evaluated as either FDLs or not. When a FDL reaches the 10% threshold, the model is retrained. C) Overview of DeepAutoCoV process: k-mers extraction, AutoEncoder, prediction and filtering [21]

3.5. Evaluation metrics: lead time, PPV, prioritization

The evaluation of the proposed anomaly detection framework relied on three primary metrics designed to quantify its effectiveness in identifying FDLs under realistic genomic surveillance conditions.

The first metric, **lead time (LT)**, measures the temporal advantage provided by the system in detecting an FDL prior to its dominance in the circulating viral population. It is defined as:

$$LT = t_{threshold} - t_{detection} \quad (7)$$

where $t_{detection}$ is the week of the first true positive identification of the FDL by the model, and $t_{threshold}$is the week in which the same lineage reaches the predefined dominance threshold (10% of all sequences collected in that week). A larger value of LT indicates a greater anticipation capacity, enabling earlier public health action.

The second metric, **positive predictive value (PPV)**, quantifies the precision of the system's predictions by calculating the proportion of correctly identified FDLs among all sequences flagged as anomalies:

$$PPV = \frac{TP}{TP + FP} \tag{8}$$

where TP represents the number of true positives (flagged sequences that indeed belong to an FDL detected before it reaches the dominance threshold) and FP represents the number of false positives (flagged sequences not corresponding to an FDL or detected after it has already reached the threshold). Higher PPV values indicate that a greater fraction of flagged sequences is genuinely relevant, thereby minimizing unnecessary downstream validation efforts.

The third metric, **prioritization (P)**, assesses the model's capability to rank true positive FDLs among the top candidates in each evaluation batch. It is computed as:

$$P = \frac{1}{N_w} \sum_{i=1}^{N_w} r_i \tag{9}$$

where N_w is the number of evaluation weeks containing at least one true positive (i.e., at least one sequence that meets the FDL criteria in that week), and r_i is the ranking position of the first true positive in week i within the list of sequences sorted in descending order of anomaly score (mean squared reconstruction error). Lower values of P correspond to better prioritization, as the most relevant sequences appear earlier in the ranked list.

All the model performance was evaluated on the top 100 sequences sorted by MSE. The restriction to the top 100 sequences per week was motivated by practical constraints in genomic surveillance workflows, where only a limited number of candidates can undergo further phenotypic or in vitro validation

Together, these metrics provide a comprehensive framework for evaluating the timeliness, precision, and operational relevance of the proposed method in a real-world genomic surveillance setting, avoiding

reliance on general classification accuracy, which may be less informative in the context of emerging variant detection.

3.6. Comparison with alternative detection methods

The comparative evaluation of DeepAutoCoV against alternative detection approaches was conducted to assess its relative performance in forecasting FDLs within a realistic genomic surveillance framework. Both supervised and unsupervised models were considered as benchmarks, each implemented following established methodologies and applied to the same datasets under identical preprocessing conditions.

The first supervised baseline was Logistic Regression (LR), one of the fundamental models in machine learning. LR establishes a relationship between a set of independent variables and a binary dependent variable, producing as output the probability that a given observation belongs to one of two categories, in this case, *Anomaly* or *Non-Anomaly*. The model was implemented using the Scikit-Learn library with default hyperparameters [106].

The Spike2Vec (S2V) method [111] was also used as a supervised benchmark. This approach performs Spike protein feature extraction by computing k-mer frequencies and applying the Fourier transform to capture compositional and periodic sequence patterns. Proteins are embedded into fixed-length vectors, which are subsequently projected into a low-dimensional space using an approximate kernel method. These representations are then fed into a Naive Bayes (NB) [106] classifier, implemented in accordance with the original Spike2Vec publication, to identify known SARS-CoV-2 lineages through supervised classification.

A third supervised approach, Spike2Signal (S2S) [112], transforms Spike protein sequences into numerical signals by mapping each amino acid to a specific number and computing the cumulative distribution of these values. The resulting time series is used as the input to a deep residual neural network (ResNet) designed to classify known SARS-CoV-2 (sub)lineages. This representation enables the capture of structural patterns in protein sequences that may be relevant to variant classification.

As an unsupervised baseline, a Distance-Based (DB) anomaly detection model was implemented. This method ranks sequences according to the mean squared error (MSE) calculated between each sequence in the evaluation week and all sequences contained in the training set, based on

their k-mer representations. Sequences with higher distances are considered more likely to be anomalous. This approach does not rely on lineage labels and is therefore theoretically capable of identifying novel variants; however, its reliance on raw distance metrics often limits robustness in high-dimensional feature spaces.

For all models, evaluation was conducted on a weekly basis, with performance restricted to the top 100 sequences ranked as anomalous or high-priority by each method. Metrics used for the comparison included lead time, positive predictive value (PPV), and prioritization, as previously defined.

The unified evaluation protocol ensured that performance differences among methods could be attributed to their intrinsic algorithmic characteristics rather than to variations in data availability or preprocessing. In this context, DeepAutoCoV's deep AutoEncoder architecture, which learns non-linear, high-level features of Spike protein *k*-mer distributions, provided a methodological advantage for detecting subtle sequence deviations that precede lineage dominance, a capability not matched by the evaluated alternative methods, as reported in the following subchapter.

3.7. Results

Anticipating the Future: Lead Time in FDL Detection

Across the global dataset (Figure 41, Table 7, Table A3), the system achieved a median lead time of 17 weeks, with an interquartile range (IQR) of 11 to 32 weeks. At the national level (Figure 41, Table A4-A7) lead times varied according to sequence volume, ranging from 5 weeks in Denmark (IQR 2–11) to 12 weeks in the United States (IQR 7–37). Per FDLs lead time details are provided in Table A3-A7. These results indicate that DeepAutoCoV consistently provides substantial advance notice of emerging dominant lineages, even in countries with smaller sequencing volumes.

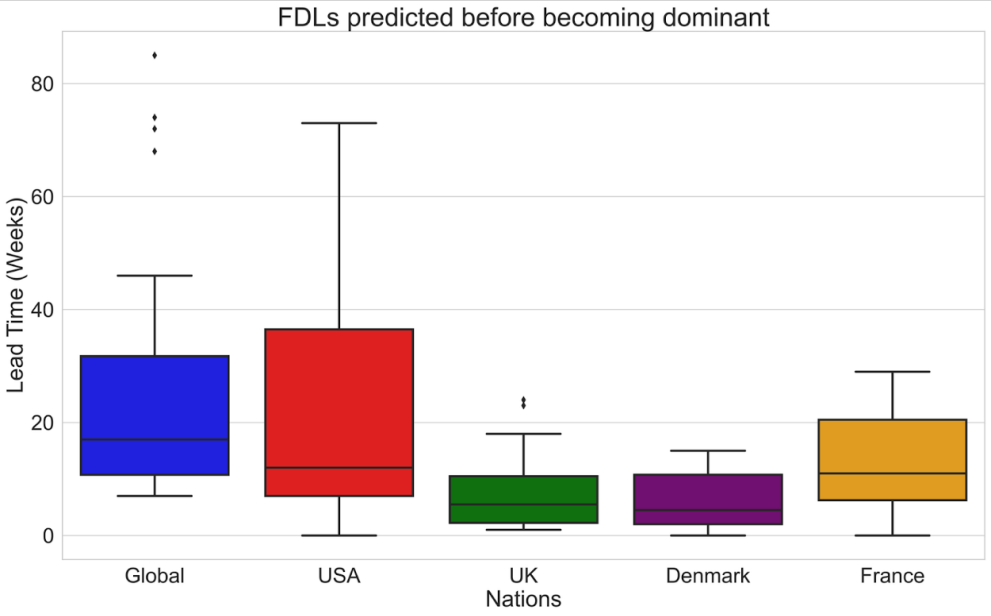
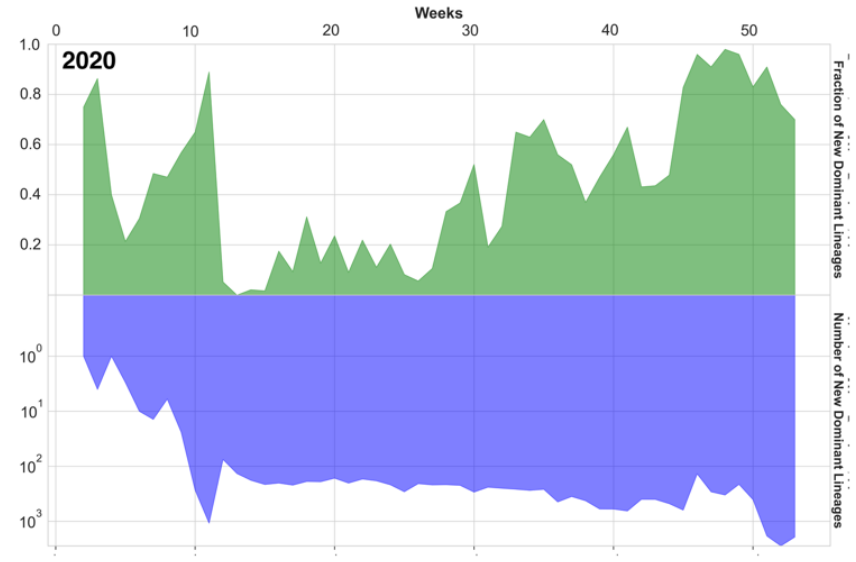
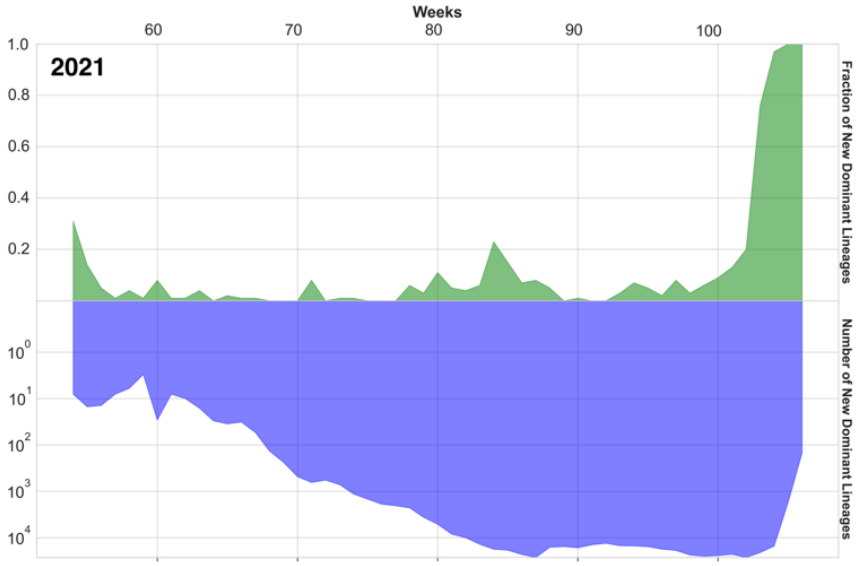


Figure 41: Lead time (in weeks) for Future Dominant Lineages (FDLs) predicted before reaching dominance. The boxplots show the distribution of lead times across different datasets: Global, USA, UK, Denmark, and France. The global dataset exhibits a median lead time of approximately 15 weeks, while the USA shows greater variability with predictions extending up to 70 weeks. European nations (UK, Denmark, France) present shorter and more consistent lead times, with France displaying intermediate variability compared to the UK and Denmark. Outliers indicate cases where FDLs were identified exceptionally early relative to their dominance [21].

Precision Under Pressure: Positive Predictive Value (PPV)

Median PPV values varied across datasets, with the highest observed in the United Kingdom (0.57, IQR 0.24–0.92) and lower values in France (0.47, IQR 0.17–0.70) and Denmark (0.49, IQR 0.02–0.90). The global dataset achieved a median PPV (Figure 42) of 0.30 (IQR 0.06–0.42). PPV trends were observed to be inversely correlated with dataset entropy, reflecting that the model performs more accurately when the genomic landscape is characterized by clusters of closely related sequences rather than high levels of diversity.


Weeks
0
10
20
30
40
50
2020
1.0
0.8
0.6
0.4
0.2
10^0
10^1
10^2
10^3
Fraction of New Dominant Lineages
Number of New Dominant Lineages


Weeks
60
70
80
90
100
2021
1.0
0.8
0.6
0.4
0.2
10^0
10^1
10^2
10^3
10^4
Fraction of New Dominant Lineages
Number of New Dominant Lineages

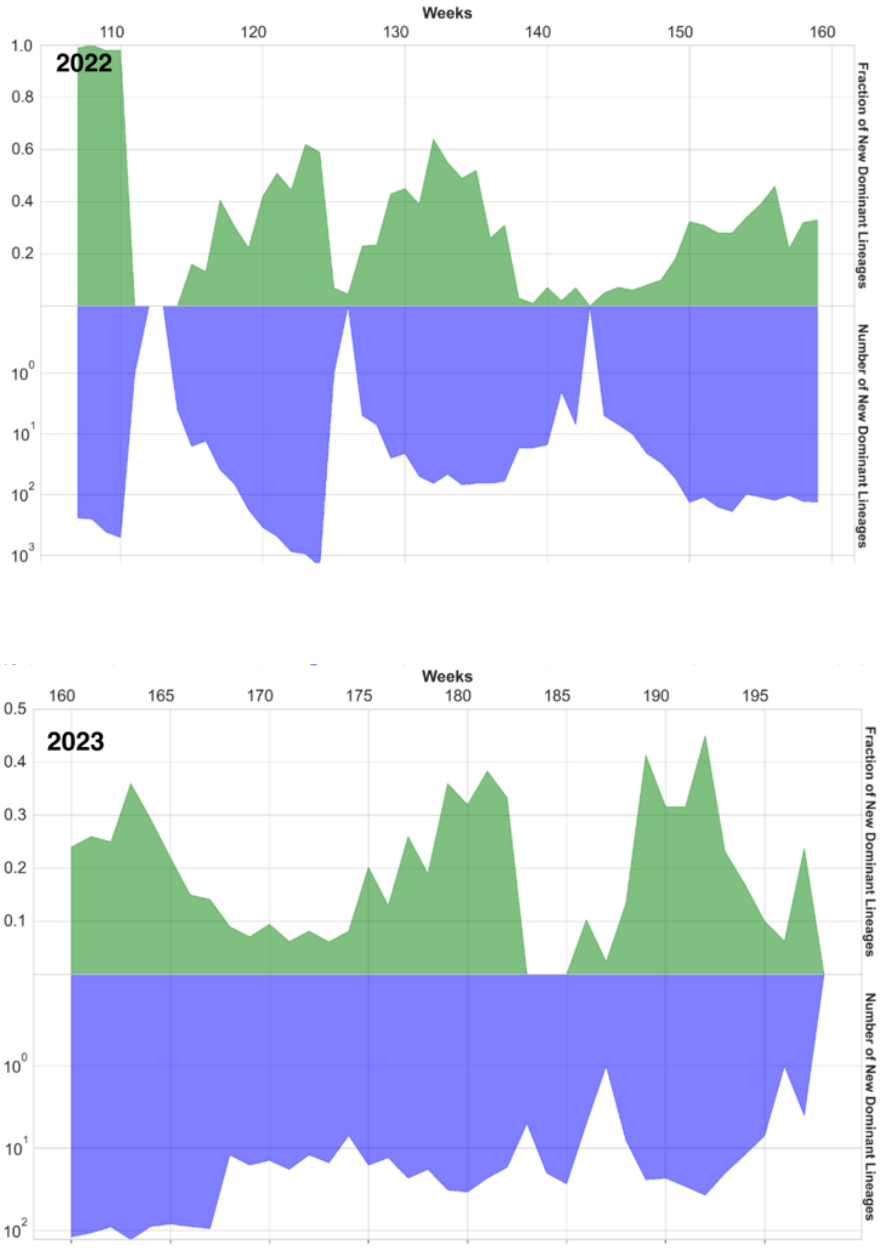


Figure 42: PPV on Global dataset. X-axis: simulation weeks; y-axis top side: PPV (green); bottom side: number of new FDLproteins in semi-logarithmic scale, each week (blue) [21].

Ranking What Matters: Prioritization of True Positives

The ability of the system to prioritize true positive FDLs within the top-ranked flagged sequences is critical, given the limited resources available for laboratory validation in a real world scenario. At the global level (Table 7), the first FDL was detected on average in the sixth position,

whereas in national datasets (Table 7) such as the USA and UK, true positives typically appeared within the top three or four positions. This prioritization capacity underscores the practical utility of DeepAutoCoV in directing attention to the most relevant sequences for downstream validation

Mutation Analysis

In addition to predictive performance, DeepAutoCoV provides interpretable results by surfacing key Spike FDL mutations. The mechanism relies on a binary input encoding (presence/absence of specific k-mers) and an AutoEncoder trained to reconstruct these k-mer patterns. For sequences flagged as potential FDLs, we inspect the reconstruction residuals and extract the k-mers that the model fails to reproduce (i.e., highest per–k-mer reconstruction error after compression). By aligning these high-residual k-mers to the Spike protein, we localize the underlying amino-acid changes and thus pinpoint mutations that likely contribute to fitness advantages. Using this procedure, the model recovered, for example, k-mers mapping to the P681R substitution in Delta (AY.4), [113] a substitution at the furin cleavage site known to enhance viral replication and transmissibility. Similarly, in Omicron BA.2, the model highlighted the Y505H mutation [114] in the receptor-binding domain, which strengthens binding affinity with the ACE2 receptor. For the KP lineage, the model identified FLiRT mutations [115] (R346T, F456L), which have been associated with both transmissibility and immune escape. These results demonstrate the capacity of DeepAutoCoV not only to detect FDLs early but also to provide mechanistic insights into their evolutionary advantages.

Table 1: Summary of Results. The lead time is reported as the median number of weeks together with the interquartile range (IQR). The frequency of FDLs at discovery corresponds to the median (IQR) proportion of the identified FDL proteins over the total number of Spike proteins present in the dataset during the week of detection. The Positive Predictive Value (PPV) is expressed as the median (IQR). Prioritization refers to the average ranking position of the first true positive identified. Finally, statistical significance was assessed using the Wilcoxon paired test, as described in the Methods section.

Dataset	Lead time (weeks)	FDL frequency at discovery	PPV	Prioritization	p-value
Global	17 (11, 32)	0.1% (0.02%, 0.7%)	0.30 (0.06, 0.42)	6° place	4e-06
USA	12 (7, 37)	0.2% (0.05%, 3%)	0.48 (0.10, 0.76)	3°/4° place	4e-07
UK	6 (2, 11)	3% (0.1%, 8%)	0.57 (0.24, 0.92)	3°/4° place	7e-16
Denmark	5 (2, 11)	0.8% (0.2%, 5%)	0.49 (0.02, 0.90)	4° place	4e-03
France	11 (6, 20)	1.5% (0.4%, 7%)	0.47 (0.17, 0.70)	4° place	7e-12

Comparison with competing methods

The evaluation of DeepAutoCoV against both supervised and unsupervised baseline methods (Table 8) demonstrated a consistent superiority across all datasets. When compared with Logistic Regression, Spike2Vec, and Spike2Signal, none of these supervised approaches was able to correctly flag any FDLs, resulting in a PPV of zero. This outcome reflects the inherent limitation of supervised classification strategies in anticipating lineages that, by definition, are not yet represented in the training set.

In contrast, the unsupervised distance-based anomaly detection model achieved non-zero results but remained markedly inferior to DeepAutoCoV. In the Global dataset, the distance-based model obtained a median PPV of only 0.05, substantially lower than the 0.30 achieved by DeepAutoCoV under the same conditions.

Beyond PPV, DeepAutoCoV also demonstrated superior prioritization capacity, consistently ranking true FDLs among the top positions across datasets. On average, the first true positive was detected within the top 3–6 ranked sequences. Competing methods either failed to provide meaningful prioritization, as in the case of supervised classifiers, or produced lower consistency in ranking, as observed for the distance-based model.

Taken together, these results highlight the advantage of DeepAutoCoV in the early identification of emerging lineages (Table 8). By outperforming both supervised and unsupervised alternative methods in terms of predictive value and prioritization, DeepAutoCoV emerges as the most reliable approach for incorporation into genomic surveillance frameworks.

Table 2: PPV comparison. DeepAutoCoV was compared to other unsupervised (DB = Distance-based linear), and supervised (LR = Logistic regression, S2V = Spike2Vec, S2S = Spike2Signal, NB = Naïve Bayes) methods. Median PPV (IQR) calculated over the entire time interval (2019-2023)

Method	Global	USA	UK	Denmark	France
DeepAutoCoV	0.30 (0.06, 0.42)	0.48 (0.10, 0. 72)	0.57 (0.24, 0.92)	0.49 (0.02, 0.90)	0.47 (0.17, 0.70)
DB	0.05 (0, 0.09)	0.03 (0, 0.04)	0	0.04 (0, 0.07)	0.02 (0, 0.08)
LR	0	0	0	0	0
S2V NB, S2VT NB, S2S RESNET	0	0	0	0	0

3.8. Discussion and Future Directions

This chapter presented DeepAutoCoV, an unsupervised deep-learning framework based on an autoencoder architecture designed to identify future dominant lineages (FDLs) of SARS-CoV-2 from weekly genomic surveillance data. The model was trained on Spike protein k-mer profiles and evaluated in a rolling simulation that emulated real-world surveillance conditions. By using reconstruction error as an anomaly signal, DeepAutoCoV was able to flag sequences likely to belong to lineages that would later cross a dominance threshold. The evaluation included multiple datasets at both global and national scales (United States, United Kingdom, France, and Denmark), and performance was quantified using lead time, positive predictive value (PPV), and prioritization metrics. Comparative analyses against supervised and unsupervised baselines indicated that DeepAutoCoV can anticipate emerging lineages and rank them among the most relevant candidates for further validation.

Strengths and contributions

Overall, DeepAutoCoV demonstrated the ability to detect emerging lineages at extremely low frequencies, often providing lead times from several weeks to months before dominance. This capability is particularly relevant for public-health preparedness, where early signals may support proactive responses such as vaccine updates, therapeutic re-evaluation, and targeted laboratory characterization. Importantly, the anomaly-detection paradigm does not require prior assumptions about specific mutation sites, which makes the method adaptable to novel evolutionary trajectories and potentially transferable to other rapidly evolving pathogens.

A further advantage is the interpretability of the anomaly signal: by inspecting poorly reconstructed k-mers, the framework can highlight sequence patterns associated with novelty and facilitate the localization of candidate mutations for downstream annotation. This bridges predictive modeling and biological insight, enabling hypothesis generation for subsequent experimental validation and surveillance prioritization

Limitations

Despite these strengths, several limitations should be made explicit. DeepAutoCoV is ultimately constrained by the quality and

representativeness of the genomic surveillance stream on which it relies. In real-world settings, sequences are often uploaded with non-negligible delays, and sampling can be strongly non-random (for example, prioritizing outbreak investigations, travelers, or hospital-based cases). These operational choices may distort apparent lineage frequencies and can produce both false positives—when targeted sequencing artificially inflates the presence of unusual genomes—and false negatives, when emerging lineages circulate but remain under-sampled or appear only after delayed reporting.

A further limitation is that the anomaly signal is defined relative to the training distribution. As SARS-CoV-2 evolves, background diversity expands and the "normal" sequence space shifts, introducing concept drift. Under these conditions, reconstruction error can increase because the population baseline has moved, not necessarily because a truly concerning lineage is emerging. This implies that real-world deployment would require careful decisions about retraining cadence, drift monitoring, and threshold calibration to avoid misinterpreting expected evolutionary change as an alarm.

The choice to model Spike-only k-mer profiles also reflects an explicit trade-off between biological completeness and scalability. While Spike is central to transmissibility and immune escape, restricting the representation to Spike may miss fitness determinants outside this protein (e.g., replicase or nucleocapsid changes) and cannot fully capture long-range epistasis across genomic regions. Moreover, k-mer features can be sensitive to sequence quality and preprocessing decisions: ambiguous bases, incomplete coverage, translation artifacts, or inconsistent filtering may introduce artificial novelty that the autoencoder interprets as anomalous. Although stringent quality control mitigates these risks, residual artefacts can still influence the anomaly distribution.

Interpretation of performance is also shaped by protocol design choices. The dominance threshold adopted here (e.g., 10% weekly prevalence) is a practical heuristic and may not transfer cleanly to different pathogens, epidemiological contexts, or surveillance regimes. Likewise, lead time and PPV depend on the retrospective validation window and the operational definition of "dominance," meaning that these metrics may require contextual re-tuning for other applications.

Finally, novelty in sequence space does not necessarily imply phenotypic relevance. A lineage can be anomalous yet epidemiologically neutral, while variants that become important may arise through incremental changes that do not generate large reconstruction errors early on. For this reason, DeepAutoCoV should be viewed primarily as a

prioritization tool rather than a definitive classifier of threat, and its outputs are best interpreted alongside complementary evidence such as lineage growth rates, antigenic assays, clinical severity indicators, and structural or biophysical analyses. In addition, the current evaluation is retrospective; prospective deployment would face evolving reporting practices, changing testing policies, and shifting sequencing priorities, all of which could alter the balance between early detection and false-alarm rates.

Future Directions

Several extensions can strengthen both generalizability and practical utility. Methodologically, DeepAutoCoV could be enhanced through **continual learning** with explicit drift detection to stabilize anomaly thresholds over time. Incorporating additional viral proteins or whole-genome representations may capture broader determinants of fitness, while hybrid approaches that combine anomaly detection with explicit models of lineage growth or mechanistic proxies of viral fitness may improve discrimination between merely novel and truly expansion-prone lineages.

From a translational perspective, integrating DeepAutoCoV into genomic monitoring pipelines could enable routine generation of prioritized candidate lists for rapid laboratory follow-up. Linking anomalies to structural modeling, molecular simulations, or binding-affinity predictors may further improve biological interpretability and support actionable prioritization. Beyond SARS-CoV-2, the same principles could be adapted to other rapidly evolving viruses (e.g., influenza, RSV, or future emerging pathogens), provided that representation choices, dominance criteria, and retraining strategies are calibrated to the pathogen's evolutionary tempo and the surveillance ecosystem.

In summary, DeepAutoCoV provides a proof-of-concept for AI-driven early warning from genomic surveillance: it can surface weak signals of potential future dominance, offer interpretable sequence-level cues, and operate under diverse sequencing intensities. At the same time, its outputs must be interpreted in light of surveillance biases, concept drift, representation constraints, and the non-equivalence between genomic novelty and phenotypic impact—motivating integration with complementary epidemiological and experimental evidence in future deployments.

Chapter 4

4 SARITA: Predictive Generation of SARS-CoV-2 Spike Using LLMs

In the previous chapters, methods were presented to identify the key epidemiological variables to be considered during a pandemic, as well as how viral variants can influence the restrictions adopted by public health authorities to mitigate viral spread. Both quantitative analyses and an anomaly-detection framework were described to early highlight lineages with the potential to spread in the population, thereby informing preventive measures aimed at limiting their dissemination. This chapter addresses a complementary challenge within genomic surveillance.

Whereas approaches such as DeepAutoCov enable the detection of dominant lineages once they have already been sequenced, the question arises as to whether it is possible to move a step further and predict viral sequences that have not yet emerged, but are nonetheless plausible, based on past evolutionary data. The rapid development of transformer architectures and the introduction of the attention mechanism [116] have laid the foundation for Large Language Models (LLMs). These models, characterized by billions of parameters, have demonstrated remarkable generative capabilities, producing coherent text sequences from a simple instruction (i.e. prompt). While LLMs primary applications have been in natural language processing (NLP) tasks, such as text generation and question answering, LLMs have increasingly been investigated in biomedical contexts, including clinical decision support [117]. However, their potential in the biological domain, and more specifically in virology, remains largely underexplored [118]. For this reason, the possibility of

leveraging LLMs to learn from historical viral sequences and to generate evolutionarily plausible synthetic variants has been investigated. Such an approach would not merely react to viral evolution, but could anticipate it, offering a proactive strategy in genomic surveillance. Importantly, the unprecedented availability of millions of high-quality SARS-CoV-2 spike protein sequences in the GISAID repository during the COVID-19 pandemic provides the extensive dataset required for effective training of these models. To address this opportunity, this chapter introduces SARITA (SARS-CoV-2 RITA), a transformer-based generative LLM specifically designed to model the SARS-CoV-2 spike protein. SARITA is built on the GPT-3 architecture [119] and leverages continual learning to specialize on spike sequences. Unlike previous approaches, it is capable of generating full-length, biologically realistic synthetic sequences of the S1 subunit, starting from a short prompt. By doing so, SARITA aims to anticipate viral mutations that may arise in the future, thereby providing a valuable tool for genomic surveillance and vaccine development.

4.1. Motivation: toward proactive viral surveillance through sequence generation

The COVID-19 pandemic has underscored the need for timely and effective strategies to track and mitigate the spread of emerging viral variants. Traditional genomic surveillance frameworks are reactive in nature, as they rely on the sequencing of viral genomes after their emergence in the population (Figure 38, Figure 43). Although such strategies have proven crucial for detecting and classifying VOCs and informing public health responses, the inherent time lag between the appearance of a novel variant and its recognition as a potential threat limits the scope of preventive interventions [21].

This delay has significant implications. Vaccines, therapeutic agents, and containment measures are often adapted only after the widespread diffusion of new variants, which diminishes the ability to contain outbreaks at an early stage. Consequently, there is a pressing need to shift from reactive to proactive surveillance strategies capable of anticipating viral evolutionary trajectories.

To do so, reliable synthetic sequences could support the implementation of such proactive surveillance strategies. LLMs have demonstrated remarkable generative capabilities, which could be applied in virology [118]: instead of generating text, LLMs can be trained to

generate realistic viral protein sequences. By learning from the vast repositories of SARS-CoV-2 genomic data collected during the pandemic, LLMs may capture underlying evolutionary patterns and extrapolate plausible future variants that have not yet been observed.

The ability to generate synthetic but biologically realistic sequences would mark a paradigm shift in genomic surveillance. Such models could anticipate potential mutations before they are detected in the population, providing critical lead time for vaccine design, therapeutic development, and the implementation of targeted interventions. Moreover, the exploration of sequence space beyond the currently observed variants would allow public health systems to prepare for evolutionary scenarios that may never arise naturally, but which, if they did, could pose substantial risks.

This proactive approach to viral surveillance, grounded in generative modeling, thus represents a novel research direction at the intersection of artificial intelligence, genomics, and public health.

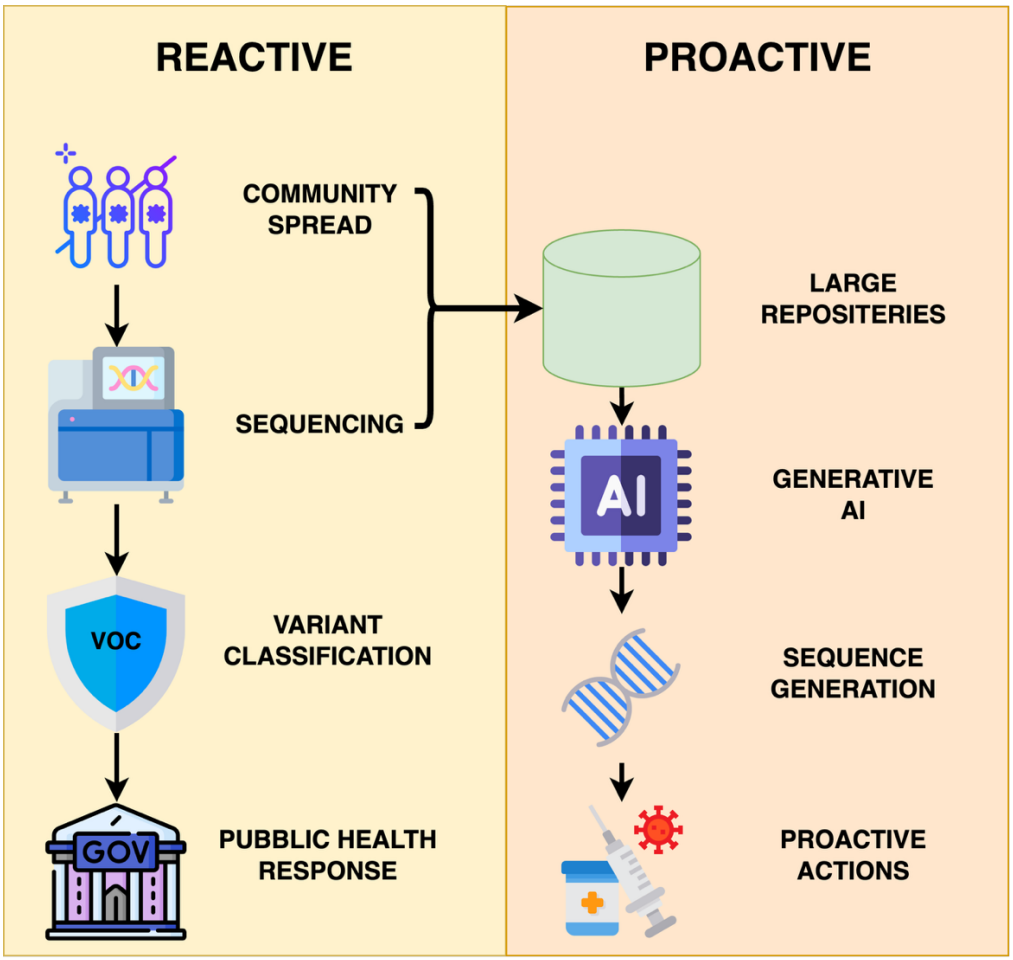


Figure 43: Comparison between reactive and proactive approaches in viral surveillance. The reactive framework (left) relies on community spread followed by sequencing, variant classification (e.g., designation as a Variant of Concern, VOC), and subsequent public health responses. In contrast, the proactive framework (right) leverages large repositories and generative AI to simulate and generate potential viral sequences in advance, enabling the development of vaccines or therapeutics and the implementation of proactive measures before widespread transmission occurs [2] .

4.2. SARITA architecture: decoder-only transformer and continual learning

The SARITA model has been conceived as a decoder-only transformer (Figure 44, Table 9), a design choice motivated by the autoregressive nature of protein sequence generation and the need to efficiently capture long-range dependencies within amino acid chains. In contrast to encoder–decoder models, which process both input and output sequences, a decoder-only transformer predicts each new token exclusively on the basis of the previously generated ones. This autoregressive paradigm is particularly suitable for protein design, as it closely reflects the biological process of peptide chain elongation, where each amino acid is appended sequentially and contextually to the growing polypeptide.

The architecture draws inspiration from GPT-3 [120, 119] and other generative protein models such as ProGen2 [121] and RITA [119], which have demonstrated the ability to produce structurally coherent and functionally plausible sequences. Within SARITA, protein sequences are first tokenized at the single–amino acid level using a Unigram tokenizer. Each amino acid residue is treated as an independent token, a strategy that preserves the granularity of biological information and avoids the potential distortions introduced by frequency-based segmentation methods such as Byte Pair Encoding. This choice ensures that essential biochemical properties and positional dependencies are maintained during training. To guarantee consistency across datasets, additional special tokens are included, such as **<PAD>** for sequence padding and **<EOS>** for termination, thereby standardizing sequence lengths and facilitating batch processing [119].

Each token is subsequently embedded into a high-dimensional vector space through a learned embedding matrix. These embeddings, initially inherited from the general-domain RITA model [119], are continuously refined during training through backpropagation, allowing the model to

encode rich contextual representations of amino acids within their evolutionary and structural environments. Positional information is incorporated using Rotary Positional Embeddings (RoPE) [122], which enhance the ability to represent relative distances between aa residues. This mechanism improves the encoding of both local motifs, such as glycosylation sites or loop regions, and long-range dependencies that are fundamental for protein folding and conformational stability.

The transformer decoder layers process these embeddings through self-attention mechanisms that operate across the full length of the sequence. This enables the model to learn global dependencies, including covariation patterns between distant aa residues, which are often indicative of structural constraints or evolutionary pressures [116]. In this way, SARITA captures not only the statistical properties of amino acid distributions but also the functional signatures embedded in viral evolution, such as recurrent mutations in immune-evasive regions of the receptor-binding domain.

To accommodate different computational requirements, we have developed SARITA in four scalable configurations, ranging from 85 million to 1.2 billion parameters (Figure 44). The smallest configuration (SARITA-S) comprises 12 transformer layers with an embedding dimension of 768, while the largest (SARITA-XL) includes 24 layers with an embedding size of 2048. Larger models consistently demonstrate superior performance in generating novel, biologically realistic sequences, owing to their enhanced capacity to capture complex evolutionary patterns. All configurations are capable of generating sequences of up to 1024 tokens, sufficient to cover the entire S1 subunit of the SARS-CoV-2 Spike protein.

A distinctive feature of SARITA lies in its training paradigm, which integrates continual learning (Figure 45, Table A8). Rather than training from scratch, SARITA builds upon RITA [119], a large protein language model pre-trained on hundreds of millions of protein sequences across different species. Continual learning (Figure 45) is then applied to adapt the general model to the specific evolutionary landscape of SARS-CoV-2. Over 100,000 high-quality Spike protein sequences from the GISAID database, collected up to early 2021, were curated and used for fine-tuning. Training was conducted over 10 epochs with a batch size of 8 for both training and evaluation. Checkpoints were saved at the end of each epoch, with validation performed at every step to monitor convergence. Optimization employed the Adam optimizer (β_1=0.9, β_2=0.999, ε=1e–8) together with a linear learning rate scheduler initialized at 5×10^{-5}. To balance efficiency and precision, mixed-precision training (fp16) was

applied through PyTorch's native Automatic Mixed Precision (AMP), taking advantage of CUDA-enabled GPUs. All experiments were performed on the University of Florida HiPerGator computing infrastructure, which provided six NVIDIA A100 GPUs, 100 GB of RAM, and 10 CPUs per task. This incremental training process enables the model to progressively acquire a specialized "lexicon" of viral mutations, shifting its latent space from generic protein representations to a highly specialized embedding of Spike S1 variability.

The continual learning strategy ensures that the model does not simply memorize previously observed patterns but develops the ability to generalize to unseen mutational contexts. As training progresses, SARITA becomes increasingly capable of generating realistic substitutions in highly variable regions, while maintaining conservation in functionally constrained motifs. This balance between variability and conservation is critical for generating synthetic sequences that are both biologically plausible and evolutionarily meaningful. Importantly, SARITA demonstrates the ability to reproduce mutational trends that were not present in its training data but emerged later in circulating variants, suggesting that the model has internalized latent evolutionary constraints and selection pressures.

The integration of decoder-only autoregressive modeling with domain-adaptive continual learning thus provides SARITA with the ability to simulate evolutionary processes with high fidelity. By combining scalable architectures, fine-grained tokenization, context-aware embeddings, and progressive specialization, SARITA represents an advancement in the application of large language models to viral genomics. Its generative capabilities highlight not only its utility as a predictive tool for anticipating future variants but also its potential role in informing vaccine design, therapeutic development, and proactive genomic surveillance.

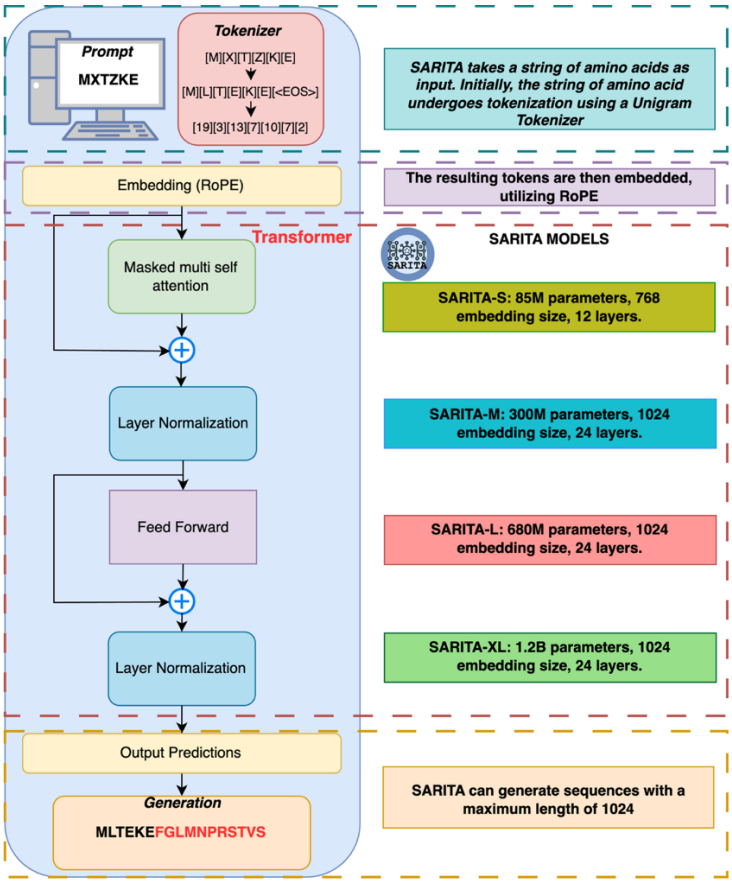


Figure 44: SARITA model architecture for protein sequence generation. Input amino acid strings are first tokenized using a unigram tokenizer and then embedded with rotary positional embeddings (RoPE). The embedded tokens are processed through a transformer architecture consisting of masked multi-head self-attention, feed-forward layers, and layer normalization. The model generates amino acid sequences as output, with a maximum sequence length of 1024. SARITA is available in different model sizes: SARITA-S (85M parameters, 768 embedding size, 12 layers), SARITA-M (300M parameters, 1024 embedding size, 24 layers), SARITA-L (680M parameters, 1024 embedding size, 24 layers), and SARITA-XL (1.2B parameters, 1024 embedding size, 24 layers) [24].

4.3. Training and Test Set Design: Filtering, Balancing, and Selection

The design of training and test sets (Figure 45) represents a critical step for ensuring the reliability, fairness, and generalizability of predictive models in viral genomics. Spike protein sequences were retrieved from the GISAID [123]. A total of 16,187,950 Spike protein sequences collected between December 2019, and November 2023 were considered. To

guarantee data quality, strict filtering criteria were applied in line with established standards for genomic curation. Only sequences meeting the following requirements were retained: (i) high-coverage genomes, defined as containing less than 1% undetermined nucleotides (Ns) and less than 0.05% unique amino acid mutations not observed in other database entries, and no insertions or deletions unless validated by the submitter; (ii) protein length of 1,273 ± 30 amino acids, consistent with a full-length Spike sequence; (iii) human host origin; (iv) assigned lineage using the Pango algorithm [95]; (v) absence of non-standard amino acid codes such as “X”, “B”, “Z”, “J”, “U”, and “O”; and (vi) availability of complete submission dates including year, month, and day [21]. After filtering, 4,952,389 sequences were retained as high-quality data.

To construct the training set, a temporal window was defined covering December 2019 to February 2021. This period was chosen to provide a sufficiently large yet controlled dataset for training while ensuring the exclusion of major Variants of Concern (VOCs) such as Delta and Omicron, which would later serve as unseen variants for validation. The resulting set comprised 793,732 sequences. Since the distribution of unique sequences was highly skewed, with certain variants represented by thousands of identical copies, a balancing procedure was applied. Specifically, the mean copy number of unique sequences (n = 25) was imposed as an upper threshold. Sequences occurring at higher frequencies were downsampled to this threshold, thereby preventing overrepresentation of dominant strains and reducing potential model bias toward their mutational signatures. This subsampling procedure reduced the Gini index of the sequence copy number distribution from 0.93 to 0.07, achieving a more uniform representation of genetic diversity [124]. After balancing, 107,017 sequences, corresponding to 32,376 unique ones, were selected for training. Ten percent of this dataset was reserved for internal validation during model optimization [125].

The test set was constructed from sequences collected between March 2021 and November 2023, encompassing 145,059 unique entries. Importantly, only Alpha, Beta, and Gamma lineages were present in the training set, while major subsequent VOCs such as Delta and Omicron, as well as Variants of Interest (VOIs) including Kappa, Iota, and Zeta, were exclusively allocated to the test set. This deliberate separation enabled the evaluation of the model’s capacity to generalize beyond its training data and to anticipate evolutionary trajectories of lineages that emerged after the training cutoff [21]. Such an approach aligns with the broader objective of forecasting Future Dominant Lineages (FDLs), whereby predictive frameworks are tasked not only with modeling past

evolutionary dynamics but also with anticipating mutational profiles that may acquire selective advantages in the population [21].

By constructing a training set representative of early viral evolution and a test set enriched in later, more divergent VOCs and VOIs, the model could be systematically assessed for its ability to predict the emergence of novel variants,a prerequisite for supporting proactive genomic surveillance and public health preparedness [47, 126].

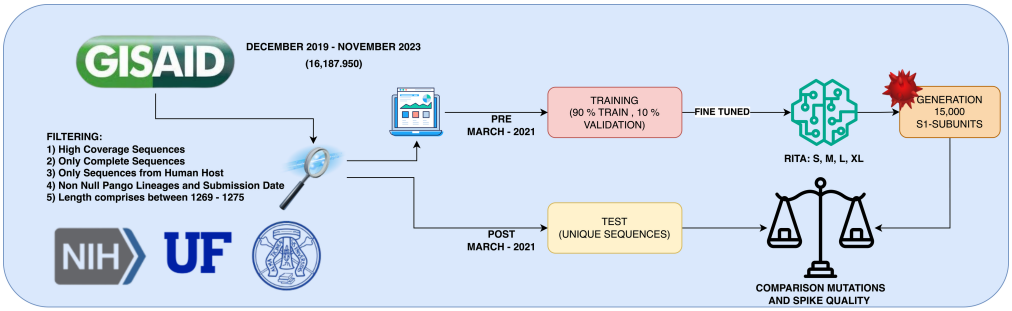


Figure 45: Data processing and model training workflow for SARS-CoV-2 sequence analysis. Sequences collected from GISAID between December 2019 and November 2023 (16,187,950 entries) were filtered to retain only high-coverage, complete sequences from human hosts with valid Pango lineages, submission dates, and lengths between 1269–1275 amino acids. The dataset was split into pre-March 2021 sequences for model training (90% training, 10% validation) and post-March 2021 sequences for testing with unique entries. Fine-tuned RITA models (sizes S, M, L, XL) were then used to generate 15,000 synthetic S1 subunits. Generated sequences were compared with real sequences to assess mutational patterns and spike quality [24].

4.4. Sequence Generation and Evaluation

Sequence generation (Figure 46) is formulated as an autoregressive completion task in which the model receives, as a fixed prompt, the 14–amino acid signal peptide that precedes the S1 subunit and is required to synthesize a full-length S1 sequence. The choice to generate only the S1 subunit is motivated by its markedly higher mutational plasticity compared to the more conserved S2 fusion machinery [127, 128]. All distinct signal-peptide prompts observed in the training data are enumerated (113 unique 14-mers), and for each prompt the model generates a cohort of sequences of target length 686 amino acids (positions 15–686 inclusive when counted from the start of Spike), thereby enabling prompt-wise and aggregate assessment under a controlled input regime

[127, 128] Generation proceeds by iterative sampling from the learned conditional distribution over amino acids until the prescribed length is reached or an end-of-sequence token is produced; the prompt is included in the reported sequence to preserve biological context. This protocol is applied uniformly to all SARITA configurations (Figure 46).

The Evaluation of synthetic S1 subunit is designed to probe three complementary facets of generative behavior: (i) sequence quality (syntactic and length validity and similarity to the Wuhan S1 reference), (ii) evolutionary plausibility measured as similarity to real-world sequences withheld from training, and (iii) biological plausibility of single–amino acid substitutions assessed through positional statistics and structure-based energetics. The train/test temporal split ensures that VOC/VOIs emerging after the training cutoff are represented only in the evaluation data, supporting a forward-looking assessment of generalization.

Quality control adopts two criteria. First, sequences are screened for invalid symbols to verify adherence to the 20 canonical amino acids. Second, a length filter centered on the S1 target length (686 amino acids, with a 10 aa tolerance) is applied. To contextualize realism with respect to the ancestral reference, global alignments against the Wuhan S1 are computed using Needleman–Wunsch, and similarity is quantified with PAM30 and BLOSUM62 scores [110], which are suited to short evolutionary distances and emphasize conservative substitutions. PAM30 is optimized for very short evolutionary distances (i.e., recent substitutions), making it sensitive to small numbers of changes close to the Wuhan reference, whereas BLOSUM62 provides a widely used, more general-purpose substitution model derived from conserved protein blocks and remains robust across a broader range of divergence. Using both matrices therefore serves as a sensitivity check: concordant results indicate that similarity estimates are not driven by a single scoring scheme, while discrepancies can flag sequences enriched for rare or non-conservative substitutions. These operations provide a baseline indication that synthetic sequences remain within a credible mutational neighborhood of the reference.

Evolutionary plausibility with respect to circulating diversity is assessed by comparing each generated sequence to all sequences in the separated test set (March 2021–November 2023). For each synthetic sequence, the Levenshtein distance (LD) is computed against the entire test set to capture best-match similarity and overall dispersion. The LD is defined as a string metric that computes the minimal number of single-

character edits (insertions, deletions, or substitutions) required to transform one sequence into another.

In parallel, alignment-based similarity (best PAM30/BLOSUM62 against the test set) is computed to complement edit-distance analyses with substitution-aware scoring. To quantify spurious divergence, the False Mutation Rate (FMR) is defined as the fraction of mutations present in the synthetic sequence (S) that are not shared with its best-aligned (M) test counterpart, normalized by the total mutations of M relative to Wuhan; this metric penalizes unrealistic or unsupported amino acid changes (Equation 10).

$$FMR = \frac{Number\ of\ mutation\ in\ S\ -\ Number\ of\ mutation\ shared\ between\ S\ and\ M}{Total\ number\ of\ mutation\ in\ M} \quad (10)$$

Single-point mutation evaluation examines whether the model places substitutions at biologically plausible positions and with realistic positional frequencies. For each S1 position, mutation counts are tallied across synthetic sequences and summarized via Gaussian kernel density estimation (bandwidth set to 0.1) to obtain a smooth positional mutation density; base-10 logarithmic transformation is applied to stabilize heavy-tailed distributions [126]. Concordance between synthetic and empirical mutation landscapes (training-era and post-cutoff test data) is quantified with Pearson correlation, enabling a position-wise comparison of mutational pressure before and after the training horizon.

To further interrogate functional plausibility, structure-based analyses estimate the change in binding free energy (ΔΔG) induced by single substitutions within the receptor-binding domain (RBD) for two protein–protein interactions of interest: Spike–ACE2 and Spike–neutralizing antibody [129]. Predictions are obtained with mmCSM-PPI [130] using experimentally determined complexes (ACE2-bound Spike, PDB 7DF4 [131] ; RBD bound to the tixagevimab/cilgavimab antibody cocktail, PDB 8D8Q [132]). Mutations predicted to increase ACE2 affinity and/or reduce antibody binding are flagged as high-priority from a surveillance perspective. Molecular visualization is performed in PyMOL [133] to map substitutions onto the three-dimensional interface context, aiding qualitative interpretation of energetic predictions.

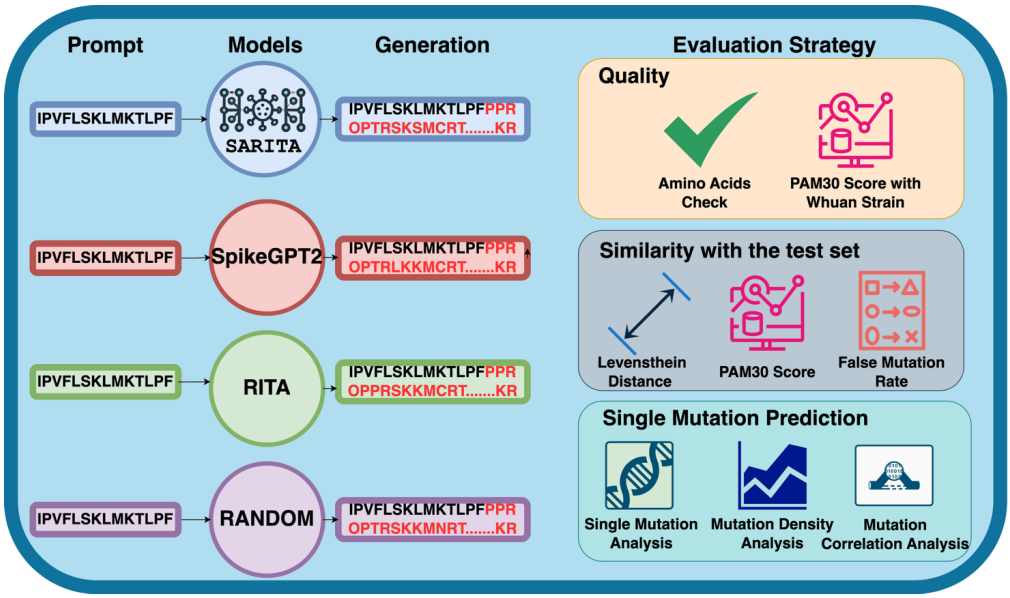


Figure 46: Comparative framework for evaluating generative models of SARS-CoV-2 spike protein sequences. A fixed amino acid prompt was used as input for SARITA, SpikeGPT2, RITA, and a Random baseline model, with each model generating novel sequence continuations. Generated outputs were assessed through a multi-step evaluation strategy: (i) Qualityassessment by verifying amino acid validity and computing PAM30 similarity scores with the Wuhan reference strain, (ii) Similarity with the test set via Levenshtein distance, PAM30 score, and false mutation rate, and (iii) Single mutation prediction including single mutation analysis, mutation density analysis, and mutation correlation analysis. This pipeline enabled a systematic comparison of generative models in terms of biological plausibility, sequence similarity, and mutational relevance [24].

4.5. Model Comparison

For benchmarking, SARITA was contrasted with two classes of generative baselines (Figure 46): (i) three non-parametric random generators designed to contextualize task difficulty, and (ii) state-of-the-art decoder-only protein language models commonly used for sequence generation.

Non-parametric random generators

To contextualize task difficulty and calibrate performance in the absence of learned structure, three random baselines were employed. A fully random sampler ("Rand") draws each residue uniformly from the 20

canonical amino acids, providing a composition-agnostic floor. A frequency-matched sampler ("Rand-Fr") samples each position according to empirical amino-acid marginals estimated on the training corpus, preserving global composition while discarding positional and higher-order dependencies. A PAM-guided sampler ("Rand-PAM30") proposes substitutions in proportion to PAM30 substitution probabilities, approximating short-range evolutionary drift without any contextual conditioning. These baselines establish conservative lower bounds for realism and context awareness against which learned models can be meaningfully compared.

Decoder-only protein language models

To enable a coherent and fair comparison across SARITA and general-purpose and SARS-CoV-2–specialized protein language models, this evaluation considers RITA, SpikeGPT2, and ProGen2-Base. The RITA family [119] constitutes a strong domain-agnostic baseline: four autoregressive, decoder-only transformers (S/M/L/XL; up to ~1.2B parameters) trained on >280M sequences drawn from UniRef100 [134], MGnify [135] and Metaclust [136]. Sequences are tokenized at single-residue resolution with a Unigram tokenizer, and generation proceeds left-to-right under causal (masked) self-attention. RITA is used both as a competitive baseline and as the initialization backbone for SARITA, isolating the incremental benefit of domain adaptation over a strong generalist prior. Complementing this broad pretraining, SpikeGPT2 [126],derived from ProtGPT2 pre-trained on UniRef50 [134] and comprising ~738M parameters,is fine-tuned on ~20,000 Spike sequences to align its generative distribution with coronavirus-specific regularities; although prior work often evaluates it on the receptor-binding domain (RBD), the fine-tuning spans the full-length Spike, providing a focused comparator shaped by SARS-CoV-2 sequence statistics. Finally, ProGen2-Base ([137]; ~764M parameters) represents the ProGen2 family of decoder-only models that scale to 6.4B parameters and are trained on >1B sequences aggregated from UniRef90 [134], BFD30 [138], and immune-repertoire resources (e.g., OAS). ProGen2-Base offers a high-capacity yet computationally tractable comparator with demonstrated cross-domain protein design performance. Taken together, these choices bracket the spectrum from broad protein-language pretraining (RITA, ProGen2-Base) to virus-specific adaptation (SpikeGPT2), enabling a unified and fair comparison between SARITA and the state-of-the-arts of general protein-language models and domain-specific language models.

4.6. Results

Sequence quality assessment: validity, length, biological realism

Sequence quality was first assessed in terms of syntactic validity and length correctness on the full set of synthetic S1 subunits generated from the 113 distinct 14–amino-acid signal peptide prompts (100 continuations per prompt, 11,300 sequences per model). A sequence was deemed valid if it contained only canonical amino acids (i.e., no "X", "B", "Z", "J", "U", or "O"), and length-correct if it matched the expected S1 span of 686 ± 10 aa residues. Under these criteria, SARITA models produced high-quality sequences in at least 97% of generations, while the RITA baselines achieved 100% validity and length compliance. In contrast, only 3% of SpikeGPT2 outputs met the quality definition; moreover, none of its sequences matched the S1 target length, with a median length of 2,003 aa residues (~192% longer than S1), indicating systematic length drift in that model's continuations. These outcomes confirm that SARITA reliably preserves basic biochemical syntax and architectural constraints at inference time (Table 9; Figure A5).

Biological realism (Figure 47) was evaluated by alignment-based similarity to the Wuhan reference S1 using Needleman–Wunsch with PAM30 and BLOSUM62 scoring matrices, which are suited to short-distance evolutionary changes in closely related proteins (Subchapiter 4.4) . Across SARITA variants, median PAM30 scores ranged from 5,191 (SARITA-S) to 5,278 (SARITA-M), substantially exceeding those of all competitors (e.g., 661 for Rand-PAM30; 2,062 for SpikeGPT2). Higher scores indicate closer evolutionary plausibility relative to the reference backbone and, consequently, a lower incidence of implausible substitutions or indels. Results with BLOSUM62 mirrored these trends, reinforcing that SARITA maintains realistic substitution patterns while avoiding excessive divergence from known biology (Figure 47; Figure A6).

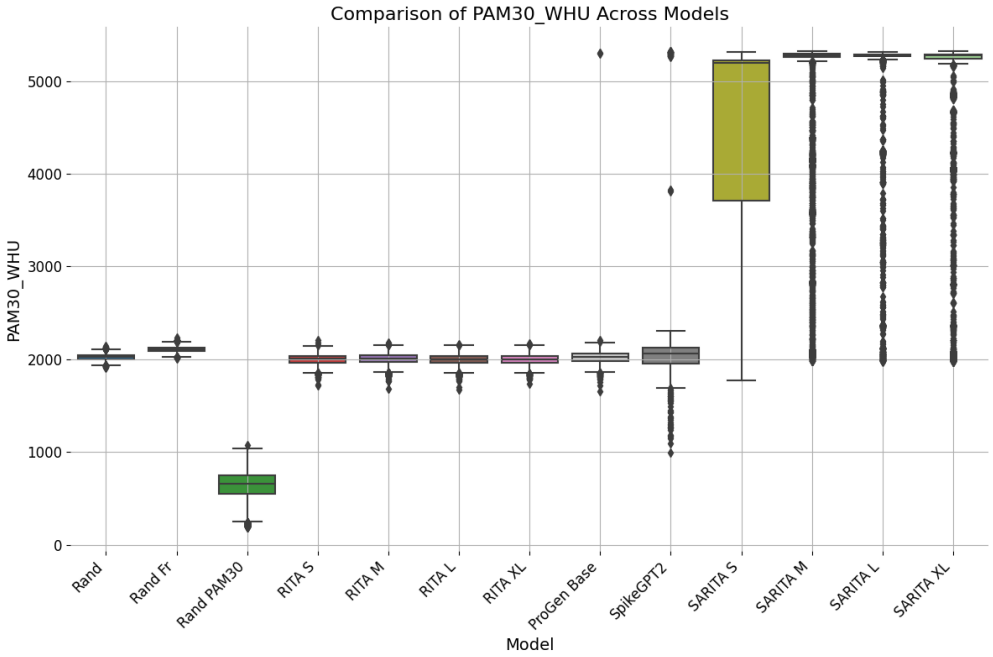


Figure 47: Comparison of PAM30 similarity scores with the Wuhan reference strain (PAM30_WHU) across different generative models. Boxplots summarize the distribution of PAM30 scores for Random baselines (Rand, Rand Fr, Rand PAM30), RITA models (S, M, L, XL), ProGen Base, SpikeGPT2, and SARITA models (S, M, L, XL). SARITA variants, particularly the larger models (M, L, XL), consistently achieve higher PAM30 scores, indicating greater biological realism and closer alignment to the Wuhan strain compared to baseline and alternative models. Outliers reflect individual generated sequences with divergent similarity levels [24].

Table 3: Comparison between SARITA and the competing models, in terms of i) number of model parameters, ii) computational resources used for prediction, and iii) number (percentage) of high-quality generated sequences

Model	# of parameters	Computational resources	# High quality sequences
Rand	-	1 Core, 1.41MB	11,300 (100%)
Rand FR	-	1 Core, 1.41MB	11,300 (100%)

Rand PAM30	-	1 Core, 243MB	11,300 (100%)
RITA S	85 million	1 A100, 1 Core, 6.06 GB	11,300 (100%)
RITA M	120 million	1 A100,1 Core, 6.20 GB	11,300 (100%)
RITA L	300 million	1 A100,1 Core, 6.07 GB	11,300 (100%)
RITA XL	1.2 billion	1 A100,1 Core, 6.05 GB	11,300 (100%)
SpikeGPT2	738 million	1 A100,1 Core, 6.07 GB	339 (3%)
SARITA S	85 million	1 A100,1 Core, 6.07 GB	11,074 (98%)
SARITA M	300 million	1 A100,1 Core, 6.22 GB	11,074 (98%)
SARITA L	600 million	1 A100,1 Core,	10,961 (97%)

		6.05 GB	
SARITA XL	1.2 billion	1 A100,1 Core, 6.05 GB	11,187 (99%)

Evolutionary similarity to real-world data: Levenshtein, PAM30, FMR

Evolutionary proximity between synthetic sequences and real-world SARS-CoV-2 evolution was quantified against a held-out test set comprising unique GISAID S1 sequences collected from March 1, 2021 to November 8, 2023, including major VOCs (Delta, Omicron) and VOIs (e.g., Kappa, Iota, Zeta). For each of the 11,300 continuations per model (113 distinct 14-aa prompts × 100 generations), similarity was assessed using three complementary metrics: the Levenshtein distance (LD) as an edit-based measure of minimal sequence changes, the PAM30 alignment score as an evolutionary similarity proxy at short distances, and the False Mutation Rate (FMR) as the fraction of spurious substitutions not supported by the best-matching test sequence. The LD analysis retained both minimum and distributional summaries per generated sequence, PAM30 was computed by global alignment to the best test match, and FMR was defined with respect to Wuhan-referenced mutations in the generated and best-aligned test sequences.

By LD (Table 10), SARITA produced sequences that are routinely very close to post-cutoff real data. Minimum LD values below 10 were observed for the vast majority of outputs across model sizes (e.g., 93.2% for SARITA-S; ~98% for SARITA-M/L/XL), and numerous generations achieved LD < 4 (e.g., 31.7% for SARITA-S, rising further in larger models). In contrast, random baselines and RITA variants never reached a best LD < 500 against the test set, and only 1.8% of SpikeGPT2 generations crossed that threshold, underscoring the difficulty of matching real evolutionary trajectories without domain adaptation. Qualitatively, SARITA generated S1 subunits differing by only 1–2 amino acids from Delta/Omicron sublineages in the test set, indicating a capacity to anticipate variant-like constellations beyond the training horizon (Table 10).

Table 4: Percentage of generated sequences with LD < 4, 10, 100, 500, 700, 1500, and 2000 against the best aligned test sequence. SARITA outperforms other competive approaches.

Model	LD<4	LD< 10	LD<100	LD<500	LD<700	LD<1500	LD<2000
Rand	0%	0%	0%	0%	100%	100%	100%
Rand Fr	0%	0%	0%	0%	100%	100%	100%
Rand PAM30	0%	0%	0%	0%	100%	100%	100%
RITA S	0%	0%	0%	0%	100%	100%	100%
RITA M	0%	0%	0%	0%	100%	100%	100%
RITA L	0%	0%	0%	0%	100%	100%	100%
RITA XL	0%	0%	0%	0%	100%	100%	100%
ProGen2	2.94%	2.94%	2.94%	2.94%	100%	100%	100%
SpikeGPT2	0%	0%	0%	1.8%	5.6%	75%	100%
SARITA S	31.7%	93.2%	99.3%	100%	100%	100%	100%
SARITA M	86.%	97.9%	98.9%	100%	100%	100%	100%
SARITA L	63.5%	98.4%	98.4%	100%	100%	100%	100%
SARITA XL	53.7%	98.6%	98.7%	100%	100%	100%	100%

Alignment-based similarity corroborated these findings. When comparing each synthetic sequence to its best test counterpart, SARITA achieved the highest median PAM30 scores among all methods

(approximately 5,210–5,302 across sizes), whereas baselines yielded substantially lower medians (e.g., ~670 for Rand-PAM30 and ~2,102 for SpikeGPT2). Because PAM30 emphasizes short-range evolutionary substitutions, higher scores indicate that SARITA's outputs adhere closely to realistic mutational neighborhoods observed in circulating viruses, rather than drifting toward implausible substitution patterns.

FMR provided a mutation-level realism check by penalizing extra substitutions absent from the best-aligned test sequence. SARITA exhibited markedly lower FMR than all competitors, with the lowest median observed for SARITA-L (0.5), while the best baseline remained far higher (≈0.98). Beyond the rate itself, SARITA maintained a restrained mutation load aligned with real data (median ~14–38 mutations per sequence across sizes) compared with SpikeGPT2's excessive burden (median ~438), indicating that SARITA balances novelty with evolutionary plausibility. Notably, SARITA-M/L/XL generated subsets of sequences with very low FMR (≈0.5–1%), closely mirroring VOC/VOI backbones in the test set.

Collectively, the LD, PAM30, and FMR analyses converge on the same conclusion: SARITA captures salient evolutionary constraints of SARS-CoV-2 S1 and generates sequences that are both edit-close to post-cutoff variants and compositionally consistent with short-distance evolutionary models, while minimizing unsupported mutations. This triad of metrics establishes the evolutionary realism of SARITA's generations and substantiates its suitability for proactive genomic surveillance.

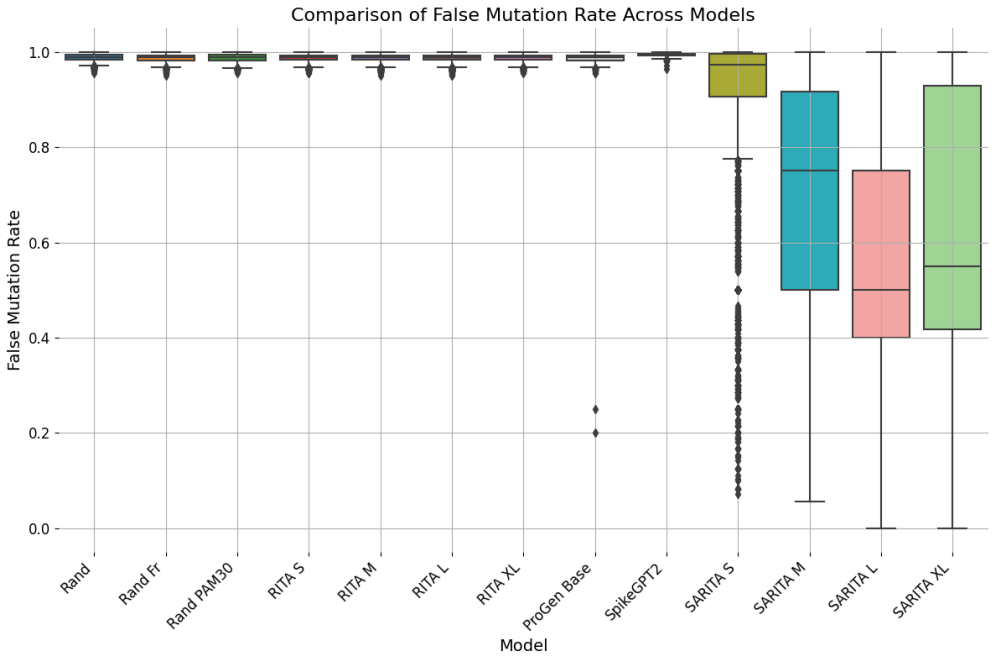


Figure 48: Comparison of false mutation rates across generative models. Boxplots show the distribution of false mutation rates for Random baselines (Rand, Rand Fr, Rand PAM30), RITA models (S, M, L, XL), ProGen Base, SpikeGPT2, and SARITA models (S, M, L, XL). While most baseline and RITA models exhibit near-maximal false mutation rates, SARITA models,especially the larger ones (M, L, XL) ,demonstrate reduced variability and significantly lower median false mutation rates, indicating improved biological plausibility in sequence generation. Outliers represent sequences with anomalous mutation patterns [24]

Reproduction of VOC/VOI Signature Mutations and Structure-Based ΔΔG Analysis of ACE2 and Monoclonal Antibody Binding

The ability of SARITA to recapitulate real-world evolutionary signatures was evaluated by inspecting whether hallmark substitutions of major Variants of Concern (VOCs) and Variants of Interest (VOIs) emerged in synthetic S1 subunits generated from pre–March 1, 2021 training. SARITA reproduced several defining or highly recurrent mutations that later characterized circulating lineages, despite their absence from the training set. Examples include L212I, a signature of Omicron lineages [139]; R158L, reported in Iota [140]; T19L and T95P, which are shared across Delta and Iota variants [141]; and E406K, associated with reduced susceptibility to therapeutic monoclonal antibodies [142]. The model also generated N501Y, a substitution known to enhance ACE2 engagement and transmission, and G446S, a class 3 antibody–evasive RBD mutation observed in BA.1/BA.2.75/XBB/BQ.1 lineages. These substitutions appeared within coherent constellations and at positions of high empirical variability, indicating that SARITA internalized latent constraints that govern adaptive change in S1.

To test functional plausibility at the structure level, single-mutation effects on protein–protein interactions were estimated with mmCSM-PPI using experimentally resolved complexes for Spike–ACE2 (PDB: 7DF4, Figure 49A) and Spike RBD–tixagevimab/cilgavimab (PDB: 8D8Q, Figure 49B). Across SARITA variants, the predicted ΔΔG distributions for Spike–ACE2 centered within experimentally supported ranges for circulating VOCs, with means of ˜0.66 kcal/mol (e.g., 0.662 [0.33, 0.90] for SARITA-L; 0.669 [0.335, 0.90] for SARITA-S). For the antibody complex, ΔΔG means were slightly higher, ˜0.85 kcal/mol (e.g., 0.852 [0.48, 1.14] for SARITA-L; 0.859 [0.48, 1.14] for SARITA-S). Importantly, restricting the analysis to novel substitutions absent from both training and test sets yielded comparable ΔΔG ranges, supporting the biological plausibility of previously unseen mutations proposed by the model. The predominance

of mildly positive ΔΔG values indicates a trend toward moderate destabilization of binding interfaces, consistent with immune-evasive or affinity-modulating effects observed in real variants.

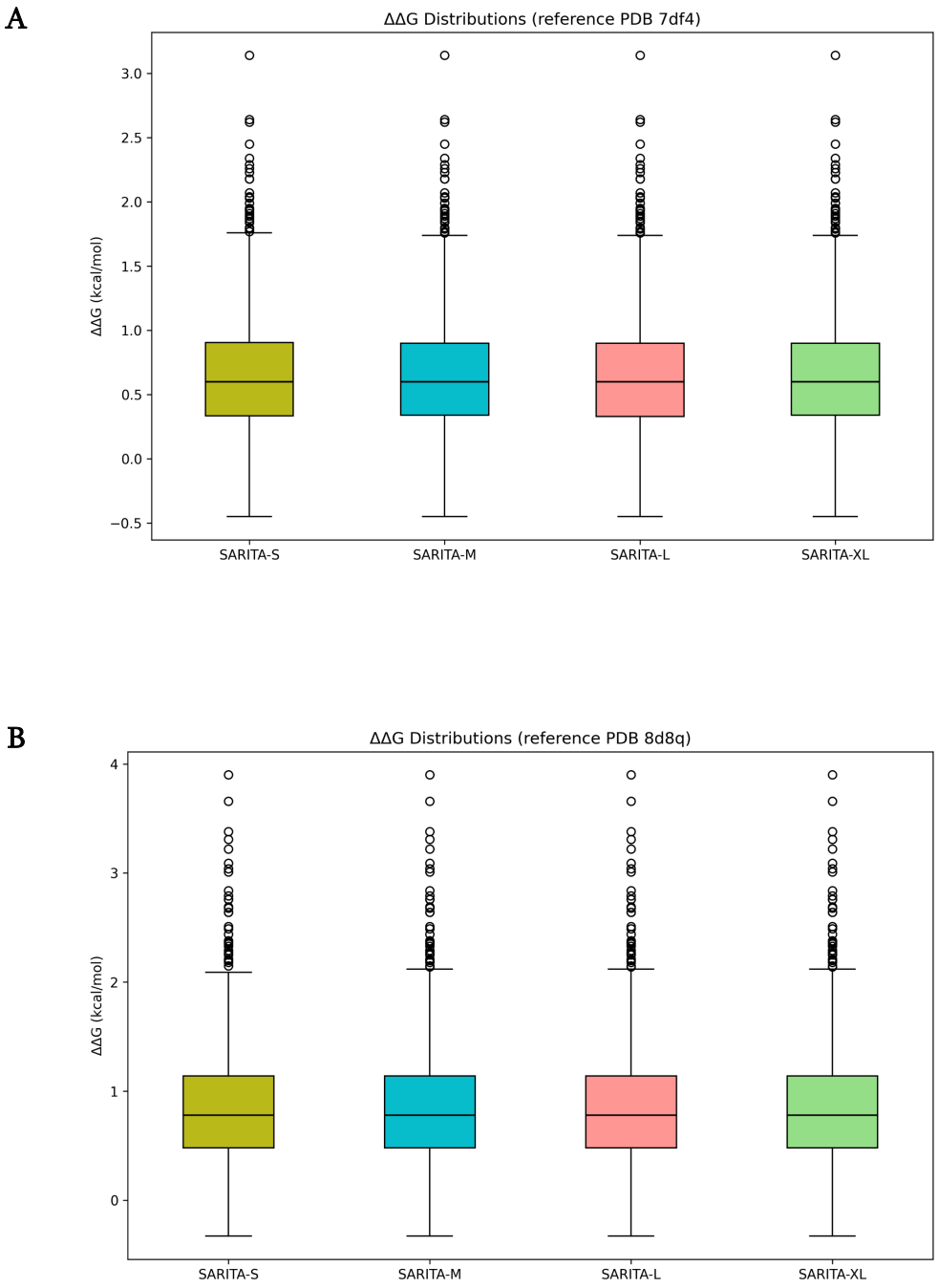


Figure 49: Boxplot comparison of predicted binding affinity changes (ΔΔG, in kcal/mol) for all mutations generated by SARITA models across four sizes (S, M, L, XL). Each distribution represents predicted ΔΔG values computed using mmCSM-PPI with two different protein complexes as

structural references. (A) Predictions based on the spike protein in complex with the human ACE2 receptor (PDB: 7df4), modeling potential changes in receptor binding affinity. (B) Predictions based on the SARS-CoV-2 spike protein complexed with the monoclonal antibody cocktail (PDB: 8d8q), modeling potential immune escape. All SARITA variants yield comparable ΔΔG distributions, suggesting consistent generation of plausible mutations across model sizes [24].

These results show that SARITA not only reconstructs lineage-defining substitutions across VOCs/VOIs from a pre-cutoff training regime, but also proposes structurally credible mutations whose predicted effects on ACE2 and therapeutic antibody binding fall within empirically observed regimes. The reproduction of key mutations and structure-level ΔΔG profiles substantiates the evolutionary realism of SARITA's generations and supports their use in proactive genomic surveillance (Figure 43) and candidate-mutation triage.

Mutation density and positional analysis: SARITA vs empirical data

Positional mutation patterns in SARITA generations were compared with empirical SARS-CoV-2 S1 data to assess whether the model reproduces where mutations occur along the sequence, not only how many. For each model, 11,300 synthetic S1 subunits were generated, and per-position mutation frequencies were computed relative to the Wuhan S1 reference, matching the analysis window used for empirical sequences (positions 15–686). To obtain smooth positional profiles, mutation frequencies were transformed on a log10 scale and mapped to continuous curves using Gaussian kernel density estimation (KDE) with a smoothing parameter of 0.1, which balances local resolution and noise [126]. The same procedure was applied to the training (pre–March 1, 2021) and test (March 1, 2021–November 8, 2023) sets, enabling direct SARITA-vs-empirical comparisons.

Visual inspection of KDE curves showed that SARITA closely recapitulates the empirical distribution of positional variability across S1. In particular, SARITA-M (Figure 50) captured both the pre-cutoff (training) and post-cutoff (test) mutation-density landscapes with high fidelity, indicating that the model internalizes positional constraints and variability hotspots that later characterized real-world evolution. Quantitatively, the mean squared error (MSE) between SARITA-generated and test-set density profiles was on the order of 10^{-7} across

model sizes (S, M, L, XL), with SARITA-M attaining the lowest deviation among the four variants (reported values: ~-8.9×10^{-8}, -6.6×10^{-8}, -7.2×10^{-8}, -8.5×10^{-8} for S, M, L, XL, respectively; small negative values reflect numerical rounding around zero). These results support a close match between SARITA's positional mutation signal and that observed in circulating viruses.

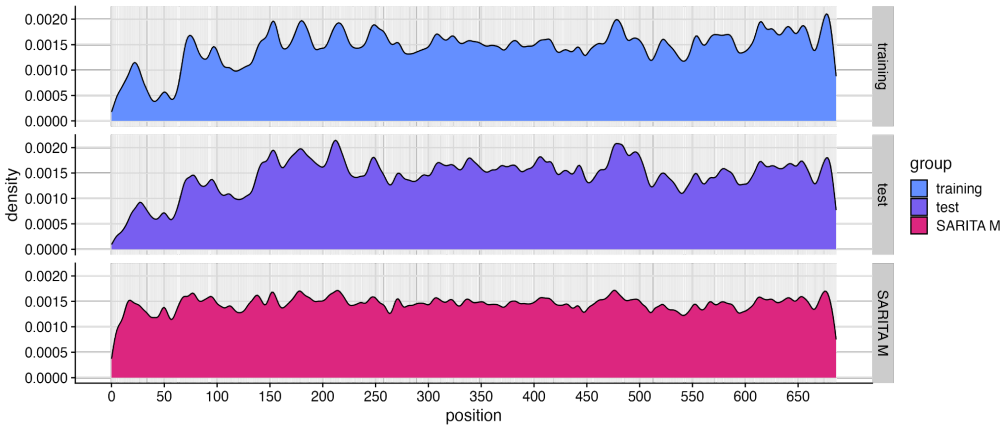


Figure 50: Gaussian kernel density estimate plot. The observed distribution of variants across positions in the SARS-CoV-2 S1 Subunit both pre- (training) and post-February 2021 (test) is well approximated by SARITA M [24].

Correlation analyses further corroborated this agreement. Pearson's correlation between per-position mutation frequencies in SARITA-M and the test set reached $r \approx 0.70$ ($p < 2.2\times10^{-16}$), indicating a strong positive association in which mutations tend to occur along S1 (Figure 51); analogous results for the other sizes are reported in the Appendix (Table A9). Together with the KDE/MSE findings, this evidence shows that SARITA reproduces the empirical spatial organization of mutational pressure across S1, rather than distributing substitutions uniformly or at implausible sites.

A B

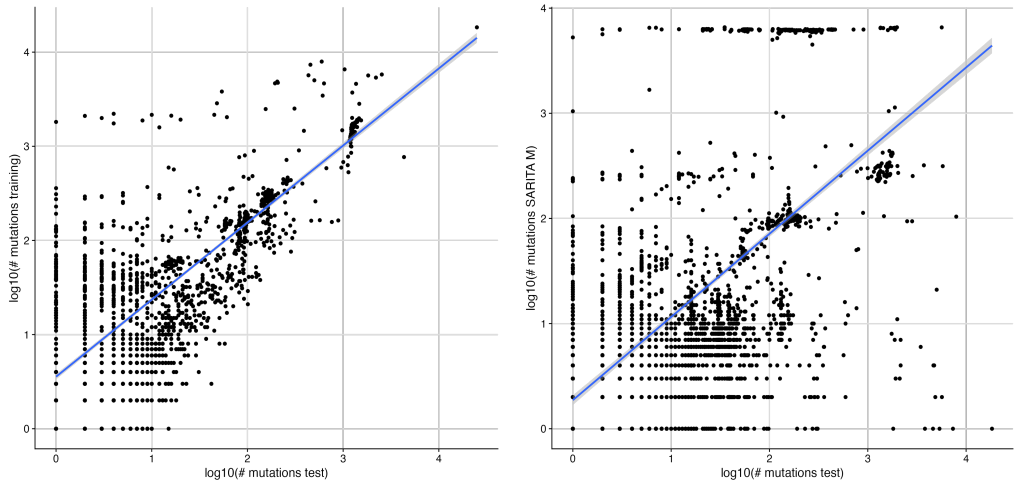


Figure 51: Correlation of S1 subunit mutations. A) Correlation of mutation frequency in logarithmic scale between training (x-axis) and test (y-axis) set. B) Correlation of mutation frequency in logarithmic scale between SARITA-M sequences (x-axis) and test (y-axis) set [24].

Complementary analyses on the composition of generated mutations (Figure 52) indicate that the fraction of **novel** substitutions (absent from both training and test sets) increases with model size (≈48% in SARITA-S, ≈55% in SARITA-M, ≈54% in SARITA-L/XL) while maintaining alignment-based plausibility: the PAM30 score distribution of generated mutations is closely matched to that of natural mutations (Figure 53) (median difference ~1 point). This pattern suggests that SARITA explores new mutational neighborhoods without sacrificing positional realism, a desirable property for proactive surveillance where novelty must remain biologically credible.

The positional density and correlation results demonstrate that SARITA not only learns how many mutations to introduce but also where to place them along the S1 subunit in ways consistent with observed viral evolution before and after the training cutoff. This positional fidelity underpins subsequent findings on evolutionary similarity and key-mutation recapitulation, strengthening the case for SARITA as a realistic generator of candidate S1 variants.

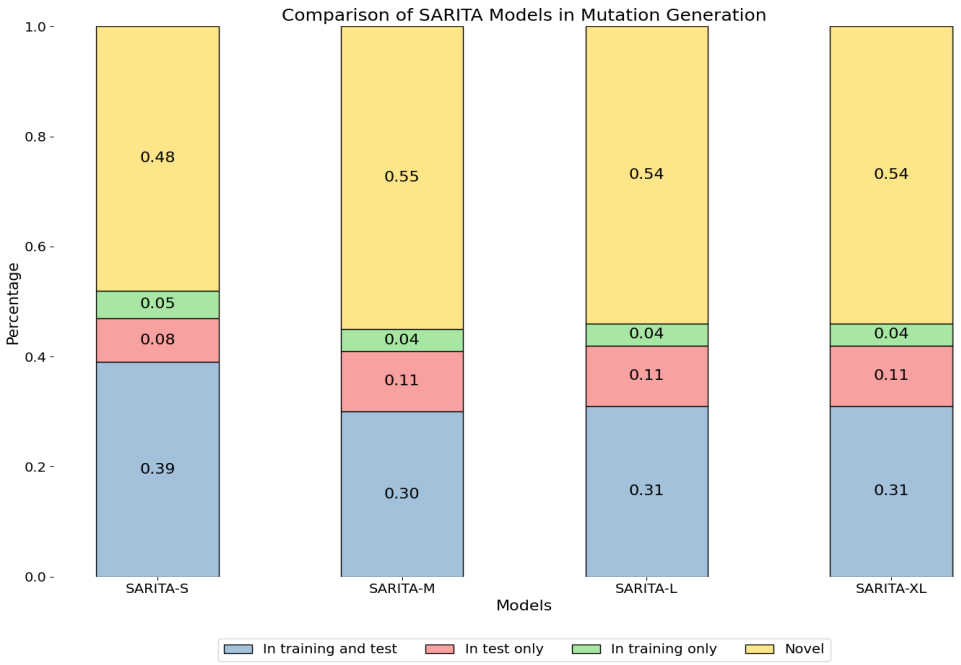


Figure 52: Comparison of SARITA models in mutation generation. The stacked bar chart represents the distribution of mutations across four SARITA models. Each bar is divided into four categories: present in both training and test sets (blue); present in test set only (red); present in training set only (green); and novel, i.e., not present in the training nor the test sets (yellow) [24].

4.7. Discussion and Conclusion

SARITA demonstrates that a domain-adapted, decoder-only transformer can generate full-length SARS-CoV-2 Spike S1 sequences that are valid, length-accurate, and evolutionarily coherent with post-cutoff viral diversity, thereby supporting the use of generative LLMs as proactive instruments for genomic surveillance and mutation triage. The model preserves biochemical syntax and architectural constraints at inference time,most generations satisfy strict validity and 686 ± 10 residue length criteria,while achieving high alignment similarity to circulating viruses by PAM30/BLOSUM scoring. Evolutionary proximity to a large, held-out test set is further evidenced by very small edit distances (Levenshtein) to best matches, together with restrained False Mutation Rates that penalize substitutions unsupported by real sequences, indicating that SARITA proposes edits within empirically traversed regions of sequence space rather than drifting toward implausible constellations. Positional analyses show that the model learns where to mutate, not only how much: mutation-density profiles along S1 closely

track both pre- and post-cutoff landscapes, with strong per-position correlations, consistent with internalization of selective constraints across NTD, RBD and CTD regions. Beyond global similarity, SARITA reproduces lineage-defining substitutions that later characterized VOCs/VOIs despite training on pre-March-2021 data (e.g., L212I, R158L, T19L, T95P, E406K), and also yields canonical changes such as N501Y and G446S associated with ACE2 affinity modulation and antibody evasion. Structure-based plausibility analyses using ΔΔG estimates on Spike–ACE2 and Spike–monoclonal-antibody complexes fall within experimentally supported regimes for circulating variants, including when restricted to novel substitutions absent from both training and test sets, suggesting that SARITA's innovations remain compatible with biophysical constraints. Two design choices appear central to these outcomes: first, an autoregressive decoder-only formulation, conditioned on the conserved 14-amino-acid signal peptide, which enables stable full-length continuation while maintaining long-range epistasis across S1; second, domain-adaptive continual learning from a general protein LLM to $>10^5$ curated S1 sequences, which shifts latent representations from generic protein regularities to SARS-CoV-2-specific evolutionary signals, concentrating novelty where natural selection permits change and preserving conservation elsewhere. Comparisons with non-learned (uniform, frequency-matched, PAM-guided) and learned baselines (RITA family, ProGen2-Base, SpikeGPT2) indicate consistent improvements across validity/length, Levenshtein/PAM30/BLOSUM similarity, and FMR, while scaling SARITA from small to larger sizes increases the share of novel yet alignment-plausible substitutions without degrading realism.

Limitations should be acknowledged. Model behavior is coupled to the quality and representativeness of available sequences; residual biases, lineage imbalances, and metadata inaccuracies in public repositories can propagate into the learned distribution despite stringent filtering and down-sampling. The current study validates sequence generation only for the Spike S1 subunit of SARS-CoV-2, a biologically motivated choice given S1's higher mutational plasticity and dominant roles in receptor binding and antigenicity, and a methodological choice compatible with the 1024-token context window. Consequently, these results do not directly establish performance for S2, non-Spike proteins, or other pathogens without further adaptation and re-validation. Likewise, the validation strategy primarily assesses sequence-level and proxy functional plausibility: realism is quantified through alignment and edit-distance proximity to observed post-cutoff diversity, positional mutation-density concordance, and ΔΔG predictions on a small number of fixed structural

complexes. These proxies cannot fully capture within-host fitness, transmission advantage, glycan shielding, conformational dynamics, or polyclonal immune pressure, and they do not substitute for experimental phenotyping or prospective deployment. Finally, the evaluation is retrospective with respect to the March 2021 cutoff; demonstrating operational utility will require time-stamped prospective evaluation within a live surveillance pipeline.

Within these bounds, the findings support a clear outlook. Incorporating immunological measurements (escape maps, epitope repertoires) could sharpen identification of high-risk substitutions and guide antigen design. Coupling sequence generation with richer structural modeling of whole complexes and conformational ensembles would strengthen functional interpretation beyond single-mutation ΔΔG estimates. Extending continual learning to S2 and other viral proteins, and to additional respiratory pathogens, would test portability and expose model assumptions. Embedding SARITA outputs into phylodynamic and epidemiological frameworks could enable early-warning scores that combine sequence-level risk with population-level spread dynamics. Taken together, the evidence indicates that a domain-adapted decoder-only LLM, trained via continual learning on curated S1 data, can generate full-length sequences that align with real evolutionary trajectories, recapitulate key variant signatures, and exhibit credible structural effects,positioning SARITA as a practical, complementary component to sequencing pipelines for anticipatory genomic surveillance and informed public-health decision-making.

Chapter 5

5 Pan-Viral Genome Embedding with Mistral

The preceding chapters examined how artificial intelligence can support pandemic response by identifying epidemiological indicators that best describe disease dynamics and inform public-health decision making (Chapter 2), detecting, at very low prevalence, viral lineages that are likely to become dominant (Chapter 3), and forecasting plausible future sequences to anticipate vaccine and therapeutic updates (Chapter 4). These investigations focused on SARS-CoV-2, responsible for the last pandemics. To move beyond this single-virus focus, a general, pan-viral representation (Figure 53) is required, one that can organize and recognize genomic sequences across diverse viruses. This chapter addresses that need by introducing MistrVirus (Figure 53), a transformer-based embedding model (built on the Mistral architecture) that maps biological sequences into a continuous vector space in which geometric proximity reflects biological relatedness. Such representations preserve semantic and similarity structure, allowing downstream machine-learning tasks,e.g., viral vs non-viral discrimination, lineage assignment for SARS-CoV-2, and differentiation between different viruses, to be performed effectively within a unified framework. It is shown that MistrVirus and in general, the propose embedding based provides an efficient representation of viral sequences and outperforms existing pan-viral embeddings across multiple benchmark tasks, supporting broader applications in genomic surveillance and diagnostics.

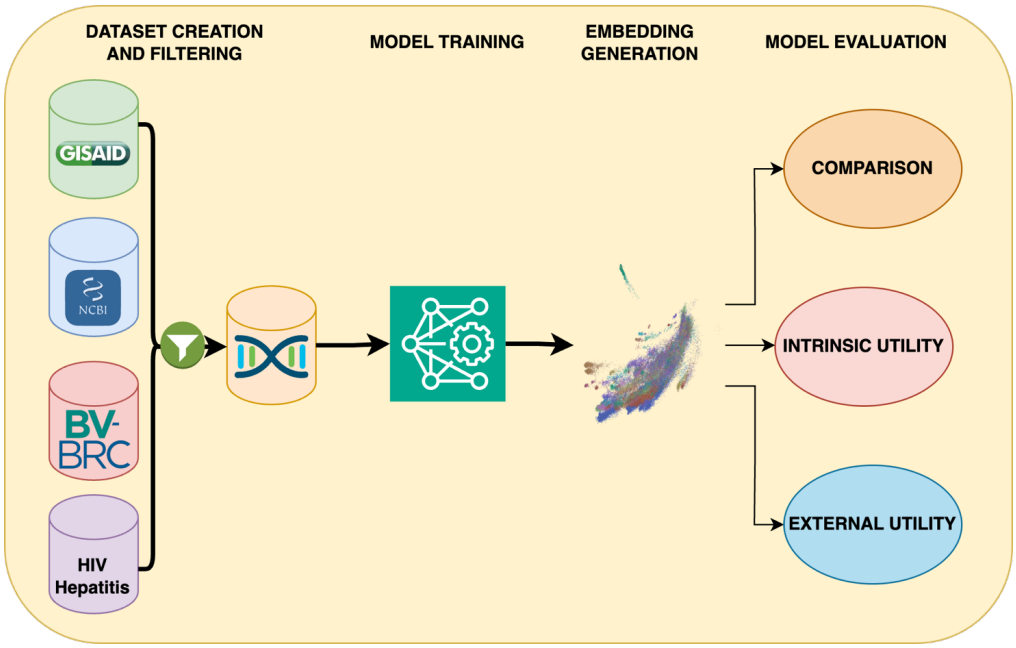


Figure 53: Schematic representation of the workflow for viral language model development and evaluation. Viral genome sequences were collected from multiple repositories (GISAID, NCBI, BV-BRC, and HIV/Hepatitis datasets) and subjected to filtering and preprocessing. The curated datasets were then used for model training, followed by the generation of embeddings representing viral sequences in a high-dimensional space. Model evaluation was conducted through three complementary strategies: direct comparison with baseline methods, intrinsic utility assessment (capturing representational quality and biological coherence), and external utility assessment (evaluating downstream applicability to virology tasks) [2].

5.1. Motivation and Scope

The rapid expansion of viral sequencing, accelerated during the COVID-19 pandemic and sustained by increasingly accessible genomics, has produced an abundance of data and a new interpretive bottleneck [143, 144]. Alignment-centric pipelines remain effective when close references exist and time permits careful curation, but performance degrades as sequences diverge, as references lag behind emerging diversity, and in metagenomic contexts where the viral signal is sparse within complex samples [145, 146, 147]. These constraints hinder early discovery, routine classification, and timely public-health decision-making, even in the presence of unprecedented volumes of genomes. In parallel, transformer architectures trained directly on nucleotide strings have matured into effective representation learners, yielding dense,

context-aware embeddings that can be reused across tasks with minimal task-specific engineering [116]. These considerations motivate a pan-viral embedding strategy: transformer models trained or adapted across diverse viruses to produce compact representations that support decisions that are rapid, dependable, and operationally useful. Although the preceding chapter examined studies centered on individual viruses, particularly SARS-CoV-2, long-term surveillance demands representations that generalize across viral taxa. While SARS-CoV-2 provided the open-sequence scale needed to prototype and validate methods, effective surveillance over time requires embeddings that organize and recognize genomic diversity across the viral kingdom. The objective, therefore, is to develop and evaluate a single, reusable embedding space in which geometric proximity mirrors biological relatedness, allowing downstream classifiers to perform robustly even when explicit alignments are unreliable or unavailable. Against this background, MistrVirus (Figure 53) is presented as a transformer-based embedding derived from the Mistral architecture and adapted to on a curated pan-viral corpus. The working hypothesis is that domain adaptation across diverse viruses enhances the embedding's intrinsic organization and its effectiveness on surveillance-relevant downstream tasks.

The focus of this Chapter is on representation learning and its impact on routine decisions, rather than on de novo sequence generation or clinical outcome modeling. Training an embedding model requires assembling a large, quality-controlled corpus by harmonizing multiple repositories, standardizing metadata and taxonomy, removing duplicates, and resolving ambiguous bases. To validate the efficacy of the proposed embedding, comparisons are made against alternative embeddings reflecting different design choices. Evaluation considers both intrinsic structure, quantified with clustering validity metrics and low-dimensional projections to assess whether species-level organization emerges, and extrinsic utility, tested on tasks aligned with surveillance practice: viral versus non-viral discrimination for metagenomic triage, DNA versus RNA classification to reflect replication strategies, and lineage-level resolution within SARS-CoV-2 for fine specificity. Where feasible, temporal holdouts are included to gauge stability as new data accrue.

5.2. Dataset Creation and Curation

Data collection and curation (Figure 54) were designed to maximize breadth across viral taxa while enforcing consistent metadata and reproducible quality control, in line with the general principle that model reliability ultimately depends on data reliability. As noted in previous Chapters, high-quality, well-annotated genomic datasets, with standardized taxonomic fields, host, collection date, and location, are a prerequisite for trustworthy ML/DL in computational virology. This motivates rigorous filtering and harmonization before any representation learning. The corpus integrates multiple repositories with complementary strengths: NCBI Virus [148]serves as the open, programmatically accessible backbone (more than 14 million viral nucleotide sequences as of May 2025); BV-BRC [149]contributes comprehensive pathogen coverage and tooling, albeit with slower bulk retrieval; [123] offers unrivalled, rapidly updated datasets under restricted terms (e.g., >17 million SARS-CoV-2 sequences by May 2025); and specialist resources (LANL-HIVdb [150], HBVdb [151]) supply richly curated entries for high-priority viruses. Together these sources balance scale, recency, and metadata depth. Retrieval follows source-specific best practices under a uniform, auditably scripted workflow. From NCBI Virus, records are obtained with the NCBI Datasets CLI by specifying the viral taxonomic identifier (TaxID 10244). In this way assembles all public viral genomes matching this identifier into a single archive that contains FASTA sequences and JSON metadata organized per genome. From BV-BRC, genomes are fetched via the FTP service by navigating the Virus directory, downloading the .fna files, and parsing their headers, where BV-BRC encodes key attributes such as completeness, into a structured table; partial entries are removed to keep only complete genomes. Because GISAID constrains bulk export and organizes content by species rather than through a unified genome browser, only a small set of epidemiologically relevant but underrepresented species is retrieved (e.g., Zika virus, Mpox, Respiratory Syncytial Virus, Dengue virus, Chikungunya virus), deliberately excluding SARS-CoV-2 and influenza which are already abundant in the other Dataset; within each species portal, the “complete genome” filter is applied prior to download, and the resulting FASTA headers are parsed to extract metadata into the harmonized table. To enrich clinically critical taxa, LANL-HIVdb and HBVdb are accessed through their web interfaces; given their modest size and the absence of programmatic endpoints, complete-genome entries are exported manually in FASTA format and processed with the same header-parsing routine used for the other sources. In particular HBVdb provides manually annotated hepatitis B virus genomes

and partial sequences with high-quality metadata (e.g., genotypes and drug-resistance profiles) and tools for automated genotyping and resistance interpretation. Instead LANL-HIVdb aggregates extensive HIV-1 (and selected HIV-2/SIV) sequences together with curated annotations and web tools for resistance mutation interpretation. Although smaller than NCBI Virus, BV-BRC, or GISAID, these databases add high-fidelity sequences and metadata that improve coverage and label quality for clinically important pathogens, strengthening both training and evaluation sets.

Once aggregated, all Metadata are normalized to an NCBI-style schema and value-level inconsistencies are resolved (e.g., GISAID's "DenV" reconciled to the canonical *Orthoflavivirus denguei*), after which duplicates are removed within and across sources, keeping the NCBI record where multiple entries exist. At this stage, the merged collection totals ~3.5 million genomes. Quality control (QC) addresses ambiguous bases explicitly: sequences are retained if the fraction of N (unknown base) is lower **than** 1% and no run of consecutive Ns > 20 is dected**,** a policy that preserves ≥99% resolved bases per genome and aligns with GISAID "high-coverage" guidance and Nextclade penalties for missing data. This approach reflects metagenomic realities, ambiguous calls are common, while avoiding dilution of transformer attention by long unresolved tracts. After uniform preprocessing and QC, ~1.9 million complete genomes remain. Because "completeness" labels vary across repositories (e.g., a subset of NCBI "complete" entries are flagged incomplete in BV-BRC), a targeted enrichment adds near-complete sequences from NCBI's *partial* set: applying the same N-filters reduces 7.9 million partials to 4.2 million high-quality candidates; comparing each record to the species- (or segment-) specific median complete-genome length and retaining entries within ±5% yields ~2.2 million near-complete genomes, more than half of which duplicate entries already present among complete genomes. The remainder (on the order of one million unique sequences) expands species coverage without compromising quality.

The final curated corpus (Table A10) contains ~2.9 million viral genomes, reflecting the pandemic-era surge in submissions and a highly skewed species distribution: SARS-CoV-2 dominates (≈1.9 million sequences, ~66% of the corpus) (Figure 55), with Alphainfluenzavirus also heavily represented (~20%) (Figure 55). This imbalance can bias downstream learners and mask poor minority-class performance [24]. It may also reduce training efficiency by overexposing models to redundant SARS-CoV-2 patterns while

underrepresenting rarer viral families, thereby inflating apparent majority-class performance and weakening generalization, necessitating a balanced subset for representation learning. Accordingly, species with ≥100 genomes are retained, while SARS-CoV-2 (~1.9 M) and Alpha-influenza (~574 k) are down-sampled to 100,000 sequences each (stratified by segment for influenza). The resulting training panel spans 295 species and 565,747 genomes, split 70/30 into 396,024 training and 169,723 test sequences, with a small validation slice from the training set for model selection. Accession IDs and harmonized metadata are preserved to ensure traceability and enable temporal hold-outs.

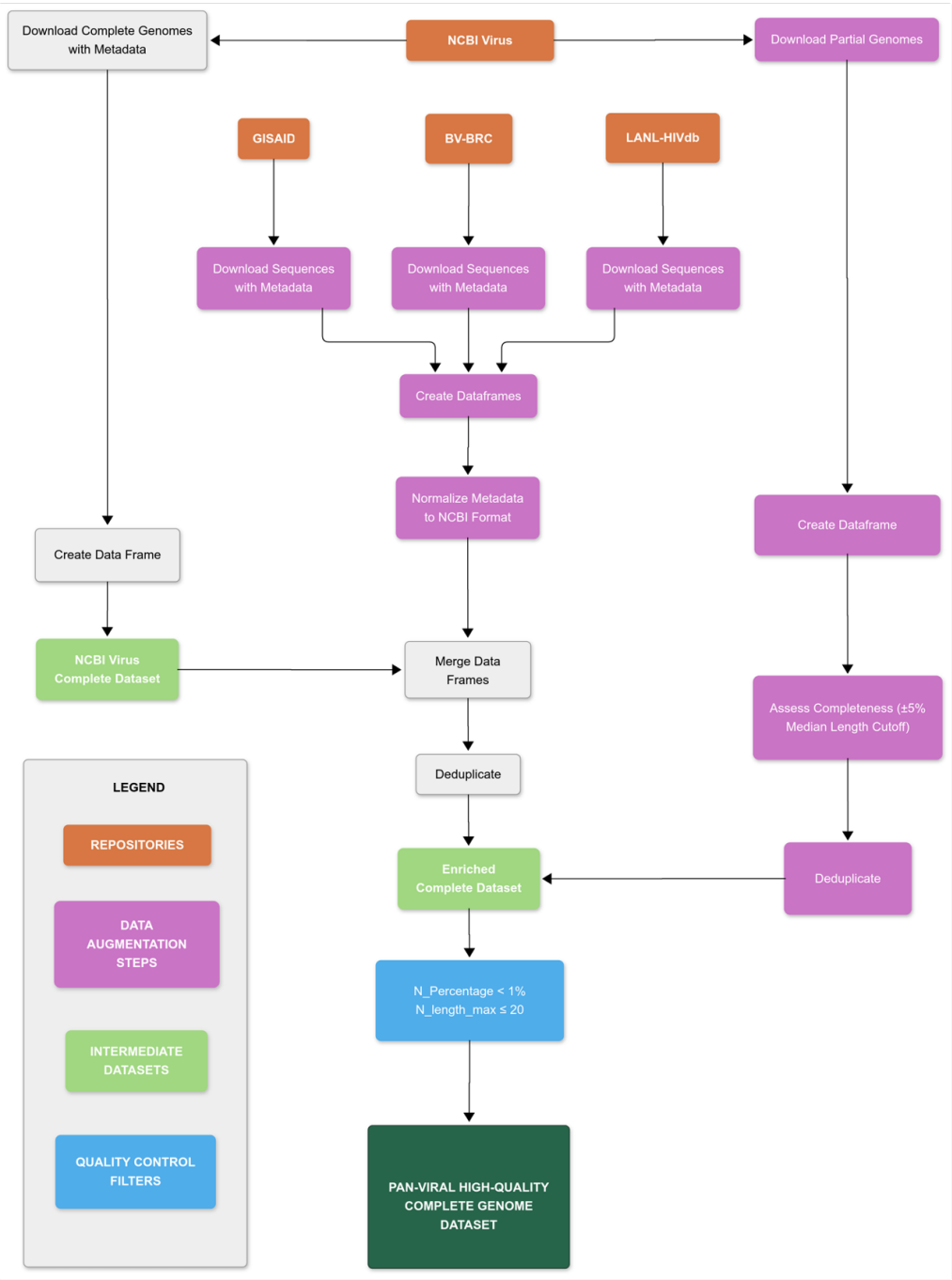


Figure 54: Workflow for the construction of the pan-viral high-quality complete genome dataset. Viral genome sequences and metadata were collected from multiple repositories (NCBI Virus, GISAID, BV-BRC, and LANL-HIVdb), including both complete and partial genomes. Metadata were normalized to the NCBI format, and individual data frames were created and subsequently merged. Completeness of sequences was assessed using a ±5% median genome length cutoff, and duplicate entries were removed. Quality control filters were applied, including thresholds on ambiguous nucleotide content (N% < 1%) and maximum consecutive ambiguous positions (N length ≤ 20). The resulting dataset constitutes a

deduplicated, enriched, and standardized resource of pan-viral high-quality complete genomes, optimized for downstream computational modeling and comparative analyses [2].

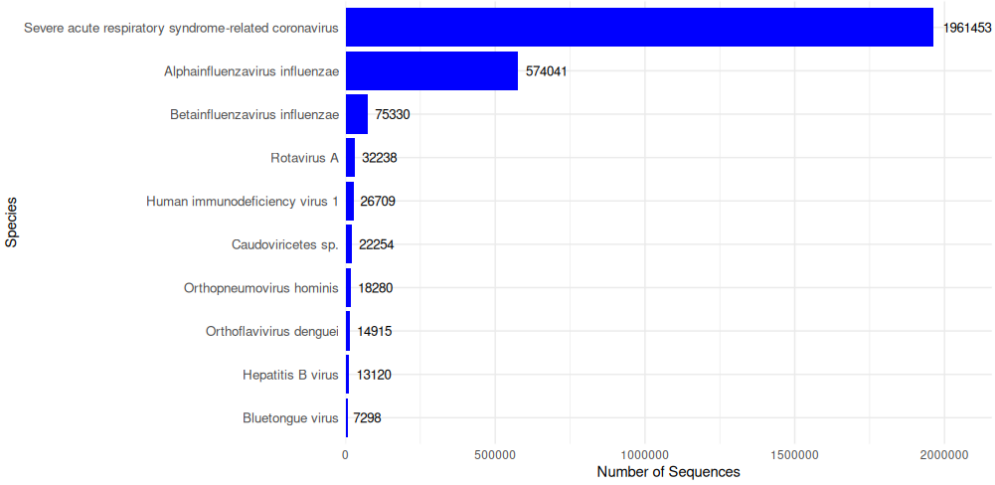


Figure 55: Top 10 Species by Sequence Count. A clear over-representation of SARS-CoV species can be seen in this bar plot; the top 1 (SARS-CoV) reaches almost 2 million sequences and the top 2 (Alphainfluenzavirus influenzae) almost at 600 thousand. From top 3 to top 10 the differences in numbers are significant.

5.3. Methods: Tokenization and Model Architecture

MistrVirus is a transformer-based embedding model obtained by adapting a medium-scale, decoder-only Mistral architecture (≈422M parameters) to a curated, pan-viral corpus [152]. The objective is to learn a single numerical representation per genome that preserves biologically meaningful relationships while remaining reusable across downstream tasks. Raw nucleotide strings (Figure 56) are first converted into variable-length DNA "subwords" using a DNABERT-2–style byte-pair tokenizer (vocabulary ≈4,096 units) [153]. Sequences are first normalized to uppercase and restricted to the canonical DNA alphabet, admitting ambiguity codes only where present (notably "N," which is frequent in surveillance data). The tokenizer's vocabulary (≈4k entries) is learned from large genomic corpora by iteratively merging the most frequent adjacent symbols, starting from the base alphabet; the result is a compact inventory that spans short motifs and longer, highly recurrent k-mer patterns. At inference time, a greedy, left-to-right segmentation applies

the longest available vocabulary match at each position, yielding an integer token sequence. This design has practical benefits over fixed k-mer tokenization: common biological motifs collapse to single tokens (reducing sequence length and improving computational efficiency), while rare or noisy regions fall back to shorter units without becoming out-of-vocabulary. Ambiguous bases are handled natively because the vocabulary includes tokens that cover "N" either alone or within merged motifs; no ad-hoc masking or imputation is required, preserving metagenomic realism. Special symbols are reserved for padding and unknowns; mask tokens exist in the vocabulary but are not used during causal training. For causal autoregressive learning, padding is placed on the left to ensure that right-aligned contexts remain valid, and batches are padded to the length of the longest sequence in the batch to avoid unnecessary truncation. When a sequence exceeds the model's context, it is processed in contiguous windows; because MistrVirus ultimately produces a single embedding per genome via fixed pooling, token vectors from all windows contribute uniformly to the final representation. Reverse-complement handling is not enforced at tokenization time (which would be non-trivial for variable-length subwords); instead, invariance is encouraged by the breadth of the pan-viral corpus and, if needed, can be strengthened downstream by evaluating both orientations and pooling. In practice, DNABERT-2–style BPE (Figure 56) reduces token counts substantially compared with fixed k-mers, stabilizes training by aligning tokens with frequent biological patterns, and ensures that occasional ambiguities do not fragment the input,properties that are particularly important for a pan-viral model like MistrVirus, which must accommodate wide sequence diversity without relying on task-specific preprocessing.

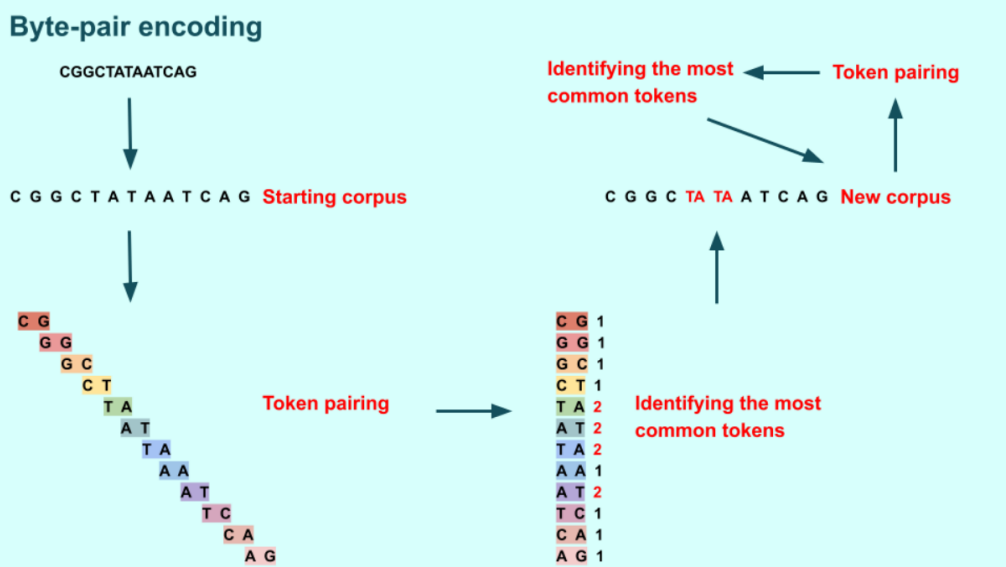


Figure 56: Illustration of the Byte-Pair Encoding (BPE) process applied to nucleotide sequences. Starting from an initial corpus of characters (e.g., "CGGCTATAATCAG"), the algorithm iteratively identifies the most frequent adjacent token pairs (e.g., "TA"), merges them into a single token, and updates the corpus. This process continues, progressively constructing new tokens and reducing sequence length while capturing recurring substructures. BPE enables efficient tokenization by balancing vocabulary size and sequence representation, making it particularly useful in genomic language models [154].

The model is trained with a standard next-token (causal language modeling) objective [105], allowing it to capture long-range dependencies and compositional regularities characteristic of viral genomes. Training is implemented with efficient attention kernels (FlashAttention-2) to handle long inputs at practical cost. Hyperparameters are intentionally conservative to prioritize stability over aggressive fine-tuning: 10 epochs, AdamW optimizer with learning rate 5×10^{-4} and weight decay 0.01, and **no mixed precision** to avoid numerical edge cases. Evaluation and checkpointing occur at the end of each epoch, with **early stopping (patience 3)** and automatic selection of the **best model by validation loss**; a small number of recent checkpoints is retained to limit storage. Data loading uses pinned memory and multiple workers to sustain throughput.

At inference time, the transformer yields a context-aware vector at every token position. To obtain a single, task-agnostic representation per genome, token-level vectors are condensed by max pooling, a simple, robust rule that is insensitive to local noise and resilient to variable sequence length. The pooling choice, together with the fixed tokenizer

and padding policy, is held constant throughout the study so that any observed differences in intrinsic organization or downstream performance can be attributed to what MistrVirus has learned from broad viral exposure rather than to preprocessing artifacts. In summary, MistrVirus retains the architectural bias of a decoder-only transformer, is taught the "language of viruses" through continual adaptation on a diverse corpus, and produces compact, comparable embeddings designed to support routine surveillance decisions with minimal task-specific engineering.

5.4. Experimental Design and Evaluation Metrics

The experimental design (Figure 57) evaluates whether an embedding space learned by MistrVirus exhibits coherent biological organization and supports surveillance-relevant decisions. All experiments are conducted under controlled preprocessing, shared DNABERT-2 tokenization, identical padding, fixed max-pooling to obtain one vector per genome, together with scripted pipelines, seeded randomness, and "load-best-at-end" selection based on validation loss to ensure reproducibility. The primary corpus is the curated pan-viral dataset described earlier; for representation learning and evaluation, a balanced dataset is constructed by retaining species with adequate support, down-sampling extreme outliers (such as SARS-CoV-2 and influenza) to fixed caps, and applying a stratified 70/30 split into development and test partitions. External evaluation cohorts are assembled to probe generalization at three levels: (i) virus vs non-virus, pairing the pan-viral test set with curated bacterial RefSeq genomes [148, 149]; (ii) DNA vs RNA, a balanced set of 10,000 complete viral genomes drawn from species absent from the pan-viral panel [148, 149]; and (iii) SARS-CoV-2 variants, balanced across major VOC/VOI and Omicron sublineages [123]. Where timestamps permit, temporal hold-outs(later submissions unseen during training) are used to assess stability over time. For each genome, token-level hidden states produced by MistrVirus are collapsed to a fixed-length sequence embedding via max pooling, a robust readout chosen for its insensitivity to local noise and variable length; this readout is held constant across all experiments so that observed effects can be attributed to the learned space rather than to downstream feature engineering.

Intrinsic structure is quantified without supervision by treating species labels as reference clusters and computing the Silhouette [155], Calinski–Harabasz [156], and Davies–Bouldin [157] indices on the raw, high-

dimensional embeddings. A higher Silhouette score indicates that, on average, each sequence lies much closer to members of its own species than to members of other species, reflecting clusters that are both coherent and well separated. The Calinski–Harabasz index increases when the variance between species is large relative to the variance within species; therefore, higher valuessignal tighter clusters with clearer separation. In contrast, the Davies–Bouldin index summarizes, for each cluster, the worst-case ratio of internal dispersion to separation from its most similar neighbor; consequently, lower values correspond to more compact clusters that are farther apart. For qualitative inspection only, PCA projections to two or three dimensions are produced to visualize cluster compactness, overlaps among related taxa, and large-scale geometry; these visualizations are used solely for interpretation and not for model selection. Extrinsic utility is then tested on tasks that mirror practical decision points. For virus vs non-virus, the classifier is trained only on viral embeddings to emulate metagenomic triage, with viral data split 70/15/15 (train/validation/test) and bacterial data split 50/50 (validation/test) to set thresholds and measure specificity. Two complementary anomaly detectors are employed,one-class SVM (a compact envelope around viral density) and Isolation Forest (partition-based isolation of outliers),and an AND-rule ensemble labels a sequence as viral only if both agree, prioritizing low false positives; hyper-parameters are tuned on the validation portion and all metrics are reported on the untouched test split. For DNA vs RNA (binary), the balanced 10k set is split 80/20 with stratification by class; logistic regression, RBF-kernel SVM, and random forest are tuned via five-fold cross-validation on the training portion, refit on the full training set with selected hyper-parameters and evaluated once on the test set to estimate generalization. The choice of an 80/20 split, rather than 70/30 as in the previous use case, was motivated by the smaller size of this dataset compared to the virus vs non-virus task, where more data were available. This ensured that a larger fraction of the data could be used for model training while still preserving a representative test set for generalization assessment. For SARS-CoV-2 variant classification (multiclass), a balanced cohort across lineages is split 80/20 with per-variant stratification; the same trio of classifiers and tuning protocol is used, followed by a temporal generalization check that scores the best model on later submissions without retraining. Confusion matrices are examined to identify systematically confused lineages. For each External utility task Evaluation Measure as F1, Accuracy, Sensitivity, Specificity fair are used. The performance obtained on Internal and External utility

task are compared with other Transformer embedding as The performance obtained on intrinsic and extrinsic utility tasks is compared against two transformer-based embedding baselines: (i) a Human-genomics, decoder-only model called Mistral-DNA with environ ~422M parameters without viral pretraining, and (ii) an encoder-only viral specialist called ModernBERT-virus having ~37M parameters trained on curated RefSeq viral genomes. All methods use the same DNABERT-2 tokenizer (Figure 56), padding policy, pooling strategy, and stratified data splits to ensure a, apples-to-apples comparison.

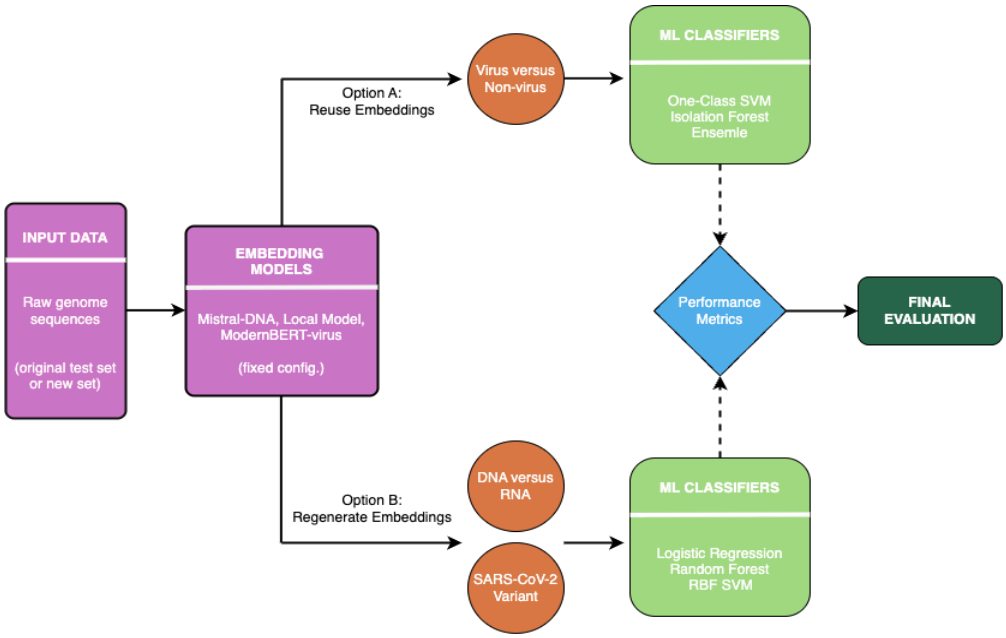


Figure 57: Workflow of genome classification using embedding models and machine learning classifiers. Raw genome sequences are processed through embedding models (Mistral-DNA, Local Model, ModernBERT-virus) to generate fixed representations. These embeddings are used for multiple classification tasks: (i) virus versus non-virus using One-Class SVM, Isolation Forest, and ensemble methods; (ii) DNA versus RNA and SARS-CoV-2 variant classification using Logistic Regression, Random Forest, and RBF SVM. Performance metrics are computed across tasks, leading to the final evaluation of model effectiveness

5.5. Results

Results indicate that the **MistrVirus** yields a representation space that is both more biologically organized and more useful for downstream decisions than the general-genomics baseline and is competitive with a compact viral specialist. Intrinsically, species structure strengthens after adaptation: when treating species as reference clusters (Table 11) in the

raw, high-dimensional space, the Silhouette score improves from –0.285 with the human-trained baseline to –0.012 with MistrVirus, while the Calinski–Harabasz index rises from 533 to 814 and the Davies–Bouldin index falls from 3.88 to 3.02 (higher Silhouette/CH and lower DB reflect tighter, better-separated clusters). ModernBERT-virus shows intermediate CH (≈600) and a higher DB (≈8.69), consistent with compact but less well-separated manifolds under these global indices. The PCA projections (Figure 58) corroborate these patterns qualitatively: the baseline distributes points diffusely with little separability; MistrVirus and the specialist both form discernible species groupings along a shared curved manifold, with MistrVirus showing broader, biologically interpretable spread within species (Figure 59).

Table 11: Intrinsic evaluation of embedding models based on clustering quality metrics. Reported values include Silhouette Score (higher values indicate better-defined clusters), Davies–Bouldin (DB) index (lower values reflect better separation and compactness), and Calinski–Harabasz (CH) index (higher values indicate improved clustering performance). Among the compared models, the MistrVirus achieved the most favorable balance across metrics, while ModernBERT-virus showed weaker separation as indicated by the high DB index

Model	Parameters	Silhouette Score	DB index	CH index
Mistral-DNA	422M	-0.2848	3.8793	532.86
MistrVirus	422M	-0.0115	3.0241	814.3
ModernBERT-virus	37M	-0.0208	8.6928	600.27

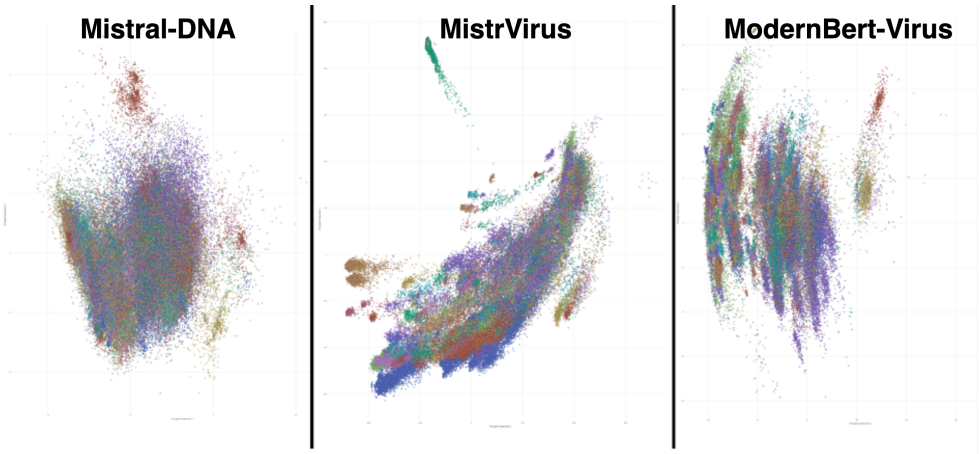


Figure 58: Two-dimensional visualization of viral genome embeddings produced by three models: Mistral-DNA (left), MistrVirus (center), and ModernBERT-virus (right). Each point represents an embedded sequence, with colors denoting viral taxa. The Local Model demonstrates relatively compact and separable clusters, while Mistral-DNA and ModernBERT-virus show more dispersed structures, reflecting differences in how models capture sequence similarity and inter-taxa boundaries.

Extrinsically, anomaly detection for **virus vs non-virus** (Figure 60) achieves the strongest and most balanced performance with MistrVirus embeddings when combining a one-class SVM and an Isolation Forest in an AND-rule ensemble: balanced accuracy 0.943, specificity 0.999, sensitivity 0.886, F1 0.940, ROC–AUC 0.945. Individually, the one-class SVM attains very high specificity on MistrVirus (0.996) by enveloping a compact viral density, whereas the Isolation Forest benefits when the viral cloud retains internal structure. The human-trained baseline underperforms (ensemble balanced accuracy 0.829) due to sparse viral geometry, and the viral specialist attains lower ensemble balance (0.813), reflecting a trade-off between recall and false alarms. Error inspection explains residual failures: bacteriophages (e.g., Caudoviricetes and "Bacteriophage sp.") concentrate near the bacterial manifold and are preferentially rejected, while a small set of outlying viral lineages (e.g., Orthopneumovirus hominis, Orthoebolavirus zairense) lie at the edges of the SVM envelope. A new-species stress test dominated by previously unseen phages confirms this geometry: the ensemble on MistrVirus maintains useful discrimination (balanced accuracy 0.823; ROC–AUC 0.903) but with the expected drop in specificity (0.777), highlighting the biological entanglement between phages and their bacterial hosts and suggesting value in phage-aware fine-tuning or multi-class (bacteria/free-phage/prophage) formulations.

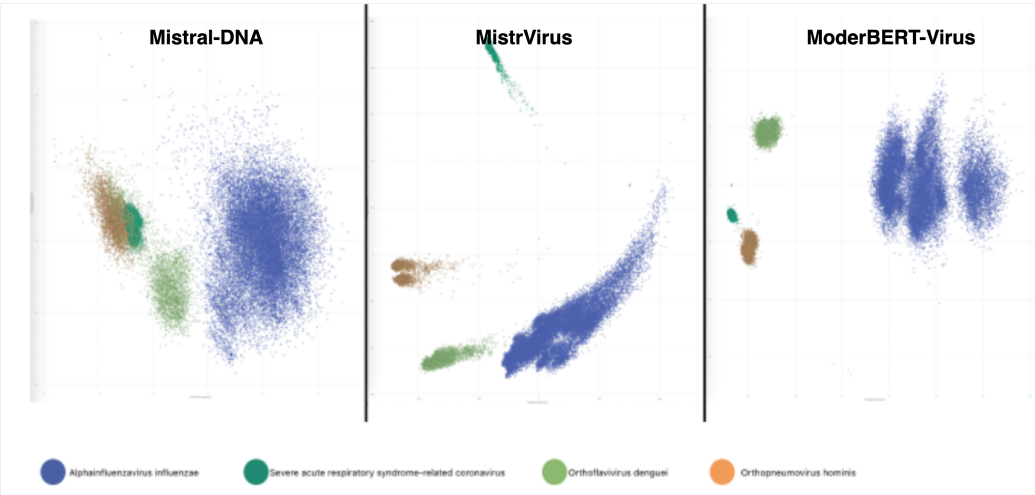


Figure 59: Visualization of viral genome embeddings projected in two dimensions, showing clustering of four representative viral taxa: Alphainfluenzavirus influenzae (blue), severe acute respiratory syndrome-related coronavirus (green), Orthoflavivirus denguei (light green), and Orthopneumovirus hominis (orange). Each panel corresponds to embeddings produced by different models, highlighting variations in the degree of separation and compactness among taxa.

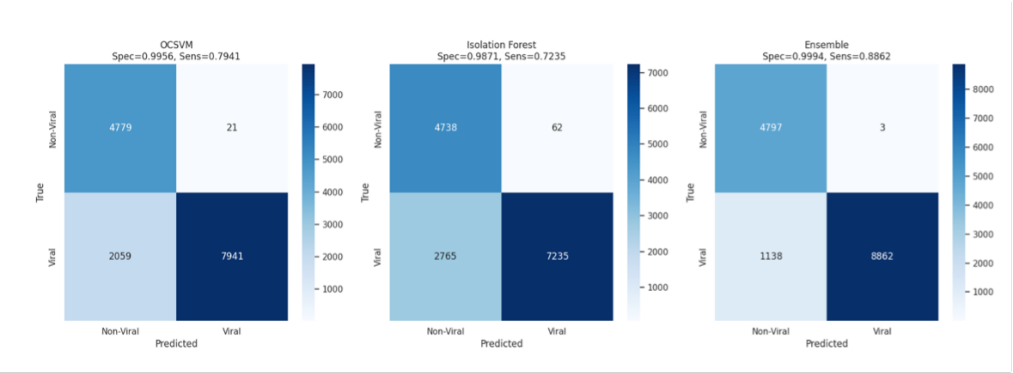


Figure 60: Confusion matrices comparing the performance of three classifiers, One-Class SVM (left), Isolation Forest (center), and an Ensemble model (right), for distinguishing viral from non-viral sequences. Metrics reported above each panel include specificity (Spec) and sensitivity (Sens), with the Ensemble model achieving the best balance (Spec = 0.9994, Sens = 0.8862) relative to individual classifiers.

On **DNA vs RNA** discrimination (Table 12. Table 13, Table 14, Figure 61), all embeddings support high performance, indicating that replication-strategy signals are strongly encoded. With MistrVirus embeddings, accuracy peaks at 0.968 using an RBF-kernel SVM (precision 0.965, recall 0.982, specificity 0.954; F1 0.968); even a simple logistic

regression reaches 0.945 accuracy, consistent with a largely linearly separable structure in the full space despite overlap in 2D PCA. The ModernBERT-virus consistently trails MistrVirus by some points, and the Mistral-DNA baseline remains serviceable (≈0.89 with SVM), suggesting that coarse biochemical distinctions are robustly recoverable once tokenized and pooled.

Table 12: Performance Metrics on DNA versus RNA Task with Logistic Regression Algorithm. The MistrVirusl achieved the highest overall performance across all metrics (Accuracy = 0.945, Precision = 0.941, Recall = 0.949, Specificity = 0.941, F1 Score = 0.945), outperforming both Mistral-DNA and ModernBERT-virus.

Model	Accuracy	Precision	Recall	Specificity	F1 Score
Mistral-DNA	0.823	0.831	0.808	0.838	0.819
MistrVirus	0.945	0.941	0.949	0.941	0.945
ModernBERT-virus	0.870	0.877	0.857	0.882	0.867

Table 13: Performance Metrics on DNA versus RNA Task with Random Forest Algorithm. The MistrVirusl demonstrated the best balance across all evaluation metrics (Accuracy = 0.917, Precision = 0.926, Recall = 0.909, Specificity = 0.925, F1 Score = 0.922). In contrast, Mistral-DNA and ModernBERT-virus achieved higher recall (0.978 and 0.966, respectively) but at the expense of lower precision and specificity.

Model	Accuracy	Precision	Recall	Specificity	F1 Score
Mistral-DNA	0.871	0.805	0.978	0.766	0.883
MistrVirus	0.917	0.926	0.909	0.925	0.922
ModernBERT-virus	0.880	0.822	0.966	0.794	0.889

Table 14: Performance Metrics on DNA versus RNA Task RBF SVM Algorithm. The MistrVirus achieved the highest overall performance (Accuracy = 0.968, Precision = 0.965, Recall = 0.982, Specificity = 0.954, F1 Score = 0.968), indicating both strong sensitivity and robustness. Mistral-DNA and ModernBERT-virus also performed well, with Mistral-DNA showing balanced recall and F1 score (0.941 and 0.894, respectively), while ModernBERT-virus maintained consistent accuracy and precision (0.914 and 0.895, respectively)

Model	Accuracy	Precision	Recall	Specificity	F1 Score
Mistral-DNA	0.889	0.851	0.941	0.837	0.894
MistrVirus	**0.968**	**0.965**	**0.982**	**0.954**	**0.968**
ModernBERT-virus	0.914	0.895	0.937	0.892	0.915

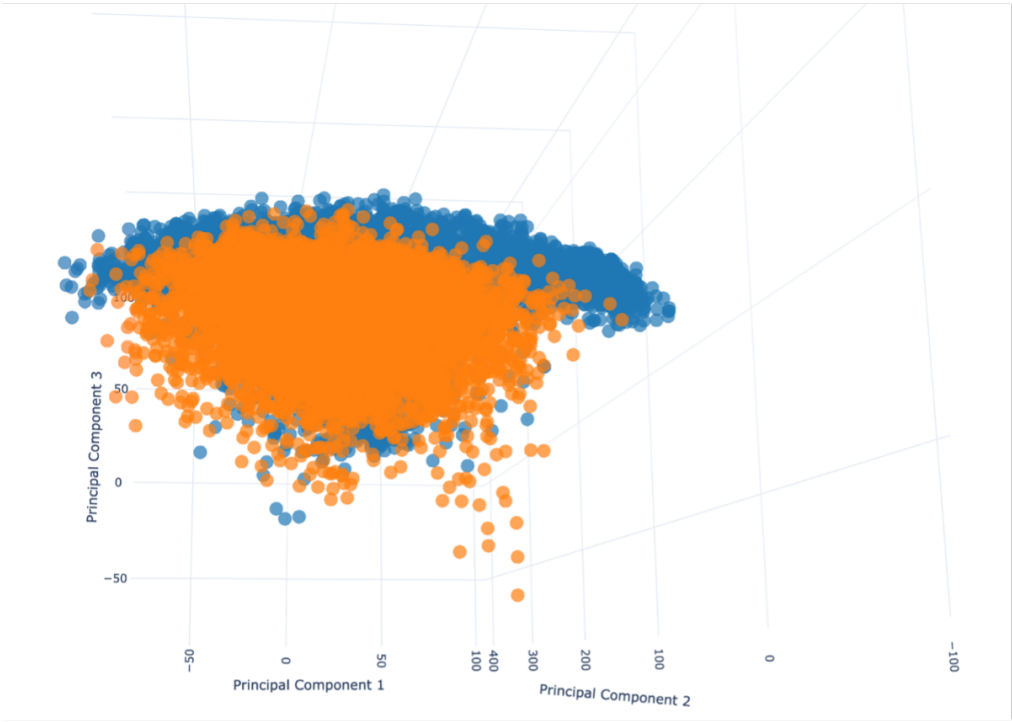


Figure 61: DNA/RNA Embeddings Generated by the Local Model on 3D PCAs. The plot illustrates the separation achieved between DNA and RNA classes (blue vs. orange) in the reduced feature space

For **SARS-CoV-2 variant classification** (18 balanced lineages, Table 15), all model–classifier pairs exceed ~0.87 balanced accuracy. The best run is observed with the ModernBERT-virus plus RBF-SVM (balanced accuracy 0.954), closely followed by MistrVirus with logistic regression (0.947) and RBF-SVM (0.929). The major errors that the models made are concentrated within recent Omicron sublineages (e.g., XBB family vs BA.2.86/JN.1), in line with minimal genomic distances and nomenclature aliasing; earlier VOC/VOI (Alpha–Delta, Lambda, Mu) are separated almost perfectly. A **temporal generalization** check on later submissions (April–May 2025) sustains high accuracy without retraining (balanced accuracy 0.959 after merging BA.* and JN.1 labels to avoid aliasing ambiguities), indicating stability of the learned decision surfaces as new data arrive.

Table 15: Performance Metrics for SARS-CoV-2 Variant Classification. Logistic Regression (LR) achieved the highest balanced accuracy with ModernBERT-virus (0.9501), while RBF SVM slightly outperformed LR with ModernBERT-virus (0.9537). Random Forest (RF) consistently

yielded lower performance compared to LR and RBF SVM across all models.

Balanced Accuracy	Mistral-DNA	MistrVirus	ModernBERT-virus
LR	0.9062	0.9470	0.9501
RBF SVM	0.9065	0.9291	0.9537
RF	0.8741	0.9124	**0.9426**

Taken together, these findings support three conclusions. First, domain adaptation over a broad viral repertoire markedly improves the intrinsic geometry of a medium-scale decoder-only model and translates into strong, specificity-focused anomaly detection in metagenomic settings. Second, coarse biological signals (viral vs cellular; DNA vs RNA) are reliably encoded and accessible to simple classifiers, while fine granularity (SARS-CoV-2 lineages) is achievable across embeddings, with minor advantages for reference-anchored specialists in single-virus settings. Third, the principal residual weakness lies at the virus–bacteria interface where phage genomes overlap the bacterial manifold; results suggest targeted augmentation or explicit multi-class formulations as practical next steps. Overall, MistrVirus offers the best trade-off across tasks,enhancing intrinsic organization and delivering high downstream utility with strong specificity and stable temporal behavior,thus meeting the intended objectives of a reusable, pan-viral embedding space for surveillance.

5.6. Conclusion and Limitations

This chapter demonstrates that a pan-viral, transformer-based embedding can convert raw genomic sequences into a reusable numerical space that is both biologically organized and operationally useful for surveillance. The study spans the full pipeline, from large-scale, multi-repository data assembly and harmonization to intrinsic analysis of the embedding geometry and extrinsic evaluation on tasks that mirror real

decision points, so conclusions rest on a consistent and reproducible foundation rather than on a single benchmark.

Several insights emerge from the data layer. Public repositories now contain an enormous volume of viral genomes, yet abundance is uneven: pandemic-era submissions dominate, especially for SARS-CoV-2, while many taxa remain under-represented. Curating a pan-viral corpus therefore requires more than bulk download; it demands rigorous deduplication of near-identical records, reconciliation of heterogeneous "complete/partial" flags, explicit handling of ambiguous bases, and resolution of inconsistent taxonomy. In practice, dataset construction becomes a substantive big-data task whose quality directly conditions representational performance.

On representation learning, results indicate that domain adaptation on a broad viral repertoire matters. Embeddings produced by a human-genome–trained model tend to be diffuse when applied to viruses, whereas exposure to viral corpora encourages a geometry in which species occupy more compact and separable regions. Biologically, this organization can be interpreted as the model internalizing a mixture of signals that co-vary with viral identity: compositional biases (e.g., GC content and k-mer usage), gene-content and coding constraints, and lineage-specific sequence motifs shaped by replication strategy and host adaptation. In this sense, the embedding does not represent explicit functional annotations (such as a particular protein domain) but rather a distributed summary of sequence regularities that correlate with taxonomic and ecological differences. Low-dimensional projections such as PCA provide only an illustrative slice of this structure and may look noisier than the actual separability implied by high-dimensional classifiers; nevertheless, both intrinsic indices and qualitative plots point in the same direction: pan-viral training improves biological organization.

Importantly, this also clarifies the biological scope of MistrVirus. The embedding is learned from nucleotide sequences as input, and its strongest biological meaning is therefore at the level of genome-scale signatures (taxonomic relatedness, broad replication class, and coarse evolutionary constraints). It should not be interpreted as a mechanistic model of within-host fitness, virulence, transmissibility, or immune escape, which depend on phenotype, host context, and protein-level structure–function relationships not directly observable from raw genomic strings alone. Rather, MistrVirus provides a compact representation that is useful for triage, grouping, and classification, and that can act as a front-end for downstream analyses that incorporate

additional biological evidence (metadata, host information, proteomic/structural features, or epidemiology).

For virus-versus-non-virus discrimination, anomaly detectors operating on pan-viral embeddings achieve high specificity against bacteria, which is desirable in metagenomic triage. The main residual errors concentrate in bacteriophages, reflecting their biological proximity to bacterial hosts and their shared ecological niche; adding phage-rich data improves recall, suggesting that a phage-aware specialization (or a multi-class bacteria/free-phage/prophage formulation) would further reduce confusions. The DNA-versus-RNA task confirms that coarse replication-strategy signals are strongly encoded; even general-genomics pretraining provides a usable baseline, while viral adaptation delivers additional gains. The SARS-CoV-2 lineage task, although fine-grained, proves comparatively tractable: when other organisms are absent, classifiers need to detect a small set of token changes, and all embedding spaces support high balanced accuracies with stable performance on later submissions, an indication that the learned decision surfaces are not brittle to moderate temporal drift.

From a biological-interpretation perspective, these task results suggest a useful hierarchy: the embedding reliably captures broad, conserved genome signatures (virus/non-virus; DNA/RNA), and it can also separate closely related lineages when the signal reduces to consistent sets of sequence changes within a constrained genomic background (SARS-CoV-2). Conversely, tasks that would require distinguishing viruses with convergent genome composition, strong recombination, or mosaic architectures are expected to be more challenging and may require segment-aware encoders, explicit modeling of recombination, or integration of protein-level constraints.

Important limitations motivate caution. Inference quality ultimately depends on metadata fidelity and taxonomic stability, both of which vary across repositories and over time. Class imbalance remains a structural feature of public genomics and can bias both representation learning and evaluation if not explicitly controlled. Architectural and implementation choices—fixed pooling, lack of enforced reverse-complement invariance, reliance on conventional classifiers—were intentionally conservative to ensure interpretability and fairness, but they are not the only plausible choices; different pooling rules, equivariant designs, or sequence-segment models might capture signal lost here. Finally, practical deployment must contend with licensing constraints, privacy considerations, and the need for continual refresh as new genomes arrive.

To conclude, the evidence supports the use of transformer-based, pan-viral embeddings as a practical interface between today's genomic data streams and the decisions required for surveillance and diagnostics. MistrVirus should be viewed as a general-purpose representation layer whose biological meaning is primarily taxonomic and genome-architectural, enabling fast organization and triage rather than direct mechanistic inference. By strengthening data governance, advancing architecture and training objectives, improving reliability under distribution shift, and packaging models for routine use, these representations can help move AI-driven genomics from promising experiments to dependable, cost-effective practice, supporting earlier detection, more resilient health systems, and a broader scientific understanding of the viral world.

Chapter 6

6 Conclusion and Future Directions

The preceding chapters examined how artificial intelligence can support pandemic response by (i) identifying epidemiological indicators that best capture disease dynamics and inform public-health decision making (Chapter 2), (ii) detecting, at very low prevalence, viral lineages that are likely to become dominant (Chapter 3), and (iii) forecasting plausible future sequences to anticipate vaccine and therapeutic updates (Chapter 4). However, those studies were virus-specific, focusing on SARS-CoV-2. To address this, Chapter 5 introduces a multispecies tool designed for diverse tasks, such as distinguishing viruses from non-viruses and classifying DNA versus RNA viruses, so that similar capabilities can be applied beyond a single pathogen. Across these contributions, the unifying objective has been to anticipate and support response during public-health emergencies to reduce transmission and accelerate the development of vaccines and therapeutics, in line with World Health Organization public-health goals. This chapter discusses the limitations of the proposed approaches and delineates their practical implications, including how they can be integrated into genomic surveillance pipelines to strengthen public-health protection.

6.1. Summary of key contributions

This thesis advances AI-enabled genomic surveillance along four complementary axes. First, it links epidemiology and policy by combining a multilevel vector autoregressive (mlVAR) analysis across nine European

countries (2020–2022) with foundation-model forecasting (TimeGPT). The mlVAR is specified in a one-model-per-year design with lag selection by Akaike Information Criterion (optimal 1-week lag in 2020), treating countries as random effects to capture shared dynamics with national heterogeneity; results show hospitalizations as the primary driver of stringency in 2020, cases gaining importance in 2021, and vaccination doses reducing cases and hospitalizations with dose- and year-specific patterns (e.g., second doses associated with reduced stringency in 2021). In parallel, a sequencing-focused analysis finds Weekly Sequencing Rate to be the most informative exogenous regressor for next-month cases in 54% of countries, with markedly variable accuracy (median RMSE ≈ 78.8 cases/100k; best RMSEs in Ireland and Bulgaria) and country-specific correlation profiles with ICU, hospitalizations, cases, or deaths over 2020–2023. Together, these findings provide an evidence base for proactive, data-aware public-health decisions and clarify when sequencing intensity carries forecasting value beyond clinical indicators.

Second, it proposes DeepAutoCoV, an anomaly-detection pipeline that flags Future Dominant Lineages (FDLs) well before they cross a 10% prevalence threshold. The system trains a k-mer (k=3)–based AutoEncoder on high-quality Spike S1 sequences filtered for biological realism; anomalies are calls whose reconstruction error exceeds 1.5 SD above the median, followed by a PAM30 filter to down-weight near-neighbors. A weekly simulation ingests new sequences, retraining once a lineage becomes dominant to prevent re-flagging incumbents. Lead-time, PPV, and prioritization metrics quantify utility: globally, the median lead time is 17 weeks (IQR 11–32), with country-level medians of 12 (USA), 6 (UK), 5 (Denmark) and 11 (France) weeks; PPV peaks in the UK (0.57, IQR 0.24–0.92) and true positives typically appear within the top 3–6 ranked anomalies,outperforming both supervised baselines (LR, Spike2Vec, Spike2Signal; PPV=0) and a distance-based unsupervised detector. Importantly, the reconstruction-error saliency over k-mers surfaces fitness-linked substitutions (e.g., Delta P681R at the furin site; Omicron BA.2 Y505H in RBD; FLiRT R346T/F456L in KP), offering mechanistic clues alongside early alerts.

Third, it explores SARITA, a domain-adapted, decoder-only protein LLM trained with continual learning to generate plausible full-length Spike S1 sequences. SARITA uses amino-acid Unigram tokenization with RoPE, scales from ~85M to ~1.2B parameters, and is adapted from a general protein LLM (RITA) on >100k curated Spike sequences (pre-2021) using mixed-precision training on A100 GPUs. Benchmarking against non-learned (uniform/frequency/PAM-guided) and learned (RITA,

ProGen2-Base, SpikeGPT2) baselines shows consistent gains across validity/length (686 ± 10 aa), edit-distance and substitution-matrix (PAM30/BLOSUM) similarity, and False Mutation Rate, with larger SARITA variants increasing the share of novel yet alignment-plausible substitutions. Beyond global similarity, SARITA recapitulates lineage-defining signatures (e.g., L212I, R158L, T19L, T95P, E406K; canonical N501Y and G446S), and ΔΔG screens against ACE2 and monoclonal-antibody complexes fall within empirically supported regimes, even when restricted to novel substitutions, indicating biophysical plausibility. Design choices driving these outcomes include an autoregressive decoder-only formulation conditioned on the conserved 14-aa signal peptide (stabilizing full-length continuation and epistasis) and domain-adaptive continual learning that concentrates novelty where selection permits change.

Fourth, it develops MistrVirus, a pan-viral embedding built by adapting a ~422M-parameter Mistral decoder-only model to a rigorously curated corpus. From ~2.9M genomes assembled across repositories,with explicit handling of duplicate/near-duplicate entries, heterogeneous "complete/partial" flags, ambiguous bases, and inconsistent taxonomy,the training set is balanced by down-sampling over-represented taxa (SARS-CoV-2 and Alphainfluenza to 100k each), yielding 295 species and ~565,747 genomes with preserved accessions for traceability and temporal hold-outs. Tokenization uses a DNABERT-2–style byte-pair vocabulary (~4,096 DNA "subwords") to compress common motifs while remaining robust to ambiguity. The resulting embedding organizes viral diversity and supports surveillance-relevant tasks: coarse signals (viral vs non-viral; DNA vs RNA) are linearly separable, while fine granularity (SARS-CoV-2 lineage classification) attains high balanced accuracy (e.g., 0.947/0.929 with LR/SVM for MistrVirus; 0.953–0.950 for ModernBERT-virus) and sustains temporal generalization on later 2025 submissions (balanced accuracy 0.959 after label-alias merging). These results, coupled with the corpus-level analysis, establish MistrVirus as a reusable representation that reduces task-specific engineering and accelerates deployment across surveillance pipelines.

6.2. Impact on Public Health and Biomedical Engineering

At the public-health level, the mlVAR/TimeGPT program turns sequencing effort into an actionable policy lever rather than a passive surveillance metric. By quantifying when cases, hospitalizations, vaccination, and weekly sequencing intensity most strongly shape short-term trajectories, it enables health authorities to time non-pharmaceutical interventions (NPIs) and booster campaigns with finer granularity, reducing collateral costs (school closures, elective-care delays) while preserving epidemic control. The same framework supports counterfactual "what-if" analyses (e.g., increasing sequencing by X% or accelerating second-dose coverage) and can be embedded in Public Health Emergency Operations Center (PHEOC) workflows to inform situational awareness briefings and procurement decisions (test kits, antivirals, ICU capacity).

From a surveillance operations standpoint, DeepAutoCoV's weeks-to-months of lead time shift practice from retrospective lineage labeling to proactive triage. The system's top-N prioritization aligns with real laboratory throughput, channeling scarce confirmatory assays and neutralization experiments toward candidates most likely to become dominant. Earlier, better-targeted wet-lab validation compresses the time from signal to action, improving risk communication, updating diagnostic primers, and anticipating vaccine-antigen deliberations, while reducing reviewer burden through rank-ordered, k-mer-level rationales that make alerts auditable.

On the biomedical engineering side, MistrVirus provides a compact, reusable representation layer that standardizes how viral sequences are ingested across heterogeneous pipelines. By lowering task-specific feature engineering and enabling high performance with lightweight classifiers, it shortens development cycles and facilitates deployment at the edge (e.g., low-latency triage in sequencing facilities) as well as in cloud/hybrid settings. The representation's cross-pathogen generality supports maintainable analytics for metagenomic pre-screening, host–pathogen discrimination, and lineage resolution, easing integration with LIMS/ELN systems and CI/CD-style MLOps (versioned models, data provenance, drift monitoring).

More broadly, the thesis advances translation-readiness: (i) evaluations that preserve the real-world timing of data availability, ensuring models are tested only on future data as it would arrive in practice; (ii) interpretable outputs (mutation/k-mer saliency; $\Delta\Delta G$ screens) that fit human-in-the-loop review, improving acceptability among laboratorians and clinicians; (iii) engineering patterns (data curation, traceability, hold-outs) that align with quality-management principles and future

regulatory pathways. Taken together, these contributions help transform genomic surveillance from a collection of separate analyses into a more integrated and interpretable framework, supporting stronger connections between molecular data, epidemiological forecasting, and public-health decision-making.

6.3. Methodological limitations and ethical considerations

A first limitation arises from data imbalance and curation. Pandemic-era repositories are saturated with SARS-CoV-2 (and, to a lesser extent, influenza), which means that even after careful down-sampling the model ecosystem inherits residual skew. In practice, richer taxa provide clearer signals and cleaner decision boundaries than under-sampled pathogens. This was visible in the pan-viral embedding: despite balancing and temporal hold-outs, linearly separable classes tended to be those with abundant, higher-quality submissions, while small families showed noisier cluster structure. The same issue appears upstream, in metadata heterogeneity. “Complete” vs “partial” flags are not standardized across sources, ambiguous bases inflate k-mer counts, and taxonomic aliases complicate stratification. The thesis mitigated these effects through species-level balancing, duplicate/near-duplicate filtering, explicit temporal splits, and quality screens; nevertheless, generalization to low-resource viruses remains less certain than for well-represented taxa. A related trade-off emerged in DeepAutoCoV: strict validity and biological-realism filters reduced false positives but risked discarding rare, early true positives that do not yet look “typical.” That is, curation protects training, but can inadvertently hide exactly the edge cases early-warning systems hope to find.

A second limitation concerns immunological and structural realism. Sequence-only detectors and generators learn statistical regularities that only partially reflect immune selection. DeepAutoCoV’s anomaly score highlights unusual k-mer patterns, but it does not ingest neutralization escape maps, B-/T-cell epitope repertoires, or polyclonal sera behavior; as a result, it may flag variants that are unusual yet immunologically inconsequential, and it may under-weight variants whose immune advantage is subtle but real. On the generative side, SARITA’s plausibility checks rely on $\Delta\Delta G$ computed on fixed complexes (ACE2–RBD and a limited panel of monoclonal antibodies). Such proxies are informative but narrow: they miss conformational dynamics, glycan shielding,

cooperative epistasis outside the local mutational neighborhood, and the breadth of polyclonal responses. In this thesis, ΔΔG screens were used as conservative filters, not as claims of fitness,and signal peptide conditioning was introduced to stabilize full-length continuation, but those design choices do not substitute for explicit immunological modeling. A natural extension is to fuse deep-mutational-scanning escape landscapes, IEDB-style epitope knowledge, and glycan-aware structural ensembles, and to widen the scope beyond Spike S1 to S2 and other proteins that modulate entry, replication, or immune evasion.

Evaluation design is a third constraint. Most results are retrospective, with care taken to use time-stamped splits and rolling ingestion to mimic weekly operations. Even so, retrospective back-tests are gentler than reality: labels are eventually settled, metadata are cleaner in hindsight, and hyperparameters, selected on past data, still benefit from a privileged view of the future. For DeepAutoCoV, the weekly simulation approximates live use (retraining after dominance, top-N budgets), but the ground truth for "future dominant" is resolved post hoc. For the mlVAR/TimeGPT analyses, exogenous regressors are aligned to weekly cadences, yet some policy signals (e.g., local NPIs) are only coarsely captured. The thesis therefore proposes, but does not yet deliver, a prospective, "shadow deployment" in a national pipeline: alerts pre-registered and time-stamped, budgets fixed ex ante, success measured by lead time, precision, and false-alert rate per week, and all decisions audited for data leakage. Until such a study is run, the external validity of the gains reported here should be interpreted as promising rather than definitive.

Early-warning systems can change behavior; if poorly communicated, they can fuel over-reaction or fatigue. Data governance is equally central. Much of the most valuable sequence data are shared under restrictive terms; models trained on those data should preserve provenance, respect license constraints, and avoid redistributing sensitive content. The corpus assembly and embedding work therefore emphasize traceability (accessions retained, filters logged) and reproducibility, but further safeguards, such as federated fine-tuning across jurisdictions or privacy-preserving aggregation of features rather than raw sequences, are needed for routine cross-border use. Finally, dual-use risk is real for generative modeling. SARITA was developed strictly for *in silico* plausibility studies, without wet-lab execution.

Taken together, these limitations and ethical guardrails point to a pragmatic stance: the methods in this thesis are most valuable when used as decision aids embedded in transparent, human-in-the-loop workflows,

not as autonomous arbiters. In concrete terms, anomaly scores should prioritize what to test next, not declare a variant "of concern"; ΔΔG filters should triage structural hypotheses, not claim clinical impact; and short-term forecasts should complement, not replace, local situational awareness. With prospective evaluation, immunology-aware features, and stronger governance, the same tools can become safer and more reliable, preserving their lead-time advantage while aligning with the norms of public-health practice.

6.4. Future Directions: multi-omics, federated learning, explainability

A coherent next phase is to move from sequence-only signals to immunologically informed, and prospectively validated pipelines that keep humans in the loop. On the biology side, sequence anomalies and generative hypotheses should be weighted by antigenic context: token/k-mer saliency mapped deterministically to standard mutation notation and cross-referenced with deep-mutational-scanning escape maps, curated B- and T-cell epitope repertoires, glycosylation motifs, and proteolytic sites. Structure checks should evolve beyond single-complex ΔΔG proxies toward ensembles that reflect conformational heterogeneity and glycan dynamics (e.g., MD-backed rotamer sampling, glycan shielding/epitope masking), and toward polyclonal views that summarize panels of sera and monoclonals rather than isolated binders. The scope should also widen beyond Spike S1: S2 and other structural/non-structural proteins modulate fusion, replication, and immune evasion and therefore belong in the risk calculus. For SARITA-style generators, conditioning can incorporate immunogenicity priors and hard biological constraints so that sampling is steered toward sequences that are both biophysically plausible and antigenically interpretable. The same playbook should be stress-tested across pathogens (e.g., influenza HA/NA and other respiratory viruses) to enable transfer learning where antigenic principles are shared.

To make these capabilities routinely usable across jurisdictions, models ought to be trained where the data live. A federated loop, weekly fine-tuning of anomaly detectors and embeddings at national or regional labs, can capture local lineage dynamics without centralizing raw genomes or sensitive metadata. Robust federated optimization (e.g., FedAvg/FedProx-style variants for skewed clients), secure aggregation, and optional differential privacy on updates would reduce re-identification risk; where law or policy forbids sequence sharing outright, feature-level federation

(hashed k-mers or compact embedding sketches) provides a workable compromise. Governance should be first-class: model cards, license carry-through, audit trails of filters and thresholds, per-site validation reports, and automatic drift monitoring so that changes in chemistry, sampling, or base-caller versions are visible and acted upon. In practice, DeepAutoCoV would adapt its alert thresholds to local prevalence while contributing to a global prior over anomaly patterns; MistrVirus would receive periodic federated refreshes so its representation stays current as taxa and protocols evolve.

Interpretability must become a primary output. Every alert should come with a concise, reviewer-ready explanation: which residues drive the score, what functional annotation they touch (domain/epitope/glycan motif), where they sit on a structure snapshot, and how certain the system is. Uncertainty should be quantified alongside saliency, e.g., conformal risk scores or calibrated intervals, so decision makers see both "why" and "how sure." Human oversight can then be formalized rather than ad hoc: curators accept or override the ranked shortlist; their feedback updates local thresholds and the global prior; dashboards track the precision/lead-time trade-off under fixed weekly top-N budgets. For generators, safeguards should combine ex-ante constraints (forbidden motifs, maximum edit budgets, rejection of unsafe patterns) with post-hoc rationales ($\Delta\Delta G$ contributors, clashes with known epitopes) to keep exploration safe and intelligible. Explanations for representation models should surface prototypical neighbors and simple, auditable decision boundaries, allowing lightweight classifiers on MistrVirus to be inspected rather than treated as black boxes.

All of this should be exercised prospectively, end-to-end, inside operational pipelines. A pragmatic "shadow deployment" would pre-register metrics and budgets; ingest sequences on a rolling, time-stamped schedule; fix decision thresholds ex ante; and evaluate outcomes with lead time, precision, false-alert rate, and, crucially, downstream actionability (e.g., fraction of primer updates or neutralization assays correctly anticipated). Outputs from DeepAutoCoV, MistrVirus, and SARITA can be fused with phylodynamic and epidemiological layers into a single early-warning score that blends sequence-level risk with expected spread (growth-rate/Re estimates) while correcting for sampling biases; the same score can drive resource allocation,what to sequence next, what to test in the lab, what to elevate in national briefings,and leave a complete audit trail in MLOps-style infrastructure. Extending beyond clinical samples to wastewater and sentinel networks will test robustness to noisier inputs and broaden coverage. In sum, the most promising path forward couples

richer biology with stronger privacy and standardized interpretability, then proves its worth in prospective use, so these tools not only detect and explain molecular change, but also reliably trigger timely, proportionate public-health action.

Chapter 7

7 Appendix

Table A1: The contemporaneous bivariate correlations between each index and the weekly epidemiological variables for 2021 were analyzed across all countries. Significant results ($p < 0.05$) are highlighted in bold. Correlations were calculated by combining data from all available countries into a single vector for each variable.

	SI
New Cases	-0.04
adm Hosp	0.17
adm ICU	0.10
death	0.31
First Dose	0.19
Second Dose	-0.15
Booster	-0.20
Alpha	0.71

Beta	0.34
Delta	-0.72
Zeta	0.26
Theta	0.12
Gamma	0.23
Epsilon	0.17
Eta	0.16
Kappa	0.08
Iota	0.15
Mu	0.05
Lambda	0.06
GH490R	-0.07
Omicron	-0.07

Table A2: The contemporaneous bivariate correlations between each index and the weekly epidemiological variables were computed for 2022 by pooling observations across all countries. Statistically significant correlations ($p < 0.05$) are shown in bold. For each variable, country-specific weekly series were concatenated into a single vector prior to correlation analysis.

	SI
New Cases	0.73
adm Hosp	0.41
adm ICU	0.47
death	0.53
First Dose	0.50
Second Dose	0.44
Booster	0.44
Delta	0.54
GH490R	0.11
Omicron	-0.06
Alpha	0.13
BA.2.75	-0.27
CH.1.1	-0.14

XBB	-0.11
XBB.1.5	-0.05

Table A3: FDLs identified in the Global data set. Highlighted variants: yellow, Alpha; purple, Delta (sub)lineages; green, Omicron (sub)lineages. 1. Week number for FDL first appearance. 2. Week number for FDL first detection (frequency). 3. Week number for FDL reaching the threshold and becoming dominant.

Lineage	First appearance[1]	First flagging[2]	10% threshold[3]
B.1	2 (1st week, Jan 2020)	2 (1st week, Jan 2020, 9%)	11 (2nd week, Mar 2020)
B.1.1	7 (2nd week, Feb 2020)	Unflagged	12 (3rd week, Mar 2020)
B.1.2	4 (3rd week, Jan 2020)	Unflagged	45 (1st week, Nov 2020)
B.1.177	12 (3rd week, Mar 2020)	13 (4th week, Mar 2020, 0.5%)	41 (1st week, Oct 2020)
B.1.1.7	7 (2nd week, Feb 2020)	7 (2nd week, Feb 2020, 0.7%)	53 (4th week, Dec 2020)
B.1.617.2	15(2nd week, Apr 2020)	15 (2nd week, Apr 2020, 0.01%)	87 (4th week, Aug 2021)

B.1.1.284	10 (1st week, Mar 2020)	10 (1st week, Mar 2020, 1%)	32 (1st week, Aug 2020)
AY.29	64 (2nd week, Mar 2021)	64 (2nd week, Mar 2021, 0.01%)	87 (2nd week, Sep 2021)
AY.4	34 (3rd week, Aug 2020)	34 (3rd week, Aug 2020, 0.02%)	77 (3rd week, Jun 2021)
AY.43	37 (2nd week, Sep 2020)	37 (2nd week, Sep 2020, 0.02%)	105 (4th week, Dec 2021)
AY.44	27 (2nd week, Jul 2020)	27 (2nd week, Jul 2020, 0.03%)	105 (4th week, Dec 2021)
AY.103	19 (2nd week, May 2020)	19(2nd week May 2020, 0.02%)	104 (3rd week, Dec 2021)
BA.1	100 (3rd week, Nov 2021)	100 (3rd week Nov 2021, 0.01%)	110 (3rd week, Jan 2022)
BA.2	100 (3rd week, Nov 2021)	100 (3rd week Nov 2021, 0.03%)	110 (3rd week, Jan 2022)
BA.2.9	103 (2nd week, Dec 2021)	103 (2nd week Dec 2021, 0.03%)	110 (3rd week, Jan 2022)
BA.2.12.1	105 (4th week, Nov 2021)	114 (3rd week, Feb 2021, 0.05%)	124 (2nd week, May 2022)

BA.5.1	121 (3rd week, Apr 2022)	121 (3rd week, Apr 2022, 0.02%)	137 (3rd week, Aug 2022)
BA.5	121 (3rd week, Apr 2022)	121 (3rd week, Apr 2022, 0.08%)	137 (3rd week, Aug 2022)
BA.5.2	123 (1st week, May 2022)	127 (1st week, May 2022, 0.05%)	140 (2nd week, Sep 2022)
XBB.1.5	140 (2nd week, Sep 2022)	140 (2nd week, Sep 2022, 0.1%)	163 (4th week, Feb 2023)
BR.2.1	141 (3rd week, Sep 2022)	141 (3rd week, Sep 2022, 0.1%)	153 (4th week, Dec 2022)
CH.1.1	144 (2nd week, Oct 2022)	144 (2nd week, Oct 2022, 0.3%)	167 (4th week, Mar 2023)
FK.1.1	164 (1st week, Mar 2023)	164 (1st week, Mar 2023, 0.1%)	182 (4th week, Jun 2023)
XBC.1.6	173 (2nd week, Feb 2023)	173 (3rd week, Mar 2023, 1%)	182 (4th week, Jun 2023)
XBC.1.3	142 (1st week, Oct 2022)	142 (1st week, Oct 2022, 0.3%)	185 (3rd week, Jul 2023)
DV.7.1	181 (3rd week, Jun 2023)	181 (3rd week, Jun 2023, 1%)	192 (2nd week, Sep 2023)
HW.1.1	181 (3rd week, Jun 2023)	181 (3rd week, Jun 2023, 1%)	195 (1st week, Oct 2023)

EG.5.1	180 (2nd week, Jun 2023)	Unflagged	197 (3rd week, Oct 2023)

Table A4: FDLs identified in the USA data set. Highlighted variants: yellow, Alpha; purple, Delta (sub)lineages; green, Omicron (sub)lineages. 1. Week number for FDL first appearance. 2. Week number for FDL first detection (frequency). 3. Week number for FDL reaching the threshold and becoming dominant.

FDLs	First appearance[1]	First flagging[2]	10% threshold[3]

B.1	8 (4th week, Feb 2020)	11 (3rd week, Mar 2020, 12%)	11 (3rd week, Mar 2020)
B.1.2	2 (2nd week, Jan 2020)	12 (4th week, Mar 2020, 1.2%)	29 (3ht week, Jul 2020)
B.1.429	4 (4th week, Jan 2020)	4 (4th week, Jan 2020, 7%)	54 (2nd week, Jan 2021)
B.1.1.7	9 (1st week, Mar 2020)	9 (1st week, Mar 2020, 0.2%)	61 (2nd week, Mar 2021)
B.1.526	16 (2nd week, Apr 2020)	16 (2nd week, Apr 2020, 0.1%)	64 (1st week, Apr 2021)
AY.3	35 (1st week, Sep 2020)	35 (1st week, Sep 2020, 0.04%)	86 (3rd week, Sep 2021)
AY.25	17 (3rd week, Apr 2020)	17 (3rd week, Apr 2020, 0.04%)	86 (3rd week, Sep 2021)
AY.103	11 (3rd week, Mar 2020)	11 (3rd week, Mar 2020, 0.06%)	83 (4th week, Aug 2021)
AY.44	15 (1st week, Apr 2020)	15 (1st week, Apr 2020, 0.04%)	79 (4th week, Jul 2021)
B.1.617.2	28 (2nd week, Jul 2020)	28 (2nd week, Jul 2020, 0.2%)	80 (1st week, Aug 2021)

BA.1	102 (2nd week, Dec 2021)	102 (2nd week, Dec 2021, 0.05%)	110 (2nd week, Feb 2022)
BA.2.9	105 (1st week, Jan 2022)	108 (4th week, Jan 2022, 0.2%)	117(1st week, Apr 2022)
BA.2	105 (1st week, Jan 2022)	105 (2nd week, Jan 2022, 0.04%)	112 (4th week, Feb 2022)
BA.2.3	105 (1st week, Jan 2022)	109 (1st week, Feb 2022, 1%)	117(1st week, Apr 2022)
BA.2.12.1	102 (2nd week, Dec 2021)	102 (2nd week, Dec 2021, 0.03%)	120 (4th week, Apr 2022)
BA.5.2.1	127 (1st week, May 2022)	127 (1st week, May 2022, 0.1%)	136 (4th week, Aug 2022)
BA.5	128 (2nd week, Jun 2022)	128 (2nd week, Jun 2022, 0.3%)	136 (4th week, Aug 2022)
BA.5.1	127 (3rd week, Jun 2022)	127 (3rd week, Jun 2022, 0.6%)	136 (4th week, Aug 2022)
BA.5.2	127 (3rd week, Jun 2022)	129 (1st week, Jul 2022, 0.1%)	136 (4th week, Aug 2022)
BQ.1.1	142 (1st week, Oct 2022)	142 (1st week, Oct 2022, 1%)	157 (1st week, Jan 2023)
XBB.1.5	138 (1st week, Sep 2022)	138 (1st week, Sep 2022, 0.8%)	162 (2nd week, Feb 2023)

CH.1.1	144 (3rd week, Oct 2022)	144 (3rd week, Oct 2022, 0.8%)	169 (3rd week, Apr 2023)
FK.1.2.2	169 (3rd week, Apr 2023)	169 (3rd week, Apr 2023, 4%)	170 (4th week, Apr 2023)
FK.1.5	169 (3rd week, Apr 2023)	176 (2nd week, Jun 2023, 11%)	176 (2nd week, Jun 2023)
DV.1.1	164 (3rd week, Apr 2023)	164 (3rd week, Apr 2023, 1%)	176 (2nd week, Jun 2023)
FK.1.1	167 (2nd week, Apr 2023)	167 (2nd week, Apr 2023, 3%)	181(3rd week, Jul 2023)
XBB.1.16	180 (2nd week, Jul 2023)	184 (2nd week, Aug 2023, 3%)	188 (2nd week, Sep 2023)

Table A5: FDLs identified in the UK data set. Highlighted variants: yellow, Alpha; purple, Delta (sub)lineages; green, Omicron (sub)lineages. 1. Week number for FDL first appearance. 2. Week number for FDL first detection (frequency). 3. Week number for FDL reaching the threshold and becoming dominant.

FDLs	First appearance[1]	First flagging[2]	10% threshold[3]
B.1.177	18 (3rd week, May 2020)	18 (3rd week, May 2020, 0.3%)	31 (3rd week, Aug 2020)
B.1.1.7	34 (4th week, Aug 2020)	34 (4th week, Aug 2020, 0.05%)	45 (1st week, Nov 2020)
B.1.1	3 (3rd week, Feb 2020)	5 (1st week, Mar 2020, 8%)	6 (2nd week, Mar 2020)
B.1	4 (4th week, Feb 2020)	4 (4th week, Feb 2020, 7%)	6 (4th week, Mar 2020)
AD.2	13 (2nd week, Apr 2020)	13 (2nd week, Apr 2020,0.1%)	31 (3rd week, Aug 2020)

AY.4	54 (3rd week, Jan 2020)	54 (3rd week, Jan 2020, 0.01%)	69 (3rd week, Mar 2021)
AY.4.2	71 (2nd week, Apr 2021)	71 (2nd week, Apr 2021, 0.1%)	95 (1st week, Oct 2021)
AY.4.2.2	78 (2nd week, Jun 2021)	79 (2nd week, Jun 2021, 0.01%)	101 (1st week, Nov 2021)
BA.2	103 (3rd week, Nov 2021)	103 (2nd week, Nov 2021, 0.4%)	108 (2nd week, Feb 2022)
XE	104 (4th week, Jan 2022)	104 (4th week, Jan 2022, 5%)	108 (4th week, Feb 2022)
BA.5.2.1	128 (2nd week, Jun 2022)	128 (2nd week, Jun 2022, 9%)	132 (2nd week, Aug 2022)
BA.1	101 (1st week, Nov 2021)	101 (1st week, Nov 2021, 0.02%)	108 (2nd week, Feb 2022)
BA.2.12.1	119 (2nd week, Apr 2022)	119 (2nd week, Apr 2022, 0.8%)	123 (2nd week, May 2022)
BA.2.75	126 (1st week, Jun 2022)	126 (1st week, Jun 2022, 4%)	131 (1st week, Aug 2022)
BA.2.76	129 (4th week, Jun 2022)	129 (4th week, Jun 2022, 8%)	131 (1st week, Aug 2022)
BN.1	140 (2nd week, Oct 2022)	140 (2nd week, Oct 2022, 8%)	142 (4th week, Oct 2022)

CH.1.1	145 (1st week, Dec 2022)	145 (1st week, Dec 2022, 9%)	147 (3rd week, Dec 2023)
DV.6	155 (1st week, Feb 2023)	155 (1st week, Jan 2023,2%)	161 (4th week, Mar 2023)
DV.7	159 (2nd week, Mar 2023)	159 (2nd week, Mar 2023, 2%)	168 (2nd week, May 2023)
XBB.1.5	172 (2nd week, Jun 2023)	172 (2nd week, Jun 2023, 9%)	174 (4th week, Jun 2023)
XBB.1.16	172 (2nd week, Jun 2023)	172 (2nd week, Jun 2023, 9%)	174 (4th week, Jun 2023)
EG.5.1	177 (3rd week, Jul 2023)	177 (3rd week, Jul 2023, 8%)	183 (1st week, Sep 2023)
DV.7.1	187 (1st week, Oct 2023)	187 (1st week, Oct 2023, 9%)	190 (4th week, Oct 2023)
EG.5.1.3	182 (4th week, Aug 2023)	182 (4th week, Aug 2023, 8%)	190 (4th week, Oct 2023)

Table A6: FDLs identified in the DENMARK data set. Highlighted variants: yellow, Alpha; purple, Delta (sub)lineages; green, Omicron (sub)lineages. 1. Week number for FDL first appearance. 2. Week number for FDL first detection (frequency). 3. Week number for FDL reaching the threshold and becoming dominant.

FDLs	First appearance[1]	First flagging[2]	10% threshold[3]
B.1.177	26 (2nd week, Aug 2020)	26 (2nd week, Aug 2020, 4%)	29 (1st week, Sep 2020)

B.1.1	2(2nd week, Feb 2020)	3 (3rd week, Feb 2020, 4%)	18 (4th week, Jul 2020)
B.1.160	22 (2nd week, Aug 2020)	26 (1st week, Sep 2020, 0.2%)	29 (1st week, Sep 2020)
B.1.1.7	37 (1st week, Nov 2020)	37 (1st week, Nov 2020, 0.4%)	49 (1st week, Feb 2021)
AY.43	67 (4th week, Apr 2021)	71(4rd week, May 2021, 3%)	75 (4th week, Jun 2021)
AY.4	67 (4th week, Apr 2021)	67 (4th week, Apr 2021, 0.2%)	73 (2nd week, Jun 2021)
AY.122	67 (4th week, Apr 2021)	67 (4th week, Apr 2021, 0.4%)	75 (4th week, Jun 2021)
AY.7.1	68 (2nd week, May 2021)	72 (1st week, Jun 2021,0.3%)	75 (4th week, Jun 2021)
AY.4.6	75 (4th week, Jun 2021)	75 (4th week, Jun 2021, 0.3%)	81 (2nd week, Aug 2021)
BA.2	94 (3rd week, Nov 2021)	95 (4st week, Nov 2021, 0.1%)	98 (3rd week, Dec 2021)
BA.2.9	94 (3rd week, Nov 2021)	95 (4st week, Nov 2021, 0.07%)	98 (3rd week, Dec 2021)
BA.2.12.1	111 (4th week, Mar 2022)	113 (2nd week, Apr 2022, 0.8%)	120 (1st week, Jun 2022)

BA.5	117 (2nd week, Mar 2022)	121 (1st week, Jun 2022, 0.2%)	128 (4th week, Jul 2022)
BA.5.1	121 (1st week, Jun 2022	121 (1st week, Jun 2022, 0.1%)	131 (3rd week, Aug 2022)
BA.5.2.1	113 (2nd week, Feb 2022)	113 (2nd week, Feb 2022, 0.03%)	128 (4th week, Jul 2022)
BF.7	121 (2nd week, Jun 2022)	121 (2nd week, Jun 2022, 0.2%)	130 (2nd week, Aug 2022)
BQ.1.1	135 (3rd week, Sep 2022)	135 (3rd week, Sep 2022, 0.05%)	148 (1st week, Jan 2023)
CH.1.1	143 (1st week, Oct 2022)	143 (3st week, Oct 2022, 0.9%)	155 (4th week, Feb 2023)
XBB.1.5	148 (1st week, Jan 2023)	148 (1st week, Jan 2023, 0.6%)	160 (1st week, Apr 2023)
FK.1.2.1	155 (4th week, Feb 2023)	161(2n week, Apr 2023, 5%)	161(2n week, Apr 2023)
DV.6	150 (3rd week, Jan 2023)	150 (3rd week, Jan 2023, 1%)	163 (4th week, Apr 2023)
EG.5	181 (2nd week, Sep 2023)	181 (2nd week, Sep 2023, 30%)	181 (2nd week, Sep 2023)
DV.7.1	181 (2nd week, Sep 2023)	181(2nd week, Sep 2023, 4%)	187 (1st week, Nov 2023)

Table A7: FDLs identified in the FRANCE data set. Highlighted variants: yellow, Alpha; purple, Delta (sub)lineages; green, Omicron (sub)lineages. 1. Week number for FDL first appearance. 2. Week number for FDL first detection (frequency). 3. Week number for FDL reaching the threshold and becoming dominant.

FDLs	First appearance [1]	Firs flagging [2]	10% threshold [3]
B.1.160	2 (2nd week, Jan 2020)	2 (2nd week, Jan 2020, 3%)	28 (2nd week, Jul 2020)
B.1	8 (3rd week, Feb 2020)	4 (4th week, Feb 2020, 7%)	13 (4th week, Mar 2020)
B.1.367	30 (1st week, Aug 2020)	30 (1st week, Aug 2020, 6%)	34 (4th week, Aug 2020)
B.1.177	18 (1st week, May 2020)	18 (1st week, May 2020, 0.5%)	45 (1st week, Nov 2020)
B.1.1.7	45 (2nd week, Nov 2020)	45 (2nd week, Nov 2020, 0.3%)	55 (3rd week, Jan 2021)
B.1.617.2	56 (4st week, Jan 2021)	56 (4st week, Jan 2021, 0.5%)	76 (3rd week, May 2021)
AY.98.1	72 (3rd week, Apr 2021)	72 (3rd week, Jun 2021, 0.2%)	84 (4th week, Jul 2021)
AY.42	48 (4th week, Nov 2022)	48 (4th week, Nov 2022, 0.04%)	82 (2nd week, Jul 2021)
AY.43	53 (1st week, Jan 2021)	53 (1st week, Jan 2021, 0.2%)	82 (2nd week, Jul 2021)

AY.4	70 (1st week, Apr 2021)	70 (1st week, Apr 2021, 0.6%)	100 (2nd week, Dec 2021)
AY.5	75 (2nd week, May 2021)	80 (3rd week, Jun 2021, 0.08%)	100 (2nd week, Dec 2021)
AY.122	70 (1st week, Apr 2021)	78 (1st week, Jun 2021, 1%)	94 (4th week, Oct 2021)
AY.125	78 (3rd week, Jun 2020)	79(3rd week, Jun 2021, 1%)	94 (4th week, Oct 2021)
BA.2	105 (3rd week, Feb 2022)	105 (3rd week, Feb 2022, 0.2%)	112 (2nd week, Apr 2022)
BA.2.12.1	123 (1st week, Jun 2022)	124(2nd week, Jun 2022, 0.8%)	130 (1st week, Sept 2022)
BA.2.9	108 (2nd week, Mar 2022)	108 (2nd week, Mar 2022, 1%)	142 (3rd week, Nov 2022)
BA.1	103 (1st week, Feb 2022)	104 (2nd week, Feb 2022, 0.6%)	110 (4th week, Mar 2022)
BA.2.75	135 (2nd week, Oct 2022)	135 (2nd week, Oct 2022, 0.4%)	144 (1st week, Dec 2022)
BA.5	123 (1st week, Jun 2022)	123 (1st week, Jun 2022, 1%)	138 (4th week, Oct 2022)
BE.6	142 (3rd week, Nov 2022)	147 (3rd week, Dec 2022, 9%)	148 (1st week, Jan 2023)

BF.7	135 (2nd week, Oct 2022)	135 (2nd week, Oct 2022, 1.4%)	148 (1st week, Jan 2023)
BQ.1.1	145 (2nd week, Dec 2023)	145 (2nd week, Dec 2023, 0.05%)	152 (1st week, Feb 2023)
BN.1	151 (4th week, Dec 2023)	151 (4th week, Dec 2023, 4%)	161 (3rd week, Apr 2023)
CH.1.1	151 (1st week, Feb 2022)	151 (1st week, Feb 2022, 2%)	159 (1st week, Apr 2023)

Table A8: Information about training times and batches.

	Training times	**Batches**
SARITA-S	6 hours, 21 minutes and 15 seconds	8 (Training Data) 8 (Validation Data)
SARITA-M	15 hours, 30 minutes, 54 seconds	8 (Training Data) 8 (Validation Data)
SARITA-L	1 day, 1 hour, 17 minutes, 45 seconds	8 (Training Data) 8 (Validation Data)
SARITA-XL	1 day, 12 hours, 32 minutes, 27 seconds	8 (Training Data) 8 (Validation Data)

Table A9: Correlation (Spearman) of mutation frequency in logarithmic scale between SARITA sequences and test set with respective p-value.

Models	Test - Synthetic
SARITA-S	0.66 (p value 2.2 e-16)
SARITA-L	0.64 (p value 2.2 e-16)
SARITA-XL	0.60 (p value 2.2 e-16)

Table A10: Genome Counts by Source in Final Dataset.

Source	Genomes
NCBI Virus	2,212,661
BV-BRC	718,162
GISAID	19,921

LANL-HIV-db	21,114
HBVdb	3,914

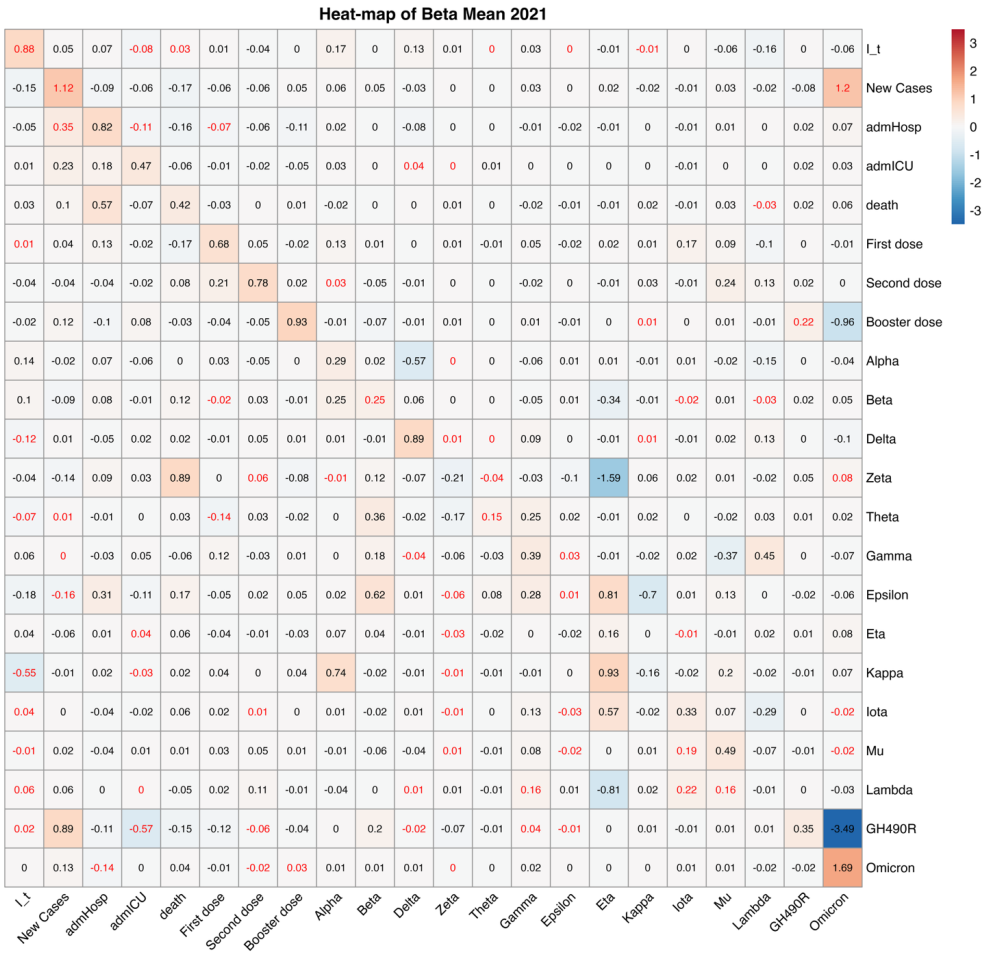


Figure A1: Heatmap of the relevant fixed effects' coefficients for the year 2021. The significant ones (P-Value < 0.05) are highlighted in red. The rows indicate the variables at time t, i.e. the outcomes, whereas the columns indicate the variables at time t – _1, i.e. the covariates. For instance, in the second row, New Cases is the outcome variable, while each column (I_t , New Cases, AdmHosp, AdmICU, Death, First Dose, Second Dose, Alpha, Beta, Delta, Zeta, Theta, Gamma, Epsilon, Eta, Kappa, Iota, Mu, Lambda, GH490R, Omicron) represents a covariate used by mlVAR to predict this outcome.

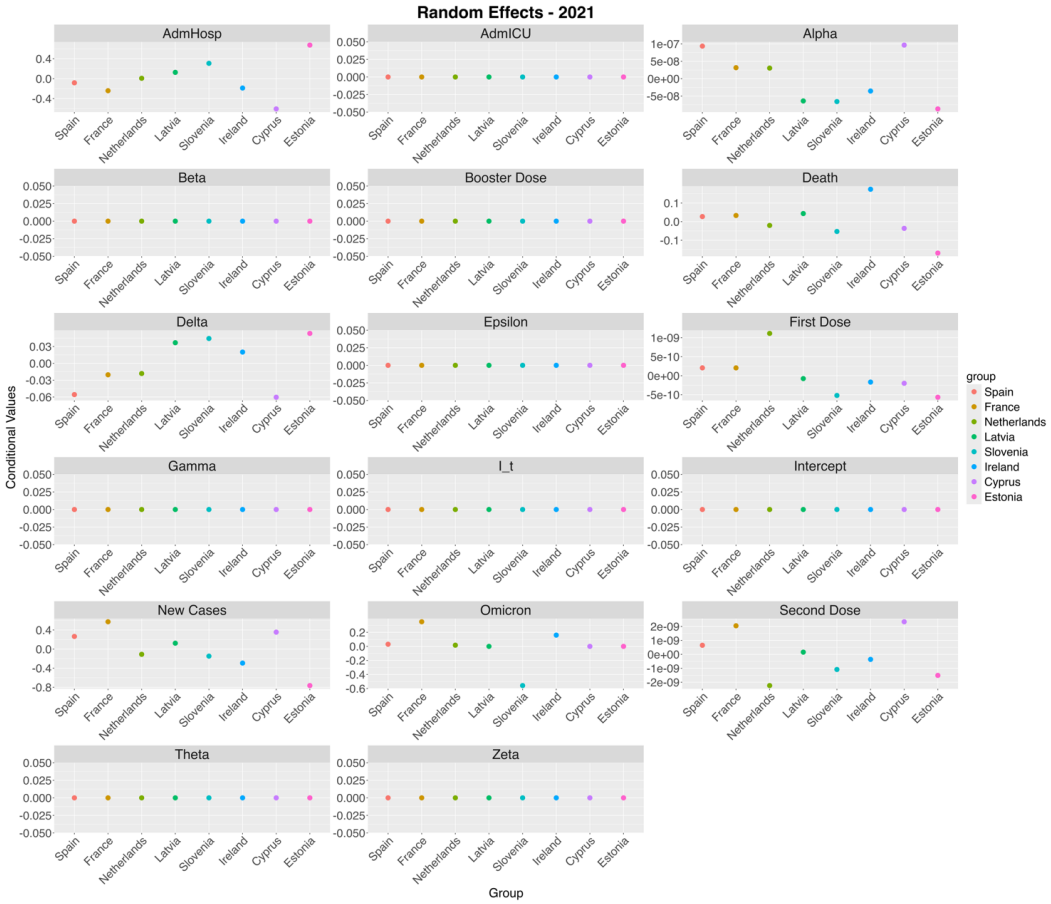


Figure A2: Random effects (country-wise variation on the model intercept) for all predictors with stringency index as outcome in 2021.

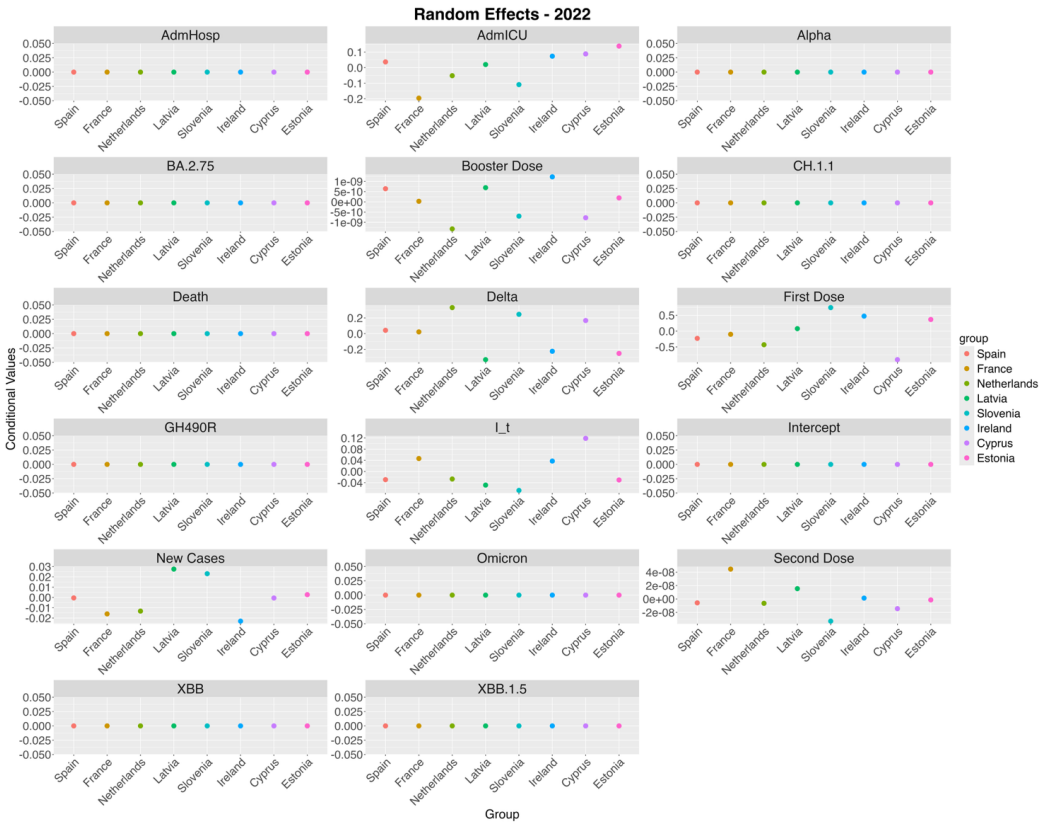


Figure A3: Random effects (country-wise variation on the model intercept) for all predictors with stringency index as outcome in 2022.

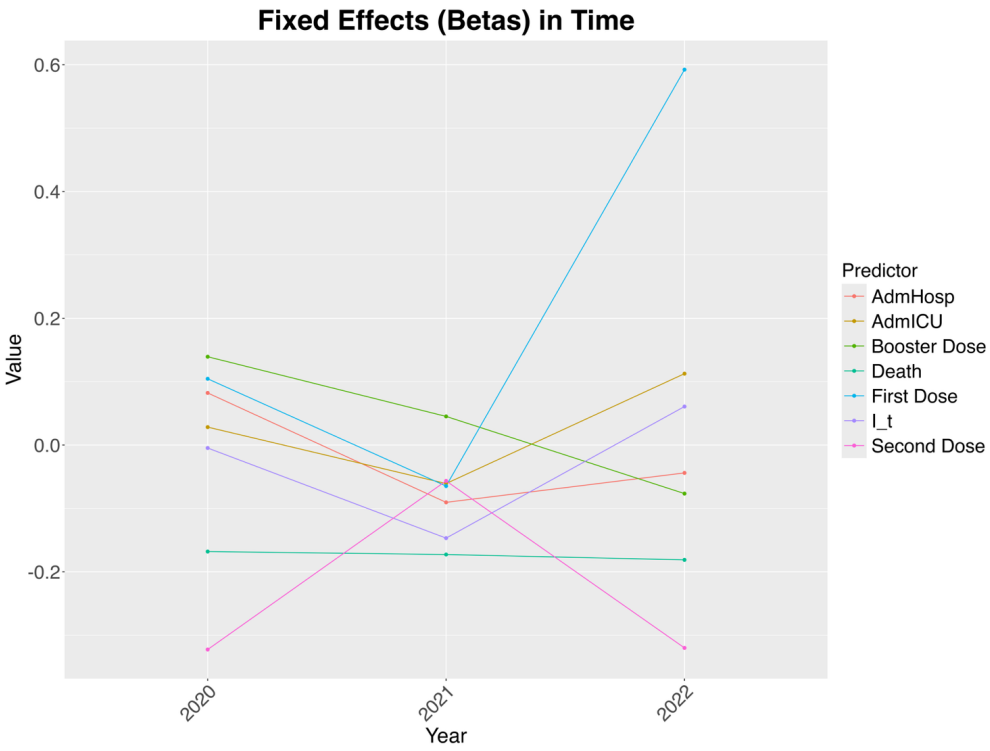


Figure A4: Variations of the fixed effects coefficients indicating the effects on new cases of the other epidemiological variables and the stringency index in the three considered years.

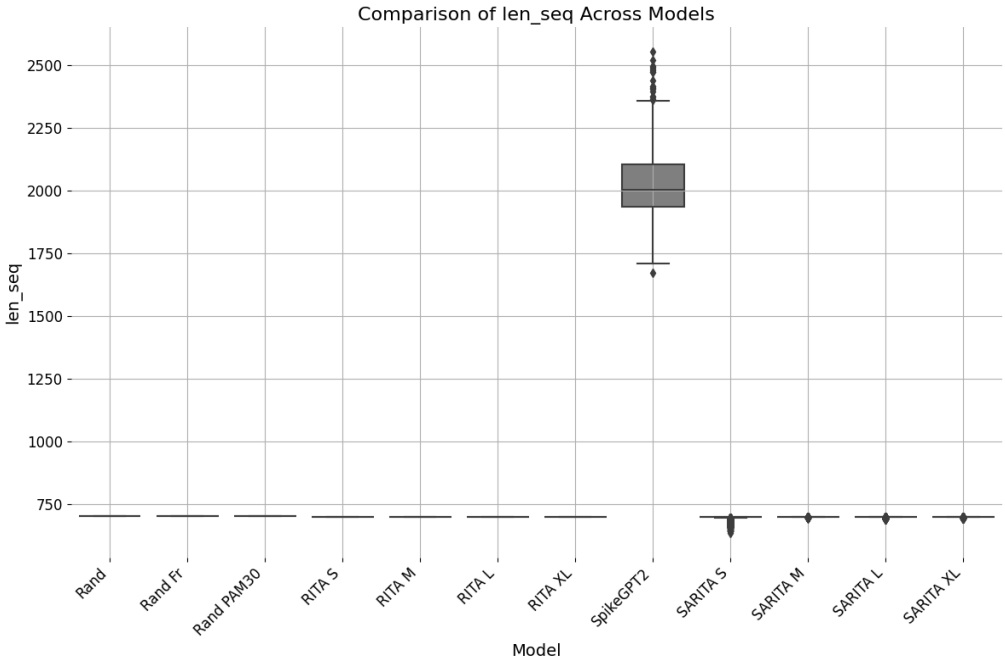


Figure A5: Boxplot illustrating the distribution of sequence lengths (len_seq) generated in synthetic sequences. Notably, SpikeGPT2 does not adhere to user-defined limits for creating sequences with a fixed length.

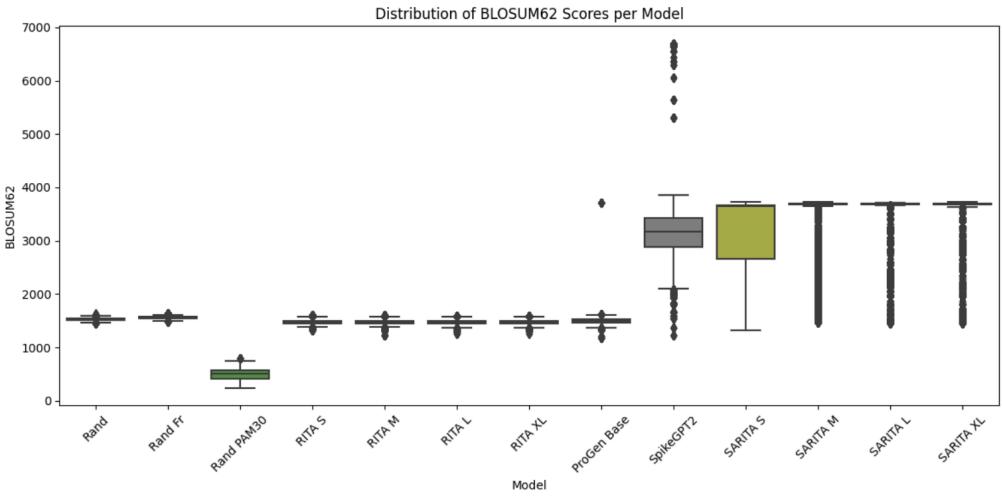


Figure A6: Boxplot showing the BLOSUM62 distribution scores computed on the synthetic sequences (higher BLOSUM62 score means that the sequences are strongly related)

References

[1] J. Piret and G. Boivin, "Pandemics throughout history," *Frontiers in Microbiology*, vol. 11, p. 631736, 2021, doi: 10.3389/fmicb.2020.631736.

[2] C. R. Harris *et al.*, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020, doi: 10.1038/s41586-020-2649-2.

[3] S. F. Fujimura, "Purple Death: The Great Flu of 1918," Pan American Health Organization (PAHO), 2003. [Online]. Available: https://www.paho.org/en/who-we-are/history-paho/purple-death-great-flu-1918. Accessed: Sep. 7, 2025.

[4] Centers for Disease Control and Prevention (CDC), "*Emerging Infectious Diseases*, vol. 12, no. 1,January 2006 (Influenza theme issue)," 2006. [Online]. Available: https://wwwnc.cdc.gov/eid/. Accessed: Sep. 7, 2025.

[5] [M. E. Nickol and J. Kindrachuk, "A year of terror and a century of reflection: perspectives on the great influenza pandemic of 1918–1919," *BMC Infectious Diseases*, vol. 19, no. 1, p. 117, 2019, doi: 10.1186/s12879-019-3750-8.

[6] E. Moon, "How the U.S. fought the 1957 flu pandemic," *Smithsonian Magazine*, Jun. 2020. [Online]. Available: https://www.smithsonianmag.com/smithsonian-institution/united-states-vaccine-1957-flu-pandemic-180974906/. Accessed: Sep. 7, 2025.

[7] ANZIC Influenza Investigators, “Critical care services and 2009 H1N1 influenza in Australia and New Zealand,” *New England Journal of Medicine*, vol. 361, no. 20, pp. 1925–1934, 2009, doi: 10.1056/NEJMoa0908481.

[8] A. Maccaro *et al.*, “Pandemic Preparedness: A Scoping Review of Best and Worst Practices from COVID-19,” *Healthcare (Basel)*, vol. 11, no. 18, p. 2572, 2023, doi: 10.3390/healthcare11182572.

[9] W. Msemburi *et al.*, “The WHO estimates of excess mortality associated with the COVID-19 pandemic,” *Nature*, vol. 613, pp. 130–137, Jan. 2023, doi: 10.1038/s41586-022-05522-2.

[10] S. Naseer *et al.*, “COVID-19 outbreak: Impact on global economy,” *Frontiers in Public Health*, 2023, doi: 10.3389/fpubh.2023.1134985.

[11] [M. E. El Zowalaty and J. D. Järhult, “From SARS to COVID-19: A previously unknown SARS-related coronavirus (SARS-CoV-2) of pandemic potential infecting humans,Call for a One Health approach,” *One Health*, vol. 9, p. 100124, 2020, doi: 10.1016/j.onehlt.2020.100124.

[12] Society of Critical Care Medicine, “Configuring the ICU in the COVID-19 era: Real-world experience,” May 7, 2020. [Online]. Available: https://sccm.org/blog/configuring-the-icu-in-the-covid-19-era-real-world-experience. Accessed: Sep. 7, 2025.

[13] E. Petersen *et al.*, “Comparing SARS-CoV-2 with SARS-CoV and influenza pandemics,” *The Lancet Infectious Diseases*, vol. 20, no. 9, pp. e238–e244, 2020, doi: 10.1016/S1473-3099(20)30484-9.

[14] A. Ghosh *et al.*, “An insight into SARS-CoV-2 structure, pathogenesis, target hunting for drug development and vaccine initiatives,” *RSC Medicinal Chemistry*, vol. 13, pp. 647–675, 2022, doi: 10.1039/D2MD00087A.

[15] J. Zhao *et al.*, "The potential intermediate hosts for SARS-CoV-2," *Frontiers in Microbiology*, vol. 11, p. 580137, 2020, doi: 10.3389/fmicb.2020.580137.

[16] S. Lytras *et al.*, "The animal origin of SARS-CoV-2," *Science*, vol. 373, no. 6558, pp. 968–970, 2021, doi: 10.1126/science.abh0117.

[17] J. W. LeDuc and B. McNabb, "SARS, the First Pandemic of the 21st Century," *Emerging Infectious Diseases*, vol. 10, no. 11, p. e26, 2004, doi: 10.3201/eid1011.040797.

[18] A. K. D. Akhilesh and K. H. Dhanasekaran, "COVert IDentities of current worldwide pandemic," *International Reviews of Immunology*, vol. 40, no. 1–2, pp. 1–4, 2021, doi: 10.1080/08830185.2021.1874362.

[19] T. P. Peacock *et al.*, "The furin cleavage site in the SARS-CoV-2 spike protein is required for transmission in ferrets," *Nature Microbiology*, vol. 6, pp. 899–909, 2021, doi: 10.1038/s41564-021-00908-w.

[20] G. Giwangkancana *et al.*, "Lunar New Year and Eid al-Fitr: the circle of COVID-19," *IJID Regions*, vol. 6, pp. 127–129, 2023, doi: 10.1016/j.ijregi.2023.07.003.

[21] Wikipedia, "Coronavirus spike protein," last updated Jul. 10, 2025. [Online]. Available: https://en.wikipedia.org/wiki/Coronavirus_spike_protein. Accessed: Sep. 7, 2025.

[22] L. Duan *et al.*, "The SARS-CoV-2 Spike Glycoprotein Biosynthesis, Structure, Function, and Antigenicity: Implications for the Design of Spike-Based Vaccine Immunogens," *Frontiers in Immunology*, vol. 11, p. 576622, 2020, doi: 10.3389/fimmu.2020.576622.

[23] S. Rancati *et al.*, "SARITA: a large language model for generating the S1 subunit of the SARS-CoV-2 spike protein," *Briefings in Bioinformatics*, 2025, in press.

[24] A. M. Almehdi, G. Khoder, A. S. Alchakee, A. T. Alsayyid, N. H. Sarg, and S. S. M. Soliman, "SARS-CoV-2 spike protein: pathogenesis, vaccines, and potential therapies," *Infection*, vol. 49, no. 5, pp. 855–876, 2021, doi: 10.1007/s15010-021-01677-8.

[25] A. M. Carabelli *et al.*, "SARS-CoV-2 variant biology: immune escape, transmission and fitness," *Nature Reviews Microbiology*, vol. 21, no. 3, pp. 162–177, 2023, doi: 10.1038/s41579-022-00841-7.

[26] P. Mistry *et al.*, "SARS-CoV-2 Variants, Vaccines, and Host Immunity," *Frontiers in Immunology*, vol. 12, p. 809244, 2022, doi: 10.3389/fimmu.2021.809244.

[27] S. Ramesh *et al.*, "Emerging SARS-CoV-2 Variants: A Review of Its Mutations, Its Implications and Vaccine Efficacy," *Vaccines (Basel)*, vol. 9, no. 10, p. 1195, 2021, doi: 10.3390/vaccines9101195.

[28] A. K. Chakraborty, "Hyper-Variable Spike Protein of Omicron Corona Virus and Its Differences with Alpha and Delta Variants: Prospects of RT-PCR and New Vaccine," *Journal of Emerging Diseases and Virology*, vol. 7, no. 1, pp. 1–13, 2022. [Online].
Available: https://sciforschenonline.org/journals/virology/article-data/JEDV166/JEDV166.pdf. Accessed: Sep. 7, 2025.

[29] L. Fang *et al.*, "The effects of amino acid substitution of spike protein and genomic recombination on the evolution of SARS-CoV-2," *Frontiers in Microbiology*, vol. 14, p. 1228128, 2023, doi: 10.3389/fmicb.2023.1228128.

[30] A. Saha *et al.*, "Beyond the Pandemic Era: Recent Advances and Efficacy of SARS-CoV-2 Vaccines Against Emerging Variants of Concern," *Vaccines (Basel)*, vol. 13, no. 4, p. 424, 2025, doi: 10.3390/vaccines13040424.

[31] S. Khare et al., "GISAID's role in pandemic response," China CDC Weekly, vol. 3, no. 49, pp. 1049–1051, Dec. 2021, doi: 10.46234/ccdcw2021.255.

[32] UKHSA Genomics Public Health Analysis Team; UKHSA Infectious Disease Modelling Team; UKHSA TARZET Technical Secretariat, "SARS-CoV-2 genome sequence prevalence and growth rate update: 1 May 2024," GOV.UK, Aug. 8, 2024. [Online]. Available: https://www.gov.uk/…/sars-cov-2-genome-sequence-prevalence-and-growth-rate-update-1-may-2024. Accessed: Sep. 7, 2025.

[33] M. Roser, "What is the COVID-19 Stringency Index?," Our World in Data, Dec. 24, 2021. [Online]. Available: https://ourworldindata.org/metrics-explained-covid19-stringency-index. Accessed: Sep. 7, 2025. Our World in Data

[34] P. R. Martins-Filho et al., "Dynamics of hospitalizations and in-hospital deaths from COVID-19 in northeast Brazil: a retrospective analysis based on the circulation of SARS-CoV-2 variants and vaccination coverage," Epidemiology and Health, vol. 44, e2022036, 2022, doi: 10.4178/epih.e2022036.

[35] M. Van Beusekom, "COVID-19 death, hospitalization rates fell from 2020 to 2022," CIDRAP News, May 26, 2023. [Online]. Available: https://www.cidrap.umn.edu/covid-19/covid-19-death-hospitalization-rates-fell-2020-2022. Accessed: Sep. 7, 2025.

[36] S. Dhami et al., "Data-enabled responses to pandemics: policy lessons from COVID-19," Nature Medicine, vol. 28, no. 11, pp. 2243–2246, Nov. 2022, doi: 10.1038/s41591-022-02054-0.

[37] Our World in Data, "COVID-19 Data Explorer," Jul. 13, 2025. [Online].
Available: https://ourworldindata.org/explorers/covid.
Accessed: Sep. 7, 2025.

[38] X. Ding, B. Shang, C. Xie, J. Xin, and F. Yu, "Artificial intelligence in the COVID-19 pandemic: balancing benefits and ethical challenges in China's response," Humanities and Social Sciences Communications, vol. 12, art. 245, 2025, doi: 10.1057/s41599-025-04564-x.

[39] K. Norwood et al., “In silico genomic surveillance by CoVerage predicts and characterizes SARS-CoV-2 variants of interest,” Nature Communications, vol. 16, art. 6281, 2025, doi: 10.1038/s41467-025-60231-4.

[40] K. Gangavarapu et al., “Outbreak.info genomic reports: scalable and dynamic surveillance of SARS-CoV-2 variants and mutations,” Nature Methods, vol. 20, no. 4, pp. 512–522, Apr. 2023, doi: 10.1038/s41592-023-01769-3.

[41] G. Nicora et al., “Predicting emerging SARS-CoV-2 variants of concern through a One Class dynamic anomaly detection algorithm,” BMJ Health & Care Informatics, vol. 29, no. 1, e100643, 2022, doi: 10.1136/bmjhci-2021-100643.

[42] X. Zhuang et al., “Early detection of emerging SARS-CoV-2 variants from wastewater through genome sequencing and machine learning,” Nature Communications, vol. 16, art. 6272, Jul. 2025, doi: 10.1038/s41467-025-61280-5.

[43] N. Sapoval et al., “Enabling accurate and early detection of recently emerged SARS-CoV-2 variants of concern in wastewater,” Nature Communications, vol. 14, art. 2834, 2023, doi: 10.1038/s41467-023-38184-3.

[44] S. Li et al., “Machine learning early detection of SARS-CoV-2 high-risk variants,” Advanced Science, vol. 11, 2405058, 2024, doi: 10.1002/advs.202405058.

[45] S. Tosta et al., “Global SARS-CoV-2 genomic surveillance: What we have learned (so far),” Infection, Genetics and Evolution, vol. 108, 105405, 2023, doi: 10.1016/j.meegid.2022.105405.

[46] S. Nagpal et al., “Genomic surveillance of COVID-19 variants with language models and machine learning,” Frontiers in Genetics, vol. 13, 858252, 2022, doi: 10.3389/fgene.2022.858252.

[47] B. Hie, E. D. Zhong, B. Berger, and B. Bryson, “Learning the language of viral evolution and escape,” Science, vol. 371, no. 6526, pp. 284–288, Jan. 2021, doi: 10.1126/science.abd7331.

[48] N. N. Thadani et al., “Learning from prepandemic data to forecast viral escape,” Nature, vol. 622, pp. 818–825, Oct. 2023, doi: 10.1038/s41586-023-06617-0.

[49] F. Huang et al., “Identifying COVID-19 severity-related SARS-CoV-2 mutation using a machine learning method,” Life (Basel), vol. 12, no. 6, 806, 2022, doi: 10.3390/life12060806.

[50] Q. Sun et al., “VarEPS: an evaluation and prewarning system of known and virtual variations of SARS-CoV-2 genomes,” Nucleic Acids Research, vol. 50, no. D1, pp. D888–D897, Jan. 2022, doi: 10.1093/nar/gkab921.

[51] D. Panteli et al., “Artificial intelligence in public health: promises, challenges, and an agenda for policy makers and public health institutions,” The Lancet Public Health, vol. 10, no. 5, pp. e428–e432, 2025, doi: 10.1016/S2468-2667(25)00036-2.

[52] K. Nixon et al., “Real-time COVID-19 forecasting: challenges and opportunities of model performance and translation,” The Lancet Digital Health, vol. 4, no. 10, pp. e699–e701, Oct. 2022, doi: 10.1016/S2589-7500(22)00167-4.

[53] J.-W. Wu et al., “Assessment of the benefits of targeted interventions for pandemic control in China based on machine learning method and web service for COVID-19 policy simulation,” Biomedical and Environmental Sciences, vol. 35, no. 5, pp. 412–418, 2022, doi: 10.3967/bes2022.057.

[54] G. Nicora et al., “Bayesian networks in the management of hospital admissions: A comparison between explainable AI and black box AI during the pandemic,” J. Imaging, vol. 10, no. 5, 117, 2024, doi: 10.3390/jimaging10050117.

[55] S. Epskamp, M. K. Deserno, and L. F. Bringmann, mlVAR: Multi-Level Vector Autoregression. R package v0.5.2, CRAN, 2025. [Online]. Available: https://CRAN.R-project.org/package=mlVAR. Accessed: Sep. 7, 2025.

[56] University of Oxford, Blavatnik School of Government, "COVID-19 Government Response Tracker (OxCGRT)," Project page, 2025. [Online]. Available: https://www.bsg.ox.ac.uk/research/covid-19-government-response-tracker. Accessed: Sep. 7, 2025.

[57] European Centre for Disease Prevention and Control (ECDC), "Homepage," 2025. [Online]. Available: https://www.ecdc.europa.eu/en. Accessed: Sep. 7, 2025.

[58] China CDC Weekly, "GISAID's role in pandemic response," Feb. 10, 2025. [Online]. Available: https://weekly.chinacdc.cn/en/article/doi/10.46234/ccdcw2021.255. Accessed: Sep. 7, 2025.

[59] A. Garza, C. Challu, and M. Mergenthaler-Canseco, "TimeGPT-1," arXiv, Oct. 2023. [Online]. Available: https://arxiv.org/abs/2310.03589. Accessed: Sep. 7, 2025.

[60] C. Podrecca et al., "A comparative analysis of the effects of containment policies on the epidemiological manifestation of the COVID-19 pandemic across nine European countries," Scientific Reports, vol. 13, 11631, 2023, doi: 10.1038/s41598-023-37751-4.

[61] S. Rancati et al., "A comparative analysis of the role of containment policies, vaccination strategies and virus variants in the COVID-19 pandemic across nine European countries," Sci. Rep., vol. 15, no. 1, Art. no. 29109, Aug. 2025, doi: 10.1038/s41598-025-10132-9.

[62] S. Rancati et al., "Sequencing Efforts and Epidemiological Trends: Analyzing SARS-CoV-2 Dynamics Across European Nations," in Proc. 2024 IEEE Int. Conf. Bioinf. Biomed. (BIBM), Lisbon, Portugal, Dec. 2024, pp. 6183–6188. doi: 10.1109/BIBM61966.2024.10823311.

[63] R. S. Kubinec et al., “Statistically Validated Indices for COVID-19 Public Health Policies,” SocArXiv, Apr. 2021, doi: 10.31235/osf.io/rn9xk.

[64] S. Rancati, “Forecasting dominance of SARS-CoV-2 lineages by anomaly detection using deep AutoEncoders,” Brief. Bioinform., 2024, bbae145, doi: 10.1093/bib/bbae145.

[65] D. Ferrari et al., “The death of laparoscopy,” Surg. Endosc., vol. 38, no. 5, pp. 2677–2688, 2024, doi: 10.1007/s00464-024-10774-2.

[66] R. G. Moulder, N. D. Duran, and S. K. D’Mello, “Assessing Multimodal Dynamics in Multi-Party Collaborative Interactions with Multi-Level Vector Autoregression,” in Proc. ACM Int. Conf. Multimodal Interaction (ICMI ’22), Bengaluru, India, Nov. 2022, pp. 615–625, doi: 10.1145/3536221.3556595.

[67] G. Zhou, R. G. Moulder, C. Sun, and S. K. D’Mello, “Investigating Temporal Dynamics Underlying Successful Collaborative Problem Solving Behaviors with Multilevel Vector Autoregression,” in Proc. 15th Int. Conf. Educational Data Mining (EDM 2022), Durham, UK, Jul. 24–27, 2022. [Online].
Available: https://educationaldatamining.org/edm2022/proceedings/ (Accesso: 7 set. 2025).

[68] S. Epskamp, M. K. Deserno, and L. F. Bringmann, mlVAR: Multi-Level Vector Autoregression, R package version 0.5.2, 2024. [Online]. Available: https://CRAN.R-project.org/package=mlVAR (Accesso: 7 set. 2025).

[69] D. Bates et al., lme4: Linear Mixed-Effects Models using ‘Eigen’ and S4, R package version 1.1-37, Mar. 26, 2025. [Online]. Available: https://cran.r-project.org/package=lme4 (Accesso: 7 set. 2025).

[70] S. A. Lauer et al., "The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application," Ann. Intern. Med., vol. 172, pp. 577–582, 2020, doi: 10.7326/M20-0504.

[71] S. W. Park et al., "Inferring the differences in incubation-period and generation-interval distributions of the Delta and Omicron variants of SARS-CoV-2," Proc. Natl. Acad. Sci. USA, vol. 120, no. 22, e2221887120, 2023, doi: 10.1073/pnas.2221887120.

[72] C. J. Courtemanche, J. Garuccio, A. Le, J. Pinkston, and A. Yelowitz, "Strong social distancing measures in the United States reduced the COVID-19 growth rate," Health Aff., vol. 39, no. 7, pp. 1237–1246, 2020, doi: 10.1377/hlthaff.2020.00608.

[73] G. P. Guy Jr. et al., "Association of State-Issued Mask Mandates and Allowing On-Premises Restaurant Dining with County-Level COVID-19 Case and Death Growth Rates , United States, March 1–December 31, 2020," MMWR Morb. Mortal. Wkly. Rep., vol. 70, no. 10, pp. 350–354, 2021, doi: 10.15585/mmwr.mm7010e3.

[74] S. Hsiang et al., "The effect of large-scale anti-contagion policies on the COVID-19 pandemic," Nature, vol. 584, pp. 262–267, 2020, doi: 10.1038/s41586-020-2404-8.

[75] A. Lasry et al., "Timing of Community Mitigation and Changes in Reported COVID-19 and Community Mobility , Four U.S. Metropolitan Areas, February 26–April 1, 2020," MMWR Morb. Mortal. Wkly. Rep., vol. 69, no. 15, pp. 451–457, 2020, doi: 10.15585/mmwr.mm6915e2.

[76] H. Joo et al., "Decline in COVID-19 Hospitalization Growth Rates Associated with Statewide Mask Mandates , 10 States, March–October 2020," MMWR Morb. Mortal. Wkly. Rep., vol. 70, no. 6, pp. 212–216, 2021, doi: 10.15585/mmwr.mm7006e2.

[77] M. Arabi et al., “Severity of the Omicron SARS-CoV-2 variant compared with the previous lineages: A systematic review,” J. Cell. Mol. Med., vol. 27, no. 11, pp. 1443–1464, 2023, doi: 10.1111/jcmm.17747.

[78] A. Garza, C. Challu, and M. Mergenthaler-Canseco, “TimeGPT-1,” arXiv:2310.03589, 2023. [Online]. Available: https://arxiv.org/abs/2310.03589 (Accesso: 7 set. 2025).

[79] S. Reiter and S. W. R. Werner, “Interpolatory model reduction of dynamical systems with root mean squared error,” IFAC-PapersOnLine, vol. 59, no. 1, pp. 385–390, Mar. 2025, doi: 10.1016/j.ifacol.2025.03.066.

[80] M. Antonini, A. Fouda, M. Hinwood, A. Melia, and F. Paolucci, “The interplay between global health policy and vaccination strategies in the shift towards COVID-19 endemicity,” Health Policy Technol., vol. 13, no. 1, Art. 100854, 2024, doi: 10.1016/j.hlpt.2024.100854.

[81] Á. O’Toole, O. G. Pybus, M. E. Abram, E. J. Kelly, and A. Rambaut, “Pango lineage designation and assignment using SARS-CoV-2 spike gene nucleotide sequences,” BMC Genomics, vol. 23, no. 1, p. 121, 2022, doi: 10.1186/s12864-022-08358-2.

[82] CDC, “SARS-CoV-2 variant classifications and definitions,” 2024. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-classifications.html (Accesso: 7 set. 2025).

[83] B. Zhou et al., “TEMPO: A transformer-based mutation prediction framework for SARS-CoV-2 evolution,” Comput. Biol. Med., vol. 152, Art. 106264, 2023, doi: 10.1016/j.compbiomed.2022.106264.

[84] G. Wang et al., “Deep-learning-enabled protein–protein interaction analysis for prediction of SARS-CoV-2 infectivity and variant evolution,” Nat. Med., vol. 29, no. 8, pp. 2007–2018, 2023, doi: 10.1038/s41591-023-02483-5.

[85] M. C. Maher et al., “Predicting the mutational drivers of future SARS-CoV-2 variants of concern,” Sci. Transl. Med., vol. 14, no. 633, eabk3445, 2022, doi: 10.1126/scitranslmed.abk3445.

[86] H. Zhao et al., “VOC-alarm: mutation-based prediction of SARS-CoV-2 variants of concern,” Bioinformatics, vol. 38, no. 14, pp. 3549–3556, Jul. 2022, doi: 10.1093/bioinformatics/btac422.

[87] J. M. Taft et al., “Deep mutational learning predicts ACE2 binding and antibody escape to combinatorial mutations in the SARS-CoV-2 receptor-binding domain,” Cell, vol. 185, no. 21, pp. 4008–4022.e14, Oct. 2022, doi: 10.1016/j.cell.2022.08.024.

[88] S. Marini et al., “Optimizing viral genome subsampling by genetic diversity and temporal distribution (TARDiS) for phylogenetics,” Bioinformatics, vol. 38, no. 3, pp. 856–860, 2022, doi: 10.1093/bioinformatics/btab725.

[89] S. Marini, C. Boucher, N. Noyes, and M. Prosperi, “The K-mer antibiotic resistance gene variant analyzer (KARGVA),” Front. Microbiol., vol. 14, 1060891, 2023, doi: 10.3389/fmicb.2023.1060891.

[90] K. Berahmand, F. Daneshfar, E. S. Salehi, Y. Li, and Y. Xu, “Autoencoders and their applications in machine learning: A survey,” Artif. Intell. Rev., vol. 57, Art. 28, 2024, doi: 10.1007/s10462-023-10662-6.

[91] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: https://www.deeplearningbook.org (accessed Sep. 7, 2025).

[92] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011. [Online].
Available: https://jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

[93] M. Abadi et al., “TensorFlow: A system for large-scale machine learning,” in Proc. 12th USENIX Symp. Operating Systems Design and Implementation (OSDI’16), 2016, pp. 265–283. [Online]. Available: https://arxiv.org/abs/1605.08695.

[94] Y. Wang, H. Yao, and S. Zhao, “Auto-encoder based dimensionality reduction,” Neurocomputing, vol. 184, pp. 232–242, Apr. 2016, doi: 10.1016/j.neucom.2015.08.104.

[95] R. Corizzo, M. Baron, and N. Japkowicz, “CPDGA: Change point driven growing auto-encoder for lifelong anomaly detection,” Knowl.-Based Syst., vol. 247, p. 108756, Jul. 2022, doi: 10.1016/j.knosys.2022.108756.

[96] R. Trivedi and N. H. (R.) Varadarajan, “Substitution scoring matrices for proteins,An overview,” Protein Sci., vol. 29, no. 11, pp. 2150–2163, 2020, doi: 10.1002/pro.3945.

[97] M. P. S. Ali, “Spike2Vec: An Efficient and Scalable Embedding Approach for COVID-19 Spike Sequences,” in Proc. IEEE Int. Conf. Big Data, 2021, pp. 4749–4758, doi: 10.1109/BigData52589.2021.9671857.

[98] M. P. S. Ali et al., “Spike2Signal: Classifying Coronavirus Spike Sequences with Deep Learning,” in Proc. IEEE 8th Int. Conf. Big Data Computing Service and Applications (BigDataService), 2022, pp. 39–46, doi: 10.1109/BigDataService55547.2022.00013.

[99] Y. Liu et al., “Delta spike P681R mutation enhances SARS-CoV-2 fitness over Alpha variant,” Cell Rep., vol. 39, no. 7, p. 110829, 2022, doi: 10.1016/j.celrep.2022.110829.

[100] M. Bhattacharya et al., “Omicron variant (B.1.1.529) of SARS-CoV-2: understanding mutations in the genome, S-glycoprotein, and antibody-binding regions,” Geroscience, vol. 44, no. 2, pp. 619–637, 2022, doi: 10.1007/s11357-022-00533-2.

[101] P. Kumar et al., “The emerging challenge of FLiRT variants: KP.1.1 and KP.2 in the global pandemic landscape,” QJM, vol. 117, no. 7, pp. 485–487, 2024, doi: 10.1093/qjmed/hcae130.

[102] A. Vaswani et al., “Attention Is All You Need,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 5998–6008. [Online]. Available: https://arxiv.org/abs/1706.03762.

[103] T. Yu et al., “Large Language Models in Biomedical and Health Informatics: A Review with Bibliometric Analysis,” J. Healthc. Inform. Res., vol. 8, pp. 658–711, 2024, doi: 10.1007/s41666-024-00166-5.

[104] M. Shamsabadi, J. D., and S. A., “Large Language Models for Scientific Information Extraction: An Empirical Study for Virology,” in Findings of ACL: EACL 2024, St. Julian’s, Malta, 2024, pp. 2193–2208. [Online]. Available: https://aclanthology.org/2024.findings-eacl.165.

[105] D. Hesslow et al., “RITA: A Study on Scaling Up Generative Protein Sequence Models,” arXiv preprint arXiv:2205.05789, 2022.

[106] T. B. Brown et al., “Language Models are Few-Shot Learners,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2020, pp. 1877–1901. [Online]. Available: https://arxiv.org/abs/2005.14165.

[107] E. Nijkamp, J. A. Ruffolo, E. N. Weinstein, N. Naik, and A. Madani, “ProGen2: Exploring the Boundaries of Protein Language Models,” arXiv preprint arXiv:2206.13517, 2022. [Online]. Available: https://arxiv.org/abs/2206.13517.

[108] J. Su, Y. Lu, S. Pan, A. M. Bai, W. Zhang, and Y. Liu, “RoFormer: Enhanced Transformer with Rotary Position Embedding,” arXiv preprint arXiv:2104.09864, rev. 2023.

[109] Y. Shu and J. McCauley, “GISAID: Global initiative on sharing all influenza data,from vision to reality,” Euro Surveill., vol. 22, no. 13, p. 30494, 2017, doi: 10.2807/1560-7917.ES.2017.22.13.30494.

[110] F. A. Farris, “The Gini Index and Measures of Inequality,” Amer. Math. Monthly, vol. 117, no. 10, pp. 851–864, 2010, doi: 10.4169/000298910X523344.

[111] J. Hoffmann et al., “Training compute-optimal large language models,” arXiv preprint arXiv:2203.15556, 2022.

[112] R. M. Dhodapkar, “A deep generative model of the SARS-CoV-2 spike protein predicts future variants,” bioRxiv, 2023, doi: 10.1101/2023.01.27.525949.

[113] Y. Huang et al., “Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19,” Acta Pharmacol. Sin., vol. 41, no. 9, pp. 1141–1149, 2020, doi: 10.1038/s41401-020-0485-4.

[114] J. Zhang, T. Xiao, Y. Cai, and B. Chen, “Structure of SARS-CoV-2 spike protein,” Curr. Opin. Virol., vol. 50, pp. 173–182, 2021, doi: 10.1016/j.coviro.2021.08.010.

[115] J. Shang et al., “Structural basis of receptor recognition by SARS-CoV-2,” Nature, vol. 581, no. 7807, pp. 221–224, 2020, doi: 10.1038/s41586-020-2179-y.

[116] C. H. M. Rodrigues, D. E. V. Pires, and D. B. Ascher, “mmCSM-PPI: predicting the effects of multiple point mutations on protein–protein interactions,” Nucleic Acids Res., vol. 49, no. W1, pp. W417–W424, 2021, doi: 10.1093/nar/gkab273.

[117] C. Xu et al., “Conformational dynamics of SARS-CoV-2 trimeric spike glycoprotein in complex with receptor ACE2 revealed by cryo-EM,” Sci. Adv., vol. 7, no. 1, eabe5575, 2021, doi: 10.1126/sciadv.abe5575.

[118] E. M. Parzych et al., “DNA-delivered antibody cocktail exhibits improved pharmacokinetics and confers prophylactic protection against SARS-CoV-2,” Nat. Commun., vol. 13, no. 1, p. 5886, Oct. 2022, doi: 10.1038/s41467-022-33309-6.

[119] S. Yuan and H. Zheng, "Using PyMOL as a platform for computational drug design," WIREs Comput. Mol. Sci., vol. 7, e1298, 2017, doi: 10.1002/wcms.1298.

[120] B. E. Suzek, H. Huang, P. McGarvey, R. Mazumder, and C. H. Wu, "UniRef: comprehensive and non-redundant UniProt reference clusters," Bioinformatics, vol. 23, no. 10, pp. 1282–1288, 2007, doi: 10.1093/bioinformatics/btm098.

[121] A. L. Mitchell, A. Almeida, M. Beracochea, M. Boland, J. Burgin, G. Cochrane, et al., "MGnify: the microbiome analysis resource in 2020," Nucleic Acids Research, vol. 48, no. D1, pp. D570–D578, 2020, doi: 10.1093/nar/gkz1035.

[122] M. Steinegger and J. Söding, "Clustering huge protein sequence sets in linear time," Nature Communications, vol. 9, article 2542, 2018, doi: 10.1038/s41467-018-04964-5.

[123] E. Nijkamp, J. A. Ruffolo, E. N. Weinstein, N. Naik, and A. Madani, "ProGen2: Exploring the boundaries of protein language models," Cell Systems, vol. 14, no. 11, pp. 968–978.e3, 2023, doi: 10.1016/j.cels.2023.10.002.

[124] Q. Hou, P. F. P. F. Fan, X. Fu, R. M. Rao, and F. Qiu, "Using metagenomic data to boost protein structure prediction and discovery," Computational and Structural Biotechnology Journal, vol. 20, pp. 434–442, 2022, doi: 10.1016/j.csbj.2021.12.030.

[125] R. Kumar, S. Yadav, M. Pathak, S. Tripathi, T. Ghosh, A. Sharma, et al., "Understanding mutations in human SARS-CoV-2 spike glycoprotein: A systematic review & meta-analysis," Viruses, vol. 15, no. 4, p. 856, 2023, doi: 10.3390/v15040856.

[126] S. Ramesh, M. Govindaraj, P. R. Parikh, N. R. Kumar, L. M. Nadarajah, S. K. et al., "Emerging SARS-CoV-2 variants: A review of its mutations, its implications and vaccine efficacy," Vaccines (Basel), vol. 9, no. 10, p. 1195, 2021, doi: 10.3390/vaccines9101195.

[127] K. Guruprasad, "Mutations in human SARS-CoV-2 spike proteins, potential drug binding and epitope sites for COVID-19 therapeutics development," Current Research in Structural Biology, vol. 4, pp. 41–50, 2022, doi: 10.1016/j.crstbi.2022.01.002.

[128] A. Addetia, Y.-J. Park, T. N. Starr, A. J. Greaney, K. R. Sprouse, J. E. Bowen, et al., "Structural changes in the SARS-CoV-2 spike E406W mutant escaping a clinical monoclonal antibody cocktail," Cell Reports, vol. 42, no. 6, p. 112621, 2023, doi: 10.1016/j.celrep.2023.112621.

[129] S. Knyazev, K. Chhugani, V. Sarwal, R. Ayyala, H. Singh, S. Karthikeyan, et al., "Unlocking capacities of genomics for the COVID-19 response and future pandemics," Nature Methods, vol. 19, no. 4, pp. 374–380, 2022, doi: 10.1038/s41592-022-01444-z.

[130] N. Sitharam, H. Tegally, D. de Castro Silva, C. Baxter, T. de Oliveira, and J. S. Xavier, "SARS-CoV-2 Genomic Epidemiology Dashboards: A review of functionality and technological frameworks for the public health response," Genes, vol. 15, no. 7, p. 876, 2024, doi: 10.3390/genes15070876.

[131] C. Wymant; BEEHIVE Collaboration, "Easy and accurate reconstruction of whole HIV genomes from short-read sequence data with shiver," Virus Evolution, vol. 4, no. 1, 2018, article vey007, doi: 10.1093/ve/vey007.

[132] L. Zsichla, M. Vass, D. F. Danko, Á. Mézes, É. Dombi, M. Dolejsi, et al., "Comparative evaluation of open-source bioinformatics pipelines for full-length viral genome assembly," Viruses, vol. 16, no. 12, p. 1824, 2024, doi: 10.3390/v16121824.

[133] R. Rose, B. Constantinides, A. Tapinos, D. L. Robertson, and M. Prosperi, "Challenges in the analysis of viral metagenomes," Virus Evolution, vol. 2, no. 2, 2016, article vew022, doi: 10.1093/ve/vew022.

[134] National Center for Biotechnology Information (NCBI), "NIH," 2025. [Online]. Available: https://www.ncbi.nlm.nih.gov. Accessed: Sep. 7, 2025.

[135] R. D. Olson, T. Brettin, C. N. Curtin, C. C. Davis, J. D. Dempsey, D. A. Dickerman, et al., "Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC, IRD and ViPR," Nucleic Acids Research, vol. 51, no. D1, pp. D678–D689, 2023, doi: 10.1093/nar/gkac1003.

[136] M.-W. Tang and R. W. Shafer, "The HIVdb system for HIV-1 genotypic resistance interpretation," Intervirology, vol. 55, no. 2, pp. 98–101, 2012, doi: 10.1159/000331998.

[137] J. Hayer, J. P. V. Liefting, C. C. Hu, A. S. Lemonidis, F. S. Leenen, M. S. Avarre, et al., "HBVdb: a knowledge database for Hepatitis B Virus," Nucleic Acids Research, vol. 41, Database issue, pp. D566–D570, 2013, doi: 10.1093/nar/gks1022.

[138] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, et al., "Mistral 7B," arXiv, arXiv:2310.06825, 2023, doi: 10.48550/arXiv.2310.06825.

[139] Z. Zhou, Y. Ji, W. Li, P. Dutta, R. Davuluri, and H. Liu, "DNABERT-2: Efficient foundation model and benchmark for multi-species genome," arXiv, arXiv:2306.15006, 2023.

[140] D. Kassab, DNABERT, a linguistic approach for sequential predictions within Biology and Health – Bachelor's Thesis, University of Groningen, 2024. [Online]. Available: https://fse.studenttheses.ub.rug.nl/31828/1/bBIO_2024_KassabDB.pdf. Accessed: Sep. 7, 2025.

[141] K. R. Shahapure and C. Nicholas, "Cluster quality analysis using silhouette score," in Proc. 2020 IEEE 7th Int. Conf. on Data Science and Advanced Analytics (DSAA), Sydney, Australia, 2020, pp. 747–748, doi: 10.1109/DSAA49011.2020.00096.

[142] T. Caliński and J. Harabasz, “A dendrite method for cluster analysis,” Communications in Statistics,Theory and Methods, vol. 3, no. 1, pp. 1–27, 1974, doi: 10.1080/03610927408827101.

[143] D. L. Davies and D. W. Bouldin, “A cluster separation measure,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-1, no. 2, pp. 224–227, 1979, doi: 10.1109/TPAMI.1979.4766909.

[144] S. L. Stenton et al., “Critical assessment of variant prioritization methods for rare disease diagnosis within the Rare Genomes Project,” Human Genomics, vol. 18, no. 1, p. 44, 2024, doi: 10.1186/s40246-024-00604-w.