From Mechanism to Practice: Evolutionary Forecasting for SARS-CoV-2

Marlin Figgins

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2024

Reading Committee:
Ivana Bozic, Chair
Trevor Bedford, Advisor
Mark Kot

Program Authorized to Offer Degree:
Applied Mathematics

University of Washington

Abstract

From Mechanism to Practice: Evolutionary Forecasting for SARS-CoV-2

Marlin Figgins

Chair of the Supervisory Committee:
Associate Professor Ivana Bozic
Applied Mathematics

Novel genetic variants often arise due to mutations in circulating viral populations. These mutations can sometimes provide fitness advantages to members of the population allowing them to out-compete other variants through mechanisms such as partial immune escape and increased transmissibility. This interplay of mutation, transmission, and selection leads to evolution in the population. Therefore, understanding the genetic composition of viral populations and its relation to virus phenotype can be useful for understanding the current and future epidemic potential of viral variants.

This dissertation develops several theoretical ideas, statistical methods, and software tools that enable evolutionary forecasting for SARS-CoV-2 and other rapidly evolving pathogens using concepts from population genetics, mathematical epidemiology, and statistics.

We begin by developing a Bayesian method for estimating the effective reproduction number of genetic variants using counts of variant sequences and measures of incidence such as case counts.

To evaluate this method among others, we develop a workflow to compare frequency-based forecast models in a live forecasting environment, quantifying the short-term accuracy of such methods and suggesting a minimal sequencing capacity to ensure high quality forecasts.

Next, we develop a larger theory for how mechanistic models of transmission constrain how variant frequencies change over time. This leads to theoretical results for the trade-

off between immune escape and increased transmissibility and suggests new methods for modeling variant fitness using approximate Gaussian processes as well as latent pseudo-immune factors.

We then apply these ideas to incorporate molecular data on immune escape and phylogenetic structure into relative fitness estimates to enable out-of-sample prediction of relative fitness from sequence-level predictors.

Our focus then shifts to the operational problem of evolutionary forecasting, where we develop open-source software and visualization tools that can be used to implement, automate, and interpret evolutionary forecasts.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to begin by thanking my family, friends, lab members, and other collaborators. None of this work would be possible without your patience, kindness, compassion, and collaboration. Science cannot be done in isolation. This work is the sum of all of our efforts.

I would like to thank my PhD advisor Trevor Bedford. His guidance and belief in me and my work has allowed us to do science that I can be truly proud of. His dedication to ensuring that I maintained academic freedom and his encouragement to pursue science that is impactful and practical has completely changed my vision as a scientist, and I am eternally grateful for this.

I would also like to thank the rest of my committee members: Ivana Bozic, Mark Kot, Erick Matsen, and Betz Halloran. Your scientific input and expertise has allowed me to see so many approaches and philosophies of science. I could not be writing this without your support and inspiration.

To all the members of the Bedford lab, thank you for providing a scientific home. Though I may not use the words often, you are special to me and I'm honored to work alongside you all. Your collective brilliance inspires me daily.

DEDICATION

To our best efforts, to our failures, to our successes however marginal.

Chapter 1

INTRODUCTION

The emergence of SARS-CoV-2, the virus behind COVID-19, has drastically changed our world, affecting health, economies, and daily life. First identified in late 2019, this virus quickly spread globally, triggering a pandemic that has impacted millions.

The pandemic had been characterized by recurrent waves of infection, driven by genetic variants such as the Alpha, Beta, Delta, and Omicron variants. [79, 85, 84, 27]

Even following the end of the pandemic, there have been multiple infection waves caused by seasonality in transmission as well as evolution of novel SARS-CoV-2 variants that can evade existing immune responses in individuals with past exposure.

Immune escape driven by antigenic evolution has motivated annual updates to the COVID-19 vaccine to better protect individuals with weakened immune systems and those without past or recent exposure.

Evolution in SARS-CoV-2 has produced genetic variants that exhibit strain-specific immune responses in humans as measured through serology, neutralization titers, and deep mutational scanning assays. [9, 45, 24, 86]

Therefore, understanding both the virus's evolution and transmission dynamics are crucial for managing the pandemic and crafting effective public health strategies. As SARS-CoV-2 continues to spread, the virus will mutate and new variants with unique traits emerge, influencing their transmissibility, severity, and ability to evade immunity.

Novel genetic variants often arise due to mutations in currently circulating virus populations. These mutations can sometimes provide fitness advantages to members of the population allowing them to out-compete other variants through mechanisms such as partial immune escape, increased transmissibility, among others. This interplay of mutation, trans-

mission, and selection leads to evolution in the population. Therefore, understanding the genetic composition of viral populations and its relation to virus phenotype can be useful for understanding the current and future epidemic potential of viral variants.

Forecasting these variants is essential for predicting future infection waves, guiding public health measures, and informing vaccine development. However, this task is challenging due to the virus's rapid mutation rate and the dynamic nature of human immunity and behavior.

Accurate models are needed to predict how these variants will evolve and spread, to monitor SARS-CoV-2 evolution, and to mitigate future outbreaks.

Objectives and problem statement The primary objective of this dissertation is to develop and evaluate models that accurately forecast the evolution and spread of SARS-CoV-2 variants. This involves integrating concepts from epidemiology, evolutionary biology, and applied mathematics to create a comprehensive framework for understanding viral dynamics.

One of the fundamental challenges in forecasting SARS-CoV-2 variants is accounting for the virus's rapid mutation rate and its interaction with the human immune system. Mutations can lead to new variants with increased transmissibility or the ability to evade immunity. To address this, it is essential to incorporate mathematical models that can capture the complexities of viral evolution and spread.

In order to contextualize this work, I will begin with an overview of existing methods for understanding infectious disease transmission, relevant epidemiological quantities, and their relationships to one another.

I will then discuss existing methods for understanding population turnover from population genetics and their relationship to some of the related quantities in the epidemiological context.

I will then provide an overview of Bayesian inference, model development, and optimization methods that will be used throughout this dissertation to estimate quantities from data and make predictions while quantifying uncertainty.

Lastly, I will end this introduction by discussing the larger problem of evolutionary fore-

casting: its components, its aims, and the benefits of developing data-driven evolutionary forecasts on a global scale.

1.1 Epidemic models

I will give a basic introduction of the ordinary differential equation (ODE) formulation of epidemic models. We begin with the classic SIR model [47].

We will assume that there are three types of individuals in the population: susceptible (S), infected (I), and recovered (R). Susceptible individuals can become infected, infected individuals eventually recover (or die), and recovered individuals cannot become infected once again. This gives us the system of equations

$$\frac{dS}{dt} = -\beta SI, \tag{1.1}$$

$$\frac{dI}{dt} = \beta SI - \gamma I, \tag{1.2}$$

$$\frac{dR}{dt} = \gamma I, \tag{1.3}$$

where β is an effective infectious contact rate, γ is a recovery rate. This form of the model lends itself to a couple of analyses that will become important as we extend these models to more complicated infection dynamics.

Effective reproduction number An important quantity that is derived from epidemic models is the effective reproduction number R_t. We can see that when $R_t = \beta S(t)/\gamma > 1$, the number of infected individuals will be increasing, i.e., $\frac{dI}{dt} > 0$. This effective reproduction R_t controls the direction of the epidemic and reflects the average number of secondary infections caused by a single infection. In the special case where the population is fully susceptible ($S(t) = 1$), this quantity is called the basic reproduction number R_0, and it is interpreted as the average number of secondary infections caused by a single individual in an otherwise completely susceptible (naive) population.

1.2 Multistrain models

We'll now dive further into the evolutionary component of these models. In order for evolution to occur, we need a balance of mutation (generating heritable diversity) and selection (differential survival and reproduction based on that diversity).

For the goal of evolutionary forecasting of virus strains, we need to develop a notion of a multistrain model. These models will describes population with different types of infecting strains whose differences are heritable. This enables selection for different strains over time.

A two-strain epidemic model We first consider a standard ODE-based model for infectious diseases with multiple variants: an SIR model in which there are a wild-type wt and a variant var. We assume that the variant virus can differ in two ways from the wild type: it may differ in innate transmissibility by a factor of η_T, or it may differ in its infectious period by a factor η_G. This gives us the system of ordinary differential equations

$$\frac{dS}{dt} = -\beta S I_{\text{wt}} - \beta \eta_T S I_{\text{var}}, \tag{1.4}$$

$$\frac{dI_{\text{wt}}}{dt} = \beta S I_{\text{wt}} - \gamma I_{\text{wt}} = r_{\text{wt}}(t) I_{\text{wt}}, \tag{1.5}$$

$$\frac{dI_{\text{var}}}{dt} = \beta \eta_T S I_{\text{var}} - \gamma \eta_G I_{\text{var}} = r_{\text{var}}(t) I_{\text{var}}. \tag{1.6}$$

In this model, we now have two different R_0 values, $R_0^{\text{wt}} = \frac{\beta}{\gamma}$ and $R_0^{\text{var}} = \frac{\eta_T}{\eta_G} R_0^{\text{wt}}$. This shows that there is a dependence between transmissibility and the timing of infections. If $\eta_T > \eta_G$, the variant will spread faster than the wildtype, leading to increase in the relative proportion of the variant within the population.

Though this model is not appropriate for infectious diseases with non-trivial competition dynamics via strain-specific immunity for example, it gives us an essential result for epidemiological modeling of genetic variants. [36]

1.3 Population genetics, viral fitness, and selection

As viruses spread from individual to individual, they generate genetic variation through mutation, drift, and selection. In the context of ordinary differential equations, we've shown that multiple strains can co-circulate and may dominate one another over time. However, we still need to develop an understanding of how this co-circulation and transmission affects the underlying population's genetic diversity.

When dealing with pathogens with serial replacement, causing large waves of infection, we're primarily interested in selection. Selection acts as a "force" on the genetic diversity of the population favoring particular genetic variants and causing the population to evolve in response to particular environmental pressures. We'll now define selection and evolution.

- *Selection* is the process by which individuals produce more offspring in certain environments.

- *Evolution* is the change in the genetic composition of the population over time due to selection, genetic drift, and heritable variation.

Often, we're interested in not just the presence of selection but its magnitude. When quantifying selection in population genetics, the selection coefficient or relative fitness are often discussed. In the following section, we'll introduce both these quantities and their relationship to one another.

Suppose that we have alleles A and B in a population with x_A individuals having allele A and x_B having allele B. We assume that individuals with alleles A and B produce on average W_A and W_B offspring respectively. These W_A and W_B are the Wrightian fitness of these alleles. Assuming that the number of offspring has Poisson distribution with means W_A and W_B respectively, we can write the count of offspring with allele A in the next generation x'_A as

$$x'_A \sim \text{Multinomial}\left(N, \frac{W_A x_A}{W_A x_A + W_B x_B}\right), \tag{1.7}$$

where we've fixed the total number of offspring at N. Using this equation, we can write the expected frequency of allele A, $p_A = \frac{x_A}{x_A+x_B}$, and its variance as

$$\mathbb{E}[p'_A] = \frac{W_A p_A}{W_A p_A + W_B p_B} = \frac{W_A}{\bar{W}} p_A, \tag{1.8}$$

$$\mathrm{Var}[p'_A] = \frac{p'_A(1-p'_A)}{N}, \tag{1.9}$$

where we have defined the mean fitness $\bar{W} = W_A p_A + W_B p_B$. This provides a baseline for how fitness changes under fitness dynamics.

To measure the magnitude of selection for a particular allele, we can define the selection coefficient s so that the fitness of A is $W_A = (1+s)W_B$ per generation. Using our earlier derivation, we can write the expected frequency of the allele A assuming selection coefficient s after n generations of length τ days

$$p_A(n) = \frac{(1+s)^n p_A(0)}{(1+s)^n p_A(0) + (1-p_A(0))}. \tag{1.10}$$

There are alternative ways to quantify selection. The methods above used the absolute fitness W_A and W_B to derive selection coefficients s in order to quantify selection. Alternative models focus on exponentially growing populations. These approaches have been developed for and applied to modeling SARS-CoV-2 variant turnover. These models often estimate the relative fitness of variants based on changes in variant frequencies. *Relative fitness* is the relative capacity for individuals to reproduce in a population. In exponentially growing populations, we can think about this as the difference between the growth rates. Where absolute fitness tells us expected number of offspring, relative fitness tells us how variants differ in their capacity to spread.

One model in this vein that has gained popularity for analyzing SARS-CoV-2 variant frequencies and estimating relative fitness is multinomial logistic growth (MLR). Assuming that each variant has a fixed fitness relative to all others, we can derive a model for variant frequencies over time

$$f_v(t) = \frac{p_v \exp(r_v t)}{\sum_u p_u \exp(r_u t)}, \tag{1.11}$$

where f_v is the frequency of variant v in the virus population at time t. Here, r_v and p_v can be thought of as variant-specific growth rates per day and initial prevalence. Due to the fact that adding a constant to the growth rates does not affect the frequencies, r_v is not uniquely identifiable for all variants. Additionally, we have the constraint that the initial prevalence of each variant must add to one. Therefore, we often instead work with a model of the form

$$f_v(t) = \frac{\exp(\lambda_v t + \alpha_v)}{\sum_u \exp(\lambda_u t + \alpha_u)}, \tag{1.12}$$

where we've set some variant as a pivot with $\lambda_{\text{pivot}} = \alpha_{\text{pivot}} = 0$. By choosing to constrain $\lambda_{\text{pivot}} = 0$, we make the model the identifiable. We've also incorporated the logarithm of the initial prevalence $\log p_v$ into the exponential function as α_v, mirroring the form of a generalized linear model. Using these MLR models, the relative fitness of variant v relative to the pivot is often interpreted as being λ_v.

By using the mean generation time τ, we can approximate the relative effective reproduction number of variants as $R_t^v / R_t^{\text{pivot}} = \exp(\lambda_v \tau)$, which is equivalent to assuming that the generation time distribution is a point mass at τ. This approximation allows us to convert our estimated relative fitness λ_v from time in days to a per-generation growth advantage. We can relate this model of relative fitness to our earlier model of selection by noting that we can convert from number of generations n to number of days t using the generation time, $t = n\tau$. Substituting the above into our model of selection, we compare the selection coefficient s to the relative fitness λ,

$$(1+s)^{t/\tau} = \exp(\lambda_v t) \implies \lambda_v \tau = \log(1+s) \approx s. \tag{1.13}$$

This approximation holds for small $s \approx 0$ using a linear approximation of $\log(1+x)$. This provides a simple relationship between the traditional selection coefficient s used in population genetics and the relative fitness λ that we'll use throughout this dissertation.

1.4 Bayesian inference

For this general project of relating epidemic dynamics and population genetics, we need to take these models and connect them to what is observed or observable for statistical analysis. In our case, we rely primarily on Bayesian inference methods. We'll shortly review several features of Bayesian inference that are useful for this kind of modeling.

In short, Bayesian inference is focused on the properties of the joint probability distribution $p(\mathbf{x}, \boldsymbol{\theta})$ of data $\mathbf{x}$ and parameters $\boldsymbol{\theta}$ under a given model m.

Posterior distribution. Given a model and data, we can provide estimates of a possible distribution of parameters conditional on observed data using Bayes rule

$$p(\boldsymbol{\theta} \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})}{\int p(\mathbf{x} \mid \boldsymbol{\theta}')p(\boldsymbol{\theta}')d\boldsymbol{\theta}'}. \tag{1.14}$$

This distribution is called the *posterior distribution of θ under model m*. Notice, this distribution depends on both the likelihood of the data given the model parameters $\boldsymbol{\theta}$ $p(\mathbf{x} \mid \boldsymbol{\theta})$ and the prior probability of the parameters $p(\boldsymbol{\theta})$ before observing data. These priors are often set as a part of model building and can be useful for regularization, smoothing, and modeling data hierarchically.

Posterior samples allow us to estimate expectations. Most often, we approximate this distribution using samples from the posterior distribution using methods like Markov Chain Monte Carlo or Variational Inference.

These samples allow us to get expectations of function under the posterior distribution, so that we can compute

$$\mathbb{E}_{p(\boldsymbol{\theta}|\mathbf{x})}[f(\boldsymbol{\theta})] = \int_{\Theta} f(\boldsymbol{\theta})p(\boldsymbol{\theta} \mid \mathbf{x})d\boldsymbol{\theta}. \tag{1.15}$$

These expectations can be used to compute the probability of events under the posterior distribution including distributions for missing or future data.

Using these objects in Bayesian inference allows us to incorporate prior information about a system of interest, estimate parameters accounting for uncertainty in the model and observation, and build models that can be informed by various data sources simultaneously.

1.5 Genetic diversity and variant classification

The classification of viral populations into variants, as seen in systems such as WHO variant labeling, Nextstrain clades, and Pango lineages, plays a central role in making viral evolution interpretable and simplifying communication of infection risk or evolution. [40, 4, 68] These classifications exist not only to monitor viral diversity but also to guide public health decisions, ensuring that responses are grounded in structured simplifications of genetic data. In this context, variant assignment allows us to collapse the continuous landscape of viral genetic diversity into discrete categories, making it easier to perform population-level analyses and quantify key evolutionary forces, such as selection pressure.

This section will explore how variant assignment provides both a technical framework and a practical method for simplifying viral diversity. While mathematically related to clustering problems in statistics and machine learning, the utility of variant assignment lies in how it reduces complexity in order to inform viral forecasting, selection quantification, and public health interventions. We will demonstrate how this approach helps us predict viral evolution and guide real-world decisions.

Variant assignment for simplifying genetic diversity In viral populations, genetic diversity is vast and continuous, with sequences differing from one another by a spectrum of mutations. This continuous diversity creates challenges when analyzing population-level traits, such as fitness or transmissibility. Assigning sequences to variants simplifies this diversity by grouping sequences with similar biological properties, allowing for easier analysis without losing critical biological information. Variant assignment can be formalized using phenotypic metrics.

Let G represent a general metric that captures a relevant trait of sequence x_i such as

fitness, transmissibility, or immune escape potential. By grouping sequences based on these traits and their similarity, we reduce the dimensionality of the viral population, making the population easier to analyze. Importantly, the purpose of these classifications is to provide meaningful categorizations of genetic diversity that can be used for scientific analysis and public health decision-making.

Minimizing within-variant variance Variant assignment relies on minimizing the variability within each variant for the selected phenotypic metric $G(x)$. This process can be understood as minimizing the within-variant sum of squares (WCSS), which measures how close the sequences within each variant are to the variant's average trait value. For a set of variants $\{V_1, V_2, \ldots, V_K\}$, the WCSS is given by

$$\text{WCSS} = \sum_{k=1}^{K} \sum_{x \in V_k} (G(x) - \mu_{V_k})^2 \tag{1.16}$$

$$= \sum_{k=1}^{K} |V_k| \operatorname{Var}_{x \sim V_k}[G(x)], \tag{1.17}$$

where μ_V is the centroid of variant metric, representing the average value of the metric for sequences in that variant.

Minimizing WCSS ensures that the variance within each variant is as small as possible, meaning that sequences within a variant are phenotypically similar. This reduction of genetic diversity into variants can provide a simplified representation of viral populations, facilitating analysis while retaining the critical biological differences needed to understand viral dynamics. It is important to note that minimizing phenotypic variance within a cluster does not necessarily mean that sequences form a proper clade. In practice, we may want to produce evolutionarily meaningful clades that also minimize phenotype variance.

Simplifying Population-Level Expectations Using Variants One of the key advantages of variant assignment is that it simplifies population-level questions, such as estimating

the average fitness or transmissibility of the population. Directly analyzing every sequence in the population is computationally impractical, particularly as viral diversity grows. Assigning sequences to variants allows us to approximate population-level statistics using variant-level summaries.

For example, to estimate the expected value of a phenotypic trait $G(x)$ (such as fitness) across the population, the exact expectation is given by

$$\mathbb{E}_{X_t}[G] = \int_{\mathcal{X}} G(x) p_{X_t}(x) dx, \tag{1.18}$$

where $p_{X_t}(x)$ is the probability density of sequences at time t.

By assigning sequences to variants, we can approximate this expectation using variant frequencies $f_v(t)$ and an estimate of average trait value $\overline{G}_v$ within variant v. The average trait value approximates $\mathbb{E}_{X_t|V=v}[G]$ and can be estimated with a statistical model or taken from sample averages. This gives us an approximation

$$\mathbb{E}_{X_t}[G] \approx \sum_u f_u(t) \overline{G}_u, \tag{1.19}$$

for the expected phenotype across the population, which reduces a potentially complex integral into a weighted sum of variant means. This allows us to estimate expectations of variant phenotypes without sequence-level phenotype predictions or a generative sequence model $p_{X_t}(x)$. Instead, we operate at the level of variant.

Error Bounds and Limitations How well this approximation works depends on how similar the sequences within the variant are with respect to the trait G. We can see this by developing a bound

$$\epsilon = \left| \mathbb{E}_{X_t}[G] - \sum_u \overline{G}_u f_u(t) \right|^2 \tag{1.20}$$

$$= \left| \sum_{u=1}^{K} (\mathbb{E}_{X_t|V=u}[G] - \overline{G}_u) f_u(t) \right|^2, \tag{1.21}$$

on the error of this approximation.

We can introduce a per-variant mean squared error as

$$\epsilon_v = \mathbb{E}_{X_t|V=v}[(G(x) - \overline{G}_v)^2] \tag{1.22}$$

$$= \mathbb{E}_{X_t|V=v}[\big(G(x) - \mathbb{E}_{X_t|V=v}[G] + \mathbb{E}_{X_t|V=v}[G] - \overline{G}_v\big)^2] \tag{1.23}$$

$$= \mathrm{Var}_{X_t|V=v}[G] + (\mathbb{E}_{X_t|X=v}[G] - \overline{G}_v)^2, \tag{1.24}$$

which depends on the intrinsic variance of the phenotype within the variant and the squared bias in the estimated mean phenotype of the variant. Using that the probability of each variant is equal to its frequency $f_v(t)$, we can bound the total error with the Cauchy-Schwarz inequality

$$\epsilon = \left| \sum_u (\mathbb{E}_{X_t|V=u}[G] - \overline{G}_u) f_u(t) \right|^2 \tag{1.25}$$

$$\leq \underbrace{\left(\sum_u f_u(t) \right)}_{=1} \left(\sum_u \underbrace{(\mathbb{E}_{X_t|V=u}[G] - \overline{G}_u)^2}_{\epsilon_u} f_u(t) \right) \tag{1.26}$$

$$\leq \sum_u \epsilon_u f_u(t). \tag{1.27}$$

This shows our approximation exhibits a bias-variance trade-off. For unbiased estimates of the mean variant phenotype, this approximation will be most accurate when the within-variant variance is minimized and the variants are well-separated in the phenotypic space. If the variance of the phenotype within a variant is too large, the mean phenotype may no longer represent all sequences within the variant. Bounding the error of this approximation, by comparing the variant-level approximation to the exact population-level calculation, provides a deeper understanding of when the simplification works well and when it may fail. In practice, reassigning sequences or refining the metric might be necessary to maintain the effectiveness of the variant assignment.

Application to Viral Forecasting and Selection The broader significance of variant assignment lies in how it supports viral forecasting and the quantification of selection pres-

sures. By grouping sequences into variants, we can more easily track how traits such as fitness or transmissibility evolve over time and predict future changes in the viral population.

For example, in the context of viruses like SARS-CoV-2 or influenza, where selection plays a significant role in shaping the population, we may choose variants so that we minimize fitness variation within a variant. This allows us to quantify how selection acts on different variants by tracking changes in the variant frequencies over time. Variants with higher fitness values will increase in frequency, allowing us to forecast which variants are most likely to dominate future populations. Moreover, this framework can support public health interventions. By identifying the variants that are under the strongest selective pressure or are most likely to spread, we can make informed decisions about vaccine updates, mitigation strategies, and resource allocation.

Variant assignment, then, provides a powerful tool for simplifying the analysis of viral populations. By collapsing genetic diversity into variants, we can approximate population-level expectations using variant-level summaries, significantly reducing the complexity of the analysis. This approach also supports the broader goals of viral forecasting and public health, as it allows us to track how selection pressures shape the viral population over time at scale without requiring the entire sequence for each sample. In practice, phenotypes can be taken as molecular measurements such as neutralization titers against human sera, immune escape computed from deep mutational scanning assays, ACE-2 binding, numbers of mutations in particular regions of the genome, or multi-dimensional sequence-based embeddings. [45, 24, 39, 59] For pathogens like SARS-CoV-2 or influenza, where selection plays a critical role, using fitness as a metric for assessing the fidelity of variant assignment helps simplify evolutionary dynamics and predict future changes in the viral population while minimizing approximation error.

1.6 Evolutionary forecasting and its components

With these preliminaries, we now identify several steps towards our goal of mechanism-informed evolutionary forecasting of viruses.

Understanding how viral populations evolve over time is essential for public health interventions, vaccine development, and pandemic preparedness. Evolutionary forecasting provides the tools to predict how variant frequencies will change in response to selective pressures, mutation, and genetic drift.

I'll begin by describing the major targets of evolutionary forecasting mathematically using models of frequency change. This requires describing the population genetic forces that underlie frequency change and evolution. Though we've focused on selection thus far, it is important to note that selection coexists with genetic drift and mutation.

Genetic drift quantifies the randomness in frequency change due to demographic stochasticity impacting reproduction between generations. This is represented in our early model of allele frequency change by the variance of the frequency of allele A in the next generation p'_A. In equation 1.9, we can see the variance between generation decreases as the population size N increases. In general, the magnitude of the genetic drift depends on the effective population size N_e, which acts similarly to N, but accounts for factors like unequal contribution to the next generation, overlapping generations, and changes in the population size over time, which are extremely relevant for viral populations. This effective population size N_e is an important feature of population genetic models and is often used as a proxy for prevalence in coalescent-based phylodynamic analyses, though assumption on the generation time τ are necessary to isolate N_e. [57]

Simplifying to variant-level dynamics We begin by considering a population of viral sequences over time X_t. Through genomic surveillance, we can sample from the distribution of viruses to observe sequences $x_t^{(1)}, \ldots, x_t^{N_t} \sim X_t$.

We use stochastic differential equations (SDEs) motivated by our earlier results to model the evolution of this virus. If we assume that genotype x has frequency $f(x,t)$, we can model the change in frequency of that genotype as

$$df_x(t) = \underbrace{\left[\lambda_x(t) - \bar{\lambda}(t)\right] f_x(t)dt}_{\text{selection}} + \underbrace{\sigma(x,t)d\mathbf{W}_t}_{\text{drift}} + \underbrace{\mu\frac{\partial^2}{\partial x^2}f(x,t)dt}_{\text{mutation}}, \tag{1.28}$$

where $\lambda_x(t)$ is the fitness of genotype x at time t, $\bar{\lambda}(t) = \int_{\mathcal{X}} \lambda_z(t) \cdot f_z(t)dz$ is the mean fitness in the population at time t. We also have $\sigma(x,t) = \sqrt{\frac{f(x,t)(1-f(x,t))}{N_e}}$, which captures genetic drift, i.e., the random fluctuations in the frequency due to randomness in reproduction. Lastly, we have a diffusion term, $\mu\frac{\partial^2}{\partial x^2}f(x,t)dt$, which describes the diffusion of frequency across genotype space due to mutation with μ as the mutation rate.

By leveraging the variant assignment framework introduced earlier, we can simplify our models by tracking the dynamics among variants to predict future changes in viral populations. Our SDEs are reduced to a system of V equations representing each variant

$$df_v(t) = \left[\lambda_v(t) - \bar{\lambda}(t)\right] f_v(t)dt + \mathbf{\Sigma}d\mathbf{W}_t, \tag{1.29}$$

where I've now dropped the mutation term. This reduced system of equations depends on the variant level fitnesses over time, the mean fitness $\bar{\lambda}(t) = \sum_u \lambda_u(t) \cdot f_u(t)$, and a $V \times V$ covariance matrix Σ for genetic drift. The covariance matrix Σ for the genetic drift between variants has elements

$$\Sigma_{ij} = \begin{cases} \frac{f_i(1-f_i)}{N_e}, & \text{if } i = j, \\ -\frac{f_i f_j}{N_e}, & \text{if } i \neq j. \end{cases} \tag{1.30}$$

This captures the variance of each variant ($i = j$) and the negative covariance between variants ($i \neq j$) due to the finite population constraint.

This suggests that, instead of working with continuous genotypes, we may be able to model factors like selection and genetic drift by looking at variant-level data. Thinking about frequency change in the population in terms of variant frequency drastically simplifies analysis since we only need labels of variant status over time instead of full sequences, enabling analysis at a much larger scale.

The components of evolutionary forecasting Understanding and forecasting viral evolution requires integrating genetic diversity, fitness estimation, and population-level dynamics into a cohesive framework. We'll now outline the essential components of this process, highlighting how these elements interact to enable forecasts of variant dynamics.

The first step in evolutionary forecasting is summarizing genetic diversity in a way that captures key patterns while enabling fitness estimation. Genetic diversity reflects the interplay of selection, drift, and mutation in shaping frequency.

When quantifying genetic diversity in practice, we typically work with variant classifications and the frequencies of these variants as summaries of this genetic diversity. These variant frequencies are the foundation of these forecasts as they reflect the fitness of these variants, the stochastic effects of genetic drift, and diversification within a variant due to mutation.

Relative fitness estimation is central to evolutionary forecasting and a central task of this dissertation. Beyond merely estimating relative fitness, we seek to contextualize relative fitness with the mechanisms that drive it and use these ideas to predict the population-level impact of evolution.

Understanding the biological basis for fitness differences is essential for forecasting. Mediated by mutations in viral proteins that enable evasion of host immunity or enhance binding affinity, viral characteristics like immune escape or transmissibility underpin relative fitness advantages and can cause large waves of infection.

The relative fitness is also context-dependent, influenced by past infection and exposure, spatial heterogeneity, and waning immunity within a population. These forms of population structure introduce complex spatial and temporal dynamics into evolutionary forecasts. To account for this, we need to assess how population structure affects relative fitness in real-time and how this will carry into the future.

Evolutionary forecasting is inherently uncertain, complicated by mutation, genetic drift, and changing population structure. By combining these elements, we can potentially improve our estimates and forecasts of relative fitness and project changes in variant frequencies over

time, capturing both short-term and long-term dynamics.

As we continue to improve our methods for forecasting, what remains is to develop evolutionary forecasting as a practice. This requires ensuring that we have the data, methods, and software tools needed to produce forecasts in real-time, quantify their uncertainty, and communicate these results.

1.7 Overview of chapters and contributions

This dissertation develops fitness-based models by integrating epidemiological data, population genetics, and Bayesian methods for understanding, modeling, and forecasting viral evolution.

The chapters progress from mechanistic models that jointly estimate transmission and evolutionary dynamics to data-driven, mechanism-informed models incorporating external data for long-term forecasting. Each chapter contributes to building a comprehensive framework for understanding and predicting the spread of viral variants in pathogens such as SARS-CoV-2.

Chapter 2: Inferring variant-specific effective reproduction numbers from combined case and sequencing data In this chapter, we develop a model that jointly estimates variant-specific reproduction numbers (R_t) and relative fitness. The integration of case and sequencing data allows us to make statements about both transmission rates and evolutionary dynamics simultaneously. This mechanistic framework provides deeper insights into the joint behavior of evolution and transmission.

Key contributions:

- We integrate case data and sequencing data to model evolution and transmission jointly, offering real-time insights into how variants spread and evolve under different conditions.

- We develop a model that jointly estimates both variant-specific effective reproduction

numbers and relative fitness capturing the interaction between evolutionary dynamics and transmission rates.

- We set the stage for models in later chapters that will refine the understanding of transmission and immunity by linking these processes to fitness.

Chapter 3: Fitness models provide accurate short-term forecasts of SARS-CoV-2 variant frequency This chapter develops a framework and pipeline for evaluating fitness-based models of short-term evolutionary forecasts like those developed in Chapter 2. The emphasis is on assessing the performance of statistical models and the importance of data quality and sequencing capacity in improving forecast accuracy.

Key contributions:

- We develop a framework and pipeline for evaluating MLR models and other fitness-based models in the context of short-term forecasting, focusing on forecast accuracy and reliability.

- We show the limitations of statistical models for longer-term forecasts, emphasizing the need for more comprehensive approaches.

- We identify the role of data quality and sequencing capacity in enabling accurate forecasts, suggesting that a minimum of 1,000 sequences per week is necessary for reliable short-term predictions.

- We set the stage for models in later chapters that will integrate external data to improve forecast reliability by better capturing transmission mechanisms and immune escape.

This chapter provides a rigorous evaluation of fitness-based forecasting models and emphasizes the importance of data collection and integration in producing accurate short-term forecasts.

Chapter 4: Frequency dynamics predict viral fitness, antigenic relationships and epidemic growth Building on the insights from Chapters 2 and 3, this chapter develops a comprehensive framework that combines epidemiological theory and population genetics to estimate time-varying relative fitness. This chapter introduces several contributions, including a selective pressure metric, a latent factor model, and a focus on immune escape as a critical factor for explaining fitness dynamics and viral evolution.

Key contributions:

- We develop a Gaussian Process model for estimating time-varying relative fitness, to track how variants evolve in real-time regardless of underlying mechanism.

- We derive a selective pressure metric, which predicts epidemic growth rates without the need for case data, based solely on genetic data.

- We develop a latent factor pseudo-immune model, which constructs a pseudo-antigenic space, allowing for comparisons of immunity differences between populations and informing forecasts about antigenic evolution.

This chapter provides a deeper understanding of how transmission mechanisms and immune escape influence viral evolution, offering critical insights for improving the forecasting of variant success.

Chapter 5: Forecasting SARS-CoV-2 lineage success from molecular data Building on the constraints and limitations identified in Chapters 3 and 4, this chapter develops a novel framework that integrates molecular phenotypes with relative fitness innovations to forecast the success of SARS-CoV-2 variants. By leveraging deep mutational scanning data, these models provide a mechanism-informed approach to disentangle the effects of shared ancestry and quantify the contributions of immune escape and transmissibility-related traits to lineage success.

Key contributions:

- We develop a framework that integrates molecular phenotype data with fitness innovations to quantify and predict lineage success.

- We introduce a regression-based prior for fitness that enables out-of-sample forecasting of relative fitness for unseen variants, extending models introduced in earlier chapters.

- We demonstrate that immune escape is the dominant phenotypic driver of fitness, while transmissibility-related traits such as ACE2 binding affinity and RBD expression also contribute.

- We validate our framework's predictive accuracy, highlighting its potential for real-time monitoring and broader application to other rapidly evolving pathogens.

This chapter advances the field by providing a scalable, data-driven framework that combines molecular phenotypes with evolutionary modeling to improve forecasting of lineage success. Its applications extend beyond SARS-CoV-2, offering insights into antigenic evolution and supporting public health decision-making through proactive variant surveillance.

Chapter 6: Operationalizing evolutionary forecasts: `evofr` and `forecasts-ncov`
Using the theoretical and statistical grounding developed in previous chapters, this chapter develops and describes software that takes evolutionary forecasting from a series of one-off analyses to a practice. The software tools `evofr` and `forecasts-ncov` move evolutionary forecasts from a series of bespoke analyses on relative fitness and variant frequency to a tool kit for rapidly implementing these analyses and a reproducible workflow for real-time monitoring and forecasting SARS-CoV-2 evolution in practice.
Key contributions:

- We develop a Python package `evofr` that implements many of the tools needed to analyze and forecast variant frequency change and estimate relative fitness.

- We develop an automated and reproducible workflow `forecasts-ncov` for producing and visualizing forecasts of SARS-CoV-2 variant evolution.

This chapter operationalizes evolutionary forecasting, transforming the discipline from a series of scientific analyses to a dynamic process for predicting pathogen evolution using genomic surveillance data.

Chapter 7: Conclusions and Future Work This chapter synthesizes the contributions of the previous chapters and explores potential directions for future research. It highlights how the development of data-driven and mechanism-informed models sets the stage for further advances in viral forecasting, including future integration of external data sources and the application of generative sequence models to evolutionary forecasting.

I summarize how the work in this dissertation advances viral forecasting by developing models that integrate epidemiological data, population genetics, and Bayesian inference. Next, I propose future research directions, including the integration of external data sources to refine and evaluate forecasting models, the application of these models to other viral families, and the development of generative sequence models for evolutionary forecasting. This concludes the dissertation by offering a vision for the future of evolutionary forecasting of viruses.

Chapter 2

INFERRING VARIANT-SPECIFIC EFFECTIVE REPRODUCTION NUMBERS FROM COMBINED CASE AND SEQUENCING DATA

2.1 Abstract

Accurately estimating relative transmission rates of SARS-CoV-2 variants remains a scientific and public health priority. Recent studies have used the sample proportions of different variants from genetic sequence data to describe variant frequency dynamics and relative transmission rates, but frequencies alone cannot capture the rich epidemiological behavior of SARS-CoV-2. Here, we extend methods for inferring the effective reproduction number of an epidemic using confirmed case data to jointly estimate variant-specific effective reproduction numbers and frequencies of co-circulating variants using cases and sequences across states in the US from January 2021 to March 2022. Our method can be used to infer structured relationships between effective reproduction numbers across time series allowing us to estimate fixed variant-specific growth advantages. We use this model to estimate the effective reproduction number of SARS-CoV-2 Variants of Concern and Variants of Interest in the United States and estimate consistent growth advantages of particular variants across different locations.

2.2 Introduction

As SARS-CoV-2 evolves, variants may emerge that increase in their ability to transmit and escape acquired immunity [77]. Quantifying the observed growth advantages of SARS-CoV-2 variants allows us to better understand biological differences between circulating viruses [79, 25]. Relating genomic data of SARS-CoV-2 variants to epidemic surveillance data is

difficult. Although it is typical to use phylodynamic methods to analyze genetic sequence data from epidemics, the sheer amount of data as well as challenges to describing fitness effects in phylodynamic models make these methods hard to apply to potential differences in transmission rate among circulating variants. In order to deal with the limitations of phylodynamic inference, previous studies have estimated the growth of variants using observed frequencies in sequenced SARS-CoV-2 samples [5, 31, 60, 44]. Such methods often model the frequency of variants using multinomial logistic regression [5, 60], which generally assumes that genetic variants have a fitness advantage over one another that is fixed in time and acts as an estimate for the selective advantage of different variants at the level of frequencies. Although a consistent increase in frequency of one variant over another is expected to reflect differences in transmission rate, these models do not directly account for the complicated infection and transmission dynamics that influence which variants lead to local and regional epidemics. When dealing with competition between variants, variants that are declining in frequency can still lead to an increasing number of infections. Similarly, growth in frequency does not necessarily entail an increase in absolute infections.

To more fully capture epidemiological dynamics, there are methods that describe the growth in number of infections using confirmed case, hospitalization, or death data to estimate changes in the effective reproduction number R_t, the average number of infections a single infectious individual generates at a given point of time t. Although these methods are excellent for describing overall epidemic growth rates, they cannot capture the evolutionary dynamics and fitness changes between different variants since they generally assume the population dynamics are described by a singular R_t trajectory [22, 1], which internally is unrelated to the genetic and phenotypic composition of the population. This is of particular importance in the analysis of an epidemic in which a dominant variant may be declining overall, but a minor variant is rapidly increasing in frequency and absolute prevalence, creating the potential for a secondary wave of infections that may go unnoticed at first glance. To overcome this we require models that partition case counts into contributions from different variants to estimate variant-specific effective reproduction numbers.

Ongoing SARS-CoV-2 evolution serves as an important example of this phenomenon. After initial emergence in late 2020, over the course of 2021, Variant of Concern (VOC) and Variant of Interest (VOI) viruses spread throughout the world and replaced existing viral diversity. Multiple WHO designated [52] VOC and VOI viruses circulated in spring and early summer 2021, but this diversity was largely replaced by Delta variant viruses, which became globally dominant in late summer 2021. Subsequently, Delta variant viruses were rapidly eclipsed by Omicron variant viruses after Omicron's emergence in October 2021 [83]. Although it is now clear that Delta's spread was driven by greater transmissibility than other co-circulating variants and Omicron's rapid spread was primarily driven by escape from existing population immunity, rigorous estimates of the relative fitness of circulating variant viruses are of interest. Here, we develop a joint epidemiological and population genetic model of SARS-CoV-2 to assess the growth of different variants over time and infer differences in the effective reproduction numbers of SARS-CoV-2 variants as well as underlying frequency of variants under noisy sampling. We apply this model to sequence data and case count data from the United States between January 2021 and March 2022 to estimate differences in transmissibility between circulating VOC and VOI viruses.

2.3 Results

Model Overview We implement two models of variant-specific effective reproduction number based on a renewal equation framework of epidemic spread (see Methods), a fixed growth advantage model and a time-varying growth advantage model (growth advantage random walk — GARW). These models assume that new infections are determined by two essential parameters: the effective reproduction number, which determines the average number of secondary infections generated over the course of a primary infection, and the generation time, which determines length of infection and relative transmissibility over the course of the infection. In both models, variants generate infections independently of one another, but the sum of infections across variants is observed through surveillance data like case counts or hospitalizations. In order to disaggregate infections by variant, we rely on frequency

estimates that are informed by counts of sequenced samples using a Dirichlet-multinomial likelihood.

The transmission of each variant is modeled using a deterministic renewal equation that allows for realistic delay distributions between infection, transmission, and detection as a case. With this approach, we need only to determine the initial number of infections and the variant-specific effective reproduction numbers to estimate the frequency of each variant in the population over time. Due to this, the differences between the two models are determined in how each parameterizes variant-specific effective reproduction numbers.

Each variant in the fixed growth advantage model has its own multiplicative growth advantage that acts as a scaling to a single non-variant R_t trajectory (Fig. 2.1). With this fixed growth advantage model, we parameterize fitness of variants at the level of transmission by inferring variant-specific effective reproduction numbers. This differs from previous work on variant effective reproduction numbers, which often parameterize these differences by assuming logistic growth of frequencies [27, 84]. Though, in general, our method allows one to estimate variant growth in the frequency domain in terms of effective reproduction number differences, we find that assuming a fixed advantage for variants results in estimates that are qualitatively similar to the aforementioned models, which assume fixed growth advantages in frequency growth. This model provides the benefit of the inferred parameters being interpretable as scaling the effective reproduction number.

In cases where a singular fixed growth advantage is insufficient to describe the data, we extend our model to allow time-varying growth advantages (Fig. 2.2). In the GARW model, we introduce a variant R_t that infers the effective reproduction number of each variant as having a time-varying growth advantage relative to a base variant to allow for more complex relationships between the growth rates of different variants over time. Each variant effective reproduction number is parameterized using an exponentiated spline basis, so that the log effective reproduction numbers are described by a linear basis expansion. Therefore, we can use smoothing priors on the coefficients of these basis expansions to regularize the inferred time-varying growth advantages of each variant.

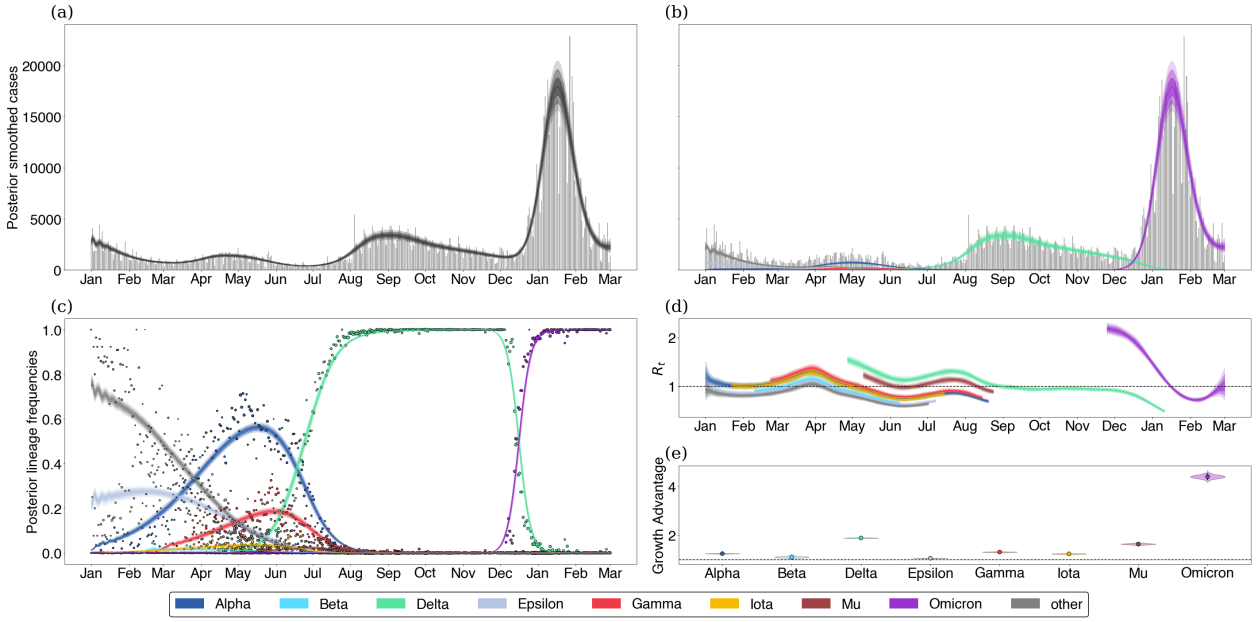

Figure 2.1: **Fitting the fixed growth advantage model to Washington state data.** (a) Posterior expected cases without weekly seasonality in reporting rate. Gray bars are observed daily case counts, while blue lines are model inferences with 50%, 80% and 95% credible intervals. (b) Posterior expected cases by variant. Each colored line is a different variant with intervals of varying opacity showing 50%, 80% and 95% credible intervals. (c) Posterior variant frequency against observed sample frequency. Dots represent observed weekly frequencies in sequence data and each colored line is a different variant with shaded CIs. (d) Variant-specific effective reproduction numbers. (e) Posterior growth advantage by variant.

We demonstrate these models on data from Washington State with results from the fixed growth advantage model shown in Figure 2.1 and results from the GARW model is shown in Figure 2.2. Further example model output for California, Florida, Michigan and New York is provided in the supplemental appendix in Figures A.5–A.12.

Estimating growth advantages in the United States We estimate the effective reproduction numbers of SARS-CoV-2 Variant of Concern and Variant of Interest viruses in the United States using daily confirmed case counts obtained from the US CDC and sequence counts annotated by variant obtained from the Nextstrain-curated 'open' dataset [40] (see Data and code accessibility). Each sequence is labeled with a Nextstrain clade [40], and we partition clades into variants based on WHO VOC/VOI designation [52]. Nextstrain clades annotated in the fashion correspond to a subset of lineages designated by Pango [68]. We consider the following 8 variants that have been flagged as variants of interest or concern and which circulated in the US during 2021 and early 2022: Alpha (Pango lineage B.1.1.7, Nextstrain clade 20I), Beta (lineage B.1.351, clade 20H), Gamma (lineage P.1, clade 20J), Delta (lineage B.1.617.2, clade 21A), Epsilon (lineage B.1.427/429, clade 21C), Iota (lineage B.1.526, clade 21F), Mu (lineage B.1.621, clade 21H) and Omicron (lineage B.1.1.529, clade 21M). We use a cutoff of 2000 sequences from a particular variant across states to determine threshold of circulation. This eliminates Eta, Theta, Kappa and Lambda from consideration and groups these variants along with ancestral 'non-variant' viruses into a single 'other' category. We use a cutoff of 12,000 sequences from a particular state as basis for including the state in the dataset. This cutoff left 34 states available for inference.

In order to inform our estimates of the frequency of genetic variants, we divide sequences from each state into daily sample counts for each of the 8 variants above and a single 'other' category. We then use these counts alongside the daily case counts in each state to estimate the effective reproduction number for individual variants using the GARW R_t model. We find that overall there appears to be consistent trends in the effective reproduction numbers of variants across the United States (Fig. 2.3). We see that early VOCs Alpha and Gamma initially had $R_t > 1$, but saw R_t decline below one across most states in April and May respectively. Upon arrival in May, Delta shows significantly higher values of R_t that don't decline below 1 until September. Initial Omicron R_t in November and December is significantly greater than earlier variants, but declines below 1 in late January and early February after driving large epidemics across states.

In order to transform these observed trends to a variant-specific growth advantage, we rely on our fixed growth advantage model, which infers a fixed variant-specific growth advantage as a multiplicative scaling of the effective reproduction number. Using the fixed growth advantage model, we find that most variants identified share some positive growth advantage except for Epsilon (Fig. 2.4). Further, these growth advantages appear to be consistent between the states analyzed. These results from the fixed growth advantage model are consistent with a multinomial logistic growth analysis (Fig. A.1). Alpha, Beta, Gamma and Iota show modest growth advantage over largely ancestral 'other' viruses, while Mu and Delta show larger growth advantages. Mu has previously been associated with increased neutralization resistance to convalescent serum [81], and its advantage of 1.2–1.8 across states is perhaps partially driven by immune escape. Despite this, Mu's growth advantage, whether from immune escape or otherwise, was insufficient to outcompete Delta in any of the states analyzed. Delta's advantage of 1.6–2.0 across states is particularly significant. Given that this large growth advantage was evident in May (Fig. 2.3), Delta's rapid rise in frequency and sizable epidemic should have been clear at the time. The significant growth advantage observed in Delta is recapitulated in other studies including Obermeyer et al. [60] and Vöhringer et al. [84]. In the case of Omicron, we see significant variability in the growth advantage, which spans 2.0–4.4. This large variability could be motivated by multiple factors including state-level variation in population immunity.

To better address the potential for change in variant growth advantage over time, we use our GARW model on the same data set to assess how variants increased or decreased in their growth advantages over time. We see growth advantages that are overall consistent with our fixed growth advantages, but are clearly able to discern time periods of variable growth advantage (Fig. 2.5). We observe oscillations in Delta's growth advantage from June to January. For Omicron, we also observe a large variability in the time-varying growth advantage, though there appears to be an upward trend after December.

2.4 Discussion

We find that a model that partitions case count data based on variant frequency in sequence data works well to describe SARS-CoV-2 variant dynamics in the United States from January 2021 to March 2022. In each state, spring waves in 2021 were primarily driven by the arrival of Alpha, Beta, Gamma, and Iota variants. However, as these waves subsided, the arrival of Delta with a significantly greater growth advantage, drove a large wave in summer 2021. Omicron's arrival in November 2021 drove a much larger wave in January/February 2022 due to significant immune escape of the variant. Importantly, we can directly estimate a variant-specific R_t, which for example, shows that Delta was a growing rapidly sub-epidemic across states in May, before its impact was noticeable in overall case counts, and that Omicron's initial R_t was estimated to be between 2 and 3 in December 2021, presaging a substantial Omicron driven wave.

We imagine that this approach could provide early warning of imminent epidemics driven by low-frequency but highly transmissible variants and generally serve to identify newly arising variants that show significant transmission advantages and that may drive epidemics. Indeed, we have continually updated estimates of spread of Omicron and Omicron sublineages BA.2, BA.2.12.1, BA.4 and BA.5 using this method and shared results in real-time online at github.com/blab/rt-from-frequency-dynamics. As an example, we estimate the growth advantages of Omicron sublineages BA.2, BA.2.12.1, BA.4, and BA.5 during their rise in the United States in Figure 2.6. These real-time estimates have served as a basis for reporting to public health, policy makers and the general public.

With this mind, this work is not without limitations. The underlying transmission model is deterministic and does not account for demographic stochasticity and over-dispersion in transmission, which has been documented in SARS-CoV-2 transmission [88]. As with all methods that depend on parameterizations of the generation time, misspecification of the generation time can lead to biased estimates of the effective reproduction number or growth advantages [38]. In order to quantify this source of error, we derive an equation relating

our inferred growth advantages, the epidemic growth rates, and the mean and standard deviation of the generation time distribution. This source of error can be partially combated by converting effective reproduction numbers to their corresponding epidemic growth rates under the generation time assumption (see Supplement Appendix). There is also a general need to account for biases in the case data that may not faithfully describe the infection dynamics of SARS-CoV-2 due to changes in case ascertainment rate, as possibly caused by differences in testing intensity and infection severity, among other reasons. However, we suspect that case ascertainment remained largely consistent from January to ∼Dec 2021, even if it declined with the advent of widespread circulation of Omicron.

We do not explicitly model additional introductions of variants outside a fixed seeding period that can play an important role in variants establishing themselves in different geographies at low infection counts and could bias our estimates of the effective reproduction number if not properly accounted for [22, 57]. However, we expect that once local transmission is predominant that estimated R_t will reflect characteristics intrinsic to the variant in the local geography. Using hierarchical models of variants to jointly estimate growth advantages and pool estimates across locations could be a useful approach for analyzing consistency between growth advantages of variants geographically and beginning to combat the issue of multiple introduction events. That said, fully combating this issue would likely involve incorporating demographic stochastic into the model at the level of transmission and likely reduce the speed of inference, scalability, and limit available inference options.

Although there are several ways to improve these methods and expand their applicability, our current model does have utility as a way of assessing early claims of variant advantages and is able to show there is evidence of consistent variant advantages shared between different geographies. Additional work is needed to attribute these inferred advantages to biological mechanisms like immune escape and transmissibility [77]. Modeling the effect of changes in other factors such as contact patterns or non-pharmaceutical interventions can be done with the current formulation of the model by including quantities of interest as features in the R_t model as in Sharma et al. [72].

In general, the development of methods that can account for fitness differences between genetic variants is much needed in order for proper epidemic preparedness. Our method provides one way of analyzing the growth rates of SARS-CoV-2 variants without directly parameterizing how variants grow in terms of frequency by instead focusing on differences in the effective reproduction number. In cases where the assumption of a fixed growth advantage is warranted and justified, our fixed growth advantage model provides a way of quantifying variant growth advantages at the level of transmission that allow for various delays between infection, transmission, and sampling. When a fixed growth advantage is unjustified, our GARW model can be used infer trends in variant growth advantages over time. Currently, our GARW model can be used to assess claims of growth advantages of variants and their sublineages.

This method can be extended further to analyze the role of specific constituent mutations defining a variant or lineage in changing the effective reproduction number of specific variants directly, similar to the model formulation of Obermeyer et al. [60]. With this in mind, our method potentially has use for evolutionary forecasting of variants for SARS-CoV-2 as we inform the frequency dynamics of co-circulating variants by describing their population-level transmission dynamics [2]. Extending the model further towards this aim will likely require methods for quantifying various sources of population immunity as well as escape potential for circulating and emerging SARS-CoV-2 variants as a way to explain these growth advantages and their underlying mechanisms using data. With these issues in mind, surveillance of variants should be folded into standard epidemiological surveillance as knowledge of variant-specific growth advantages will be useful for forecasting growth of cases, hospitalization, deaths, vaccine effectiveness among other key metrics related to epidemic response.

Further, as case surveillance for COVID-19 has decreased in reliability after 2022, we note that this method is still applicable using other proxies for infection incidence such as hospitalization data or wastewater testing. However, even in the absence of these forms of data, our approach highlights the distinction between relative fitness of viral variants and their overall transmission rates allowing us to attribute changes in incidence to selection and

variant turnover.

2.5 Methods

Using sampled counts of sequences from different variants as well as case data, we can jointly infer the proportion of variants in the larger population and the effective reproduction number of these variants.

Modeling the infection process We estimate the effective reproduction number of competing variants using a deterministic renewal equation based framework. These equations arise as the expectation of a Bellman-Harris branching process [10], which is a type of branching process in which offspring generation depends on the age of infection.

The renewal equation framework allows one to model infection processes in a way that is mathematically equivalent to standard epidemic models like the SEIR compartment model [20], but in a way that can be more suitable for estimating the effective reproduction number and forecasting using arbitrary generation times. This renewal equation can be written as

$$I(t) = R_t \int_0^t I(\tau) g_{t-\tau} d\tau, \tag{2.1}$$

where g is the generation time. In addition, we also include onset distribution o for symptoms, which allows us to compute the prevalence, or the number of active infections, as

$$P(t) = \int_0^t I(\tau) o_{t-\tau} d\tau. \tag{2.2}$$

We bin the generation time g and the onset distribution o to the nearest day, so that we estimate the daily incidence $I(t)$ and prevalence $P(t)$ as

$$I(t) = R_t \sum_{\tau<t} I(\tau) g_{t-\tau}, \tag{2.3}$$

$$P(t) = \sum_{\tau<t} I(\tau) o_{t-\tau}. \tag{2.4}$$

We parameterize all variants excluding Delta and Omicron as having generation time g as having Gamma distribution with mean 5.2 and standard deviation 1.2, in line with the estimates of [35]. Due to observed shorter serial intervals for Delta and Omicron, we instead use a mean generation time of 3.6 for Delta and a mean of 3.2 for Omicron [6, 71, 74]. For all variants, we parameterize the onset time o as having LogNormal with mean 6.8 and standard deviation 2.0 in line with [21]. We note that the choice of generation time can have strong effects on the inferred effective reproduction number and growth advantage under our renewal equation model. The effect of generation time choice is quantifiable as shown in Figures A.2, A.4 and Supplemental Appendix (see Relating epidemic growth rates to relative effective reproduction numbers). Though converting the posterior effective reproduction numbers to epidemic growth rates may be more robust to changes in generation time as can be seen in Figure A.3.

This method of using delays to represent lags between infection and observation can be extended to use multiple delays to better fit other data sources such as hospitalization or deaths.

Modeling variant frequencies In the case of V variants co-circulating in a population, we denote incidence of variant v at time t as $I_v(t)$ and prevalence as $P_v(t)$. In this case, we can compute the frequency of variant v in the population at time t under the infection process outlined above as

$$f_v(t) = \frac{P_v(t)}{\sum_{1 \leq v \leq V} P_v(t)}. \tag{2.5}$$

Since we've defined the frequency in terms of the transmission dynamics, the variant-specific effective reproduction numbers $R_{t,v}$ and initial infections $I_v(0)$ determine the frequency dynamics directly. Therefore, we do not need to impose a parametric form on $f_v(t)$ directly as in other models of variant frequency.

Observation process for cases As most case time series in the United States and elsewhere exhibit day-of-the-week biases, we estimate a reporting rate, which varies by day of

the week, so that $\rho = (\rho_1, \ldots, \rho_7)$ as in [1]. We then define the observation likelihood using a negative binomial distribution as follows

$$Y_t \sim \text{NegBinom}(\rho_{[t]} P(t), \alpha), \tag{2.6}$$

where $[t] = t \mod 7 + 1$, α is an over-dispersion parameter relative to the Poisson distribution, and $\text{NegBinom}(\mu, \alpha)$ is the negative binomial distribution with mean μ and variance $\mu + \alpha\mu^2$. In the case of multiple variants, we use $P(t) = \sum_{1 \leq v \leq V} P_v(t)$. The negative binomial likelihood is often used for modeling observation noise for count data such as epidemic time series which are often over-dispersed relative to a Poisson distribution. In order to account for zero-counts due to a lack of observations, we also include zero-inflation on the case counts.

Observation process for variant annotations If we suppose that we are tracking the growth of V variants, our data for a given day t takes the form of daily counts $C_t = (C_{t,1}, \ldots, C_{t,V})$ of sequences of each variant with daily total $N_t = \sum_{1 \leq v \leq V} C_{t,v}$. We then assume that the likelihood of observing these counts of each variant is described by a Dirichlet-multinomial distribution, so that

$$C_t \sim \text{DirMultinomial}\left(N_t, f(t) \cdot \left(\frac{1-\xi}{\xi}\right)\right), \tag{2.7}$$

given variant frequencies $f(t) = (f_1(t), \ldots, f_V(t))$ and over-dispersion parameter $0 < \xi < 1$. Here, we use a Dirichlet-multinomial distribution to account for possible over-dispersion in the counts relative to the standard multinomial distribution.

Basis expansions of log effective reproduction numbers Instead of inferring R_t directly, we parameterize the logarithm of the effective reproduction number using a basis of cubic splines. Each basis spline is written as a column in the design matrix $\mathbf{X}$, so that

$$\ln R_t = \mathbf{X}\boldsymbol{\beta}, \tag{2.8}$$

where the $\boldsymbol{\beta}$ are to be estimated to parameterize the effective reproduction number. We then use locally adaptive smoothing of order one with a Laplace prior on the coefficients $\boldsymbol{\beta}$ to promote smoothness on the inferred R_t trajectory [32]. This method also allows one to use other predictors such as vaccination proportion, intervention indicators, temperature, humidity, etc.

Modeling variant-specific effective reproduction numbers To model the variant-specific reproduction numbers, we can infer independent effective reproduction number trajectories for each variant

$$\ln R_{t,v} = \mathbf{X}\boldsymbol{\beta}_v, \tag{2.9}$$

where each variant v gets its own vector of parameters $\boldsymbol{\beta}_v$ in this model. We use the same prior structure as above to promote smoothness on inferred trajectories.

Modeling variant-specific growth advantages In order to use our model to infer growth advantages for specific variants, we can instead parameterize the effective reproduction numbers as

$$\ln R_{t,v} = \mathbf{X}\boldsymbol{\beta} + \delta_v, \tag{2.10}$$

where the parameters β are shared between all variants and δ_v is the log-scale variant-specific growth advantage of variant v. We consider $\Delta_v = \exp(\delta_v)$ to be the variant-specific growth advantage, which can be seen in Figure 2.4. This model is referred to as the "fixed growth advantage model" throughout the paper.

Estimating time varying growth advantages In reality, the growth advantage of a variant may vary in time due to factors like cross-immunity between variants, overall immune escape, etc. This can additionally occur under variant generation time misspecification [62].

To combat these issues, we extend our model to allow for time-varying growth advantages. We consider a growth advantage random walk model (GARW) in which the time-varying variant growth advantage $\delta_{t,v}$ relative to a chosen "base" variant is modeled as a spline whose

coefficients $\boldsymbol{\beta}_v$ have a Laplace random walk prior,

$$\ln R_{t,\text{base}} = \mathbf{X}\boldsymbol{\beta}_{\text{base}}, \tag{2.11}$$

$$\delta_{t,v} = \mathbf{X}\boldsymbol{\beta}_v, \tag{2.12}$$

$$\ln R_{t,v} = \ln R_{t,\text{base}} + \delta_{t,v}. \tag{2.13}$$

This model is referred to as the GARW model throughout the paper and can be seen in Figure 2.2

Estimating an average effective reproduction number for an epidemic Given variant-specific effective reproduction numbers $R_{t,v}$ and the frequency of variants in the population $f_v(t)$, we define the average effective reproduction number to be

$$R_t^{\text{ave}} = \sum_{1 \le v \le V} R_{t,v} f_v(t), \tag{2.14}$$

which is the sum of the variant-specific effective reproduction numbers weighted by their frequency. This quantity can be seen in Figure 2.2.

Priors for Bayesian Inference For both models, we provide a Laplace random walk prior on the spline coefficients $\boldsymbol{\beta}$ with scale parameter γ, which itself has a HalfNormal$(0, 0.1)$ prior distribution. In the fixed growth advantage model, only a baseline R_t trajectory is parameterized by $\boldsymbol{\beta}$ and the variant advantages δ_v are given a Normal$(0, 1)$ prior. For the GARW model, the variant growth advantage spline coefficients are modeled with a Laplace random walk with scale parameter γ_δ which has HalfNormal$(0, 0.01)$ prior distribution. The initial infected individuals for each variant have a uniform prior between 0 and 300,000. The weekly reporting rates $\rho_{[t]}$ each follow a Beta$(5, 5)$ prior, and the case observation over-dispersion is given a HalfNormal$(0, 0.05)$ prior on $\alpha^{\frac{1}{2}}$. Finally, the over-dispersion parameter ξ is given a Beta$(1, 99)$ prior to penalize high levels of over-dispersion in sequencing.

Inference The model is implemented in NumPyro [64] in Python and approximate Bayesian inference was conducted using Stochastic Variational Inference [41] using the ADAM opti-

mizer [49] with a learning rate of 0.01. For the analyses presented, all models are fit using a full-rank Gaussian variational distribution / Multivariate Normal autoguide as implemented in NumPyro [64], which approximates the posterior (with appropriate constraints on the individual parameter spaces) as a multivariate normal distribution.

Models for each individual state in the United States variants data set were fit for 60,000 iterations, and 3000 posterior samples were produced under both the fixed growth advantage model and the GARW model.

Data and code accessibility

Case counts and sequence data was obtained March 26, 2022. Case count data was obtained from the US CDC using the 'United States COVID-19 Cases and Deaths by State over Time' dataset available from data.cdc.gov. Sequence data including date and location of collection as well as clade annotation was obtained via the Nextstrain-curated 'open' dataset [40] that pulls from sequences shared to NCBI GenBank. Sequence metadata is available from data.nextstrain.org. Clades in this dataset are assigned via Nextclade annotation [4]. Here, we subsetted to sequences with specimens collected from the USA between January 1, 2021 and March 1, 2022. We additionally filtered to sequences with known collection date, assigned Nextstrain clade and dropped samples that were flagged as 'bad' by Nextclade QC. This subsetting resulted in 1,906,759 sequences for analysis. However, we reduced dataset to just the 34 states with more 12,000 sequences available in this time frame. Doing so reduced the full dataset to 1,541,099 sequences for analysis.

Derived data of sequence counts and case counts, along with all source code used to analyze this data and produce figures is available via the GitHub repository github.com/blab/rt-from-frequency-dynamics.

Competing interests

The authors declare no conflicting interests.

Author contributions

MF, TB conceived the study. TB gathered sequence and case count data. MF designed and implemented inference model. MF performed the analysis. MF, TB interpreted the results. MF, TB wrote the paper.

Acknowledgements

We thank John Huddleston, Eslam Abousamra and other members of the Bedford Lab for helpful feedback. MF is an ARCS Foundation scholar and was supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1762114. TB is an Investigator of the Howard Hughes Medical Institute. This project was supported by funds from the HHMI COVID-19 Collaboration Initiative awarded to the Fred Hutchinson Cancer Research Center and the University of Washington. We thank data generators in the United States for generously sharing SARS-CoV-2 sequence data to open databases. Without open data sharing, this work would not be possible. We also thank a previous anonymous reviewer for multiple suggestions.

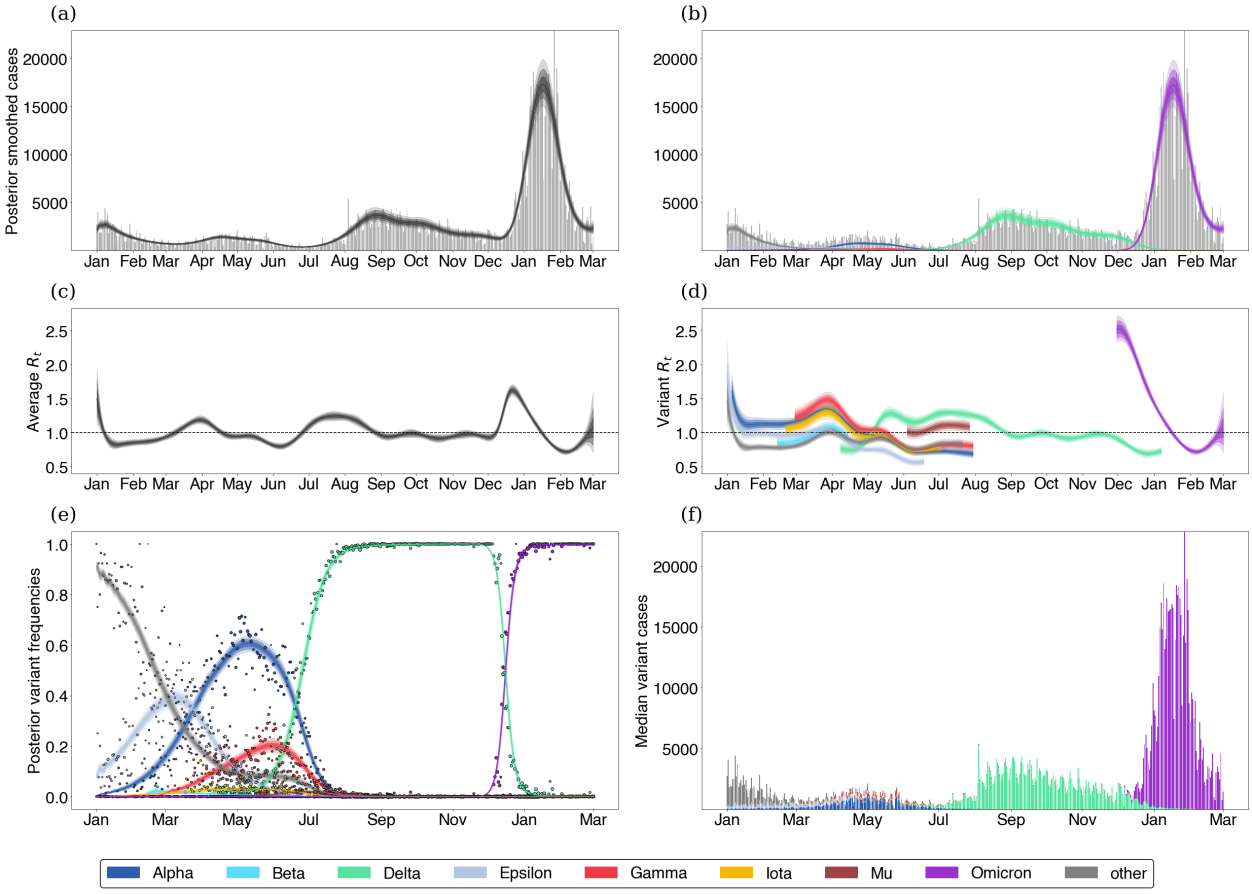

Figure 2.2: **Fitting the GARW model to Washington state data.** (a) When assessing epidemic growth rates, we often compute a single effective reproduction number trajectory which is effectively an average over the all viruses in population. We show the posterior smoothed incidence over time as well as the average effective reproduction number. Gray bars are observed daily case counts, while black intervals are the posterior 50%, 80% and 95% credible intervals. (b-d) Epidemics are made of different variants which may differ in fitness. We show the posterior variant-specific smoothed incidence (b) as well as the average and variant-specific effective reproduction numbers (c-d). (e-f) Using case counts alongside sequences of different variants allows us to understand the proportion of different variants in the infected population.

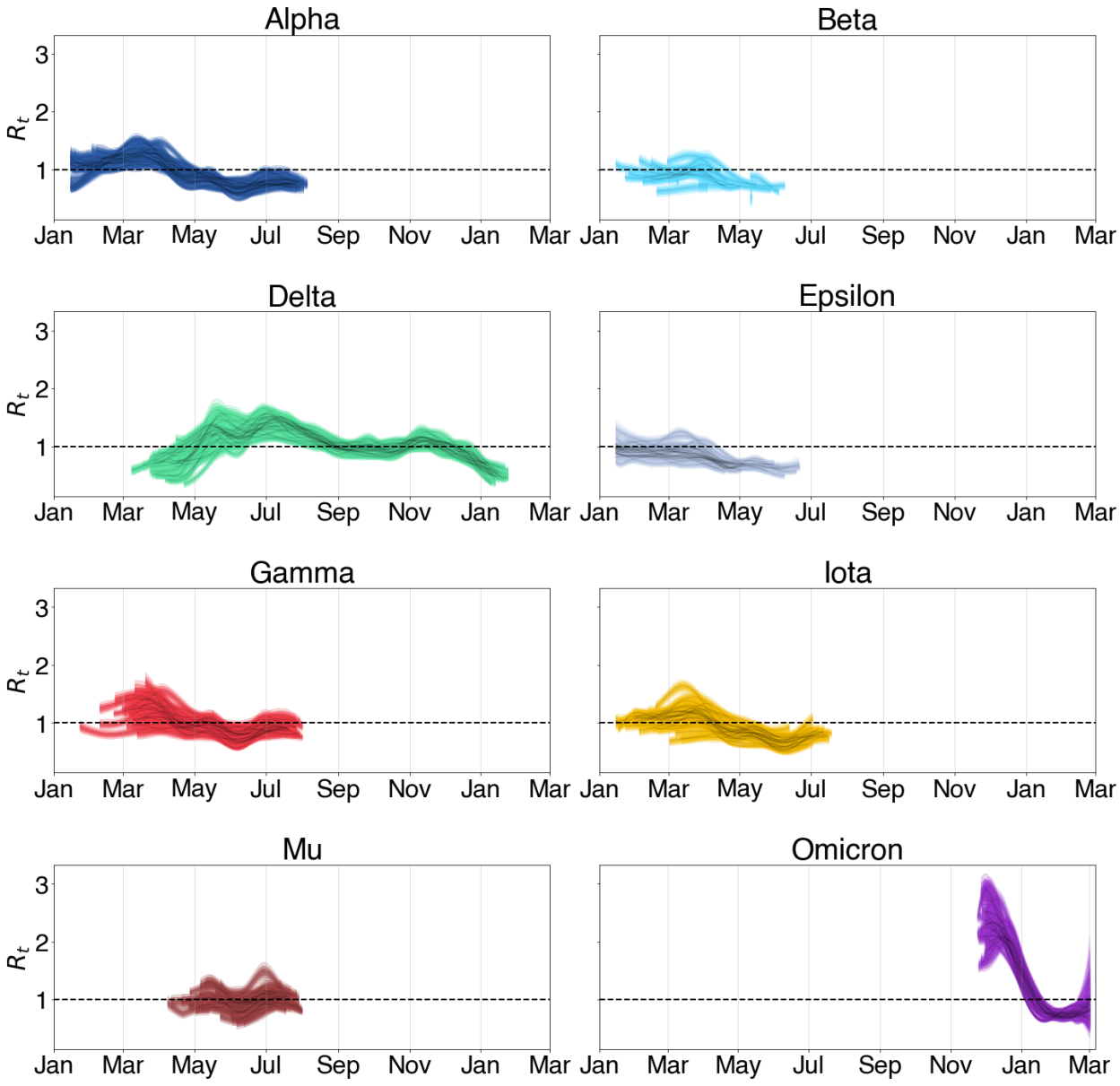

Figure 2.3: **Inferred effective reproduction numbers from GARW model in 34 states show consistent trends of variants across states.** Each panel shows a series of 34 trajectories, representing R_t through time for this variant across states. Shaded intervals show 50%, 80% and 95% credible intervals.

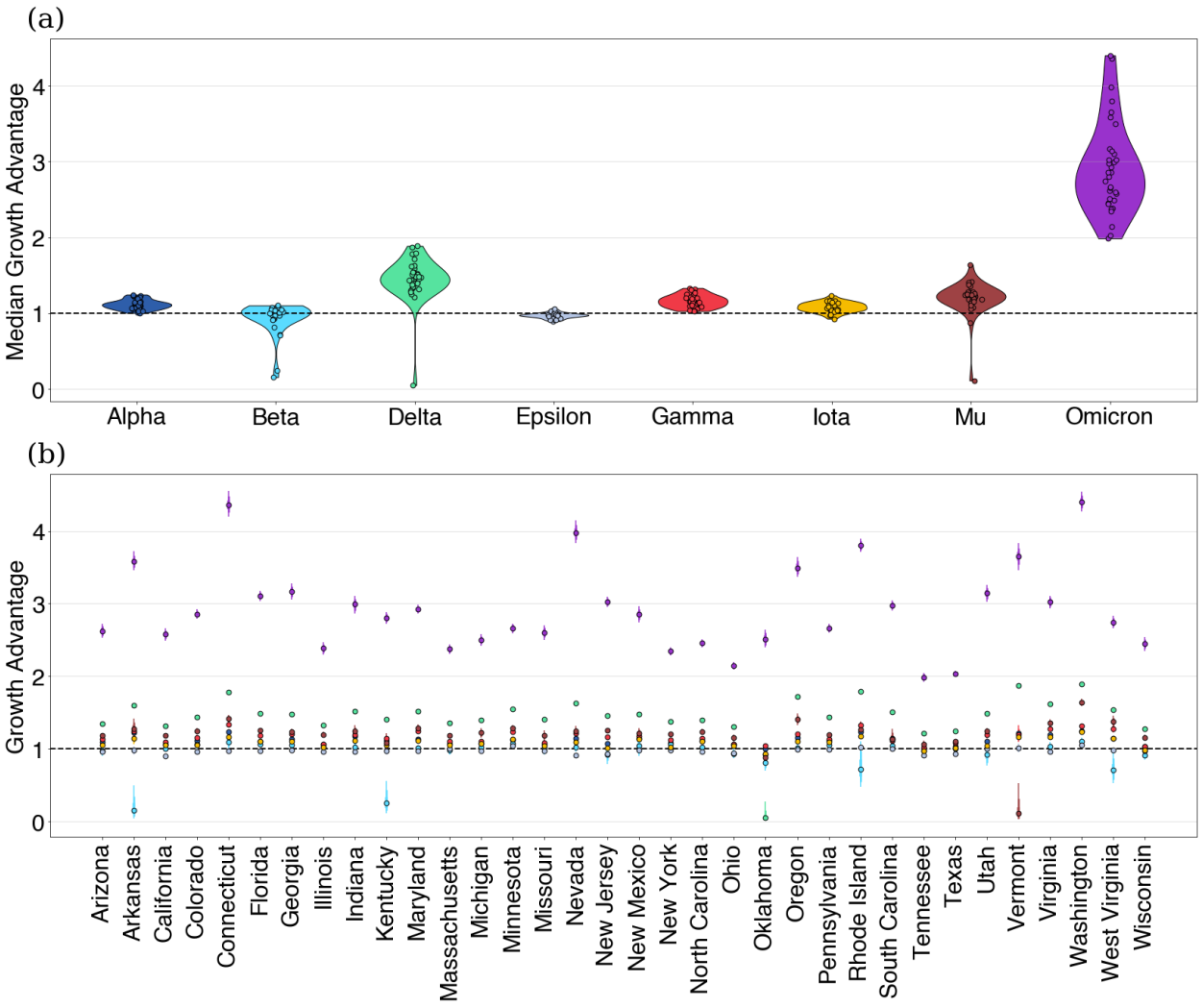

Figure 2.4: **Using fixed growth advantage model, we infer growth advantages for 8 variants in 34 US states.** (a) Growth advantages for variants of concern. Each point is the median growth advantage inferred from a single state. (b) Same as (a) but with state medians visualized by variant.

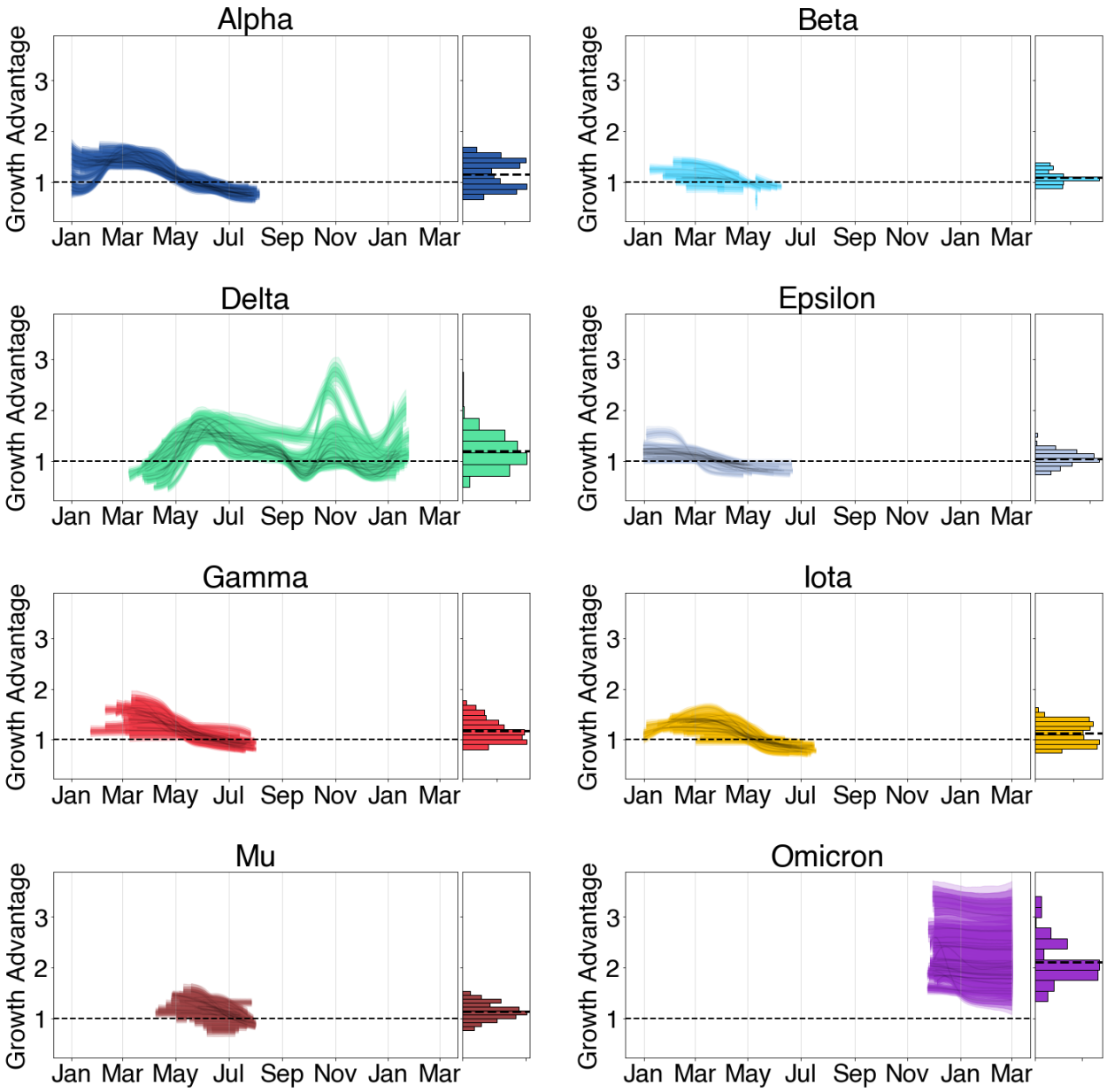

Figure 2.5: **Estimating variant growth advantages in 34 states using GARW model.** Each panel shows a series of 34 trajectories, representing Δ_v through time for variants across states. Histograms show the distribution of the variant's growth advantage over time. Shaded intervals show 50%, 80% and 95% credible intervals.

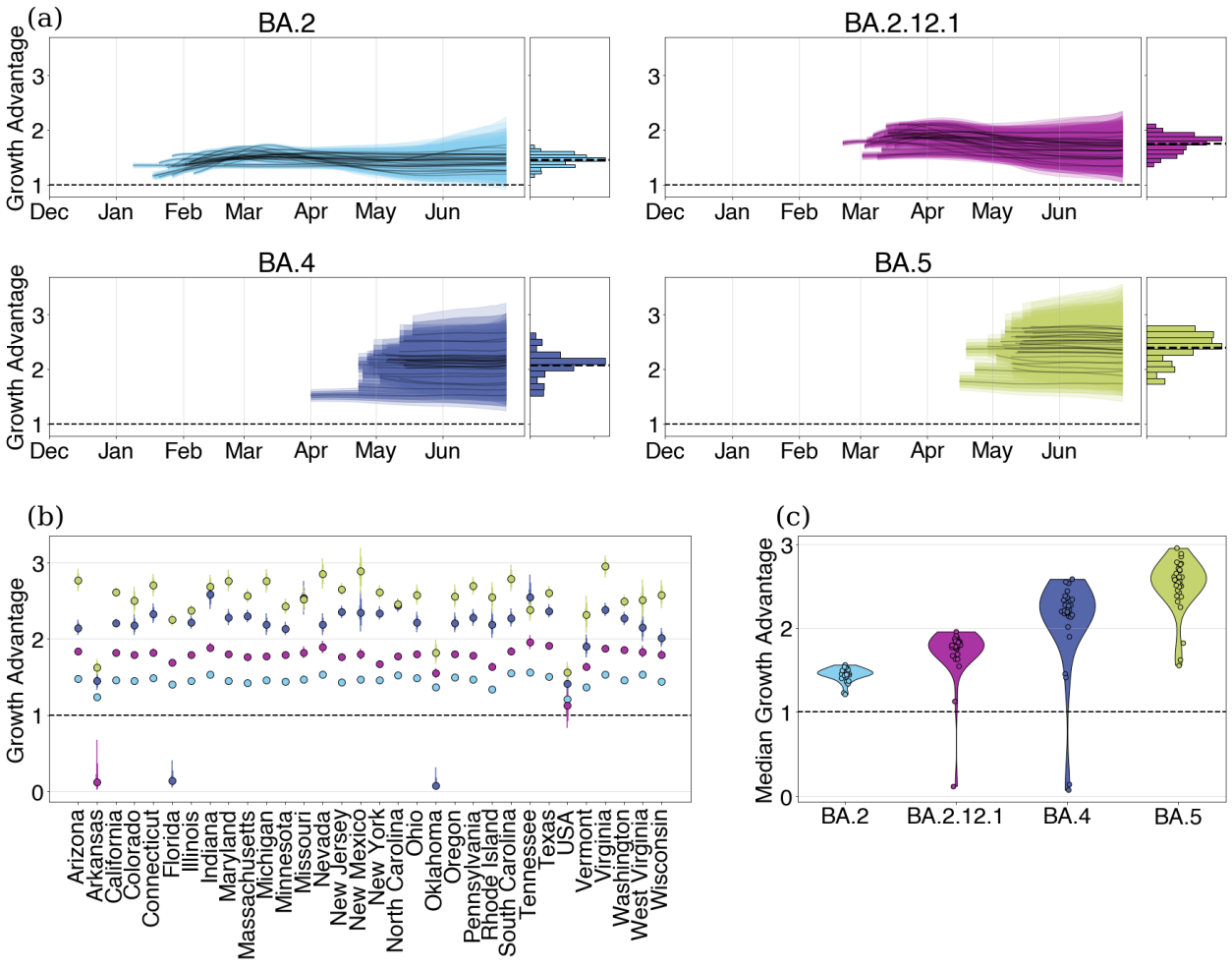

Figure 2.6: **Estimating growth advantages of Omicron sublineages relative to BA.1 in 33 US states.** (a) Time-varying growth advantages for BA.2, BA.2.12.1, BA.4, and BA.5 relative to BA.1 using the GARW model. Histograms denote the distribution of the variant growth advantages across all times. (b) Fixed growth advantages for Delta and BA.2 relative to BA.1 using fixed growth advantage model. (c) Same as (b) but with state medians visualized by variant.

Chapter 3

FITNESS MODELS PROVIDE ACCURATE SHORT-TERM FORECASTS OF SARS-COV-2 VARIANT FREQUENCY

Author summary

Over the course of the COVID-19 pandemic, SARS-CoV-2 evolved into many different genetic variants such as the well known Alpha, Beta, Gamma and Delta variants in early 2021 and the Omicron variant in late 2021. These genetic variants could more easily spread from person to person and so outcompeted previous versions of the virus. Even if they aren't being given Greek letter names, new variants are still arising with recent waves of COVID-19 caused by variants such as XBB and JN.1. Predicting which variants will increase in frequency and which variants will decrease in frequency is important for public health, particularly in terms of updating the formulation of the annual COVID-19 vaccine. In this paper, we investigate statistical models that use observed frequencies of different variants in the past weeks to estimate the frequency of different variants today and to forecast the frequency of different variants in 30 days time. We find that in countries with sufficient amounts and timeliness of genetic sequence data, these models forecast well and can be a useful tool for public health.

3.1 Abstract

Genomic surveillance of pathogen evolution is essential for public health response, treatment strategies, and vaccine development. In the context of SARS-COV-2, multiple models have been developed including Multinomial Logistic Regression (MLR), describing variant frequency growth, as well as Fixed Growth Advantage (FGA) and Growth Advantage Random Walk (GARW) parameterizations, describing variant R_t. These models provide estimates of variant fitness and can be used to forecast changes in variant frequency. We introduce a

framework for evaluating real-time forecasts of variant frequencies, and apply this framework to the evolution of SARS-CoV-2 during 2022 in which multiple new viral variants emerged and rapidly spread through the population. We compare models across representative countries with different intensities of genomic surveillance. Retrospective assessment of model accuracy highlights that most models of variant frequency perform well and are able to produce reasonable forecasts. We find that the simple MLR model provides ∼0.6% median absolute error and ∼6% mean absolute error when forecasting 30 days out for countries with robust genomic surveillance. We investigate impacts of sequence quantity and quality across countries on forecast accuracy and conduct systematic downsampling to identify that 1000 sequences per week is fully sufficient for accurate short-term forecasts. We conclude that fitness models represent a useful prognostic tool for short-term evolutionary forecasting.

3.2 Introduction

The emergence of acute respiratory virus SARS-CoV-2 causing COVID-19 disease and its subsequent circulating variants severely impacted global health and worldwide economies [61]. Due to its rapid evolution, original SARS-CoV-2 strains were replaced by derived, selectively advantageous variant lineages during 2021 [14], with Omicron, a highly transmissible and immune evasive variant becoming the dominant strain in early 2022 [83]. It has become increasingly evident that monitoring the evolution and dissemination of these variants remains crucial with SARS-CoV-2 continuing to evolve beyond Omicron [18]. Forecasting variant dynamics allows us to make informed decisions about vaccines and to predict variant-driven epidemics.

Fitness models are a key framework for forecasting changes in variant frequency through time. These models were first introduced for the study of seasonal influenza virus [55, 56, 42] and have relied on correlates of viral fitness such as mutations to epitope sites on influenza's surface proteins. In modeling emergence and spread of SARS-CoV-2 variant viruses, the use of Multinomial Logistic Regression (MLR) has become commonplace [5, 31, 60, 75]. Here, MLR is analogous to a population genetics model of a haploid population in which different

variants have a fixed growth advantage and are undergoing Malthusian growth. As such, it presents a natural model for describing evolution and spread of SARS-CoV-2 variants. Additionally, models introduced by Figgins and Bedford [34] incorporate case counts and variant-specific R_t, but still can be used to project variant frequencies while Piantham et al [66] does not incorporate them.

Here, we systematically assess the predictive accuracy of fitness models for nowcasts and short-term forecasts of SARS-CoV-2 variant frequencies. We focus on variant dynamics during 2022 in which multiple sub-lineages of Omicron including BA.2, BA.5 and BQ.1 spread rapidly throughout the world. We compare across several countries including Australia, Brazil, Japan, South Africa, Trinidad and Tobago, the United Kingdom, the United States, and Vietnam to assess genomic surveillance systems with different levels of throughput and timeliness. To assess the performance of these models, we used mean and median absolute error (AE) as a metric to compare the predicted frequencies to retrospective truth. This metric allowed us to evaluate the accuracy and reliability of the models and to identify those that were most effective in predicting SARS-CoV-2 variant frequency. We also examined aspects of country-level genomic surveillance that contribute to errors in these models and explored the role of sequence availability on nowcast and forecast errors through downsampling sequencing efforts.

3.3 Results

Reconstructing real-time forecasts

We focus on SARS-CoV-2 sequence data shared to the GISAID EpiCoV database [73]. Each sequence is annotated with both a collection date, as well as a submission date. We seek to reconstruct data sets that were actually available on particular 'analysis dates', and so we use submission date to filter to sequences that were available at a specific analysis date. We additionally filter to sequences with collection dates up to 90 days before the analysis date. We categorize each sequence by Nextstrain clade (21K, 21L, etc. . .) as such clades are

generally at a reasonable level of granularity for understanding adaptive dynamics [11]; there are 7 clades circulating during 2022 vs hundreds of Pango lineages. Resulting data sets for representative countries Japan and the USA for analysis dates of Apr 1 2022, Jun 1 2022, Sep 1 2022 and Dec 1 2022 are shown in Fig. 3.1A, while Fig. B.1 shows data sets for Australia, Brazil, South Africa, Trinidad and Tobago, the UK, and Vietnam. We see consequential backfill in which genome sequences are not immediately available and instead available after a delay due to the necessary bottlenecks of sample acquisition, testing, sequencing, assembly and data deposition. Thus, even estimating variant frequencies on the analysis date as a nowcast requires extrapolating from past week's data. Different countries with different genomic surveillance systems have different levels of throughput as well as different amounts of delay between sample collection and sequence submission [13].

We employ a sliding window approach in which we conduct an analysis twice each month (on the 1st and the 15th) and estimate variant frequencies from -90 days to $+30$ days relative to each analysis date. We illustrate our frequency predictions using the MLR model showing resulting trajectories for Japan and the US in Fig. 3.1B and showing trajectories for Australia, Brazil, South Africa, Trinidad and Tobago, the UK, and Vietnam in Figs. B.2–B.7. Sometimes we see initial over-shoot or under-shoot of variant growth and decline, but there is general consistency across trajectories. Additionally, we retrospectively reconstructed the simple 7-day smoothed frequency across variants and present these trajectories as solid black lines. We treat this retrospective trajectory as 'truth' and thus deviations from model projections and retrospective truth can be assessed to determine nowcast and short-term forecast accuracy. Consistent with less available data, we observe that the model predictions for Japan were more frequently misestimated compared to the United States with particularly large differences for clades 22B (lineage BA.5) and 22E (lineage BQ.1) (Fig. 3.1B).

Model error comparison

We utilize five models for predicting the frequencies of SARS-CoV-2 variants. The simplest of these models is Multinomial Logistic Regression (MLR) commonly used in SARS-CoV-2

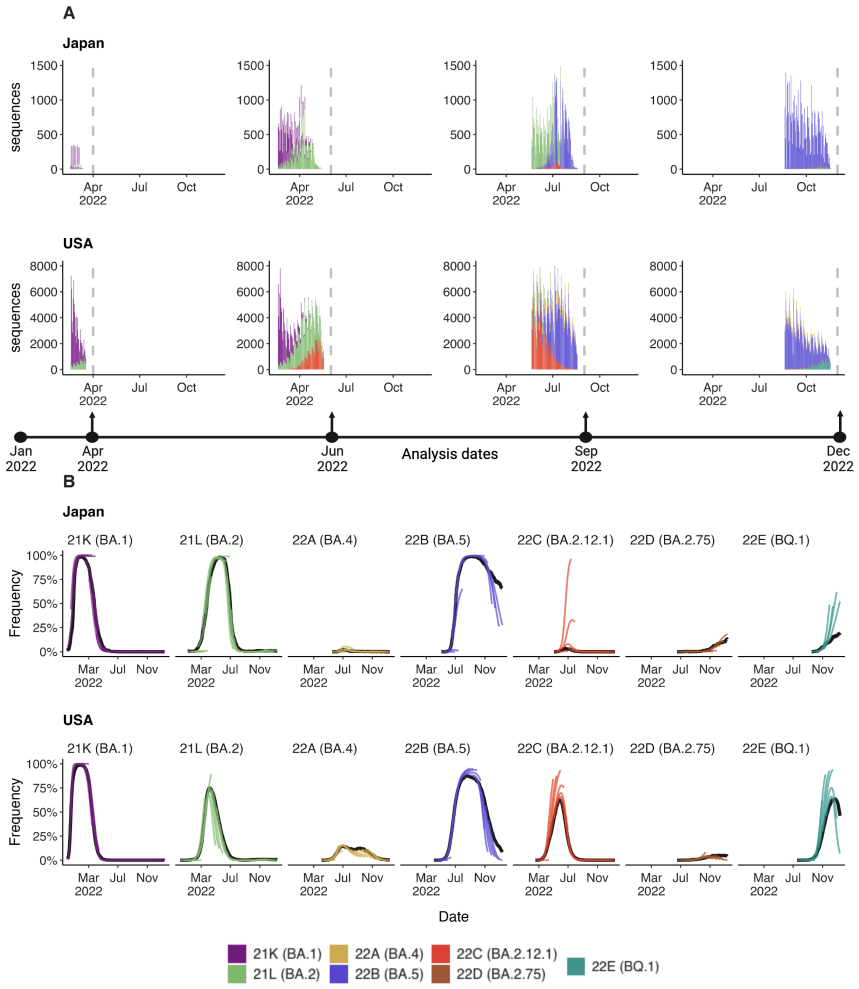

Figure 3.1: **Reconstructing available data sets and corresponding predictions for Japan and USA.** (A) Variant sequence counts categorized by Nextstrain clade from Japan and United States at 4 different analysis dates. (B) +30 day frequency forecasts for variants in bimonthly intervals using the MLR model. Each forecast trajectory is shown as a different colored line. Retrospective smoothed frequency is shown as a thick black line.

analyses [5, 31, 60, 75], which uses only variant-specific sequence counts and has a fixed growth advantage for each variant. More complex models include the Fixed Growth Advantage (FGA) and Growth Advantage Random Walk (GARW) parameterizations of the

variant R_t model introduced by Figgins and Bedford [34], which uses case counts in addition to variant-specific sequence counts. The Piantham et al. model [66] operates on a similar principle in estimating relative fitness, but differs in model details and does not use case counts. We compare these four models to a naive model to serve as a reference for comparison. The naive model is implemented as a 7-day moving average on the retrospective raw frequencies using the most recent seven days for which sequencing data is available. We compare forecasting accuracy across different time lags from −30 days back from date of analysis as hindcast, to +0 days from date of analysis as nowcast, and +30 days forward from date of analysis as forecast.

We refer to the absolute error $\mathrm{AE}_t^{m,d}$ for a given model m, data set d and time t as the difference between the retrospective 7-day smoothed frequency and the model predicted frequency (see Methods). We calculate median absolute error and mean absolute error across datasets and across time lags to assess the relative performance of the models for the eight countries (Fig. 3.2 and Table 3.1). As expected, we observe decreasing performance across models as lags increase from −30 days to +30 days. For example, median absolute error increases for the MLR model from 0.1–1.4% at −30 days, to 0.3–2.0% at 0 days and to 0.5–1.9% at +30 days. Similarly, mean absolute error increases for the MLR model from 0.4–4.2% at −30 days, to 2.2–8.6% at 0 days and to 5.8–12.0% at +30 days. All four forecasting models perform better than the naive model, with all four models exhibiting similar performance. We observe a larger decrease in performance as lags increase in terms of mean absolute error compared to median absolute error. Absolute error varies substantially across predictions for individual analysis dates and variants with most predictions having very little error, while a subset of predictions have larger error (Fig. 3.3). This skewed distribution results in the large observed differences between median and mean summary statistics. Thus, models predict frequencies well most of the time, but are occasionally incorrect and the proportion of incorrect predictions increases through time.

In addition to calculating median and mean absolute error, we estimate the coverage of 95% posterior latent frequencies (Fig. B.8A) and posterior predictive sample frequencies

Table 3.1: **Median and mean absolute error across models, countries and forecast lags** Models with the lowest error for each country / lag combination are bolded for clarity.

	Median Absolute Error					**Mean Absolute Error**				
Location	Naïve	Piantham	MLR	FGA	GARW	Naive	Piantham	MLR	FGA	GARW
-30 Lead from date of estimation										
Australia	0.80%	**0.20%**	**0.20%**	**0.20%**	**0.20%**	2.10%	**0.60%**	**0.60%**	**0.60%**	1.80%
Brazil	3.50%	0.80%	0.70%	0.80%	**0.60%**	7.60%	2.50%	**2.40%**	4.60%	3.30%
Japan	0.40%	**0.20%**	**0.20%**	**0.20%**	**0.20%**	2.90%	**1.40%**	**1.40%**	1.90%	**1.40%**
South Africa	3.70%	1.00%	0.90%	**0.80%**	**0.80%**	5.50%	2.30%	2.50%	**2.20%**	**2.20%**
Trinidad and Tobago	12.50%	1.50%	**1.40%**	**1.40%**	**1.40%**	19.90%	**4.20%**	**4.20%**	**4.20%**	**4.20%**
USA	0.20%	**0.10%**	**0.10%**	**0.10%**	**0.10%**	1.30%	0.40%	0.40%	1.80%	**0.20%**
United Kingdom	0.20%	**0.10%**	**0.10%**	**0.10%**	**0.10%**	1.50%	0.40%	0.40%	0.50%	1.20%
Vietnam	10.40%	1.50%	**1.30%**	1.40%	1.40%	21.00%	**4.00%**	**4.00%**	7.80%	6.20%
0 Lead from date of estimation										
Australia	1.80%	0.80%	**0.60%**	**0.60%**	0.70%	6.10%	3.20%	**2.80%**	**2.70%**	3.80%
Brazil	7.20%	1.10%	1.00%	**0.90%**	1.00%	18.30%	7.90%	**5.90%**	**6.10%**	6.80%
Japan	4.50%	0.50%	**0.30%**	0.50%	0.40%	10.10%	3.40%	**2.10%**	3.70%	**2.90%**
South Africa	9.30%	1.50%	1.60%	1.50%	**1.30%**	13.20%	**4.30%**	**4.30%**	4.00%	**4.30%**
Trinidad and Tobago	12.50%	1.90%	2.00%	1.70%	**1.60%**	27.50%	7.40%	7.30%	8.30%	9.50%
USA	0.60%	**0.40%**	**0.40%**	**0.40%**	**0.40%**	5.10%	**2.30%**	**2.30%**	3.00%	**1.80%**
United Kingdom	1.10%	0.50%	0.50%	0.40%	**0.20%**	5.70%	**2.30%**	**2.20%**	3.70%	**2.40%**
Vietnam	22.30%	1.50%	1.40%	**1.20%**	1.70%	25.60%	8.70%	**8.60%**	9.90%	10.30%
30 Lead from date of estimation										
Australia	6.20%	1.60%	1.50%	1.50%	**1.40%**	15.90%	6.80%	**6.20%**	6.40%	7.10%
Brazil	13.40%	**0.90%**	1.20%	1.00%	1.20%	26.80%	8.90%	**8.40%**	9.70%	
Japan	7.50%	**0.50%**	**0.50%**	**0.50%**	**0.50%**	17.50%	11.40%	7.30%	8.80%	**5.80%**
South Africa	15.80%	1.50%	1.60%	1.60%	**1.40%**	21.40%	6.90%	7.00%	**6.40%**	6.50%
Trinidad and Tobago	23.50%	2.00%	1.90%	1.60%	**1.30%**	38.60%	**11.30%**	12.00%	14.00%	16.10%
USA	2.00%	**0.60%**	0.70%	**0.60%**	**0.60%**	13.10%	6.30%	6.30%	**5.80%**	6.10%
United Kingdom	3.60%	0.80%	0.70%	**0.60%**	**0.60%**	13.90%	6.60%	**5.80%**	7.40%	**5.80%**
Vietnam	32.10%	1.60%	1.10%	**0.80%**	1.10%	33.20%	**11.00%**	11.60%	11.30%	13.30%

(Fig. B.8B) across models. We generate the posterior predictive coverage by sampling random counts for each variant using their posterior latent frequencies conditioning on the total sequences being those observed retrospectively. We find that the posterior predictive coverage is generally higher and a better fit for the models in question. Additionally, we find that the coverage is lower in countries with the highest sequencing intensity like the US and UK, suggesting that there may be over-dispersion in the sequence counts relative to binomial or multinomial sampling. We also observe that coverage is higher for the GARW model that allows for time-varying growth advantage than for the FGA or MLR models that enforce a fixed growth advantage. As clades evolve and new subclades emerge we expect clade-specific

growth advantage to change alongside.

In observing heterogeneity in prediction accuracy, we hypothesized that error is largest for emerging variants that present a small window of time to observe dynamics and where sequence count data is often rare. We investigate this hypothesis by charting how variant-specific growth advantage estimated in the MLR model varied across analysis dates (Fig. 3.4). Generally, we see sharp changes in estimated growth advantage in the first 1-3 weeks when a variant is emerging, but then see less pronounced changes. Thus, it often takes several weeks for the MLR model to 'dial in' estimated growth advantage, and accuracy will tend to be poorer in early weeks when variant-specific growth advantage is uncertain.

Genomic surveillance systems and forecast error

Using the MLR model, we find that different countries have consistently different levels of forecasting error with forecasts in Brazil and South Africa showing more error than forecasts in the UK and the USA, while Trinidad and Tobago and Vietnam show more error than the other six countries (Fig. 3.5A). We correlate broad statistics describing both quantity and quality of sequence data available at different analysis time points and in different genomic surveillance systems to forecasting error (Fig. 3.5B–E). Using Pearson correlations, we find that poor sequence quality as measured by proportion of available sequences labeled as 'bad' by Nextclade quality control [4] correlates slightly with mean AE (Fig. 3.5B). We find that good sequence quantity as measured by total sequences available at analysis has a moderate negative correlation with mean absolute error (Fig. 3.5E).

These results show that South Africa with ∼16k sequences collected in 2022 and median of 173 sequences available from the previous 30-days yields a mean absolute +30 day forecasting error of 7.0% for the MLR model (Table 3.1), which is only slightly greater than the mean absolute error of 6.3% for the US with ∼2.0M sequences collected in 2022 and of 5.8% for the UK with ∼1.2M sequences collected in 2022. However, Vietnam with ∼6k sequences collected in 2022 and median of 31 sequences available from the previous 30-delays yields a mean absolute forecasting error of 11.6% and Trinidad and Tobago with ∼2.3k sequences

collected in 2022 and median of 44 sequences available from the previous 30-delays yields a mean absolute forecasting error of 12.0%. This suggests that genomic surveillance systems with cadence and throughput greater than 50-100 sequences collected in the previous 30 days yield sufficient timely data to permit short-term forecasts.

We follow up on this across-country analysis and subsample existing sequences from the United Kingdom and Denmark to investigate what number of sequences need to be collected weekly to keep forecast error within acceptable bounds. For context, we also computed the mean weekly sequences collected for selected countries globally in 2022 (Fig. 3.6A). We select the United Kingdom due to its large counts of available sequences, relatively short submission delay, and low forecast error. Additionally, we include Denmark due to its large counts of available sequences and to explore the possibility of stochastic effects due to relative population sizes (Denmark has ∼9% the population of the UK). We simulate several downscaled data sets by subsampling the collected sequences at multiple thresholds for number of sequences per week and then fit the MLR model to each of the resulting data sets to see how forecast accuracy varies with sampling intensity. In order to properly account for variability in the subsampled data sets, we generate 5 subsamples per threshold, location and analysis date.

From this analysis, we find that increasing the number of sequences per week generally decreases the average error (Fig. 3.6B and C), as well as decreasing the proportion of out-of-bounds predictions (Fig. 3.6D and E), but there are diminishing returns. Additionally, the effect appears to saturate at different values depending on the forecast length. We find that for +14 and +30 day forecasts sampling at least 1000 sequences per week is fully sufficient to minimize forecast error, and 200 sequences per week is largely sufficient to curtail error. We arrive at a similar threshold of 1000 sequences per week for both the UK and Denmark (Figs. 3.6B-E).

Comparing country-level and hierarchical short-term forecast models

In observing poor performance in initial period of variant emergence (Fig. 3.4), as well as poor performance in countries with less intensive genomic surveillance (Fig. 3.5), we conclude lack of data results in poor fitness estimates and so poor predictive performance. Joint modeling of data from multiple countries has been proposed as a way to getting improved estimates of variant growth advantages in general and also specifically improving frequency estimates in low and middle income countries. Hierarchical or joint forecast models for short-term frequency forecasts typically operate by pooling parameters between 'groups' in a model. For our application, we pool the relative fitness of variants across countries, so that estimated relative fitnesses are informed by not just the observed relative fitness within a location, but also the relative fitnesses in other locations.

We compare the short-term forecast accuracy for individual models fit using MLR and this hierarchical MLR model in Fig. 3.7. We find that overall the hierarchical MLR matches or outperforms the single country models in all locations and at all forecast lengths. Perhaps as expected the hierarchical MLR model matches MLR performance in countries with abundant data like the US and UK, while countries with less data like Trinidad and Tobago and Vietnam show a large performance advantage to hierarchical MLR.

3.4 Discussion

In this manuscript we sought to perform a comprehensive analysis of the accuracy of nowcasts and short-term forecasts from fitness models of SARS-CoV-2 variant frequency. We observe substantial differences between median and mean absolute error (Fig. 3.2 and Table 3.1) with median errors generally quite well contained at 0.5–1.9% in the +30 day forecast, while mean errors are larger at 5.8–12.0%. This difference is due to the highly skewed distribution of model errors (Fig. 3.3) where most predictions are highly accurate, but a smaller fraction are off-target. As expected, errors increase as target shifts from −30 day hindcast to +30 day forecast, but error increases more rapidly for mean absolute error than median absolute

error. All four forecasting models explored here present a largely similar spectrum of errors.

We find that the Piantham, MLR, FGA and GARW models provide systematic and substantial improvements in forecasting accuracy relative to a 'naive' model that uses 7-day smoothed frequency at the last timepoint with sequence data (Fig. 3.2 and Table 3.1). For the MLR model, at +30 days the improvement in median absolute error over naive is 1.4–31.0% and the improvement in mean absolute error is 6.8–26.2%. This result supports the use of MLR models in live dashboards like the CDC Variant Proportions nowcast (covid.cdc.gov/covid-data-tracker/#variant-proportions) and the Nextstrain SARS-CoV-2 Forecasts (nextstrain.org/sars-cov-2/forecasts/).

We also observe improvements in accuracy for the −30 day hindcast of modeled frequency relative to naive frequency with the MLR model showing improvement in median absolute error of 0.1–11.1% and improvement in mean absolute error of 0.9–17.0%. These improvements were greatest in countries with lower cadence and throughput of genomic surveillance (Trinidad and Tobago and Vietnam). Importantly, this suggests that fitness models are useful for hindcasts in addition to short-term forecasts and that −30 day retrospective frequency should not be taken as truth, ie it takes more time than 30 days for backfill to resolve retrospective frequency.

However, we observe that coverage is generally lower than ideal with predictive coverage under 50% for countries with the most sequencing (Fig. B.8B). We believe this may be due to a combination of over-dispersion of sequence counts relative to the multinomial sampling assumption as well as clade-level growth advantages changing through time as clades evolve. The former could be addressed by including over-dispersion in the sequence observation model and the latter could be addressed by implementing growth advantages that vary through time in an auto-correlated fashion.

We find that variability in forecast errors is partially driven by data limitations. When new variants are emerging, we lack sequence counts and lack time to observe growth dynamics resulting in initial uncertainty of variant growth rates (Fig. 3.3). Relatedly, analyzing the variation in nowcast error, we find that overall sequence quality and quantity at time of

analysis are associated with model accuracy (Fig. 3.5). Thus, as expected, sequence quality, volume and turnaround time are all important for providing accurate, real-time estimates of variant fitness and frequency. Subsampling existing data in high sequencing intensity countries, we find that there are diminishing returns to increasing sequencing efforts and that maximum accuracy is achieved at around 1000 sequences per week and substantial accuracy is achieved at around 200 sequences per week (Fig. 3.6). This level of sequencing enables robust short-term forecasts of pathogen frequency dynamics at the level of a country and highlights the feasibility of pathogen surveillance for evolutionary forecasting. As observed in Susswein et al. [76], pooling data across countries using a hierarchical fitness model improves short-term forecasts for SARS-CoV-2 variant dynamics (Fig. 3.7).

In live MLR analyses at nextstrain.org/sars-cov-2/forecasts/, we have relied on the number of sequences available from samples collected in the previous 30-days as the key metric for inclusion of a country in the analysis. Along these lines, for pragmatic guidance for thresholds in which to trust MLR results, we observe that Trinidad and Tobago with 2.3k sequences collected in 2022 and a median 30-day sequence count of 43 shows a mean absolute forecasting error of 12%, that Vietnam with 6k sequences collected in 2022 and a median 30-day sequence count of 30 shows a mean absolute forecasting error of 11% and that South Africa with 16k sequences collected in 2022 and a median 30-day sequence count of 170 shows a mean absolute forecasting error of 7%. This suggests that a threshold of 50 sequences in previous 30 days should be roughly consistent with a $\sim$10% forecasting error. Keeping forecasting error under 10% seems like a reasonable target for public display of frequency forecasts and would support targeting a threshold of 50 sequences from samples collected in the previous 30 days.

In addition to differences in genomic surveillance, we expect countries may differ in variant dynamics due to differences in absolute viral prevalence. We expect that variant frequencies we more closely follow the MLR expectation when absolute prevalence is large, achieved through a large host population and/or frequent repeated infection. Rapid continued evolution of SARS-CoV-2 [51] suggests that we will continue to see widespread circulation of

SARS-CoV-2 and thus we generally expect fitness-based models to provide an adequate description of variant frequency dynamics. Of note, in comparing the UK with 67M people to Denmark with 6M people we observe similar levels of prediction error when sampling sequences at similar intensities (Fig. 3.6). This suggests that stochastic effects from low absolute viral prevalence were not strongly manifesting with a population size of 6M. However, we do expect that at some smaller population size stochastic effects and repeated importations will cause a deviation in frequency dynamics from fitness model expectations.

Although these models appear largely accurate for short-term forecasts, they may be improved by incorporating underlying biological mechanism. In general, the methods discussed here are primarily statistical in nature and do not account for much of the biological or immunological knowledge that we have or could obtain. The incorporation of such knowledge could increase the short-term and medium-term capabilities of these models. Additionally, these fitness models do not account for future mutations and can only project forward from circulating viral diversity. This intrinsically limits the effective forecasting horizon achievable by these models. Future modeling work should seek to incorporate the emergence and spread of 'adjacent possible' mutations for longer term forecasts on the order of several months or years [46]. Without empirical frequency dynamics to draw upon, the fitness effects of these adjacent possible mutations may be estimated from empirical data such as deep mutational scanning [17, 39, 23]. Continued timely genomic surveillance and biological characterization along with further model development will be necessary for successful real-time evolutionary forecasting of SARS-CoV-2.

Methods

Preparing sequence counts and case counts

We prepared sequence count data sets to replicate a live forecasting environment using the Nextstrain-curated SARS-CoV-2 sequence metadata [40], which is created using the GISAID EpiCoV database [48]. To reconstruct available sequence data for a given analysis

date, we filtered to all sequences with collection dates up to 90 days before the analysis date, and additionally filtered to those sequences which were submitted before the analysis date. These sequences were tallied according to their annotated Nextstrain clade to produce sequence count for each country, for each clade and for each day over the period of interest. Sequence counts were produced independently for the 8 focal countries Australia, Brazil, Japan, South Africa, Trinidad and Tobago, the United Kingdom, the United States, and Vietnam. We repeated this process for a series of analysis dates on the 1st and 15th of each month starting with January 1, 2022 and ending with December 15, 2022 giving a total of 24 analysis data sets for each country. Since two models (FGA, GARW) also use case counts for their estimates, we additionally prepare data sets using case counts over the time periods of interest as available from Our World in Data (ourworldindata.org/covid-cases).

Frequency dynamics and transmission advantages

We implemented and evaluated multiple models that forecast variant frequency. These models estimate the frequency $f_v(t)$ of variant v at time t, and simultaneously estimate the variant transmission advantage $\Delta_v = \frac{R_t^v}{R_t^u}$ where R_t^v is the effective reproduction number for variant v and u is an arbitrarily assigned reference variant with fixed fitness. We can interpret these transmission advantages as the effective reproduction number of a variant relative to some reference variant.

The four models of interest are: Multinomial Logistic Regression (MLR) of frequency growth, two models of variant-specific R_t: a fixed growth advantage model (FGA) parameterization and a growth advantage random walk (GARW) parameterization of the renewal equation framework of Figgins and Bedford [33], as well as another approach to estimating relative fitness by Piantham et al [66]. We provide a brief mathematical overview of these methods below.

The multinomial logistic regression model estimates a fixed growth advantage using logistic regression with a variant-specific intercept and time coefficient, so that the frequency

of variant v at time t can be modeled as

$$f_v(t) = \frac{\exp(\alpha_v + \delta_v t)}{\sum_u \exp(\alpha_u + \delta_u t)}, \tag{3.1}$$

where α_v is the initial frequency and δ_v is the growth rate of variant v, and the summation in the denominator is over variants 1 to n. Inferred frequency growth f_v can be converted to a growth advantage (or selective coefficient) as $\Delta_v = \exp(\delta_v \tau)$ assuming a fixed deterministic generation time of τ.

The model by Piantham et al [66] relies on an approximation to the renewal equation wherein new infections do not vary greatly over the generation time of the virus. This model generalizes the MLR model in that it accounts for non-fixed generation time though it assumes little overall case growth.

The fixed growth advantage (FGA) model uses a renewal equation model based on both case counts and sequence counts to estimate variant-specific R_t assuming that the growth advantage Δ_v of variant v is fixed relative to reference variant u [33]. The growth advantage random walk (GARW) model uses the same renewal equation framework and data, but allows variant growth advantages to vary smoothly in time [33].

The models used all differ in the complexity of their assumptions in computing the variant growth advantage. Growth advantages presented in this manuscript are estimated relative to the initial Omicron strain (clade 21L, lineage BA.1), providing a point of reference for competing growth advantages and how median values change over time. Further details on the model formats can be found in their respective citations. All models were implemented using the evofr software package for evolutionary forecasting (https://github.com/blab/evofr) using Numpyro for inference.

As a baseline, we compared the four models above to a naive model which that the forecast as the average of the last available frequencies.

Additionally, we implement a hierarchical variant of the model where multiple countries are fit simultaneously with a Normal prior on the relative fitness of a given variant between countries, so that $\delta_{v,g} \sim \text{Normal}(\overline{\delta}_v, \sigma)$. Similar formulations of this hierarchical model have

been used for SARS-CoV-2 frequency forecasts previously. [76]

Evaluation criteria

We calculated the 'absolute error' (AE) for a given model m and data set d as the difference between the retrospective raw frequencies and the predicted frequencies as

$$\mathrm{AE}_t^{m,d} = \frac{1}{n} \sum_{v \in V} \left| f_v^d(t) - \hat{f}_v^{m,d}(t) \right|, \tag{3.2}$$

where $f_v^d(t)$ and $\hat{f}_v^{m,d}(t)$ are the retrospective frequencies and the predicted frequencies for model m, data d, variant v and time t. The AE is the mean across individual variants for a specific model, data set and time point. Additionally, we often work with the lead time, which is defined as the difference between date of analysis for the data set and the forecast date $l = t - T_{\mathrm{obs}}$. We summarized median absolute error and mean absolute error across multiple analysis datasets in Fig. 3.2 and Table 3.1.

Throughout this study, we primarily use the median and mean absolute error to evaluate the accuracy of our point forecasts. We select the median absolute error as a measure of central tendency on our forecast errors, reducing the influence of outliers and skewed data distributions due to the contribution of forecasts which tend to diverge rapidly in forecast lead. To balance this and account for the effect of outliers and rapidly divergent forecasts, we also use the mean absolute error which is less sensitive to outliers than the mean square error and has units in terms of frequencies directly.

However, these are not the only possible choices for error metrics. Our choice of metrics is motivated by our decision to focus primarily on point forecasts of variant frequencies. To supplement this analysis, we also address the coverage of probabilistic extensions of the models discussed here.

Generating predictors of error

We explored four key variables to describe the effect of sequencing efforts on nowcast errors and estimated Pearson correlations with the mean absolute nowcast errors. These variables

are defined as proportion of bad quality control (QC) sequences according to Nextclade [4], fraction of sequences available within 14 days of the prediction time, total sequences availability within 14 days of the prediction time and median delay of sequence submission. To calculate these variables, we selected a 14-day window of data before each and every analysis date and used the collection and submission dates to determine their availability. Total sequence availability was calculated by dividing the sequences where submission date was before the date of analysis by the total collected sequences and similarly fraction of sequences at observation was estimated. Sequence submission delay was calculated by taking the difference between the submission date and the date of collection. Bad QC sequence proportion was estimated by dividing the sequences with bad QC classification by the total collected sequences. Estimates were computed for all defined dates of analysis across all countries.

Assessing coverage for short-term frequency forecasts

The main results of our analyses rely on mean and median absolute error as metrics. However, there is much to gain by using probabilistic forecasts for variant frequency. To this aim, we investigate the coverage of these different methods for forecasting variant frequency. Though not all models described initially were designed with uncertainty quantification in mind, we develop and fit Bayesian extensions of these models that are fit to the same data sets as before using stochastic variational inference.

Downscaling historical sequencing effort

We analyze the effects of scaling back sequencing efforts to assess the effect of sequencing volume on nowcast and forecast errors. Using the sequencing data from the United Kingdom and Denmark, we subsampled existing available sequences at the time of analysis at a rate of 100, 200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, and 2000 sequences per week of any submission date. We then generated datasets for the same analysis dates and study period used in the previous analyses, generating 5 replicate subsampled data sets of sequences

available at each analysis date for each eventual sequencing rate, location, and analysis date. Subsampling sequences per week before checking which sequences were available by the analysis date ensures that we respect the availability of sequences by submission date and submission delay in each country, so that countries with many sequences per week but long delays will maintain these delays. Therefore, the selected sequencing rate sets an upper limit on the number of sequences available per week at any analysis date and preserves the decline in available sequences that we typically observe in recent weeks since we only include sequences which are within our original subsample and available at the time of analysis. We then fit the MLR forecast model to each resulting data set and forecast up to 30 days after analysis date and compared these forecasts to the truth set in previous sections to compute the forecast error for each model. To better understand how the forecast error varies with sequencing intensity and forecast length, we computed the fraction of forecasts within an error tolerance (5% AE) as well as the average error at different sequence threshold and lag times.

Comparing forecasts using retrospective clade designations and real-time designations

The main analyses discussed in this manuscript rely on subsetting and filtering SARS-CoV-2 sequence metadata accessed on a particular date. However, the clade designations used throughout this manuscript may not have been the same as clade designations at the time the data was available. To understand how this affects our evaluation of forecast error, we compare the accuracy of models fit to the sequence counts from metadata at the time and using the available Nextclade reference tree to those fit on the retrospective Nextclade reference tree used in the rest of the analyses in this paper. This compares lineage designations that were available in real-time on the historical analysis date to lineage designations that are retrospectively available. In particular, we focus on the timing of the designation of lineage BQ.1 (corresponding to clade 22E) in October 2022 and show the accuracy of MLR using the different data sets at different forecast leads. We compare the resulting MAE of these analyses between Nextclade versions in Fig. B.9 and show trajectories from individual

countries in Figs. B.10–B.17.

Data and code accessibility

Sequence data including date and location of collection as well as clade annotation was obtained via the Nextstrain-curated data set that pulls data from GISAID database. A full list of sequences analyzed with accession numbers, derived data of sequence counts and case counts, along with all source code used to analyze this data and produce figures is available via the GitHub repository github.com/blab/ncov-forecasting-fit.

Acknowledgements

We thank John Huddleston for many helpful comments on the approach and on the manuscript. We gratefully acknowledge all data contributors, ie the Authors and their Originating laboratories responsible for obtaining the specimens, and their Submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based. We have included an acknowledgements table in the associated GitHub repository under `data/final_acknowledgements_gisaid.tsv.gz`.

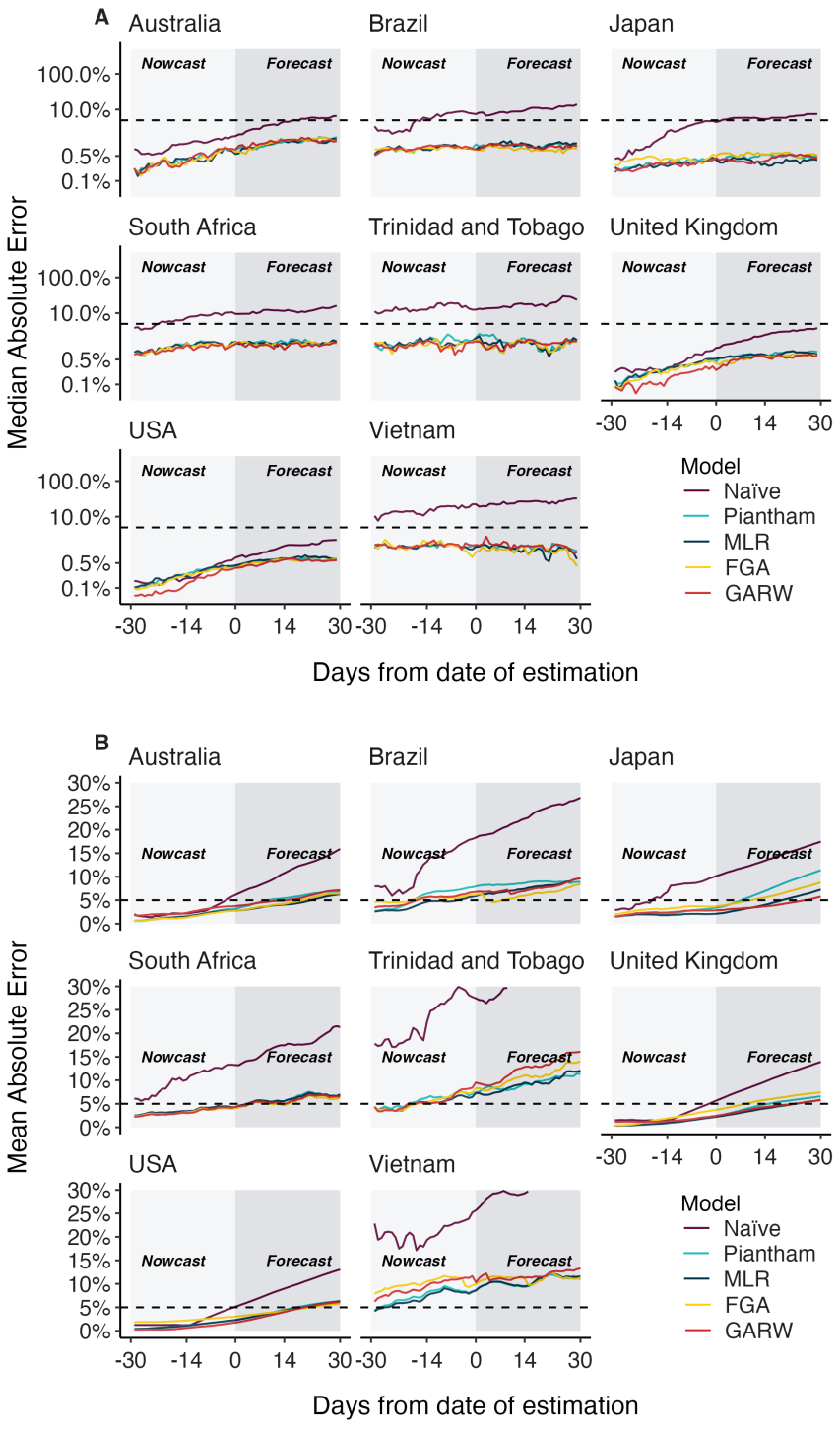

Figure 3.2: **Absolute error across models, countries and forecast lags.** (A) Median absolute error and (B) mean absolute error across countries, models and forecast lags moving from −30 day hindcasts to +30 day forecasts. For each county / model / lag combination, the median and the mean are summarized across analysis data sets. Panel A uses a log y axis for legibility while panel B uses a natural y axis.

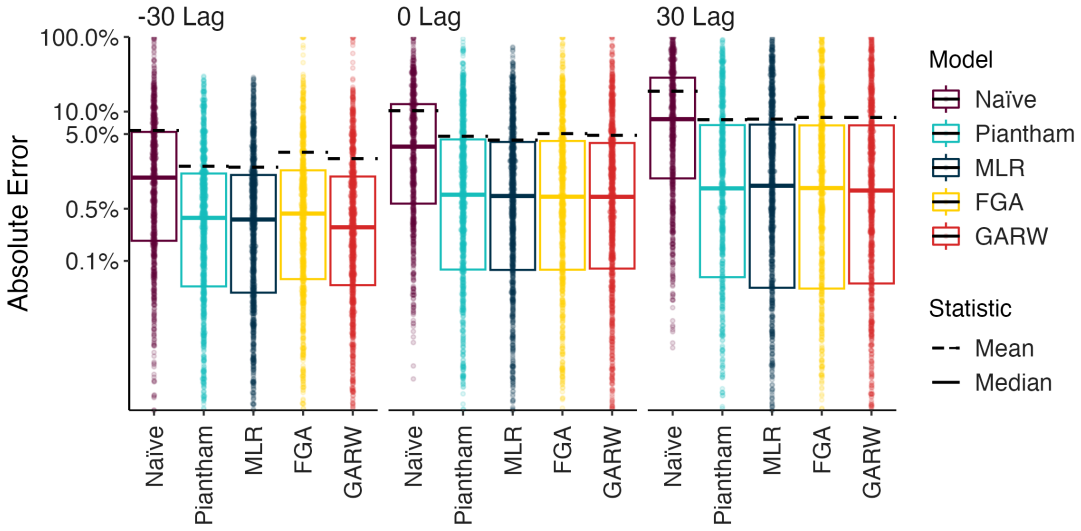

Figure 3.3: **Absolute error across models, countries and forecast lags.** Distribution of absolute error on a log scale across models and across forecast lags. Each point represents the absolute error for a data set / country combination. Solid lines show the median of these distributions and dashed lines show the means of these distributions.

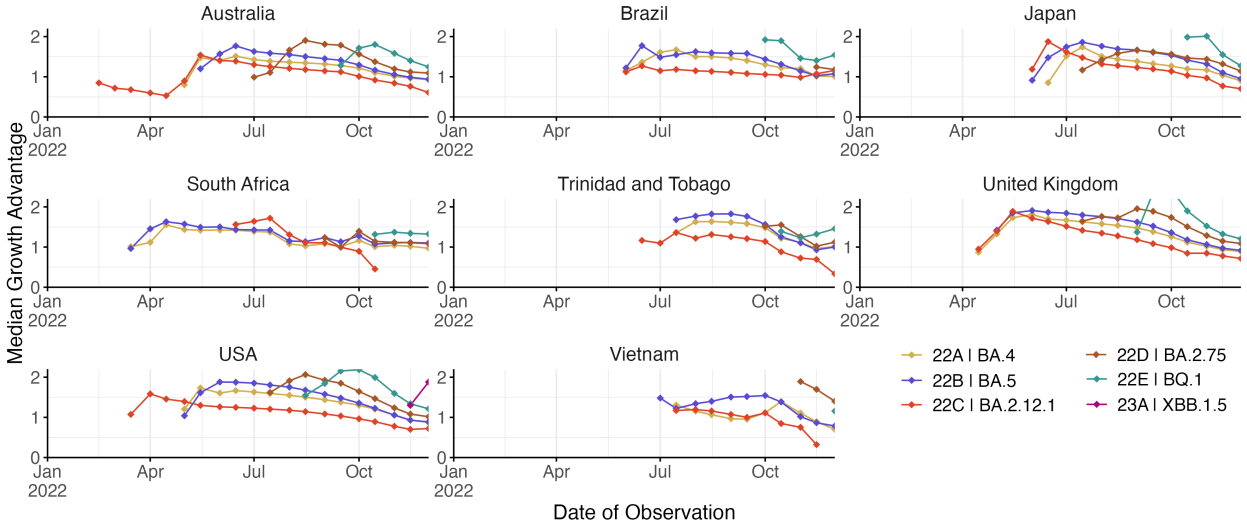

Figure 3.4: **Growth advantage of variants across analysis dates.** Growth advantage is estimated via the MLR model and is computed relative to clade 21K (lineage BA.1).

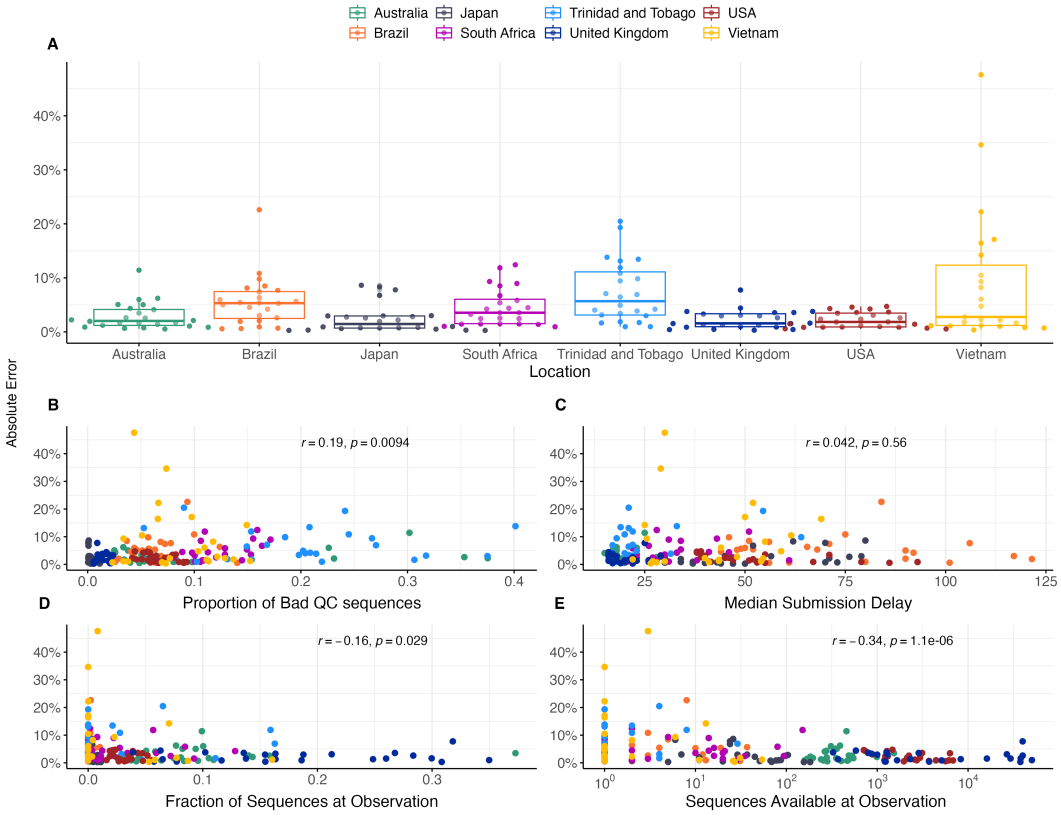

Figure 3.5: **Sequence quantity and quality influence nowcasts error.** (A) Absolute error at nowcast for the MLR model across countries. Points represent separate data sets at different analysis dates. Median and interquartile range of absolute errors are shown as box-and-whisker plots. (B-E) Correlation of sequence quality and sequence quantity metrics with absolute error. Points represent separate data sets at different analysis dates. Correlation strength and significance are calculated via Pearson correlation and are inset in each panel.

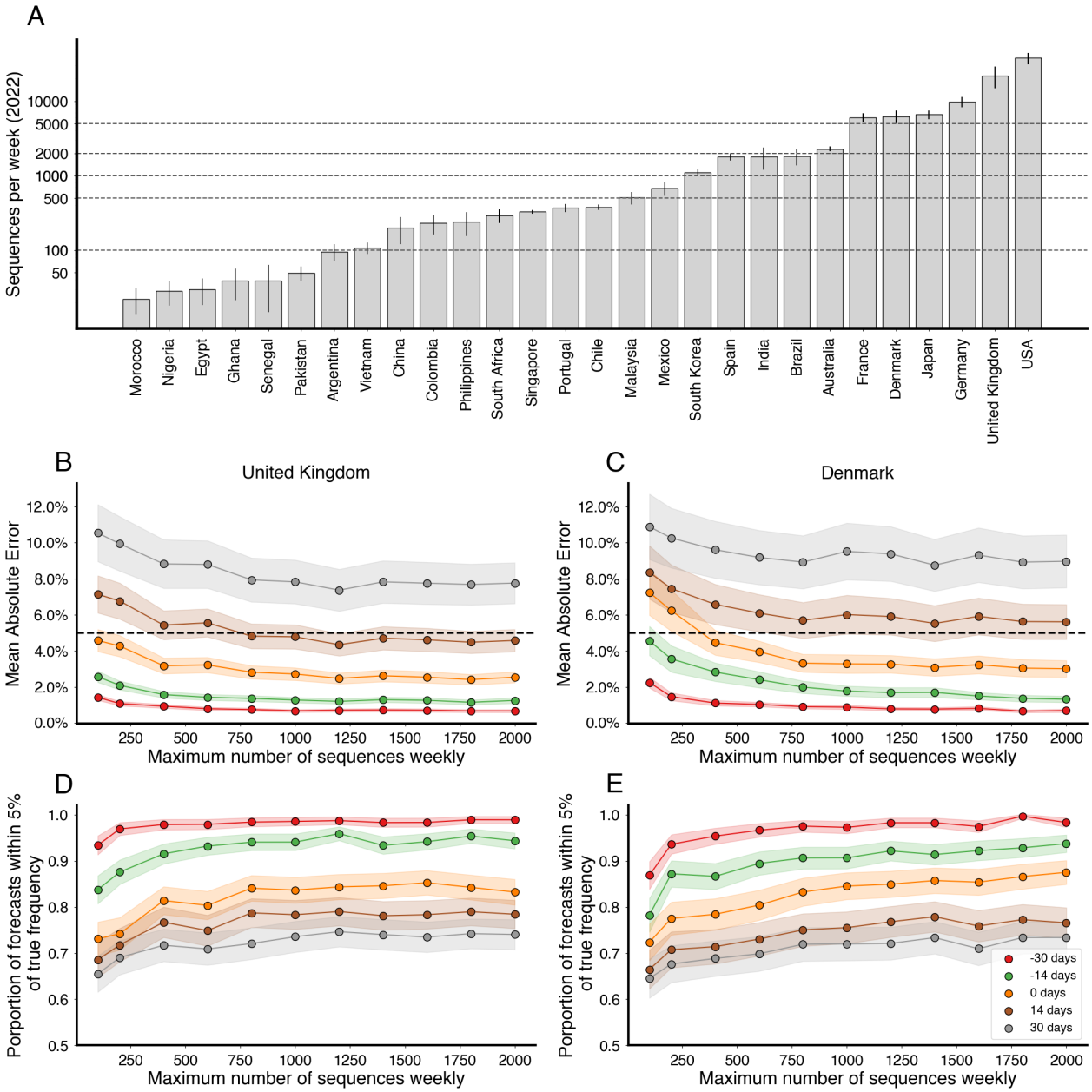

Figure 3.6: **Increasing sequencing intensity reduces forecast error** (A) Mean sequences collected per week for selected countries in 2022. Intervals are 95% confidence intervals of the mean. Dashed lines correspond to sampling rates used in (B-E). (B, C) Mean absolute error as a function of sequences collected per week colored by forecast horizon (-30 days, -15 days, 0 days, +15 days, +30 days) for the United Kingdom and Denmark. The dash line corresponds to 5% frequency error. (D, E) Proportion of forecasts within 5% of retrospective frequency as a function of sequences collected for week for the United Kingdom and Denmark.

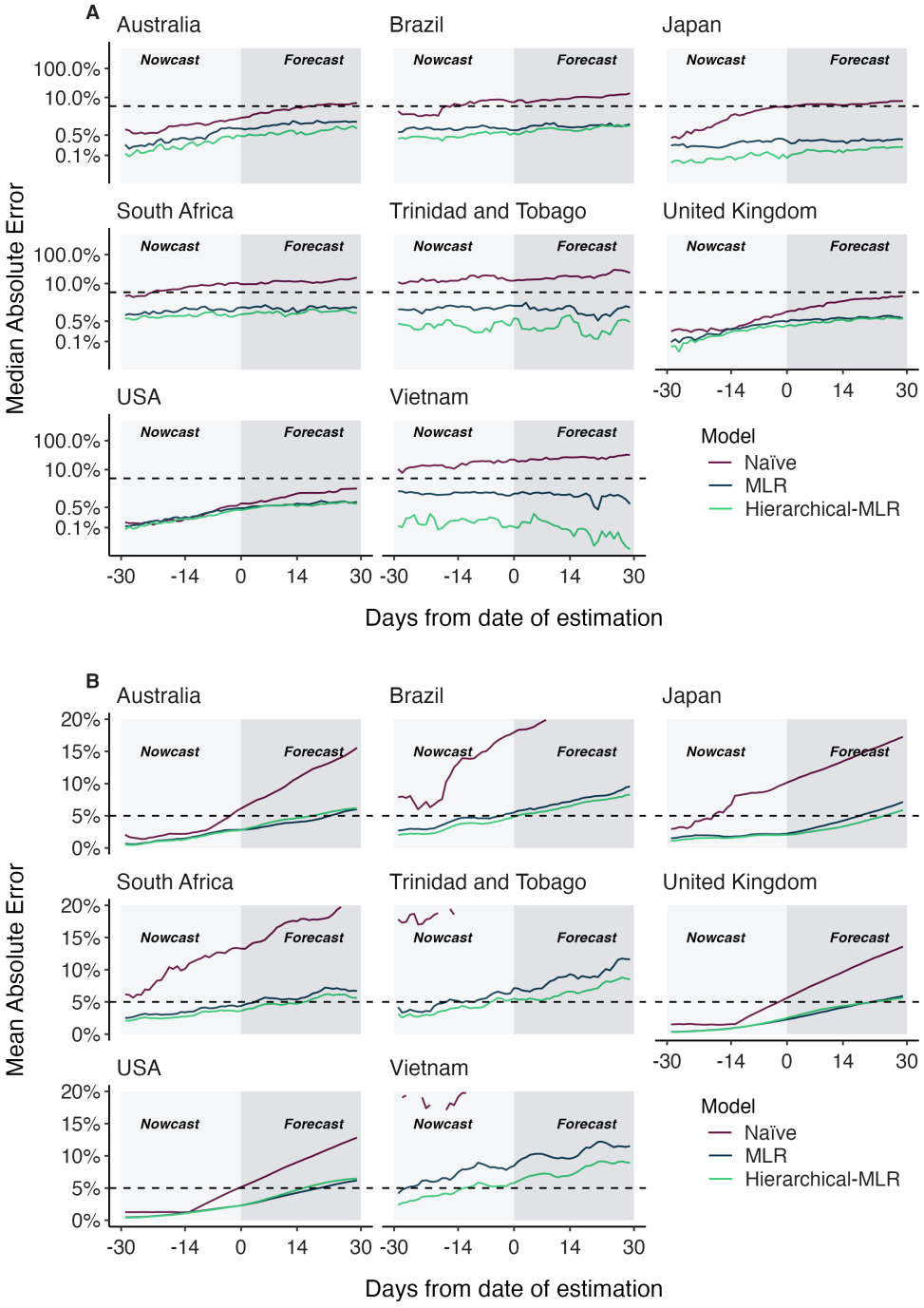

Figure 3.7: **Absolute error comparing standard MLR and hierarchical MLR across countries and forecast lags.** (A) Median absolute error and (B) mean absolute error across countries, models and forecast lags moving from −30 day hindcasts to +30 day forecasts. For each county / model / lag combination, the median and the mean are summarized across analysis data sets. Panel A uses a log y axis for legibility while panel B uses a natural y axis.

Chapter 4

FREQUENCY DYNAMICS PREDICT VIRAL FITNESS, ANTIGENIC RELATIONSHIPS AND EPIDEMIC GROWTH

4.1 Abstract

During the COVID-19 pandemic, SARS-CoV-2 variants drove large waves of infections, fueled by increased transmissibility and immune escape. Current models focus on changes in variant frequencies without linking them to underlying transmission mechanisms of intrinsic transmissibility and immune escape. We introduce a framework connecting variant dynamics to these mechanisms, showing how host population immunity interacts with viral transmissibility and immune escape to determine relative variant fitness. We advance a selective pressure metric that provides an early signal of epidemic growth using genetic data alone, crucial with current underreporting of cases. Additionally, we show that a latent immunity space model approximates immunological distances, offering insights into population susceptibility and immune evasion. These insights refine real-time forecasting and lay the groundwork for research into the interplay between viral genetics, immunity, and epidemic growth.

4.2 Main text

The COVID-19 pandemic was marked by the successive emergence of SARS-CoV-2 variant viruses, driving repeated epidemics globally [79, 85]. While these repeated large waves occurred with the emergence of novel variants, the mechanism driving these variants' success changed over time. The spread of early variants such as Alpha, Beta, Gamma and Delta were largely driven by increases in intrinsic transmissibility [19]. The Omicron variant showed substantial immune escape [19] and subsequent derived lineages within Omicron in-

cluding XBB, EG.5.1 and JN.1 appear to be driven by immune escape as evidenced through molecular studies of neutralization using human sera [16, 15, 9, 45]. Since 2022, there has been repeated replacement by subsequent Omicron-derived lineages. This rapid viral population turnover is consistent with antigenic evolution and is observed in other viruses such as seasonal influenza [8], although SARS-CoV-2 currently remains an outlier in terms of pace of its evolution [51]. This transition from transmissibility-driven to immune escape-driven success is a consequence of the interplay between population immunity and variant fitness.

With the increased temporal and geographical scale of sequencing alongside a detailed genetic nomenclature [69] and bioinformatic tools for lineage assignment [80, 4], we have gained more data for SARS-CoV-2 than for other circulating viruses giving a unique opportunity for insight into its evolution. Several models of variant frequency have been developed to estimate the fitness of emerging SARS-CoV-2 variants [5, 65, 34, 76, 54, 2]. These models estimate the relative fitness (or selective advantage) of circulating variant viruses from their frequency in sequencing data, typically represented by counts of variant sequences over time within a geographic region. Relative fitness in these models is often assumed to be constant and intrinsic to the variant of interest. However, this may be an oversimplification of the transmission process.

It has been shown that these transmission advantages differ geographically and temporally, suggesting that variant transmission advantages are not necessarily fixed and may be informed by within-region population differences [34, 82]. In fact, heterogeneity in transmission advantages may be well explained by regional differences in immune structure as Dadonaite et al. [24] show deep mutational scanning estimates of immune escape are well correlated with estimated variant growth advantages. Existing models that allow variant transmission advantages to change in time generally do not have a mechanistic underpinning for why transmission advantages exist and vary geographically and temporally [34, 76]. This lack of mechanistic grounding limits our ability to accurately predict variant dynamics, especially in diverse geographic regions with varying levels of population immunity.

In response to this gap, we introduce a novel framework that links variant dynamics

directly to transmission mechanisms using compartmental models of infectious diseases. By modeling both intrinsic transmissibility and immune escape, we explain how shifts in population immunity shape the relative fitness of viral variants and select for immune escape over intrinsic transmissibility with increasing past exposure. Furthermore, including these mechanisms suggests that relative fitness varies in time, reflecting the evolving landscape of population immunity and exposure regardless of the underlying mechanism.

Here, we present a novel non-parametric method for estimating time-varying fitness regardless of the underlying transmission mechanism. Alongside this development we introduce a "selective pressure" metric that quantifies the impact of variant turnover on population-level epidemic growth rates, as well as a latent immunity model that we use to estimate the underlying proportion of pseudo-immune groups within multiple geographies and pseudo-immune escape rates for circulating variants. Overall, our framework bridges the gap between genetic data and transmission dynamics, offering a new way to predict and manage viral outbreaks.

Variant dynamics and relative fitness in multistrain models

Multi-strain models of epidemics have been developed to understand the competition between different viral strains that exhibit different levels of cross-immunity [36, 7]. These models have typically been used to explain strain evolution in antigenically variable pathogens like seasonal influenza virus [8] and seasonal coronaviruses [50, 28].

We begin by modeling a population of V exponentially growing variant viruses each with prevalence $I_v(t)$ and time-varying growth rate $r_v(t)$. By considering the difference in these growth rates, we can define the relative fitness as $\lambda_{v,u}(t) = r_v(t) - r_u(t)$. This relative fitness determines the change in the frequencies of the variants in the population

$$f_v(t) = \frac{f_v(0) \exp\left(\int_0^t \lambda_{v,v^*}(s)ds\right)}{\sum_{u=1}^{V} f_u(0) \exp\left(\int_0^t \lambda_{u,v^*}(s)ds\right)}, \tag{4.1}$$

where v^* is a chosen pivot variant that has relative fitness zero.

In order to better understand frequency dynamics of pathogens with multiple co-circulating variants, we apply the above framework to compartmental models of epidemics, which can be written as time-varying exponential growth (detailed in Supplementary Text C.2.1). These models provide an intuition of how strain-level selection depends on the assumed transmission mechanism of the underlying epidemic model. This framework also generalizes several existing methods for relative fitness estimation and prediction (detailed in Supplementary Text C.2.2). We summarize dynamics of a three-variant mechanistic transmission model in Fig. 4.1, where we compare a transmission variant T with a 50% increase in transmissibility ($\rho = 0.5$) to an escape variant E that infects 5% of hosts possessing wildtype immunity ($\eta = 0.05$).

Our approach shows that relative fitness is often dependent on the past exposure of a population (as discussed in Supplementary Text C.2.1 and extended to full immune history models in Supplementary Text C.2.3). This suggests that serology, vaccination history, and immunological data generally can be informative of relative fitness. Additionally, when working with variant classifications, non-neutral evolution within a variant will cause the relative fitness of that variant to change in time. However, even in the absence of external data that can inform relative fitness, there is still hope.

We develop a method for using approximate Gaussian processes to model variant relative fitness. Gaussian processes are probability distributions over functions, where the structure and smoothness of these functions are defined by a kernel that encodes correlations in time. These models are flexible and allow us to encode smoothness constraints, periodicity, and other structures [37]. Gaussian processes allow us a non-parametric estimate of the relative fitness for variants through time (see Materials and Methods).

Traditional Gaussian processes, while flexible, face challenges for large time series and large data sets. Our approach overcomes this using a Hilbert Space Gaussian Process (HGSP) approximation, making the framework scalable for many variants and long time periods [70]. This enables real-time variant fitness estimation and can be applied to any frequency data regardless of the underlying mechanism. This model is used in Fig. C.1 to estimate the

relative fitnesses of different variants through time based on simulated variant sequence counts from frequencies shown in Fig. 4.1.

Later, we apply this model to empirical SARS-CoV-2 sequence data from 50 US states and England from 2021 to 2022 to estimate relative fitness for variants circulating in that period, but first we continue analytic investigation into fitness dynamics.

Determining the transmisibility-escape tradeoff

To understand the fitness trade-off between transmissibility and immune escape, we consider dynamics with a wildtype virus W with $\rho_W = 0$ and $\eta_W = 0$, an increased transmissibility variant T with $\rho_T > 0$ and $\eta_T = 0$ and an immune escape variant E with $\rho_E = 0$ and $\eta_E > 0$.

Following Equation C.22, we write relative fitnesses of the escape variant or transmissibility variant as

$$\lambda_{E,W} = \eta\beta\varphi_W(t) \tag{4.2}$$

$$\lambda_{T,W} = \rho\beta S(t). \tag{4.3}$$

In the simplest case where individuals are either susceptible or have wildtype immunity ($S(t) + \varphi_W(t) = 1$), we can compute the critical immune fraction φ^* at which $\lambda_{E,W}(\varphi^*) = \lambda_{T,W}(\varphi^*)$ as

$$\varphi^* = \frac{\rho}{\eta + \rho}. \tag{4.4}$$

For past exposure level greater than φ^* escape variants have a higher relative fitness. This trade off shows that increasing degree of escape entails that a lower proportion of past exposure is needed for escape variants to be preferred (Fig. 4.2). Additionally, this shows that when intrinsic transmissibility increases are limited escape is more likely to be a dominant mechanism for variant turnover.

Initial growth rates insufficient for predicting short-term frequency growth

One question of interest is whether knowledge of mechanism meaningfully informs our ability to forecast short-term frequency growth. The first step to addressing this is to understand

how the relative fitness may change in time to understand the predictability of relative fitness in the short-term.

We find that the mechanistic forms analyzed in this paper (Supplementary Text C.2.1) can be represented as weighted combinations of B time-varying functions $\Upsilon_b(t)$ with weights β_b. We can think of each of these functions Υ_b as an immune background and the coefficient β_b as a transmission differential, so that

$$\lambda_{v,u}(t) = \sum_{1 \leq b \leq B} \beta_b \Upsilon_b(t). \tag{4.5}$$

Even in the case of complete knowledge of the relative fitness and the underlying fitness contributions in the present and past, we have that change in the relative fitness is determined by

$$\frac{d\lambda_{v,u}}{dt} = \sum_{1 \leq b \leq B} \beta_b \frac{d\Upsilon_b}{dt}(t). \tag{4.6}$$

By considering a Taylor expansion of the relative fitness about the point of estimation t_0, we can approximate the relative fitness in the future as

$$\lambda_{v,u}(t) \approx \lambda_{v,u}(t_0) + (t - t_0) \sum_{1 \leq b \leq B} \beta_b \frac{d\Upsilon_b}{dt}(t_0). \tag{4.7}$$

This suggests small differences in the form of $\lambda_{v,u}(t)$ can lead to meaningful differences in the future relative fitnesses through changes in the underlying immune backgrounds.

We investigate whether relative fitnesses vary predictably in the short-term regardless of mechanism. To do so, we apply the two-variant model developed in previous sections for different mechanisms of immune escape and increased transmissibility. We fix the relative fitness of the novel variant at a prediction time t_0 using Equation 4.4 and assess the change in the relative fitness in the short-term. We find that although relative fitness trajectories share the same decreasing shape, they may decline at different rates depending on the mechanism (Fig. C.2). This can lead to substantial changes in the predicted incidence depending on the assumed mechanism and affects to overall rate of turnover.

Correlations insufficient for mechanism identification

Although correlations between vaccination uptake and variant growth advantage are often observed, these alone may not be sufficient to identify the mechanism behind a variant's success. A variant's fitness advantage may arise from increased transmissibility, immune escape, or a combination of both. Even in the absence of immune escape, the relative fitness of a variant depends on the proportion of the population that is susceptible to infection and therefore changes with both past exposure and vaccine uptake (Supplementary Text C.2.1). To illustrate this, we simulate the spread of a variant with increased transmissibility in populations with varying initial vaccination levels.

In populations with lower vaccination levels, the variant's prevalence peaks more sharply and its relative fitness declines quickly as immunity accumulates within the population (Fig. 4.3A-C). In contrast, higher vaccination levels constrain relative fitness, leading to a delayed peak in prevalence and more stable relative fitness as the existing immunity limits the variant's spread (Fig. 4.3A-C). Even without immune escape, estimated growth advantages for this variant decrease with increasing vaccination uptake near the beginning of an epidemic (Fig. 4.3D). Later in the epidemic, this relationship reverses with estimated growth advantages over the full period increasing with initial vaccination levels, which may be mistaken as signal for immune escape (Fig. 4.3E).

This analysis shows that correlation-based methods alone may struggle to identify the true mechanisms driving a variant's success especially under the assumption of a fixed growth advantage. By explicitly considering how immunity and transmissibility interact within populations, models that incorporate these dynamics may provide a stronger foundation for understanding why certain variants spread.

Quantifying selective pressure

Although it is useful to quantify the relative fitnesses of individual variants, we are often interested in quantifying the overall effects of selection in the population. With this in mind,

we can derive a metric of overall selective pressure

$$\psi(t) = \mathbb{E}_{f(t)}\left[\frac{d\lambda_v}{dt}\right] + \mathrm{Var}_{f(t)}[\lambda_v] \tag{4.8}$$

that describes the distribution of relative fitness in the population. This selective pressure metric serves as an indicator for high fitness variants arising in the population as change. High fitness variants rising from initially low frequency leads to large increases in the variance of the fitness distribution and therefore increases in the selective pressure.

The selective pressure metric enables us to decompose changes in the average growth rate in the population, $d\bar{r}/dt$, to an evolutionary component ψ and a residual baseline growth rate r_W following

$$\frac{d\bar{r}}{dt} = \frac{dr_W}{dt} + \psi(t). \tag{4.9}$$

This shows that increased selective pressure through emerging high fitness variants can drive waves of infection. Further, this suggests that differences between growth rates based on selective pressure alone and observed rates are attributable to changes in baseline transmission over time. This mirrors ideas of Fisher's theorem of natural selection and its later interpretations with the variance of fitness contributing directly to the change in transmission rates (or fitness) [30, 29]. This definition of selective pressure captures how relative fitness contributes to epidemic growth. This is similar to ideas quantifying rates of adaptation via fitness flux [58].

In this case, the overall growth rate $\bar{r}$ and relative incidence $I(t)/I(0)$ can be written directly

$$\bar{r}(t) = \bar{r}(0) + [r_W(t) - r_W(0)] + \Psi(t), \tag{4.10}$$

$$\frac{I(t)}{I(0)} = \exp\left(\int_0^t [r_W(s) + \Psi(s)]ds\right), \tag{4.11}$$

using the cumulative selective pressure $\Psi(t) = \int_0^t \psi(s)ds$. In addition to estimating the relative fitness, metrics derived from these models can inform us of much more.

Our "selective pressure" metric allows us to model the contribution of evolution to changes in the epidemic growth rate of a population and is independent of pivot choice for relative

fitness estimation. This metric acts as an early warning system for variant-driven outbreaks, especially in scenarios where case data are sparse or delayed. This metric can be computed using any method that estimates variant frequency and relative fitnesses and serves as a simple tool for understanding the contribution of selection to the overall population dynamics.

The full derivation of this metric and its contribution to the overall growth rate can be found in Supplementary Text C.2.4.

Predicting epidemic growth rates using selective pressure

Motivated by the relationship between epidemic growth rate and selective pressure demonstrated above, we develop a predictive model of epidemic growth rate using estimates of selective pressure. Using empirical SARS-CoV-2 case and sequence data from 50 US states between January 2021 and November 2022, we estimate epidemic growth rates through time in each state using case counts, and estimate selective pressure through time using our approximate Gaussian process model on sequence counts (Fig. 4.4 A–C.) Here we group variants at the granularity of Nextstrain clades [4] resulting in 28 distinct variants over this time period. As expected we see that relative fitness increases through time and that selective pressure corresponds to speed of clade turnover where the sweep of Omicron BA.1 (clade 21K) yields the strongest signal of selective pressure (Figs. C.3–C.7). We use these estimates to fit a gradient-boosted regressor to predict epidemic growth rates using selective pressure from the most recent 28 days, reserving data between July 2022 and November 2022 for testing (Fig. 4.4 D–I, Fig. C.8). This regressor is chosen via time series cross-validation among model architectures and grid-search parameter tuning (Fig. C.9).

We observe a strong correspondence between observed epidemic growth rate and model predictions with Pearson R^2 in the training period of 0.576 and a weaker Pearson R^2 in the testing period of 0.077. As case reporting declined over this period, we expect weaker correspondence between our predictions and epidemic growth rates computed from case data. To address this, we sought to evaluate the out-of-sample fit on case data from other countries e.g. South Africa, South Korea, and the United Kingdom, achieving an R^2 of 0.196.

To address the potential for this method under steady reporting rates, we validate this method by predicting the epidemic growth rates in England derived from the Office for National Statistics (ONS) Coronavirus Infection Survey between February 2022 and November 2022. The ONS Infection Survey represented a randomly sampled panel survey of households where nasal swabs were collected regardless of symptom status allowing for prevalence estimates despite faltering case reporting [67]. Our model is able to replicate patterns seen in epidemic growth rates in England derived from ONS data (Fig. 4.4 J–L), achieving a coefficient of variation of $R^2 = 0.329$ and mean absolute error of 0.026. Performance is significantly better for the first two subsequent waves, falling off in accuracy for the fall 2022 BQ.1 (clade 22E) wave.

Although these predictions can be biased by non-evolutionary effects on the epidemic growth, this approach provides a simple measure of epidemic growth in the absence of high quality case counts using sequence data alone.

Latent factor model of relative fitness

The representation of relative fitness using discrete immune backgrounds suggests that there may be low-dimensional structure to variant relative fitness. To generate pseudo-estimates of this latent factors, we develop and implement our method for latent factors models of relative fitness. This model assumes that variants intrinsically escape the immune responses with particular groups and that differences in a variant's relative fitness between geographies is attributable to differences in immunity between populations. This enables us to estimate a pseudo-escape rates for variants as well as pseudo-immunity groups within geographies over time.

We generate Pango lineage-level sequence counts for 18 countries and 53 variants between March 2023 and March 2024. These 18 countries were chosen based on availability of sequence data. Small lineages that do not meet a count threshold are collapsed into their parent lineages. This leaves us with a total of 53 variants, so that each variant met a threshold for number of sequences available.

Using these sequence counts, we apply our latent factor model to estimate the relative fitness of each variant over time in each country, pseudo-escape rates for each variant, and pseudo-immunity for each country simultaneously for $D = 10$ pseudo-immune groups. This model is significantly constrained relative to estimating the time-varying fitness independently in each location, resulting in a model with 2,752 parameters compared to 7,488 parameters in the independent model.

The results of this model are visualized in Fig. 4.5 for several selected variants and countries of interest. Our results show that closely related Pango lineages are often assigned similar pseudo-escape values suggesting that this is capturing some evolutionary structure to immune escape. Further, our model shows that these groups of lineages tend to target particular immune groups such as clade 24A (JN.1, JN.1.1, JN.1.4) has high pseudo-escape in dimensions 3 and 4. If immune escape is the dominant mechanism for relative fitness difference, we expect that differences in immune response between variants from serological data would mirror differences in our pseudo-escape space. Using human serological data from Jian et al [45], we compute titer distances as average log2 differences in titer values between pairs of variants. We compare these distances to distances in our pseudo-escape space (Fig. 4.5G), finding the distances between distinct pairs in the pseudo-escape space are correlated with these titer differences between variants ($R^2 = 0.402$). We bootstrap this analysis among 1,000 replicates to assess significance of this relationship (Fig. C.10, $p < 0.001$). Additionally, we subset by exposure history and find that cohorts with only very recent infection correlate more poorly than WT vaccine cohort or cohorts with more complex exposure histories (Fig. C.11).

We chose $D = 8$ for our primary analysis by noting the point at which the loss function seems to stagnate with increasing D, i.e., the "elbow" method (Fig. C.12A). Further, we observe that Bayesian Information Criterion (BIC) is minimized between 7 and 9 groups (Fig. C.12D). However, the exact choice of latent immune dimensionality is necessarily somewhat arbitrary and we observe significant correlations with empirical titer data for fewer dimensions as well, although $D = 8$ also maximizes this correlation (Fig. C.12B) and

its significance is maintained for all dimensions $D > 8$ tested. Analogous figures showing pseudo immunity and pseudo antigenic relationships across variants can be seen for $D = 2$ in Fig. C.13, $D = 4$ in Fig. C.14, $D = 6$ in Fig. C.15 and $D = 10$ in Fig. C.16.

This approach can be applied to other antigenically variable pathogens, such as influenza, making it broadly applicable beyond SARS-CoV-2. In fact, there is more utility for pathogens with larger geographic differences in immunity since this approach enables to estimate the proportion of these latent immune pools in the population and how they vary geographically and over time alongside variant difference. By approximating antigenic differences using sequence data alone, this method offers for a deeper understanding of immune dynamics and how they shape variant success in the presence of immune escape. This enables an embedding similar to those from antigenic cartography but without the need for serological data and based purely on observed variant fitness.

Conclusions, limitations, and future work

Our study demonstrates the utility of multi-strain mechanistic models in interpreting variant frequency dynamics. This enables a more detailed picture of variant success in environments with heterogeneous population immunity. Our mechanistic grounding of variant fitness allows for investigations into trade-offs between intrinsic transmissibility increase and immune escape, prediction of epidemic dynamics from sequence data alone and inference of antigenic relatedness among variants from differences in success across geographies.

Despite these advances, there are limitations to our approach. Long-term forecasts remain difficult, particularly as new variants with unknown fitness profiles emerge. This framework suggests that considering both the escape against individual immune backgrounds and the diversity in human immune escape is most useful for improving forecasts of relative fitness. Additionally, our models, while powerful in estimating short-term variant dynamics, rely on assumptions about transmission mechanisms that may not always hold across different pathogens or contexts. In fact, as we've shown, it's entirely possible for shifts in population immunity to change the dominant transmission mechanism.

Furthermore, the models considered here are deterministic in nature and do not explicitly model the emergence of variant viruses only the dynamics after their successful introduction. In reality, there are biological constraints on the types of variants that are produced in nature and even if there is a 'true' fitness boost, the chance for stochastic extinction of beneficial variants remains. These constraints present trouble for long-term forecasting as it will require a model of mutation or emergence, tying the potential for a variant to emerge with its potential to transmit in the current environment. Future work should focus on improving the integration of real-time genomic data with serological and epidemiological data, providing a more comprehensive understanding of variant dynamics over time.

In conclusion, our framework represents a significant advance in our understanding of viral evolution and transmission dynamics. By linking variant fitness to specific transmission mechanisms, we provide a more nuanced and accurate prediction of how variants will spread and impact population-level epidemic growth. The selective pressure metric and latent immunity model offer new tools for public health agencies to monitor viral evolution in real time, enabling proactive intervention and insight into the variant difference and wave potential. While our work has been applied to SARS-CoV-2, the methods developed here are broadly applicable to other evolving pathogens, offering a versatile approach for improving epidemic forecasting, variant monitoring, and overall pandemic preparedness.

Acknowledgements

We thank Ivana Bozic, Betz Halloran, Mark Kot and Erick Matsen, as well as members of the Bedford Lab for their feedback on this work. We gratefully acknowledge all data contributors, ie the Authors and their Originating laboratories responsible for obtaining the specimens, and their Submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based. We have included an acknowledgements table in the associated GitHub repository under `data/final_acknowledgements_gisaid.tsv.xz`.

Funding

This work is supported by NIH NIGMS award R35 GM119774 to TB and a Howard Hughes Medical Institute COVID-19 Collaboration Initiative award to TB. MDF is an ARCS Foundation scholar and was supported by the National Science Foundation Graduate Research Fellowship Program under grant No. DGE1762114. TB is a Howard Hughes Medical Institute Investigator.

Author contributions

MF conceived the study. MF, TB gathered sequence and case count data. MF designed and implemented the models. MF performed the analysis. MF, TB interpreted the results. MF, TB wrote the paper.

Competing interests

All authors declare no competing interests.

Data and materials availability

Source code used to generate figures, model implementations, and sequence count data are available at github.com/blab/relative-fitness-mechanisms.

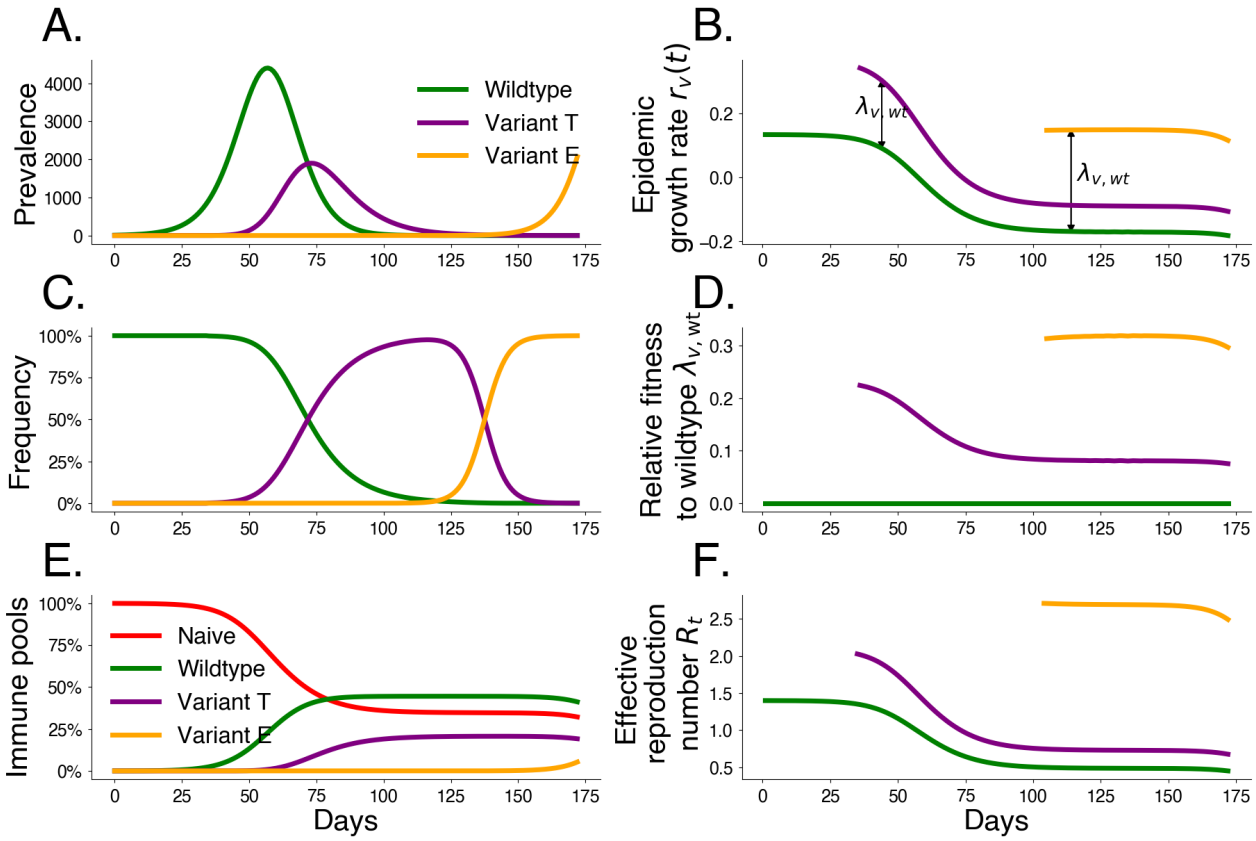

Figure 4.1: **Simulated variant dynamics in a mechanistic model.** Mechanistic transmission models constrain variant frequency dynamics by specifying a functional form for relative fitnesses. Simulations of a three-variant model including wildtype W, an intrinsic transmission variant T, and an immune escape variant E show the relationship between population-level transmission and selection. We begin the simulation with initial wildtype prevalence $I_W(0) = 1$, effective reproduction number $R_{0,W} = 1.4$, and duration of infection $1/\gamma = 3.0$ days. We introduce transmissibility variant T at $t = 20$ with frequency $f_T(20) = 10^{-5}$ and a 50% increase in transmissibility $\rho_T = 0.5$. We introduce escape variant E at $t = 70$ with frequency $f_E(70) = 10^{-6}$ that infects 5% of hosts possessing wildtype immunity $\eta_E = 0.05$. A. Prevalence I by variant. B. Exponential growth rate r by variant. C. Variant frequency f. D. Fitness relative to wildtype λ. E. Underlying immune pools. F. Effective reproduction number R_t by variant.

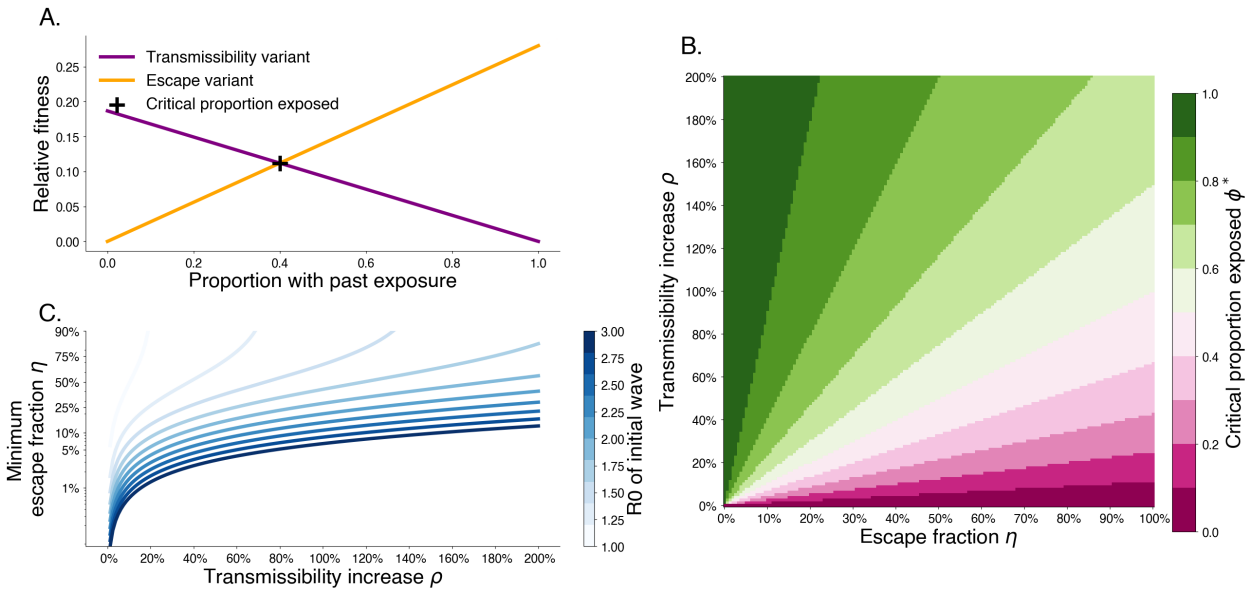

Figure 4.2: **Trade-off between degree of immune escape and increased transmissibility.** A. Relative fitness for a transmissibility increasing variant T with $\rho_T = 0.2$ and an immune escaping variant E with $\eta_E = 0.3$ for $R_{0,W} = 2.8$ and $1/\gamma = 3.0$ days. The intersection point shows that after 40% of the population has wildtype immunity, the escape variant has higher fitness. B. The critical exposure proportion is shown for various escape fraction and transmissibility increase. Above the critical exposure proportion, we expect dominance of escape variants. C. The minimum escape fraction needed for second waves to be comprised of escape variant assuming competition with transmissibility increase variants and first wave with a given R_0.

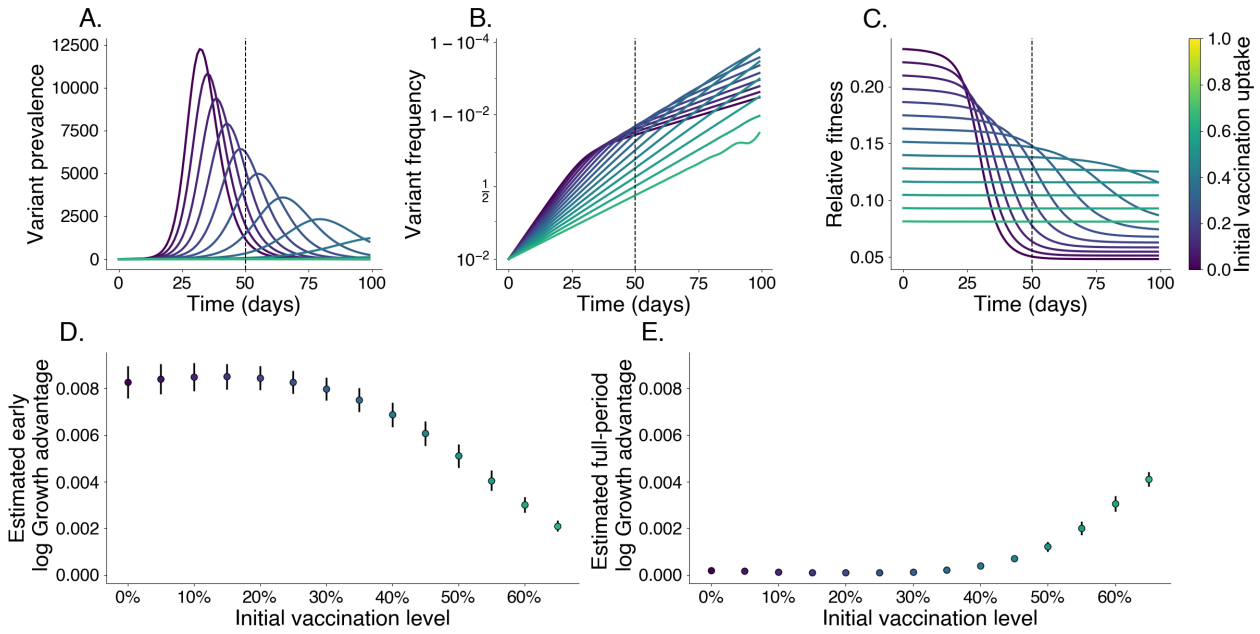

Figure 4.3: **Relative fitness is correlated with vaccination levels in the absence of immune escape.** We simulate the growth of a pure transmissibility increased variant at varying levels of vaccination. Darker colors represent lower vaccine uptake. We identify an early growth period where relative fitness is at its highest; the cutoff for this period is denoted with a vertical dashed line. A. Prevalence of variant, each line is its own simulation. B. Frequency of variant. C. Relative fitness for variant over time. D. Estimated log growth advantage using linear regression of log relative frequency of variant over wildtype using only data before the early cutoff. E. Same as D. but using data from the entire period shown.

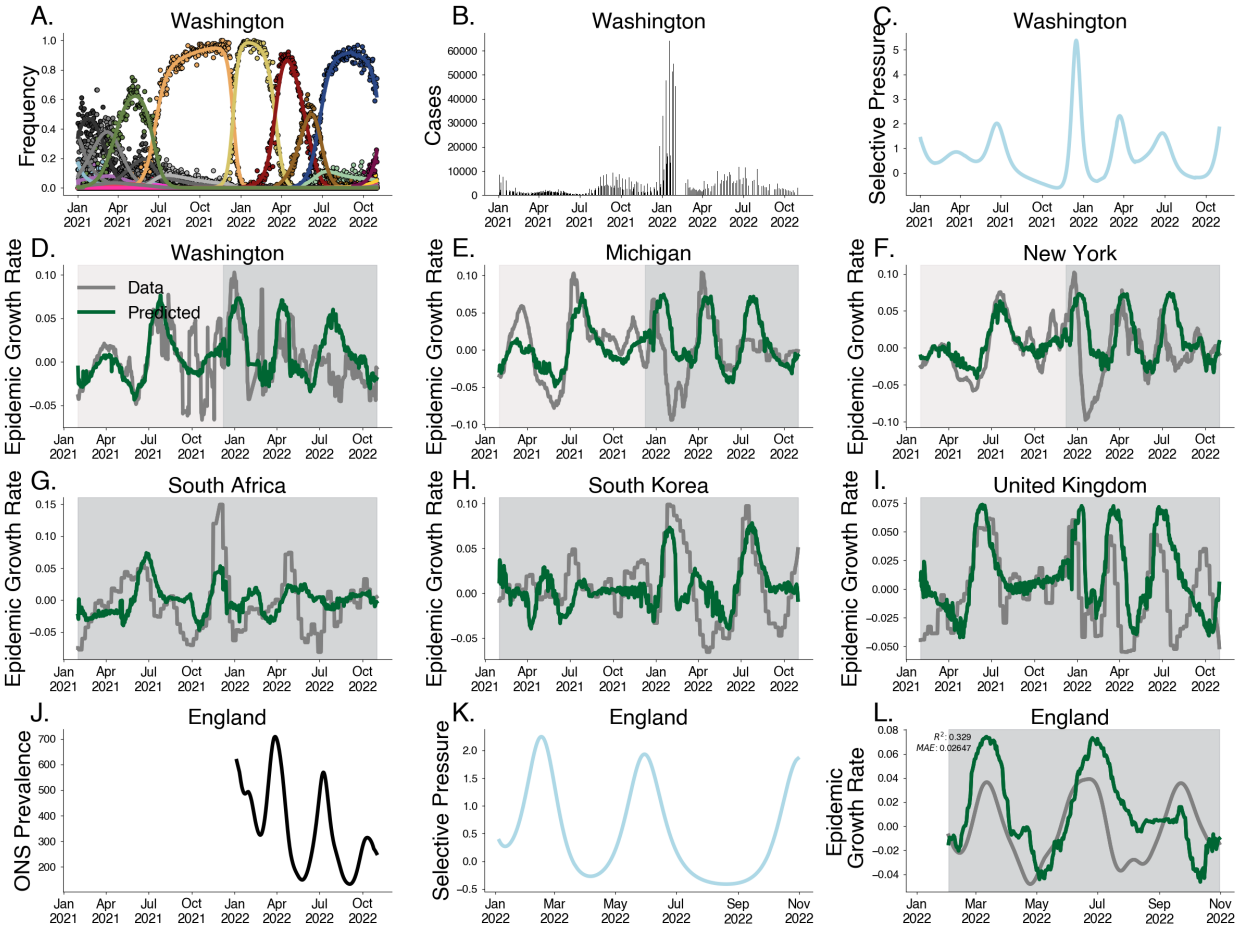

Figure 4.4: **Predicting epidemic growth rate using estimated selective pressure.** A. Variant frequency estimated using the Gaussian process relative fitness model between January 2021 and November 2022 for sequence count data from Washington state. B. Case counts from Washington state. C. Selective pressure computed using estimated variant frequencies and relative fitnesses from Washington state. D-F. Predictions for empirical growth rate from selective pressure for selected US states. The light gray period is the training period and the darker gray is the testing period. G-I. Predictions for empirical growth rate from selective pressure for countries South Africa, South Korea and the UK. J. Prevalence estimates for England from ONS Infection Survey. K. Estimated selective pressure in England. L. Empirical growth rates (gray) computed from prevalence estimates and predictions from our model (green) computed from selective pressure.

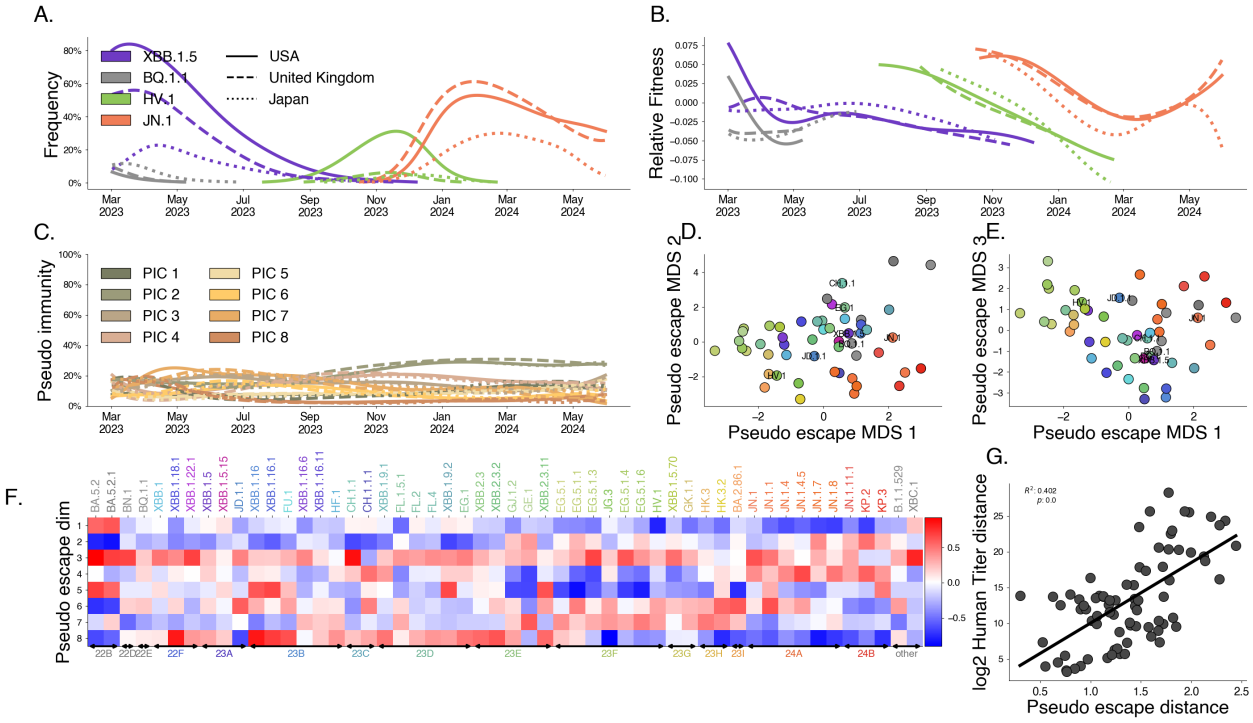

Figure 4.5: **Latent factor models of immunity describe variant dynamics.** We fit the latent immunity factor model to recent SARS-CoV-2 sequence data globally. A. Variant frequency. Lines are colored to show 4 variants of interest (of 53 total variants) with the style of the line denoting 3 countries of interest (of 18 total countries). B. Estimated relative fitness for selected variants and countries. C. Estimated pseudo-immunity cohorts (PIC) over time for multiple countries ordered by decreasing share in the first geography D, E. Dimensionality-reduced pseudo-escape rates using multidimensional scaling (MDS). F. Estimated pseudo-escape rates for each variant relative to pivot variant. H. Comparing pairwise distance between variants in the pseudo-immune space to observed distances in human titer data.

Chapter 5

FORECASTING SARS-COV-2 LINEAGE SUCCESS FROM MOLECULAR DATA

In this chapter, we develop and discuss methods for predicting relative fitness using molecular data. We intend to expand this chapter into a larger paper. This expanded paper will include analyses on how long the predictive power of molecular data lasts as well as more replicates of this idea for evolution post-BA.2 and post-JN.1.

5.1 Abstract

SARS-CoV-2 variants have driven global waves of infection, shifting from transmissibility-driven success in Alpha and Delta to immune escape in Omicron and its descendants. Models estimating relative fitness from variant frequencies provide short-term insights but often fail in long-term forecasts due to dynamic fitness landscapes and neglect of evolutionary history. Simulations reveal that this oversight inflates correlations between molecular phenotypes and fitness, misattributing patterns to ancestry rather than mechanism. To address this, we introduce a Bayesian framework that integrates molecular phenotypes, such as immune escape and receptor binding, with phylogenetic structure. By leveraging innovation-based regression priors, our method accounts for shared ancestry and fitness changes along lineages, enabling robust out-of-sample predictions. This framework advances forecasting of SARS-CoV-2 variant success and offers a generalizable tool for understanding and responding to the evolution of rapidly mutating pathogens.

5.2 Introduction

The COVID-19 pandemic has been characterized by the emergence of SARS-CoV-2 variants that have driven successive waves of infection globally. Early variants like Alpha, Beta, Gamma, and Delta achieved success largely through increases in intrinsic transmissibility. However, the emergence of Omicron marked a shift towards immune escape became the dominant driver of variant success, as demonstrated by molecular studies showing reduced neutralization by vaccine and infection-derived immunity.

Subsequent Omicron-derived lineages, including XBB, EG.5.1, and JN.1, have further exploited immune escape to drive rapid turnover in the viral population. This shift underscores the interplay between population immunity and variant fitness, where increasing exposure selects for variants with immune escape. Understanding the molecular and evolutionary factors that shape the success of these variants is critical for anticipating outbreaks and guiding vaccine design.

Currently, efforts to predict the success of viral variants rely heavily on models that estimate relative fitness from changes in variant frequencies over time. [65, 34]. These models, while valuable for short-term forecasting, often fail to generalize over longer time horizons due by repeated variant emergence or time-dependence in relative fitness [2]. Theory predicts that relative fitness should be a function of both escape against particular immune backgrounds and the distribution of these backgrounds in the population (Chapter 4). This suggests the improving forecasts of relative fitness will require integrating of molecular phenotypes to improve long-term forecasting capabilities.

Experimental methods such as neutralization titer assays and deep mutational scanning (DMS) provide valuable molecular estimates of immune escape. Although they have correlations with population-level success, it remains unclear how these data can be used to predict variant success [24]. Protein language models (PLMs) have recently emerged as tools to infer molecular properties directly from sequences and have even been used to predict relative fitness. However, both PLMs and traditional regression-based approaches share a critical

weakness: they fail to account for the shared evolutionary history of variants. This oversight inflates correlations between molecular phenotypes and fitness, with shared ancestry rather than mechanistic relationships driving the observed patterns,

Using simulations, we show that naive tip-level regressions overestimate the strength of associations between molecular phenotypes and fitness due to recent shared evolutionary history. In contrast, innovation-based approaches better isolate meaningful relationships between molecular phenotypes and fitness. These findings highlight a key limitation in existing methods and suggest a need for evolutionarily-informed approaches.

To address these challenges, we introduce a Bayesian framework that integrates molecular phenotypes, such as immune escape and receptor binding, with phylogenetic structure to estimate and predict relative fitness. By leveraging innovation-based regression priors, this method accounts for shared ancestry and fitness innovations between lineages, enabling out-of-sample predictions of relative fitness.

5.3 Results

The evolutionary success of SARS-CoV-2 lineages is influenced by a combination of molecular phenotype and population immunity. The interplay between these factor creates challenges for accurately attributing changes in lineage frequency to specific phenotypic drivers, such as immune escape. Recent shared evolutionary history additionally introduces confounding effects that can obscure relationships between molecular phenotypes and fitness, particularly when using naive regression approaches. These challenges necessitate a framework capable of disentangling the direct contributions of molecular traits from the effects of evolutionary relatedness.

We provide an overview of the evolutionary and phenotypic context used to quantify and analyze fitness innovations in Fig. 5.1.

The evolutionary relationships among SARS-CoV-2 clades are often visualized by Pango lineage and Nextstrain clade. These variant groupings allow us to easily classify and enumerate genetic variation in the population, creating nested structure in Pango lineages and

enabling us to visualize their variant-parent relationships. These variant parent relationships reflect Nextstrain clades (Fig. 5.1A). However, despite falling into the same clade, lineages may vary in their molecular phenotypes such receptor-binding domain (RBD) immune escape (Fig. 5.1B).

This phenotypic variation among lineages coexists with the temporal dynamics of selection in SARS-CoV-2. Tracking the frequencies of Nextstrain clades from January to November 2023 shows the rapid turnover of clades driven by shifts in fitness among lineages (Fig. 5.1C).

Naive regression methods show that these molecular phenotypes poorly explain relative fitness of variants ($R^2 = 0.03$), suggesting that molecular phenotypes may not be useful for predicting lineage success (Fig. 5.1D). However, accounting for evolutionary relationships using variant-parent innovation (Fig. 5.1E) in relative fitness and phenotype reveal clearer signal ($R^2 = 0.69$).

By focusing on the changes along branches, we isolate the impact of molecular phenotypes from correlations due to shared evolutionary history. These fitness innovations can be used to identify phenotypic drivers of relative fitness and to enable forecasting of relative fitness for yet unseen variants. Together, these results provide insights into the mechanisms underlying lineage success and advance the ability to forecast the emergence of novel variants.

Shared evolutionary history generates spurious correlations In naive regression methods, cumulative molecular phenotypes, such as immune escape, are directly correlated with measures of fitness. This naive approach often fails to account for the nested structure and shared evolutionary of lineages which can generate spurious correlations between fitness and phenotype in closely related lineages.

Simulating phylogenetic trees under selection, we show that fitness increases over time as more fit lineages reproduce in the population (Fig. 5.2A). Though we simulate this tree assuming selection is independent of individual mutations or their number, we find that the total number of mutations explains much of the variance in relative fitness ($R^2 = 0.76$), as

shown in Fig. 5.2B. This spurious correlation is due to shared evolutionary history creating correlations between individual lineages in the population despite branch level innovations being independent. In fact, we can see that the magnitude of the R^2 increases with the variation in fitness between branches, i.e., the strength of selection in the population (Fig. D.1A). This suggests that stronger selection amplifies the confounding effect of shared ancestry.

To address this challenge, we focus on fitness innovations, i.e., branch-specific changes in relative fitness. Using innovations or changes in fitness and mutations across branches within our regressions, we isolate the contributions of molecular phenotypes to fitness, effectively removing the confounding effects of shared ancestry (Fig. 5.2C). This relationship between fitness innovations and mutation change is preserved across varying levels of selection strength (Fig. D.1B).

We show that this approach is similar to other approaches of dealing with confounding due to shared ancestry such as phylogenetic generalized least squares in Supplementary Text D.1.

Using innovations enables us to better resolve drivers of fitness that are obscured by recent evolutionary history. By isolating the contributions of molecular phenotypes to fitness, this provides a robust foundation for identifying the specific drivers of relative fitness and improving predictions of lineage success.

Estimating fitness innovations across lineages and clades To better understand the variability in fitness changes across SARS-CoV-2 lineages, we apply the innovation idea to an inference model nested within a multinomial logistic regression framework.

This innovation model estimates each lineages' fitness relative to XBB.1.5 and its innovation using the lineage-parent mapping in Fig. 5.1A.

For this analysis, we use a Normal distribution as a prior for the relative fitness innovations. We fit this model to XBB.1.5-focused dataset spanning January to November 2023, reflecting the evolutionary dynamics during this period and visualize estimated clade frequencies in Fig. 5.1C, showing turnover in this period. These clade frequencies are generated by summing the frequency over all Pango lineages within a clade.

In Fig. 5.3C, we see that there is clear selection for clades like 23F (EG.5.1) and 24A (JN.1). Despite this, we see that the overall distribution of innovations is near 0, though it has a wide range with a slight positive skew (Fig. 5.3B). This reflects the effect of selection in driving advantageous lineages to higher frequencies, but suggests that both adaptive events and neutral evolutionary processes shaped SARS-CoV-2 evolution in this period.

When aggregated by Nextstrain clade, the fitness innovations present a granular view of the evolutionary trends in SARS-CoV-2 (Fig. 5.3C). Generally, clades which appear before 23A (XBB.1.5) such as 22B (BA.5) and 22D (BA.2.75) have a median innovation near 0 which is consistent with neutral evolution. However, clades following 23A (XBB.1.5.) tend to have positive fitness innovations over average, aligning with their observed growth and dominance during this period.

This analysis shows that fitness innovations can be used to understand patterns of adaptation within a population or clade. The variability observed across lineages and clades underscores the role of rare but impactful events in driving lineage success and driving outbreaks.

Quantifying the relationship between molecular phenotypes and fitness innovations To quantify the relationship between molecular phenotypes and fitness innovations, we integrated deep mutational scanning data to predict lineage phenotypes.

For this analysis, we use "human sera escape relative to XBB.1.5" and "ACE2 binding relative to XBB.1.5" [24] and "RBD ACE2 affinity relative to XBB.1.5", "RBD expression relative to XBB.1.5", "RBD escape relative to XBB.1.5" [78] as phenotypes.

These phenotypes were calculated for each lineage as the sum of all mutation effects relative to XBB.1.5 and lineage-parent differences in these phenotypes were computed to serve as predictors in the innovation-based analyses.

Using the relative fitness innovations estimated previously (Fig. 5.4A), we use the changes in these phenotypes across parent-child pairs to predict relative fitness with linear regression. Together, these phenotypes explain a large share of the variance in relative fitness

($R^2 = 0.79$), showing the utility of these phenotypes for predicting changes in relative fitness (Fig. 5.4B). To understand the contribution of each of these phenotypes to this regression, we estimated the partial R^2 of each phenotype (Fig. 5.4C). This is a form of the additional variance explained by adding this phenotype to the model. Using this partial R^2 approach shows that RBD yeast-display escape relative to XBB.1.5 explains most of the variance ($R^2_{\text{partial}} = 0.097$). We repeat this analysis using each predictor individually to predict the relative fitness directly (Fig. D.2) and the relative fitness innovations (Fig. D.3).

This analysis shows that molecular phenotypes are able to explain much of the variance of fitness innovations within sample. Among the phenotypes, RBD yeast-display escape relative to XBB.1.5 contributes the largest unique share to the variance. This result emphasizes the role of immune escape and its molecular basis in shaping the success of SARS-CoV-2 lineages.

This suggests these molecular phenotypes may have utility for forecasting the relative fitness of previously unseen SARS-CoV-2 lineages.

Using molecular phenotypes for out-of-sample relative fitness forecasts To assess the predictive utility of molecular phenotypes, we implemented a regression prior for the relative fitness innovation. This approach enables the prediction of fitness innovations for out-of-sample lineages by leveraging their molecular phenotypes and the lineage-parent relationships described previously. With these predicted fitness innovations, we can then forecast relative fitness using the baseline fitness of its parent (Fig. 5.5C).

We validate that the regression prior fitness model estimates relative fitnesses that are consistent with the relative fitness in the uninformed model (Fig. 5.5A), finding near perfect alignment between the two models ($R^2 = 1.0$). This suggests that our regression prior fitness model is able to capture both the effect of predictors and estimate the residual between our predictions.

We then sought to validate this model using its predictions for previously unseen variants. We repeat the same relative fitness analysis, estimating relative fitness using the normal prior model for the 6 months following the end of the original XBB.1.5 data set. Using all

lineages for which we have available phenotypes but were not included in the original data set, we predict their fitness and compare it to the estimated relative fitness in the future data set (Fig. 5.5B). Our forecasted relative fitness align closely with the future relative fitness though they differ in scale, and we find that our predictions explain a large share of the variance ($R^2 = 0.67$) in future relative fitness and are strongly positive correlated (Spearman's $\rho = 0.72$). This result highlights the ability of the regression-based fitness model to generalize fitness predictions beyond the training set, effectively utilizing molecular phenotypes as predictors.

The ability to predict fitness for previously unseen lineages provides a critical tool for evolutionary forecasting. By combining molecular phenotypes with the innovation framework, this approach offers a scalable solution for anticipating the success of novel variants in real-time.

5.4 Discussion

This study addresses the challenge of predicting the success of unseen variants in rapidly evolving populations. By integrating molecular phenotypes with relative fitness, we isolate branch-specific fitness innovations while addressing confounding due to shared ancestry. Our results suggest that molecular phenotypes, particularly immune escape, are strong predictors of lineage success. Using a regression-based prior, we extend this approach to enable out-of-sample forecasting of relative fitness, providing a scalable, mechanism-informed framework for understanding and predicting the evolutionary success of emerging variants.

Immune escape emerges as the most significant phenotypic driver of relative fitness, aligning with prior studies suggesting weaker neutralization of emerging variants in hosts with past exposure. Molecular phenotypes such as ACE2 binding affinity and RBD expression, also contribute though to a lesser extent. These results emphasize that while immune escape is a primary driver, transmissibility-related traits also play a role in shaping lineage success of SARS-CoV-2.

The ability to predict relative fitness for unseen variants has significant public health im-

plications. Linking molecular phenotypes to relative fitness enables longer-term forecasts of lineage success, which can inform vaccine strain updates by prioritizing high-fitness lineages for monitoring or use as candidate strains before they reach dominance.

Despite its strengths, this approach has several limitations. The focus on genomic data from the United States may restrict the generalizability to regions with different immune landscapes. For pathogens like influenza which have greater regional diversity, this consideration becomes extremely important. Additionally, the assumption that molecular phenotypes independently contribute to fitness innovations may oversimplify the interactions between phenotypic traits, which may motivate more complex, data-informed prior models.

Our framework also relies on the quality and relevance of deep mutational scanning data, which may not fully describe immune escape in all populations due to host-specific differences. While molecular phenotypes derived from deep mutational scanning are powerful, they may not fully capture complex interactions, such as epistasis. These limitations also suggest future directions for expanding this framework.

Future work can refine this approach by incorporating larger datasets spanning diverse regions and lineages beyond XBB.1.5, enabling evaluation of its generalizability across varying immune landscapes. Testing additional molecular phenotypes may improve prediction accuracy and reveal novel drivers of fitness.

This study establishes a framework for forecasting relative fitness for SARS-CoV-2 using molecular phenotypes. By showing the utility of these phenotypes for predicting fitness and enabling out-of-sample forecasts, our approach provides a practical tool for real-time monitoring of emerging variants and informing vaccine strain selection. While focused on SARS-CoV-2, the framework can be adapted to study other rapidly evolving pathogens, extending its relevance beyond the current pandemic. These contributions provide a foundation for advancing long-term evolutionary forecasting and guiding proactive responses to pathogen evolution.

5.5 Methods

Simulation of transmission trees with fitness-weighted offspring To investigate how evolutionary history influences correlations between molecular phenotypes and fitness, we simulated transmission trees using a discrete-generation model. Each tree begins with a single ancestral lineage at generation 0, initialized with fitness λ_0. At each generation t, the total number of offspring O_t is sampled from a Poisson distribution with mean $\overline{N}_t$. These offspring are then distributed among active lineages i are allocated using a multinomial draw based on fitness-weighted probabilities, so that the offspring by lineage are given by

$$O_{i,t} \sim \text{Multinomial}\left(O_t, p_i\right), \tag{5.1}$$

$$p_i = \frac{\exp(\lambda_i)}{\sum_j \exp(\lambda_j)}, \tag{5.2}$$

where λ_i is the fitness of lineage i. We assume that the fitness of each offspring evolves from its parent's fitness according to a Brownian motion process

$$\lambda_{\text{offspring}} = \lambda_{\text{parent}} + \psi, \quad \psi \sim \text{Normal}(0, \sigma^2), \tag{5.3}$$

where σ determines the variance in fitness evolution. Mutations accumulate along branches with their counts drawn from a Poisson distribution with rate μ.

The simulation is run over a fixed number of generations, producing trees where fitness influences both branching structure and phenotype evolution. These trees are then analyzed to compare naive tip-level regressions with branch-level contrast regressions, providing insight into how shared evolutionary history drives spurious correlations and demonstrating the importance of evolutionary-aware statistical approaches.

Generating sequence counts We prepared sequence count data sets using the Nextstrain-curated SARS-CoV-2 sequence metadata [40] which is created using the GISAID EpiCoV database [48]. These sequences were counted according to their annotated Pango

lineage [4], country of collection, and date of collection to produce sequence counts for each variant, date, and country analyzed.

Multinomial logistic growth To estimate relative rates of growth between variants of interest, we fit a multinomial logistic growth model to the sequence count data. This model can be written as

$$f_{t,v} = \frac{\exp(\alpha_v + \lambda_v t)}{\sum_{u=1}^{V} \exp(\alpha_u + \lambda_u t)}, \tag{5.4}$$

where $\alpha_V = \lambda_V = 0$.

We can then interpret λ_v as the relative fitness of variant v relative to variant V. Assuming a fixed generation time τ additionally allows us to then write the relative R_t or growth advantage for variant v over V as $\Delta_v = \exp(\lambda_v \tau)$. To fit this model, we use a multinomial likelihood with probabilities defined by the frequencies above, the vector of counts for each variant at time t S_t, and the total count of sequences at time t,

$$S_t \sim \text{Multinomial}(N_t, f_{t,\cdot}). \tag{5.5}$$

Mapping variant-parent relationships In order to estimate rough branch-specific relative fitness innovations, we develop a data set mapping variants as parent-child pairs. For each Pango lineage in our generated sequence counts, we iterate to its parent lineage until the parent lineage is either in the sequence count file or there is no parent present. If there is no parent present, we simply estimate the variant's relative fitness and growth advantage directly.

Relative fitness innovation model We can extend the previous model for the Multinomial Logistic growth to take into account for evolutionary relationships using the variant-parent lineage mapping generated in the previous section.

This updated model is instead parameterized by the relative fitness difference between variant v and its parent lineages $\psi_v = \lambda_v - \lambda_{\text{parent}_v}$ directly. Under this parameterization, we

can also define the (approximate) growth advantage innovations Ψ_v from a variant lineage v to its parent as

$$\Psi_v = \frac{\Delta_v}{\Delta_{\text{parent}_v}} = \exp(\psi_v \tau).$$

Normal prior model We consider several prior models on the relative fitness innovation ψ_v. The most basic of these models is a normal prior on the relative fitness innovations

$$\psi_v = (\lambda_v - \lambda_{\text{parent}_v}) \sim \text{Normal}(0, \sigma).$$

This induces a log normal prior on the growth advantage innovations Ψ_v.

Regression prior for growth advantage innovations To better explain the variation in relative fitness innovations and enable prediction of relative fitness for new variants given features and their parent lineage, we extend the above prior to be parameterized by various features, so that

$$\psi_v = (\lambda_v - \lambda_{\text{parent}_v}) \sim \text{Normal}\left(\sum_p \theta_p x_p, \sigma\right). \tag{5.6}$$

Using this model, we can then predict relative fitnesses and growth advantages out of sample. For a given out-of-sample variant v, we need to only specify its set of predictors x_p and its in-sample parent parent_v. First, we predict its innovation relative to its parent and its total relative fitness by sampling using Equation 5.6 with the estimated θ_p. We can compute the relative fitness of this variant by adding the predicted innovation to the estimated relative fitness of its parent. This provides an out-of-sample estimate of the relative fitness using a combination of molecular data and our knowledge of the evolutionary history of the population.

Generating features for regression prior analysis We use mutation effects estimated using pseudovirus deep mutational scanning and receptor binding domain (RBD) yeast-display deep mutational scanning to predict phenotypes at the level of Pango lineage [24, 78]. The phenotype of each lineage is computed as the sum of all mutation effects relative

to XBB.1.5. This gives us 5 molecular phenotypes: "spike pseudovirus DMS human sera escape relative to XBB.1.5", "spike pseudovirus DMS ACE2 binding relative to XBB.1.5", "RBD yeast-display DMS ACE2 affinity relative to XBB.1.5", "RBD yeast-display DMS RBD expression relative to XBB.1.", "RBD yeast-display DMS escape relative to XBB.1.5".

We also compute the variant-parent differences in the phenotypes for use as predictors in the innovation-based analyses. These variant-parent differences are computed using the variant-parent relationships described above for variants with an existing parent.

Analysing SARS-CoV-2 evolution post-XBB.1.5 emergence. To analyze selection in SARS-CoV-2 after the emergence of XBB.1.5, we begin by generating Pango lineage-level sequence counts for the United States between January 1st, 2023 and December 1st, 2023. We collapse small lineages which do not meet a count threshold into their parent lineages, forcibly including XBB, XBB.1.5, and XBB.2. Using these collapsed counts, we generate the variant-parent relationships. We, then, estimate the relative fitness of each variant and its innovation using these sequence counts and the innovation model with a normal prior on the innovations.

Using these estimated innovations, we fit a standard linear regression between the relative fitness innovation and the molecular phenotype innovations. To evaluate the contribution of adding individual phenotypes to the relative fitness innovations, we conducted a partial R^2 analysis. We compute the partial R^2 for each phenotype as the difference between the R^2 of the full model and the R^2 of the model without the phenotype of interest. This approach quantifies the proportion of variance in the dependent variable (relative fitness innovation) explained by each phenotype when adding it to a model that already uses the other phenotypes.

We also estimate the relative fitness of each variant and its innovation to its parent using the regression prior on the innovations with the 5 molecular phenotypes as predictors. This enables us to make predictors of the phenotype for out of sample lineages as long as we know their parent and their phenotypes.

Acknowledgements

We gratefully acknowledge all data contributors, i.e., the Authors and their Originating laboratories responsible for obtaining the specimens, and their Submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based.

Funding

This work is supported by NIH NIGMS award R35 GM119774 to TB and a Howard Hughes Medical Institute COVID-19 Collaboration Initiative award to TB. MF is an ARCS Foundation scholar and was supported by the National Science Foundation Graduate Research Fellowship Program under grant No. DGE1762114. TB is a Howard Hughes Medical Institute Investigator.

Competing interests

All authors declare no competing interests.

Data and materials availability

Molecular phenotype data was provisioned from https://github.com/jbloomlab/SARS2-spike-predictor-phenos on November 24, 2024. Sequence data including date and location of collection as well as clade annotation was obtained via the Nextstrain-curated data set that pulls data from GISAID database. A full list of sequences analyzed with accession numbers, derived data of sequence counts and molecular phenotypes, along with all source code used to analyze this data and produce figures is available via the GitHub repository github.com/blab/ncov-escape.

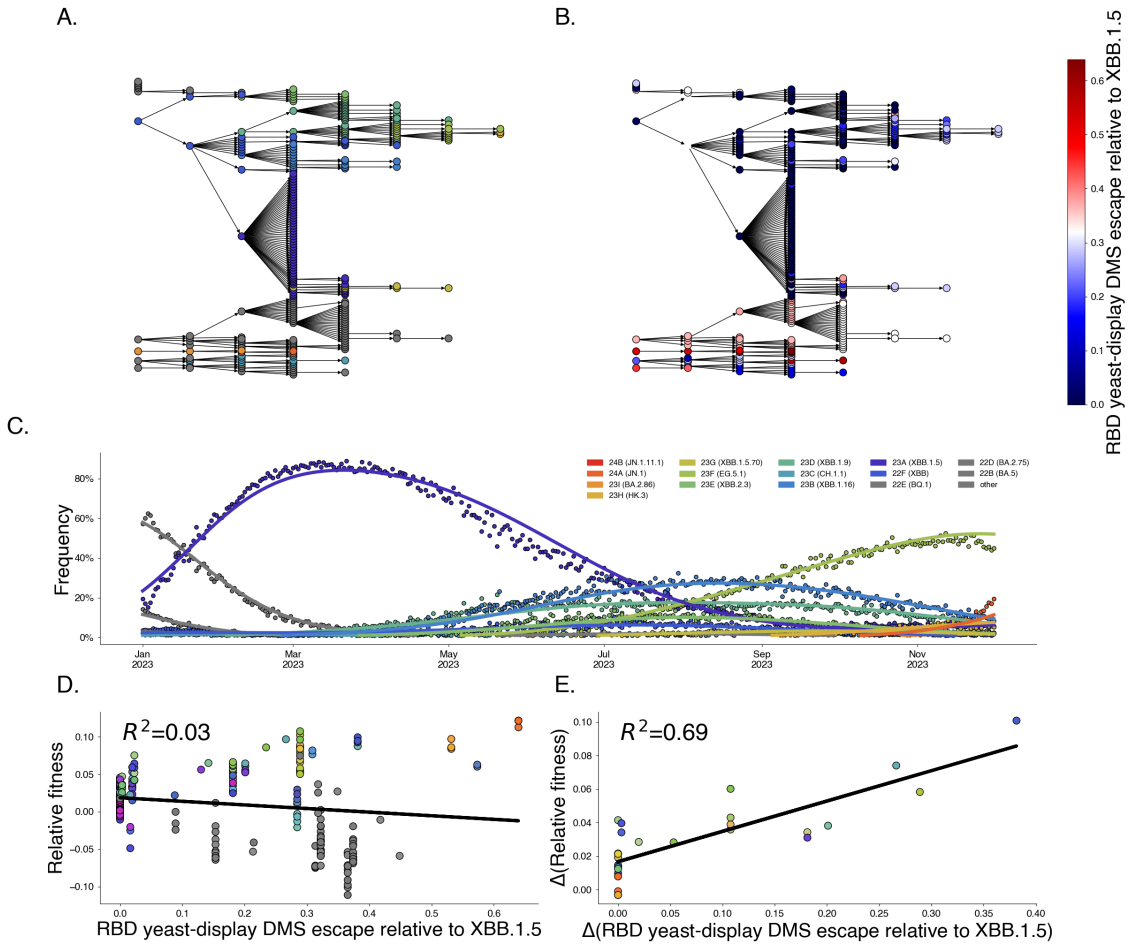

Figure 5.1: **Interplay between evolution, immune escape, and fitness among SARS-CoV-2 variants** A. Lineage and closest parent mapping of SARS-CoV-2 Pango lineages, with nodes colored by Nextsrain clade, illustrating evolutionary relationships and grouping of lineages. B. The mapping as A., now colored by receptor-binding domain (RBD) immune escape, highlighting molecular phenotypes across evolutionary lineages. C. Temporal dynamics of clade frequencies in the global SARS-CoV-2 population from January to November 2023. The figure highlights the rapid turnover of clades driven by changing fitness landscapes. D. Naive correlation between RBD immune escape (x-axis) and relative fitness (y-axis) shows weak association ($R^2 = 0.03$), reflecting the confounding effects of shared ancestry. E. Innovation-based correlation between immune escape contrasts and fitness innovations shows a significantly stronger relationship ($R^2 = 0.69$), demonstrating the importance of accounting for evolutionary context to uncover true mechanistic drivers of fitness.

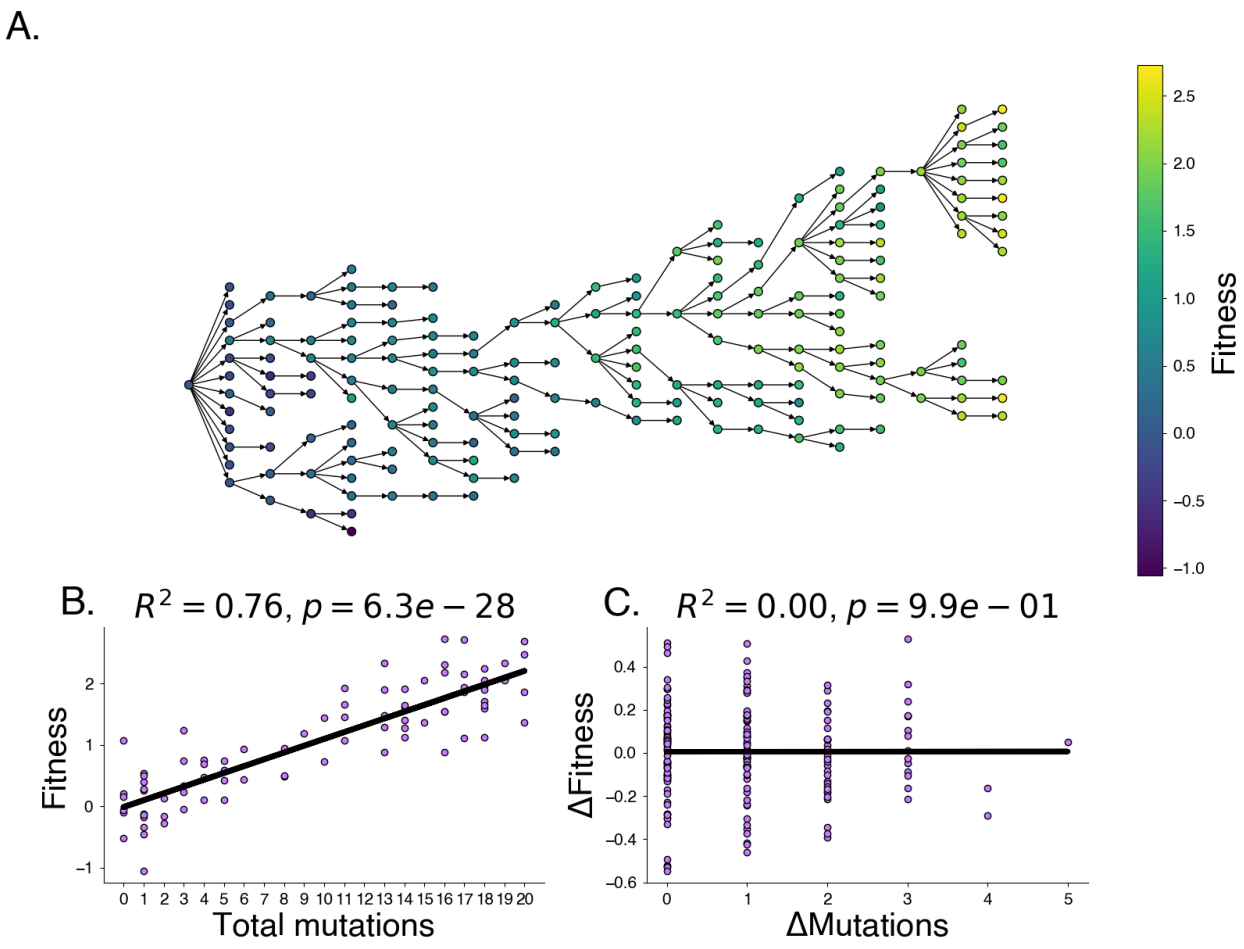

Figure 5.2: **Shared evolutionary history causes spurious correlations between predictors and fitness.** A. A transmission tree simulated with fitness-dependent branching. Each branch represents the evolutionary descent of lineages, where fitness evolves through Brownian motion. B. Naive regression of tip-level fitness on cumulative mutations. This ignores the shared evolutionary history of tips, leading to spurious correlations. C. Regression of branch-level innovations (changes in fitness versus changes in mutation counts). By analyzing innovations, we account for the shared evolutionary history and isolate the relationship between predictors and fitness, revealing the true lack of direct association.

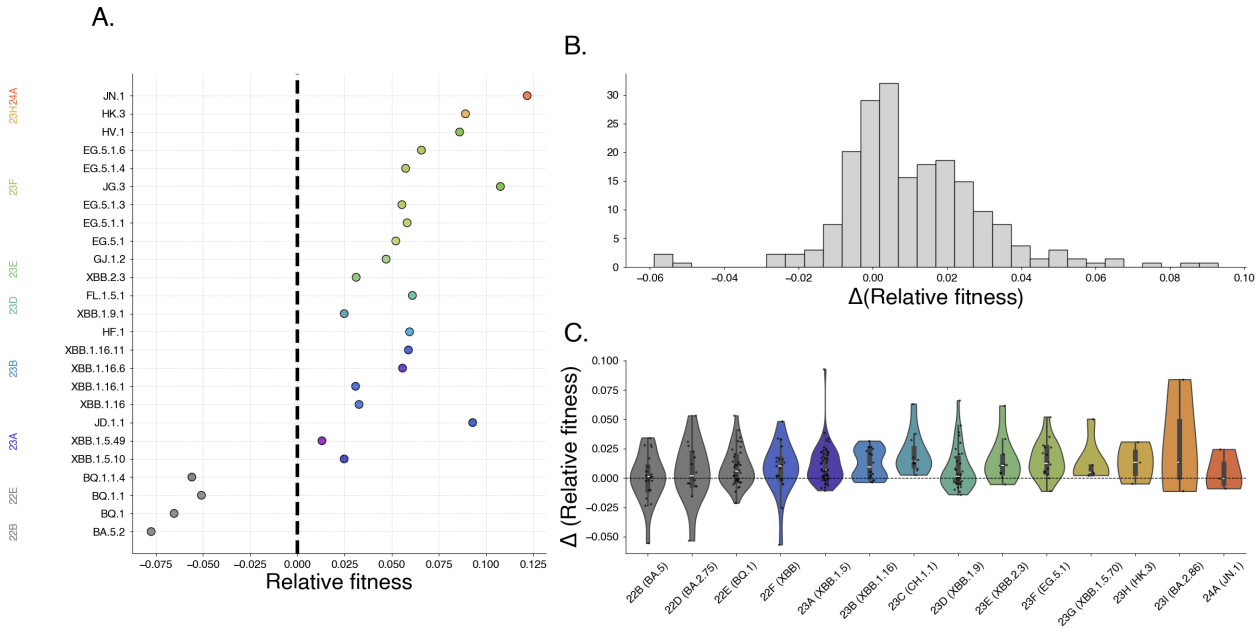

Figure 5.3: **Exploring fitness innovations across SARS-CoV-2 clades.** A. Estimated relative fitness for Pango lineages which reach at least 1% frequency. B. Histogram of overall fitness innovations across all variants. C. Estimated fitness innovations by clade.

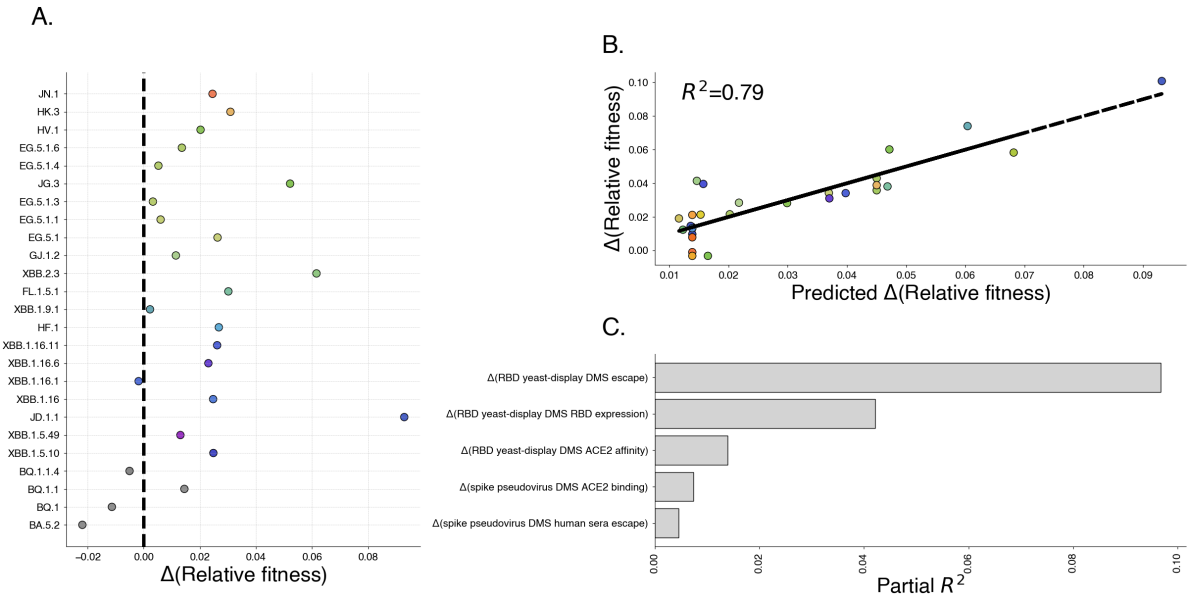

Figure 5.4: **Molecular phenotypes explain fitness innovations.** A. Relative fitness innovations for Pango lineages which reach at least 1% frequency. B. Observed versus predicted relative fitness innovations from the regression model ($R^2 = 0.72$), demonstrating the model's predictive accuracy. C. Partial R^2 values for molecular predictors, illustrating the relative contributions of features such as immune escape, receptor binding, and expression to fitness innovations.

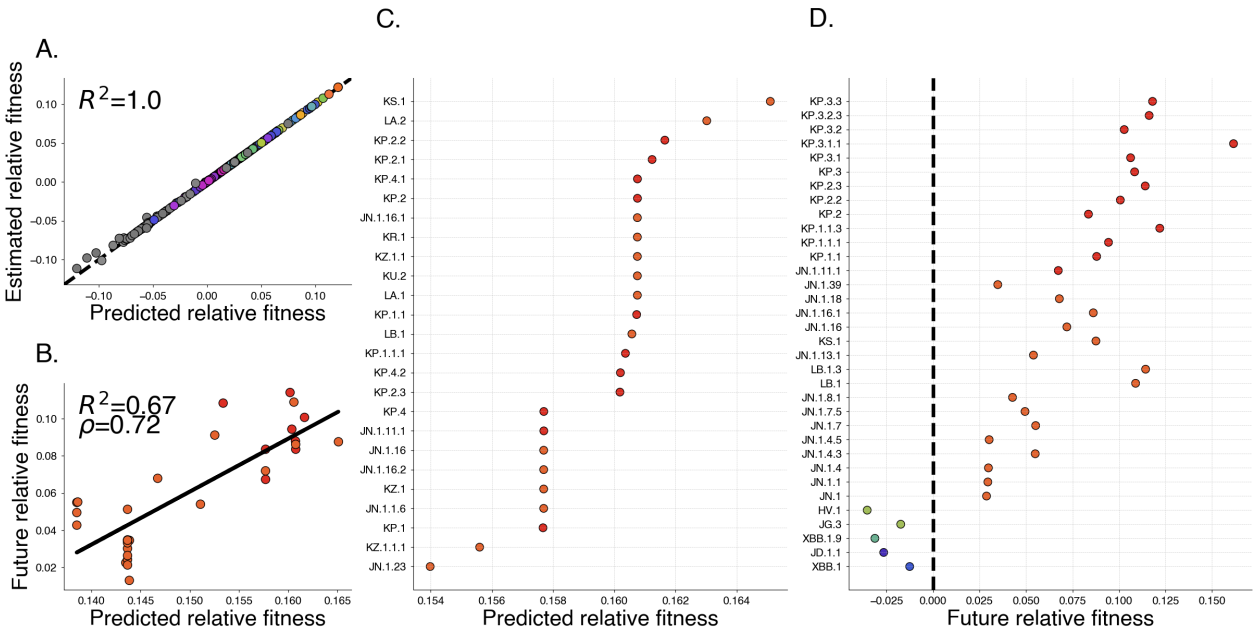

Figure 5.5: **Predicting fitness with innovations.** A. In-sample validation comparing predicted relative fitness innovations with their estimated values. Predictions were generated using the regression-based prior, incorporating molecular phenotypes as predictors. Points represent individual lineages, with the diagonal dashed line indicating perfect agreement between predicted and estimated values. B. Out-of-sample validation comparing predicted relative fitness innovations with observed values for lineages not included in the training dataset. C. Predicted relative fitness of top 25 lineages. D. Relative fitness lineages achieving at least 2% frequency, relative fitnesses are estimated with data from the 6 months following training period.

Chapter 6

OPERATIONALIZING EVOLUTIONARY FORECASTS: `evofr` AND `forecasts-ncov`

In this chapter, we develop and discuss two software tools `evofr` and `forecasts-ncov` that simplify the development, application, and communication of variant fitness dynamics and evolutionary forecasts.

6.1 Introduction

This dissertation emphasizes integrating theoretical and practical approaches and blending mechanistic insights with statistical models. While earlier chapters focused on developing, evaluating, and applying specific models, these methods were used in specific analyses motivated by existing scientific problems. However, there is a common structure in the methods described that enables us to simplify the workflow of using these kinds of models. In this way, we move from scientific analyses to reproducible, scalable, and applied analyses that can tackle broad questions in evolutionary forecasting. Having infrastructure for building, applying, and interpreting models allows us to iterate on our analyses, speeding up our responses and addressing real-world challenges.

`evofr` consolidates and expands on the methods developed in this dissertation, enabling reproducible data analysis within a modular structure that allows rapid testing and comparison between models. This package serves as the computational backbone for the analyses presented in this dissertation, integrating models from statistical tracking to mechanistic forecasting and is the bridge between the theoretical insights developed in this dissertation and their practical utility as an open-source tool.

`forecasts-ncov` incorporates `evofr` into an automated workflow and public-facing plat-

form, translating research and applied forecasts into actionable insights for public consumption. It continually retrieves publicly available sequence data to generate and visualize forecasts of SARS-CoV-2 variant frequencies for public consumption, providing forward-looking approach to the evolution of SARS-CoV-2 as a companion to Nextstrain. [40] In this way, `forecasts-ncov` serves as a first step to operationalized evolutionary forecasting in practice and at scale.

Together, these software tools provide a basis for evolutionary forecasts as a dynamic process and concrete problem, setting the stage for continued development of this practice and scientifically-informed forecasts of pathogen evolution.

6.2 `evofr`*: A toolkit for evolutionary forecasting*

Evolutionary forecasting plays a critical role in understanding genetic changes in populations over time, with direct applications in predicting the prevalence of infectious disease variants, guiding vaccine development, and informing public health interventions.

Existing tools for understanding evolution often operate at the level of phylogenetic tree estimation which is computationally expensive due to the large number of possible trees (which grows super-exponentially in terms of the number of samples). Further, these methods have shown difficulty scaling to large numbers of sequences, requiring approximation-based methods at the pandemic scale. [26] Additionally, due to these large computational demands, it can be difficult to iterate on these methods, making model development, testing, and iteration time-consuming. In general, these methods are typically employed for historical analyses, capturing past or currents trends in evolution without much concern for likely future population change. That being said, there is no need to completely replace phylogenetic analysis as it still forms an important backbone for our understanding of pathogen evolution and enables us to group continuous genetic diversity in pathogen populations to variant groupings that are evolutionarily meaningful.

This suggests a need to supplement phylogenetic analysis with fast, scalable, reproducible methods for analyzing sequence data at a coarse scale that can be employed for both historical

and forward-looking analyses. To address these gaps, we have developed `evofr`.

`evofr` is a Python package built for evolutionary forecasting of genetic variants, addressing the growing need for robust tools to analyze and predict change in genetic variation in real time. With applications in evolutionary biology, epidemiology, and public health, `evofr` integrates data pre-processing, modeling techniques, intuitive visualization tools, and a modular framework to empower researchers and scientists to understand and anticipate the dynamics of genetic variation. It seeks to simplify bespoke analyses and allow easy integration in standardized workflows, enabling reproducible and scalable analyses.

The package facilitates data preprocessing, evolutionary forecasting, and result visualization, catering to complex datasets like those generated by genomic surveillance of pathogens (e.g., SARS-CoV-2 and influenza). By leveraging modularity, `evofr` ensures scalability and extensibility, enabling users to adapt the tool for varied biological and computational challenges. The key contributions of this package include:

- **Forecasting dynamics of genetic variants**: `evofr` allows users to estimate the relative fitness and future prevalence of genetic variants using customizable modeling approaches such multinomial logistic regression (MLR), renewal-equation based models, Gaussian processes, and latent factor models among others.

- **Extensibility and Accessibility:** `evofr` provides a user-friendly, open-source framework that integrates seamlessly with Python's ecosystem, encouraging contributions and customization.

- **Modularity and Reproducibility** `evofr` promotes reproducible science by offering a modular architecture that supports continued methods development, rapid iteration, model comparison, and simple integration into forecasting workflows.

6.2.1 Design and implementation

The design and implementation of `evofr` reflect its dual purpose as a research product and public health tool. This section will describe the software's architecture and key features, highlighting its modular design, integration of diverse modeling approaches, interchangable inference methods.

Modularity `evofr` is divided into distinct modules for data preprocessing (`evofr.data`), modeling (`evofr.models`), inference (`evofr.infer`), and visualization (`evofr.plotting`). Each module performs specific tasks and is designed to be interoperable. As an example, this enables easily swapping between different models such as standard MLR, a Gaussian process-based model on the same pre-processed data. These models can then be fit with any of the inference methods in `evofr.infer` such as Markov Chain Monte Carlo (MCMC), Stochastic Variational Inference (SVI), and maximum a posteriori (MAP) estimation. The results of these models are stored in a standardized `evofr.posterior` object, which is compatible with the various plotting methods stored in `evofr.plotting`.

Reproducibility Fundamentally, `evofr` seeks to make the kind of bespoke analyses that are common in papers about SARS-CoV-2 and pathogen evolution reproducible. It contains well-documented reproductions of several methods employed in research papers, to improve the reach of these methods and enable easy replication of analyses with these papers.

While inference for these models typically involve stochastic processes, `evofr` includes mechanisms to fix random seeds, ensuring that results are deterministic and consistent across runs.

Additionally, all intermediates and outputs including preprocessed data, model configurations and posterior samples can be stored in standardized formats like JSON and CSV. This ensures that results can be reliably reproduced and shared.

Extensibility Users can extend `evofr` by adding new models, likelihoods, or priors to the models module by developing a class that inherits from `evofr.ModelSpec`. The package's backend-agnostic design ensures compatibility with diverse computational frameworks (e.g., JAX, PyMC, NumPyro) since users can simply provide a log-posterior probability function for their model to enable samplng or optimization via `evofr`'s inference methods. [12, 3, 64]

The plotting module (`evofr.plotting`) is built on top of matplotlib, but can be customized or replaced to accommodate specific visualization needs, such as alternative plot styles or integration with external tools. [43]

Generally, the open-source nature of `evofr` encourages contributions, enabling the tool to evolve with user-driven innovations.

Scalability Due to most models working at the level of variant frequency, `evofr` is designed to process genomic datasets containing thousands of sequences, with methods typically scaling with the number of variants and time-horizon for analysis. Efficient use of computational resources and data reduction strategies ensures that analyses remain feasible as data volumes grow. Further, integration with JAX enables just-in-time compilation and GPU/TPU acceleration for computationally intensive tasks.

6.2.2 Discussion

`evofr` is a Python-based toolkit designed to address challenges in evolutionary forecasting. Its modular structure and emphasis on flexible workflows enable researchers to analyze and predict genetic variant dynamics. `evofr` is well-suited for genomic datasets generated by real-time surveillance efforts, where understanding variant dynamics is essential for public health decision-making and evolutionary monitoring.

Through its modular design, `evofr` supports rapid iteration on models and analyses, offering tools for data preprocessing, modeling, inference, and visualization. By enabling reproducible workflows and fostering extensibility, `evofr` has the potential to serve a broad range of applications, from fundamental research to forecasting in practice.

The core contributions of `evofr` include the wide range models implemented within the package, which provide a robust foundation for exploring genetic variant dynamics under both statistical and semi-mechanistic frameworks. By supporting customizable likelihoods and priors, `evofr` also allows users to tailor models to specific research questions or datasets.

The package is designed to handle large genomic datasets efficiently, ensuring that `evofr` can be applied to datasets with hundreds of thousands of sequences and extended time horizons at the national and global scale, making it ideal for real-time evolutionary forecasting.

Additionally, `evofr` enables reproducible workflows with logging model configurations, and storing results in common formats. These features ensure that analyses can be reliably reproduced, shared, and extended.

While `evofr` provides many significant advantages, there are some points of limitation as is. Bayesian inference methods such as MCMC can demand significant computational resources, especially when dealing with large data sets and complex models. Though there are alternative, faster but approximate methods for inference such as stochastic variational inference and maximum a posteriori estimation available within the package, this still remains an issue since all inference methods may not be computational feasible for a given data set or model.

In general, the accuracy of these methods depends heavily on the quality and completeness of the data set of interest. This means that unforeseen biases or gaps in the data can affect the reliability of these models and their predictions. However, `evofr`'s extensibility will allow motivated researchers to iterate on these models to improve their ability to address such data issues.

This level of extensibility presents a potentially steep learning curve. Users with limited experience in probabilistic modeling may face challenges using and extended this package as is, but the documentation should provide a helpful guide for contributions looking to dive deeper.

In the future, we hope to expand the capabilities of `evofr` and address the aforementioned limitations. Further development should continue to design and implement hybrid models

which combine mechanistic insight and statistical flexibility, expanding the range of questions and forms of data that `evofr` can address and handle. This includes additional likelihood functions and priors for extending existing models as well.

There is also a need to support contributions to `evofr` from the open-source community. This includes adding more thorough documentation of the package as well as the development of extended guide for contributors that is accessible to users. Going forward, the goal should be expanding the package to meet the needs of the communities who would most benefit from it: researchers and scientists in epidemiology, evolutionary biology, and public health.

`evofr` democratizes evolutionary forecasting by turning statistical methods developed for research to a practical, scalable, and accessible tools for analysis. This positions `evofr` as a useful for real-time applications in public, such as monitoring emerging variants or informing vaccine development.

Beyond its immediate applications, `evofr` serves as a model for the development of modular, extensible tools in computational biology. Its design and implementation demonstrate the utility of combining reproducibility, scalability, and user-centered design to address complex challenges in research. By bridging the gap between theoretical models and practical applications, `evofr` sets the stage for continued advancements in evolutionary forecasting and beyond.

Code availability

`evofr` is an open source project. Its source code can be found at https://github.com/blab/evofr. Its documentation can be found at https://blab.github.io/evofr.

6.3 `forecasts-ncov`*: Automated and public-facing forecasts for SARS-CoV-2*

`forecasts-ncov` is an automated platform designed to provide real-time forecasts of SARS-CoV-2 variant frequencies. As part of the broader Nextstrain project, it extends Nextstrain's genomic surveillance capabilities by offering a forward-looking perspective. [40] While

Nextstrain's phylogenetic inference provides detailed reconstructions of evolutionary histories, `forecasts-ncov` complements this by projecting future variant trajectories. This dual approach ensures a comprehensive understanding of variant dynamics, bridging retrospective insights with predictive capabilities.

This tool operationalizes evolutionary forecasting, translating genomic data into actionable insights for public health. Its automated workflows, robust modeling framework, and accessible visualization tools make it an essential resource for addressing rapidly evolving public health challenges. By enabling real-time forecasting, `forecasts-ncov` can support early detection of emerging threats, timely policy decisions, and informed vaccine updates.

6.3.1 Design and workflow

The `forecasts-ncov` pipeline integrates data ingestion, predictive modeling, and visualization into a seamless, automated workflow. It is designed to operate at scale, updating forecasts daily to provide consistent and timely insights.

Data ingestion is handled by workflows that retrieve and preprocess genomic data from public repositories, such as GISAID and GenBank, using the Nextstrain `ncov-ingest` pipeline. These workflows ensure high-quality, curated inputs by harmonizing metadata and sequence information across multiple sources. This generates sequence count files at various levels of genetic granularity (Nextstrain clade and Pango lineage), geographic resolution (Global and the United States) from data.

Forecasting is performed using predictive models implemented in `evofr`Ṫhese models estimate key parameters such as variant frequencies, growth rates, and fitness advantages, generating forward-looking insights that complement traditional phylogenetic analyses. By focusing on coarse-scale trends, the platform prioritizes actionable predictions over detailed evolutionary reconstructions. Though compatible with any model from `evofr` live forecasts rely on an in-house hierarchical multinomial logistic regression model.

The visualization component transforms model outputs into interactive, web-based visualizations that present variant trajectories in accessible formats. This web application

allows users to explore the latest forecasts or upload local data for analysis. These visualizations provide intuitive representations of complex genomic data, making `forecasts-ncov` a valuable resource for technical and non-technical audiences alike.

6.3.2 Contributions

As a component of Nextstrain, `forecasts-ncov` fills a critical gap by adding predictive capabilities to the platform's established phylogenetic tools. It provides continuously updated forecasts that enable users to anticipate the trajectories of SARS-CoV-2 variants (Figs 6.1 and 6.2) as well as visualize the current growth advantages of variants at the level of Nextstrain clade and Pango lineage (Figs 6.3 and Figs 6.4).

This forward-looking approach complements the retrospective focus of phylogenetic inference, creating a comprehensive framework for understanding both past and future variant dynamics.

By automating the forecasting process, `forecasts-ncov` ensures the timely availability of results, which are stored publicly as JSON files on AWS S3. These outputs are stratified by data provenance (e.g., GISAID, GenBank), variant classification (e.g., Nextstrain clades, Pango lineages), and geographic resolution (e.g., global, USA). The integration of these outputs with an interactive web application further enhances their accessibility, enabling users to visualize variant trends dynamically. This functionality supports public health efforts by providing actionable insights that inform decision-making at local, national, and global levels.

6.3.3 Discussion

`forecasts-ncov` represents a significant advancement in operationalizing evolutionary forecasting, combining the predictive modeling framework of `evofr` with Nextstrain's robust infrastructure for genomic surveillance. By automating the end-to-end process—from data ingestion to real-time forecast visualization—the platform offers a forward-looking complement to the phylogenetic analyses that form the foundation of Nextstrain. This dual ap-

proach provides a more comprehensive understanding of variant dynamics, addressing both retrospective and prospective dimensions of pathogen evolution.

The platform's success lies in its ability to integrate predictive models seamlessly into a scalable pipeline that produces actionable outputs. These forecasts have proven particularly valuable for tracking SARS-CoV-2 variants, enabling public health agencies to anticipate potential surges and prioritize interventions. By presenting outputs through intuitive visualizations and accessible data formats, `forecasts-ncov` democratizes access to evolutionary insights, fostering collaboration among researchers, public health professionals, and policymakers.

However, real-time forecasting brings unique challenges that merit consideration. The reliance on external data sources, such as GISAID and GenBank, introduces dependencies that can affect the timeliness and consistency of updates. Additionally, the variability in global sequencing efforts can create gaps or biases in the underlying data, which in turn may influence forecast accuracy. While the platform's modular design supports customization and iteration to address such issues, these constraints highlight the ongoing need for robust data-sharing practices and equitable genomic surveillance.

Another area of consideration is the interpretability of forecasts. While visualizations simplify complex data, conveying the uncertainties inherent in predictive models remains challenging, particularly for non-technical audiences. Enhancing these aspects would strengthen `forecasts-ncov` as a tool for decision-making, ensuring that forecasts are not only accessible but also actionable.

Looking ahead, the scalability and extensibility of `forecasts-ncov` provide a strong foundation for future growth. The platform is well-positioned to expand its scope to other pathogens, offering similar predictive capabilities for influenza, RSV, and other emerging threats. Further development of its visualization tools could enable deeper engagement with forecast data, while continued integration with Nextstrain's ecosystem would create an increasingly unified platform for genomic epidemiology.

Ultimately, `forecasts-ncov` exemplifies the potential of evolutionary forecasting to move

beyond research applications and into the realm of real-world impact. By providing timely and actionable insights, the platform underscores the value of combining computational innovation with public health priorities, setting the stage for further advancements in genomic surveillance and predictive modeling.

Code availability

`forecasts-ncov` is an open source project. Its source code can be found at https://github.com/nextstrain/forecasts-ncov. The live SARS-CoV-2 forecasts can be found at https://nextstrain.org/sars-cov-2/forecasts.

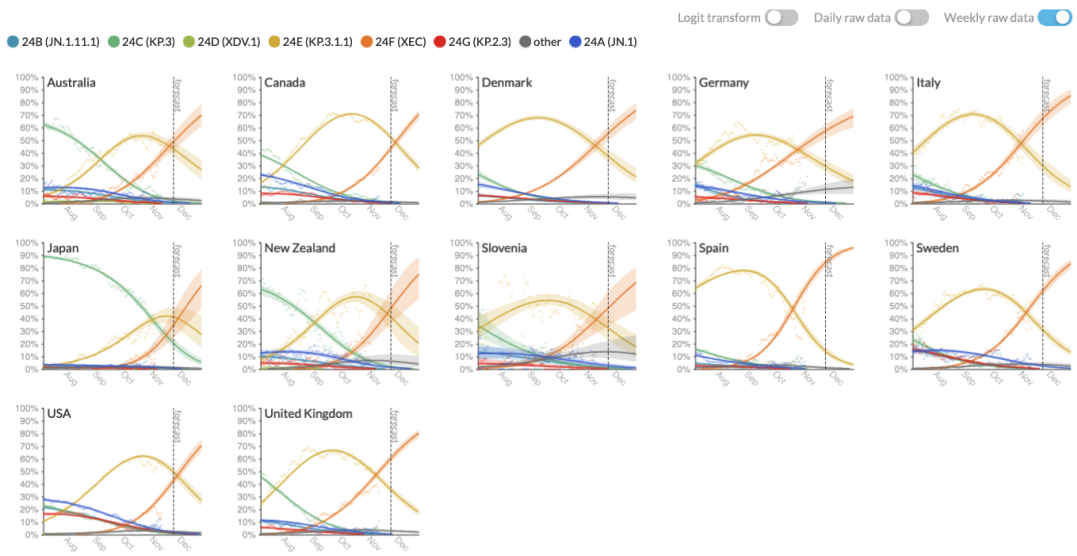

Figure 6.1: **Live global SARS-CoV-2 frequency forecasts for Nextstrain clades.**

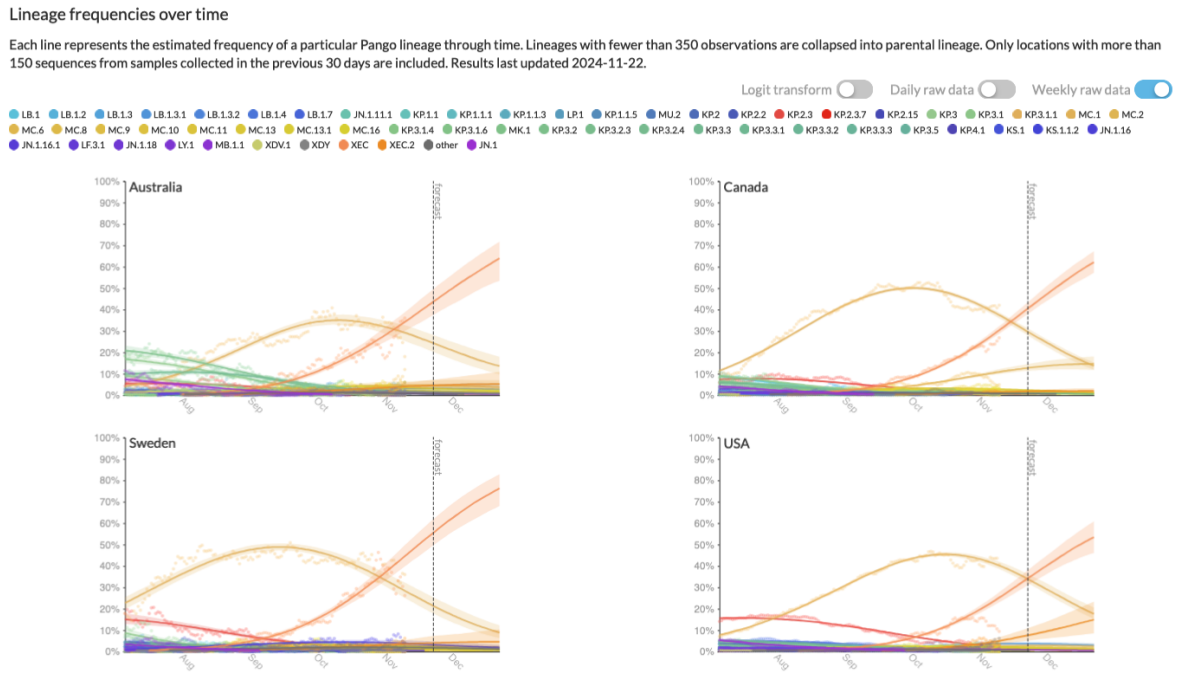

Figure 6.2: **Live global SARS-CoV-2 frequency forecasts for Pango lineages.**

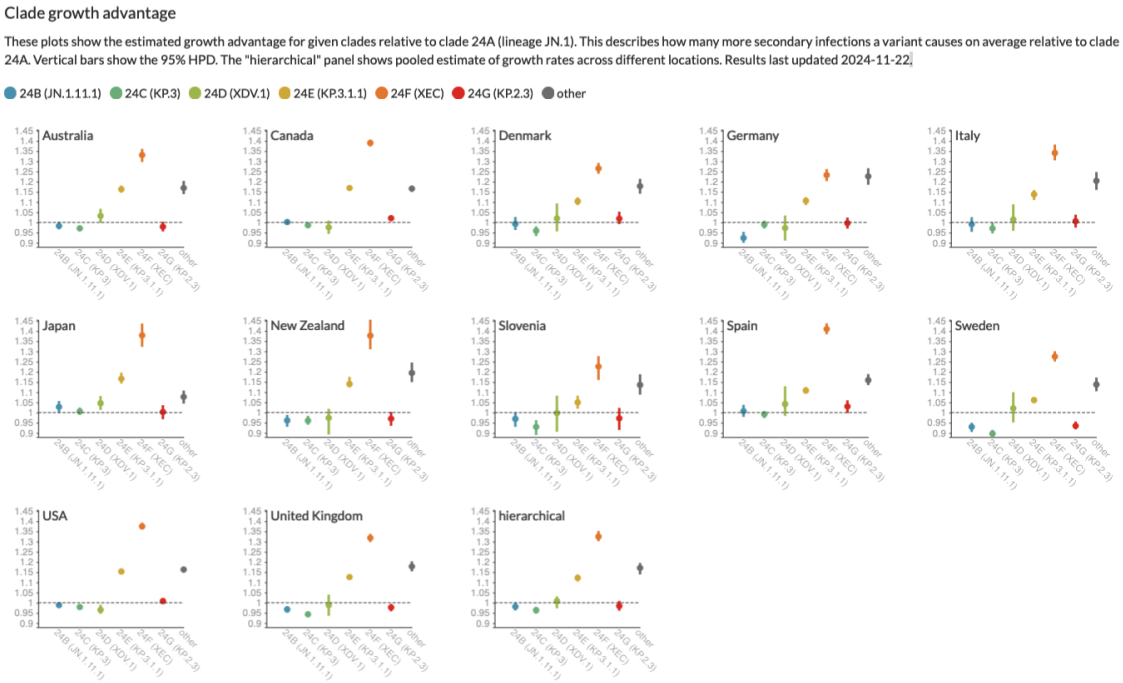

Figure 6.3: **Live global SARS-CoV-2 growth advantage estimates for Nextstrain clades.**

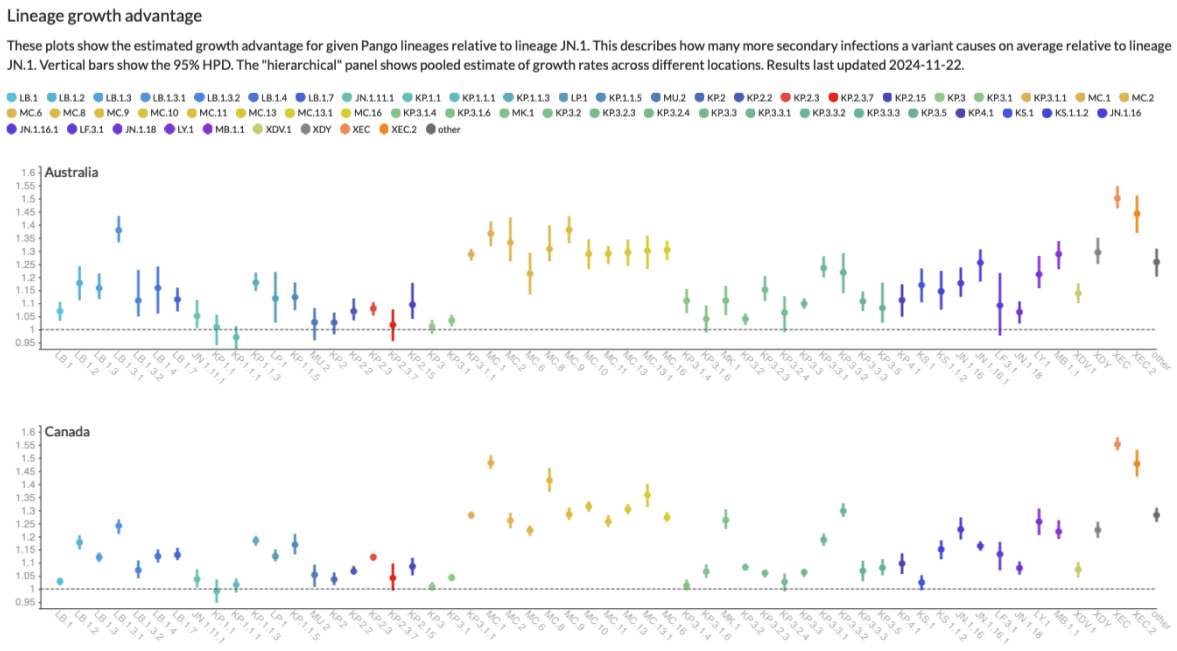

Figure 6.4: **Live global SARS-CoV-2 growth advantage estimates for Pango lineages.**

Chapter 7

CONCLUSION

This dissertation establishes evolutionary forecasting as a rigorous and dynamic process by integrating insights from population genetics, mathematical epidemiology, and statistics. Through systematic evaluation, this work identifies gaps in existing models and lays a foundation for addressing them by integrating mechanistic insight with immunological, genetic, and epidemiological data. Forecasting moves beyond tracking trends to become a dynamic process that unites real-time analysis, biological insight, and long-term evolutionary prediction. These advances rely on domain knowledge, statistical methodology, and data on the biological and environmental forces shaping pathogen evolution. They enable new kinds of predictions, such as variant-specific R_t, pseudo-immune representations, and epidemic growth rates derived from selective pressure, enhancing our ability to anticipate and respond to pathogen evolution.

This work presents several key contributions— mechanism-informed fitness models, forecasting evaluation methods, and data-informed fitness predictions— that deepen our understanding of pathogen evolution. These contributions are operationalized through tools like `evofr` and `forecasts-ncov`. `evofr` serves as a flexible toolkit for real-time analysis and forecasting while `forecasts-ncov` automates these methods to forecast in real-time, transforming methodological detail into accessible, reproducible insights for public health decision-makers. Together, these contributions interpret today's dynamics while predicting the evolution of tomorrow, turning theoretical advancements into modular, scalable systems capable of addressing diverse epidemiological and evolutionary challenges.

By emphasizing modularity and scalability, this work ensures its contributions extend beyond SARS-CoV-2 to other pathogens such as influenza or emerging zoonotic threats. This

adaptability, combined with interdisciplinary insights from genomics, immunology, epidemiology, computational biology, and statistics, strengthens our approach by ensuring it is both biologically meaningful and statistically robust. These contributions provide the scientific and practical tools scientists and public health systems need to understand, interpret, and act on the dynamics of emerging genetic variants.

This dissertation advances evolutionary forecasting by integrating mechanistic insights and empirical data into models and tools that predict the dynamics of viral evolution. While significant progress has been made, the field presents opportunities for further development, particularly in capturing the complexity of viral evolution and ensuring the practical utility of forecasts.

A critical avenue for future research is expanding the scope of data incorporated into forecasting models. While this work demonstrates the power of genomic surveillance, viral evolution is influenced by behavioral, clinical, and environmental factors that are not yet fully integrated. Incorporating these dimensions would enhance the fitness-based models developed here, offering a richer understanding of the factors shaping variant success and public health outcomes.

The methodologies presented in this dissertation, though tailored to SARS-CoV-2, suggest broader applicability to other rapidly evolving pathogens such as influenza and RSV. Generalizing these models will require adapting them to the specific evolutionary and immunological dynamics of other viruses. This extension would establish evolutionary forecasting as a universal framework for monitoring and predicting pathogen evolution.

At the same time, developing generative sequence models offers a promising direction. Such models could simulate evolutionary trajectories, predicting not only which variants may emerge but also their phenotypic traits while basing generation in our mechanistic understanding of pathogen evolution. These capabilities would build on the fitness and immune escape dynamics studied here, enabling long-term forecasting and preemptive public health strategies.

Further refinement of population structure in these models is also essential. Selective

pressures and fitness vary across populations due to differences in immunity, geography, and temporal dynamics. Building on the latent factor models developed in this dissertation, future work can better account for these variations, improving the precision of forecasts across diverse contexts.

On a practical level, this dissertation demonstrates how tools like `evofr` and `forecasts-ncov` can operationalize evolutionary forecasting. To maximize their impact, future efforts should focus on making these tools more accessible to public health practitioners through user-friendly interfaces and seamless integration into existing workflows. Collaboration between evolutionary biologists, immunologists, virologists, and public health experts will be critical for validating these models and ensuring their relevance in real-world decision-making.

Finally, as forecasting tools become increasingly influential in guiding vaccine updates and public health policies, their practical and policy consequences require careful consideration. Future work should explore frameworks to ensure forecasts are used transparently and equitably, addressing issues such as resource allocation and public communication.

The next phase of evolutionary forecasting lies in refining its scientific foundations while expanding its practical applications. By integrating diverse data sources, adapting models to new pathogens, and addressing the practical impact of these forecasts, the field can continue to evolve, meeting the needs of both science and society.

In my opinion, this work offers a clear lesson beyond advancing forecasting or any single pathogen. It has shown that pursuing a deeper understanding of dynamic systems and the structures that drive them creates new opportunities for innovation. In that understanding lies the potential to anticipate, to prepare, and to act.

BIBLIOGRAPHY

[1] Sam Abbott, Joel Hellewell, Robin N. Thompson, Katharine Sherratt, Hamish P. Gibbs, Nikos I. Bosse, et al. Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts. *Wellcome Open Res.*, 5:112, 2020.

[2] Eslam Abousamra, Marlin Figgins, and Trevor Bedford. Fitness models provide accurate short-term forecasts of SARS-CoV-2 variant frequency. *PLoS Comput. Biol.*, 20:e1012443, 2024.

[3] Oriol Abril-Pla, Virgile Andreani, Colin Carroll, Larry Dong, Christopher J. Fonnesbeck, Maxim Kochurov, et al. Pymc: a modern and comprehensive probabilistic programming framework in python. *PeerJ Comput. Sci.*, 9:e1516, 2023.

[4] Ivan Aksamentov, Cornelius Roemer, Emma B Hodcroft, and Richard A Neher. Nextclade: clade assignment, mutation calling and quality control for viral genomes. *J. Open Source Softw.*, 6:3773, 2021.

[5] Medini K. Annavajhala, Hiroshi Mohri, Pengfei Wang, Manoj Nair, Jason E. Zucker, Zizhang Sheng, et al. Emergence and expansion of SARS-CoV-2 b.1.526 after identification in New York. *Nature*, 597:703–708, 2021.

[6] Jantien A Backer, Dirk Eggink, Stijn P Andeweg, Irene K Veldhuijzen, Noortje van Maarseveen, Klaas Vermaas, et al. Shorter serial intervals in SARS-CoV-2 cases with Omicron BA.1 variant compared with Delta variant, the Netherlands, 13 to 26 december 2021. *Eurosurveillance*, 27, 2022.

[7] Trevor Bedford, Andrew Rambaut, and Mercedes Pascual. Canalization of the evolutionary trajectory of the human influenza virus. *BMC Biol.*, 10:38, 2012.

[8] Trevor Bedford, Marc A Suchard, Philippe Lemey, Gytis Dudas, Victoria Gregory, Alan J Hay, et al. Integrating influenza antigenic dynamics with molecular evolution. *eLife*, 3:e01914, 2014.

[9] Meriem Bekliz, Manel Essaidi-Laziosi, Kenneth Adea, Krisztina Hosszu-Fellous, Catia Alvarez, Mathilde Bellon, et al. Immune escape and replicative capacity of Omicron lineages BA.1, BA.2, BA.5.1, BQ.1, XBB.1.5, EG.5.1 and JN.1.1. *bioRxiv*, 2024.02.14.579654, 2024.

[10] R. Bellman and T. E. Harris. On the theory of age-dependent stochastic branching processes. *Proc. Natl. Acad. Sci. USA*, 34:601–604, 1948.

[11] Jesse D Bloom and Richard A Neher. Fitness effects of mutations to SARS-CoV-2 proteins. *Virus Evol.*, 9, 2023.

[12] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, et al. JAX: composable transformations of Python+NumPy programs. 2018.

[13] Anderson F Brito, Elizaveta Semenova, Gytis Dudas, Gabriel W Hassler, Chaney C Kalinich, Moritz UG Kraemer, et al. Global disparities in SARS-CoV-2 genomic surveillance. *Nat. Commun.*, 13:7003, 2022.

[14] Finlay Campbell, Brett Archer, Henry Laurenson-Schafer, Yuka Jinnai, Franck Konings, Neale Batra, et al. Increased transmissibility and global spread of SARS-CoV-2 variants of concern as at June 2021. *Eurosurveillance*, 26, 2021.

[15] Yunlong Cao, Fanchong Jian, Jing Wang, Yuanling Yu, Weiliang Song, Ayijiang Yisimayi, et al. Imprinted SARS-CoV-2 humoral immunity induces convergent Omicron RBD evolution. *Nature*, 614:521–529, 2022.

[16] Yunlong Cao, Jing Wang, Fanchong Jian, Tianhe Xiao, Weiliang Song, Ayijiang Yisimayi, et al. Omicron escapes the majority of existing SARS-CoV-2 neutralizing antibodies. *Nature*, 602:657–663, 2021.

[17] Yunlong Cao, Ayijiang Yisimayi, Fanchong Jian, Weiliang Song, Tianhe Xiao, Lei Wang, et al. BA.2.12.1, BA.4 and BA.5 escape antibodies elicited by Omicron infection. *Nature*, 608:593–602, 2022.

[18] Alessandro M. Carabelli, Thomas P. Peacock, Lucy G. Thorne, William T. Harvey, Joseph Hughes, Sharon J. Peacock, et al. SARS-CoV-2 variant biology: immune escape, transmission, and fitness. *Nat. Rev. Microbiol.*, 21:162–177, 2023.

[19] Alessandro M Carabelli, Thomas P Peacock, Lucy G Thorne, William T Harvey, Joseph Hughes, Sharon J Peacock, et al. SARS-CoV-2 variant biology: immune escape, transmission and fitness. *Nat. Rev. Microbiol.*, 21:162–177, 2023.

[20] David Champredon, Jonathan Dushoff, and David J. D. Earn. Equivalence of the Erlang-distributed SEIR epidemic model and the renewal equation. *SIAM J. Appl. Math.*, 78:3258–3278, 2018.

[21] Cheng Cheng, DongDong Zhang, Dejian Dang, Juan Geng, Peiyu Zhu, Mingzhu Yuan, et al. The incubation period of COVID-19: a global meta-analysis of 53 studies and a chinese observation study of 11 545 patients. *Infect. Dis. Poverty*, 10, 2021.

[22] Anne Cori, Neil M. Ferguson, Christophe Fraser, and Simon Cauchemez. A new framework and software to estimate time-varying reproduction numbers during epidemics. *Am. J. Epidemiol.*, 178:1505–1512, 2013.

[23] Bernadeta Dadonaite, Jack Brown, Teagan E McMahon, Ariana G Farrell, Daniel Asarnow, Cameron Stewart, et al. Full-spike deep mutational scanning helps predict the evolutionary success of SARS-CoV-2 clades. *bioRxiv*, pages 2023–11, 2023.

[24] Bernadeta Dadonaite, Jack Brown, Teagan E. McMahon, Ariana G. Farrell, Marlin D. Figgins, Daniel Asarnow, et al. Spike deep mutational scanning helps predict success of SARS-CoV-2 clades. *Nature*, 631:617–626, 2024.

[25] Nicholas G Davies, Sam Abbott, Rosanna C Barnard, Christopher I Jarvis, Adam J Kucharski, James D Munday, et al. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science*, 372:eabg3055, 2021.

[26] Nicola De Maio, Prabhav Kalaghatgi, Yatish Turakhia, Russell Corbett-Detig, Bui Quang Minh, and Nick Goldman. Maximum likelihood pandemic-scale phylogenetics. *Nat. Genet.*, 55:746–752, 2023.

[27] Rebecca Earnest, Rockib Uddin, Nicholas Matluk, Nicholas Renzette, Sarah E. Turbett, Katherine J. Siddle, et al. Comparative transmissibility of SARS-CoV-2 variants Delta and alpha in new england, usa. *Cell Rep. Med.*, 3:100583, 2022.

[28] Rachel T. Eguia, Katharine H. D. Crawford, Terry Stevens-Ayers, Laurel Kelnhofer-Millevolte, Alexander L. Greninger, Janet A. Englund, Michael J. Boeckh, and Jesse D. Bloom. A human coronavirus evolves antigenically to escape antibody immunity. *PLoS Pathog.*, 17:e1009453, 2021.

[29] Warren J Ewens. The fundamental theorem of natural selection: the end of a story. *Evolution*, 78:803–808, 2024.

[30] W.J. Ewens. An interpretation and proof of the fundamental theorem of natural selection. *Theor. Popul. Biol.*, 36:167–180, 1989.

[31] Nuno R Faria, Thomas A Mellan, Charles Whittaker, Ingra M Claro, Darlan da S Candido, Swapnil Mishra, et al. Genomics and epidemiology of the P. 1 SARS-CoV-2 lineage in Manaus, Brazil. *Science*, 372:815–821, 2021.

[32] James R. Faulkner and Vladimir N. Minin. Locally adaptive smoothing with Markov random fields and shrinkage priors. *Bayes. Anal.*, 13, 2018.

[33] Marlin D. Figgins and Trevor Bedford. SARS-CoV-2 variant dynamics across us states show consistent differences in effective reproduction numbers. *medRxiv*, page 2021.12.09.21267544, 2022.

[34] Marlin D. Figgins and Trevor Bedford. Inferring variant-specific effective reproduction numbers from combined case and sequencing data. *medRxiv*, page 2021.12.09.21267544, 2024.

[35] Tapiwa Ganyani, Cécile Kremer, Dongxuan Chen, Andrea Torneri, Christel Faes, Jacco Wallinga, and Niel Hens. Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, March 2020. *Eurosurveillance*, 25:2000257, 2020.

[36] Julia R. Gog and Bryan T. Grenfell. Dynamics and selection of many-strain pathogens. *Proc. Natl. Acad. Sci. USA*, 99:17209–17214, 2002.

[37] Jochen Görtler, Rebecca Kehlbeck, and Oliver Deussen. A visual exploration of Gaussian processes. *Distill*, 2019.

[38] Katelyn M. Gostic, Lauren McGough, Edward B. Baskerville, Sam Abbott, Keya Joshi, Christine Tedijanto, et al. Practical considerations for measuring the effective reproductive number, Rt. *PLoS Comput. Biol.*, 16:1–21, 2020.

[39] Allison J Greaney, Tyler N Starr, and Jesse D Bloom. An antibody-escape estimator for mutations to the SARS-CoV-2 receptor-binding domain. *Virus Evol.*, 8:veac021, 2022.

[40] James Hadfield, Colin Megill, Sidney M Bell, John Huddleston, Barney Potter, Charlton Callender, et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, 34:4121–4123, 2018.

[41] Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *J. Mach. Learn. Res.*, 14:1303–1347, 2013.

[42] John Huddleston, John R Barnes, Thomas Rowe, Xiyan Xu, Rebecca Kondor, David E Wentworth, et al. Integrating genotypes and phenotypes improves long-term forecasts of seasonal influenza A/H3N2 evolution. *eLife*, 9:e60067, 2020.

[43] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9:90–95, 2007.

[44] Kimihito Ito, Chayada Piantham, and Hiroshi Nishiura. Predicted dominance of variant Delta of SARS-CoV-2 before Tokyo Olympic Games, Japan, July 2021. *Eurosurveillance*, 26, 2021.

[45] Fanchong Jian, Leilei Feng, Sijie Yang, Yuanling Yu, Lei Wang, Weiliang Song, et al. Convergent evolution of SARS-CoV-2 XBB lineages on receptor-binding domain 455–456 synergistically enhances antibody evasion and ACE2 binding. *PLoS Pathog.*, 19:e1011868, 2023.

[46] Stuart A Kauffman. *The origins of order: Self-organization and selection in evolution*. Oxford University Press, USA, 1993.

[47] William Ogilvy Kermack, A. G. McKendrick, and Gilbert Thomas Walker. A contribution to the mathematical theory of epidemics. *Proc. R. Soc. A*, 115:700–721, 1927.

[48] Shruti Khare, Céline Gurry, Lucas Freitas, Mark B Schultz, Gunter Bach, Amadou Diallo, Nancy Akite, Joses Ho, Raphael TC Lee, Winston Yeo, et al. GISAID's role in pandemic response. *China CDC Wkly.*, 3:1049, 2021.

[49] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980v9*, 2017.

[50] Kathryn E Kistler and Trevor Bedford. Evidence for adaptive evolution in the receptor-binding domain of seasonal coronaviruses OC43 and 229e. *eLife*, 10:e64509, 2021.

[51] Kathryn E Kistler and Trevor Bedford. An atlas of continuous adaptive evolution in endemic human viruses. *Cell Host Microbe*, 31:1898–1909, 2023.

[52] Frank Konings, Mark D. Perkins, Jens H. Kuhn, Mark J. Pallen, Erik J. Alm, Brett N. Archer, et al. SARS-CoV-2 variants of interest and concern naming scheme conducive for global discourse. *Nat. Microbiol.*, 2021.

[53] Teddy Lazebnik and Svetlana Bunimovich-Mendrazitsky. Generic approach for mathematical model of multi-strain pandemics. *PLoS ONE*, 17:e0260683, 2022.

[54] Noémie Lefrancq, Loréna Duret, Valérie Bouchez, Sylvain Brisse, Julian Parkhill, and Henrik Salje. Learning the fitness dynamics of pathogens from phylogenies. *medRxiv*, page 2023.12.23.23300456, 2023.

[55] Marta Łuksza and Michael Lässig. A predictive fitness model for influenza. *Nature*, 507:57–61, 2014.

[56] Dylan H Morris, Katelyn M Gostic, Simone Pompei, Trevor Bedford, Marta Łuksza, Richard A Neher, Bryan T Grenfell, Michael Lässig, and John W McCauley. Predictive modeling of influenza shows the promise of applied evolutionary biology. *Trends Microbiol.*, 26:102–118, 2018.

[57] Nicola F. Müller, Cassia Wagner, Chris D. Frazar, Pavitra Roychoudhury, Jover Lee, and et al. Louise H. Moncla. Viral genomes reveal patterns of the SARS-CoV-2 outbreak in washington state. *Sci. Transl. Med.*, 13:eabf0202, 2021.

[58] Ville Mustonen and Michael Lässig. Fitness flux and ubiquity of adaptive evolution. *Proc. Natl. Acad. Sci. USA*, 107:4248–4253, 2010.

[59] Sravani Nanduri, Allison Black, Trevor Bedford, and John Huddleston. Dimensionality reduction distills complex evolutionary relationships in seasonal influenza and SARS-CoV-2. *Virus Evol.*, 10:veae087, 2024.

[60] Fritz Obermeyer, Martin Jankowiak, Nikolaos Barkas, Stephen F Schaffner, Jesse D Pyle, Leonid Yurkovetskiy, et al. Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness. *Science*, 376:1327–1332, 2022.

[61] Helen Onyeaka, Christian K Anumudu, Zainab T Al-Sharify, Esther Egele-Godswill, and Paul Mbaegbu. COVID-19 pandemic: A review of the global lockdown and its far-reaching effects. *Review Sci. Prog.*, 104:368504211019854, 2021.

[62] Sang Woo Park, Benjamin M. Bolker, Sebastian Funk, C. Jessica E. Metcalf, Joshua S. Weitz, Bryan T. Grenfell, et al. Roles of generation-interval distributions in shaping relative epidemic strength, speed, and control of new SARS-CoV-2 variants. *medRxiv*, 2021.

[63] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011.

[64] Du Phan, Neeraj Pradhan, and Martin Jankowiak. Composable effects for flexible and accelerated probabilistic programming in numpyro. *arXiv*, 2019.

[65] Chayada Piantham and Kimihito Ito. Predicting the time course of replacements of SARS-CoV-2 variants using relative reproduction numbers. *medRxiv*, 2022.03.30.22273218, 2022.

[66] Chayada Piantham, Natalie M. Linton, Hiroshi Nishiura, and Kimihito Ito. Estimating the elevated transmissibility of the B.1.1.7 strain over previously circulating strains in england using GISAID sequence frequencies. *medRxiv*, 2021.

[67] Koen B Pouwels, Thomas House, Emma Pritchard, Julie V Robotham, Paul J Birrell, Andrew Gelman, et al. Community prevalence of SARS-CoV-2 in England from April to November, 2020: results from the ONS Coronavirus Infection Survey. *Lancet Public Health*, 6:e30–e38, 2021.

[68] Andrew Rambaut, Edward C. Holmes, Áine O'Toole, Verity Hill, John T. McCrone, Christopher Ruis, Louis du Plessis, and Oliver G. Pybus. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.*, 5:1403–1407, 2020.

[69] Andrew Rambaut, Edward C Holmes, Áine O'Toole, Verity Hill, John T McCrone, Christopher Ruis, Louis Du Plessis, and Oliver G Pybus. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.*, 5:1403–1407, 2020.

[70] Gabriel Riutort-Mayol, Paul-Christian Bürkner, Michael R. Andersen, Arno Solin, and Aki Vehtari. Practical Hilbert space approximate Bayesian Gaussian processes for probabilistic programming. *Stat. Comput.*, 33.1:17, 2023.

[71] Sukhyun Ryu, Dasom Kim, Jun-Sik Lim, Sheikh Taslim Ali, and Benjamin J. Cowling. Serial interval and transmission dynamics during SARS-CoV-2 Delta variant predominance, South Korea. *Emerg. Infect. Dis.*, 28:407–410, 2022.

[72] Mrinank Sharma, Sören Mindermann, Charlie Rogers-Smith, Gavin Leech, Benedict Snodin, Janvi Ahuja, and et al. Jonas B. Sandbrink. Understanding the effectiveness of government interventions against the resurgence of COVID-19 in Europe. *Nat. Commun.*, 12, 2021.

[73] Y. Shu and J. McCauley. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Eurosurveillance*, 22:30494, 2017.

[74] Jin Su Song, Jihee Lee, Miyoung Kim, Hyeong Seop Jeong, Moon Su Kim, Seong Gon Kim, et al. Serial intervals and household transmission of SARS-CoV-2 Omicron variant, South Korea, 2021. *Emerg. Infect. Dis.*, 28:756–759, 2022.

[75] Zachary Susswein, Kaitlyn E Johnson, Robel Kassa, Mina Parastaran, Vivian Peng, Leo Wolansky, et al. Early risk-assessment of pathogen genomic variants emergence. *medRxiv*, pages 2023–01, 2023.

[76] Zachary Susswein, Kaitlyn E. Johnson, Robel Kassa, Mina Parastaran, Vivian Peng, Leo Wolansky, et al. Leveraging global genomic sequencing data to estimate local variant dynamics. *medRxiv*, page 2023.01.02.23284123, 2023.

[77] Kaiming Tao, Philip L Tzou, Janin Nouhin, Ravindra K Gupta, Tulio de Oliveira, Sergei L Kosakovsky Pond, et al. The biological and clinical significance of emerging SARS-CoV-2 variants. *Nat. Rev. Genet.*, 22:757–773, 2021.

[78] Ashley L. Taylor and Tyler N. Starr. Deep mutational scans of XBB.1.5 and BQ.1.1 reveal ongoing epistatic drift during SARS-CoV-2 evolution. *PLoS Pathog.*, 19:e1011901, 2023.

[79] Houriiyah Tegally, Eduan Wilkinson, Marta Giovanetti, Arash Iranzadeh, Vagner Fonseca, Jennifer Giandhari, et al. Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature*, 592:438–443, 2021.

[80] Yatish Turakhia, Bryan Thornlow, Angie S Hinrichs, Nicola De Maio, Landen Gozashti, Robert Lanfear, et al. Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nat. Genet.*, 53:809–816, 2021.

[81] Keiya Uriu, Izumi Kimura, Kotaro Shirakawa, Akifumi Takaori-Kondo, Taka-aki Nakada, Atsushi Kaneda, et al. Neutralization of the SARS-CoV-2 Mu variant by convalescent and vaccine serum. *N. Engl. J. Med.*, 385:2397–2399, 2021.

[82] Christiaan van Dorp, Emma Goldberg, Ruian Ke, Nick Hengartner, and Ethan Romero-Severson. Global estimates of the fitness advantage of SARS-CoV-2 variant Omicron. *Virus Evol.*, 8:veac089, 2022.

[83] Raquel Viana, Sikhulile Moyo, Daniel G Amoako, Houriiyah Tegally, Cathrine Scheepers, Christian L Althaus, et al. Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. *Nature*, 603:679–686, 2022.

[84] Harald S Vohringer, Theo Sanderson, Matthew Sinnott, Nicola De Maio, Thuy Nguyen, Richard Goater, et al. Genomic reconstruction of the SARS-CoV-2 epidemic in England. *Nature*, 600:506–611, 2021.

[85] Erik Volz, Swapnil Mishra, Meera Chand, Jeffrey C. Barrett, Robert Johnson, Lily Geidelberg, et al. Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. *Nature*, 593:266–269, 2021.

[86] William N. Voss, Michael L. Mallory, Patrick O. Byrne, Jeffrey M. Marchioni, Sean A. Knudson, John M. Powers, et al. Hybrid immunity to SARS-CoV-2 arises from serological recall of IgG antibodies distinctly imprinted by infection or vaccination. *Cell Rep. Med.*, 5:101668, 2024.

[87] J Wallinga and M Lipsitch. How generation intervals shape the relationship between growth rates and reproductive numbers. *Proc. R. Soc. B*, 274:599–604, 2006.

[88] Felix Wong and James J. Collins. Evidence that coronavirus superspreading is fat-tailed. *Proc. Natl. Acad. Sci. USA*, 117:29416–29418, 2020.

Appendix A

SUPPLEMENTARY MATERIALS FOR CHAPTER 2

A.1 Supplementary Text

Relationship to multinomial logistic regression

Other papers have tried to infer growth advantages of variants from sequence data alone, we show that the multinomial logistic regression model typically used in these analysis is roughly equivalent to our fixed growth advantage model, but that inferring relative effective reproduction numbers between variants using multinomial logistic regression requires additional restrictions on the generation time. Multinomial logistic regression typically models the probability of a given observation belong to class v at time t as

$$f_v(t) = \frac{p_v \exp(\beta_v t)}{\sum_{1 \leq u \leq V} p_u \exp(\beta_u t)}. \tag{A.1}$$

For our purpose, we can assume this probability is equivalent to the true frequency of variant v in the population and in this case, p_v is considered to be related to the prevalence on variant v in the population at $t = 0$ and β_v can be considered to be the growth advantage relative to a pivot class u_* which has $\beta_{k_*} = 0$. In order to see the connection between the above model and ours, we return to the original renewal equation of the form

$$I(t) = R_t \int_0^t I(t - \tau) g(\tau). \tag{A.2}$$

Assuming that g is a point mass at a mean generation time T_g, we have that

$$I(nT_g) = \left(\prod_{i=1}^{n} R_{iT_g} \right) I(0). \tag{A.3}$$

Assuming that there are several variants following these same dynamics, we have that the frequency of a given variant v can be written as

$$f_v(nT_g) = \frac{I_v(nT_g)}{\sum_{1\leq u\leq V} I_u(nT_g)}. \tag{A.4}$$

If we assume a constant growth advantage as in our model, we then have that $R_{t,v} = \Delta_v R_t$, so that

$$f_v(nT_g) = \frac{\Delta_v^n I_v(0)}{\sum_{1\leq u\leq V} \Delta_u^n I_u(0)}. \tag{A.5}$$

Writing $\Delta_v = \exp(\delta_v)$ and $t = nT_g$, allows us to see that

$$f_v(t) = \frac{I_v(0)\exp(\frac{\delta_v}{T_g}t)}{\sum_{1\leq u\leq V} I_u(0)\exp(\frac{\delta_u}{T_g}t)}. \tag{A.6}$$

By fixing one pivot class so that $I_{u_*} = 1$ and $\delta_{u_*}/T_g = 0$, we can identify our model with the multinomial logistic regression by relating the parameters as

$$\delta_v = \beta_v T_g \tag{A.7}$$

$$I_v(0) = p_v. \tag{A.8}$$

This shows that the multinomial logistic regression functions similarly to our fixed growth advantage model except with the additional assumption that the generation time is a point mass at T_g. This assumption additionally allows us to relate the epidemic growth rate r and the effective reproduction number as $R = \exp(rT_g)$ [87]. Therefore, by further assuming that the variant infections are exponentially growing with rates r_v, we can then identify $\beta_v = r_v - r_{u_*}$. This means that the relative effective reproduction number for any two variants can be written as

$$\ln\left(\frac{R_{t,v}}{R_{t,u}}\right) = (\beta_v - \beta_u)T_g.$$

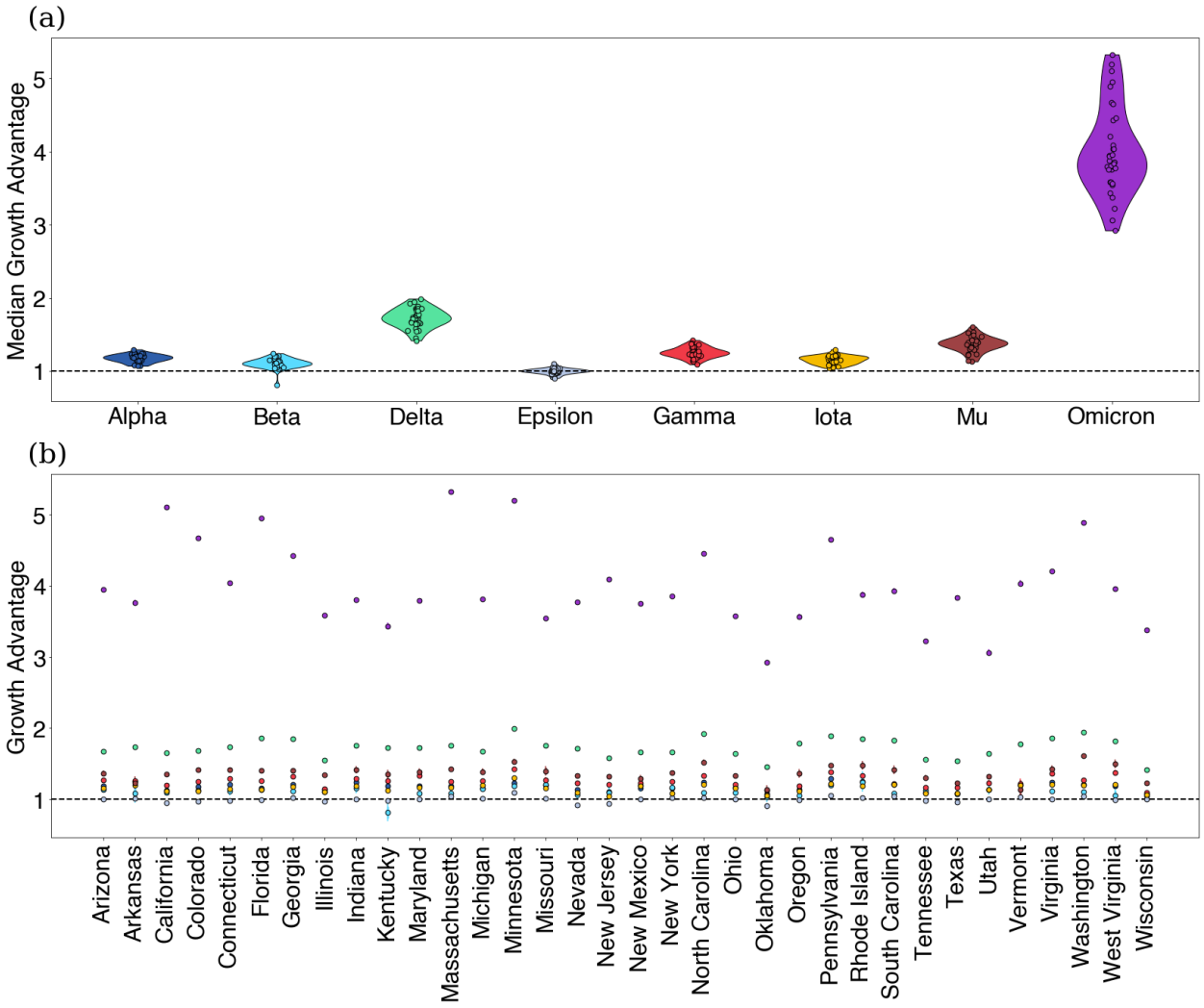

Figure A.1: **Estimating variant growth advantages in various states using Multinomial Logistic Regression** (a) Growth advantages visualized by state assuming generation time $T_g = 5.2$.. (b) Same as (a) but grouped by variant.

Relating epidemic growth rates to relative effective reproduction numbers

An important relationship of interest is between the epidemic growth rate of an epidemic and its effective reproduction number. In the case of our analysis, we are particularly interested in the effect of generation time assumptions on estimated variant-specific effective reproduction numbers. First, notice that the effective reproduction number and the epidemic growth rate of an epidemic are related by

$$R_t = \frac{1}{\int_0^\infty \exp(-r\tau)g(\tau)d\tau} = \frac{1}{M_g(-r)}$$

according to the Lotka-Euler equation [87] where r is the epidemic growth rate and M_g is the moment-generating function of the generation time g.

This allows us to write the relative reproduction number of two variants v and u as a function of their epidemic growth rates, so that

$$\frac{R_{t,v}}{R_{t,u}} = \frac{M_g(-r_u)}{M_g(-r_v)}. \tag{A.9}$$

We'll consider three common generation time assumptions. First, we consider the case where the generation time is a point mass at T_g. In which case, $M_g(-r) = \exp(-rT_g)$ and we recover the relationship

$$R_{t,v} = \exp(r_v T_g). \tag{A.10}$$

In this case, the relative effective reproduction number will depend on only the difference between the epidemic growth rates and therefore, is commonly used when converting estimated growth advantages to relative reproduction numbers in the case of logistic growth models.

Second, we consider the case where the generation time is an exponential distribution with mean T_g. This assumption is often implicit and common in models of infectious diseases such as ODEs and their stochastic variants. Using the corresponding moment-generating function, we see that

$$R_{t,v} = 1 + r_v T_g \tag{A.11}$$

Next, we consider the Gamma distributed generation times with mean T_g and standard deviation s. This is often used in models of infectious diseases via the chain trick in which multiple compartments are chained together to obtain non-exponential generation times or infectious periods. Re-parameterizing the Gamma distribution in terms of its mean and standard deviation and using its moment generating function, we have that

$$R_{t,v} = \left(1 + r_v \left(\frac{s^2}{T_g}\right)\right)^{T_g^2/s^2}. \tag{A.12}$$

From this equation, we can see that increases in the mean of the generation time of v leads to decreasing estimates of $R_{t,v}$ during epidemic decline ($r_v < 0$) and increased estimates during epidemic growth ($r_v > 0$) assuming r_v and s are fixed. Additionally, increases in the standard deviation will generally lead to lower inferred variant advantages. This effect is also visualized in Figure A.2.

Variant growth-advantages are sensitive to generation time In the case where we have two variants u, v with Gamma-distributed generation times with means T_u, T_v and standard deviations s_u, s_v respectively, we can then write the relative effective reproduction number of v over u as

$$\frac{R_{t,v}}{R_{t,u}} = \frac{\left[1 + r_v \left(\frac{s_v^2}{T_v}\right)\right]^{T_v^2/s_v^2}}{\left[1 + r_u \left(\frac{s_u^2}{T_u}\right)\right]^{T_u^2/s_u^2}}. \tag{A.13}$$

It follows that increases in the mean of the generation time of v leads to decreasing inferred variant advantages during epidemic decline and increased advantages during epidemic growth when all quantities are fixed. On the other hand, increases in the standard deviation will generally lead to lower inferred variant advantages.

Taking a logarithm, we can also evaluate the sensitivity of our inferred growth advantages from our fixed growth advantage model with respect to the generation time assuming it is Gamma distributed as

$$\delta_v = \ln\left(\frac{R_{t,v}}{R_{t,0}}\right) = \left(\frac{T_g^2}{s^2}\right) \ln\left(\frac{1 + r_v \left(\frac{s^2}{T_g}\right)}{1 + r_0 \left(\frac{s^2}{T_g}\right)}\right). \tag{A.14}$$

As the log of the relative effective reproduction number, the behavior here is analogous to that discussed above when the mean T_g and standard deviation s are changed. This effects of varying mean and standard deviation are illustrated in Figure A.4. Although the effective reproduction number and the growth advantage appear to have strong dependence on generation time parameters, we find that the epidemic growth rate r is more robust to changes in generation time (see Figure A.3).

The cases of exponential and Gamma-distributed generation times highlight that for non-deterministic generation times there is no guarantee that the relative effective reproduction number depends on only the difference in epidemic growth rates. In fact, these estimates based on the deterministic generation times correspond to the case in which the standard deviation shrinks zero, they are likely overestimates of variant advantages given the observed variation in the serial interval of SARS-CoV-2 infections.

Fixed growth advantages become time-varying under generation time misspecification We'll now consider the case where there is a true fixed-variant growth advantage. Suppose for a two-variant system that δ is the constant (log) growth advantage of the variant virus over the wildtype under the variant generation time g_T, so that $\delta = \ln\left(R_{t,v}^{g_T}/R_{t,wt}^{g_T}\right)$. Here subscripts denote the generation time used when computing R_t.

Under the misspecified variant generation time g_M, we can then write the inferred growth advantage as

$$\delta_M = \ln\left(\frac{R_{t,v}^{g_M}}{R_{t,wt}^{g_T}}\right) = \ln\left(\frac{R_{t,v}^{g_M}}{R_{t,v}^{g_T}}\right) + \delta. \tag{A.15}$$

In general, the term inside the log is non-constant meaning that fixed variant growth advantages under one generation time become non-constant under generation time specification.

A.2 Supplementary Figures

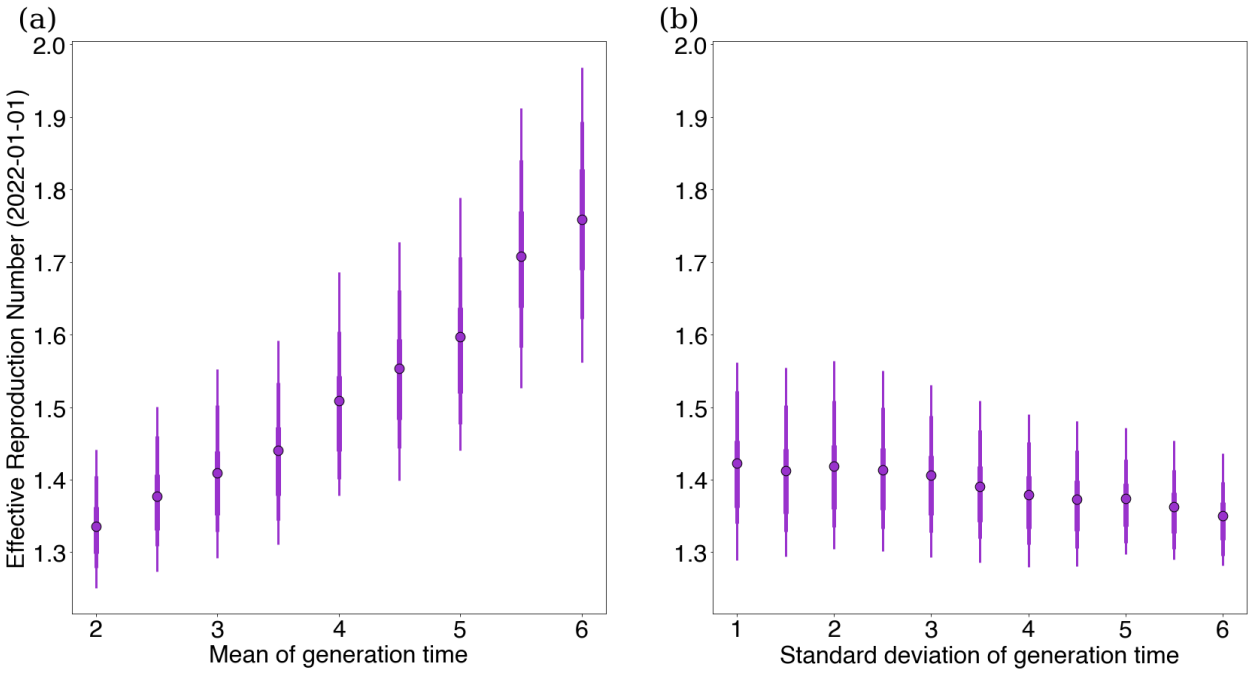

Figure A.2: **Sensitivity of effective reproduction number to changes in generation time.** (a) We vary the mean of Omicron generation time keeping a constant standard deviation 1.2 and plot against effective reproduction number estimates for Omicron in Washington state on February 1st, 2022 using our GARW model. (b) The same as (a), but we instead vary the standard deviation of Omicron generation time keeping a constant mean 3.1.

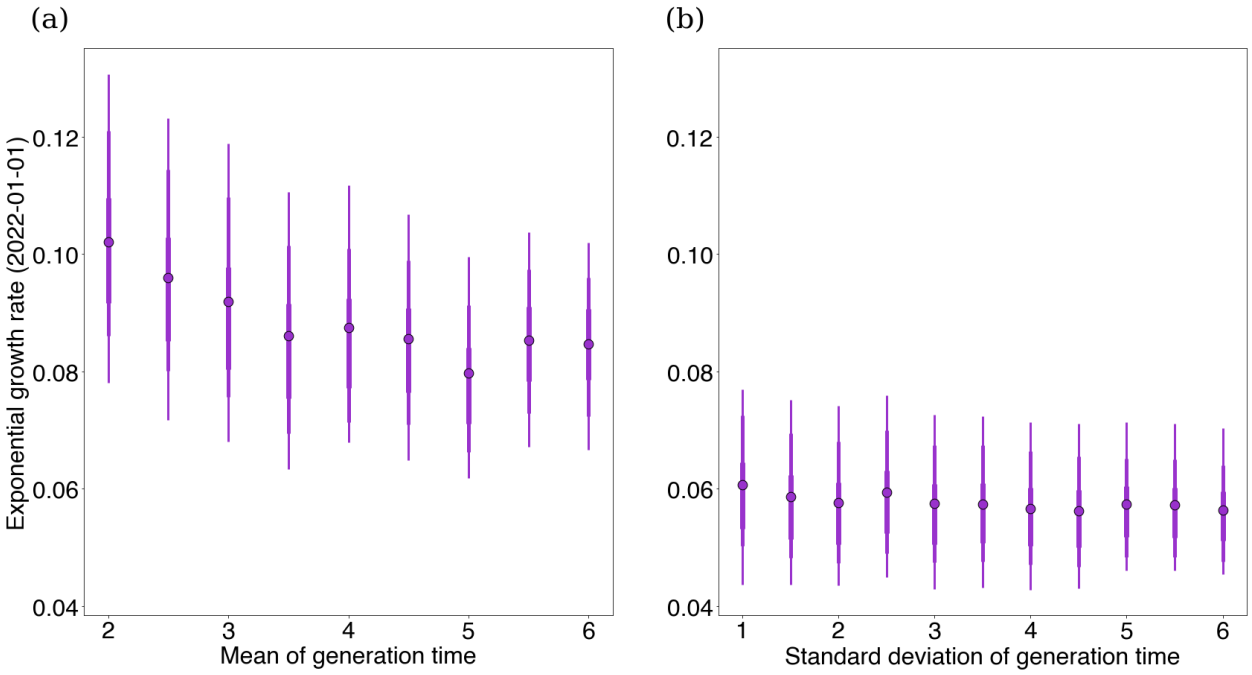

Figure A.3: **Sensitivity of epidemic growth rates to changes in generation time.** (a) We vary the mean of Omicron generation time keeping a constant standard deviation 1.2 and plot against exponential growth rates for Omicron in Washington state on February 1st, 2022 using our GARW model and assuming a Gamma-distributed generation time. (b) The same as (a), but we instead vary the standard deviation of Omicron generation time keeping a constant mean 3.1.

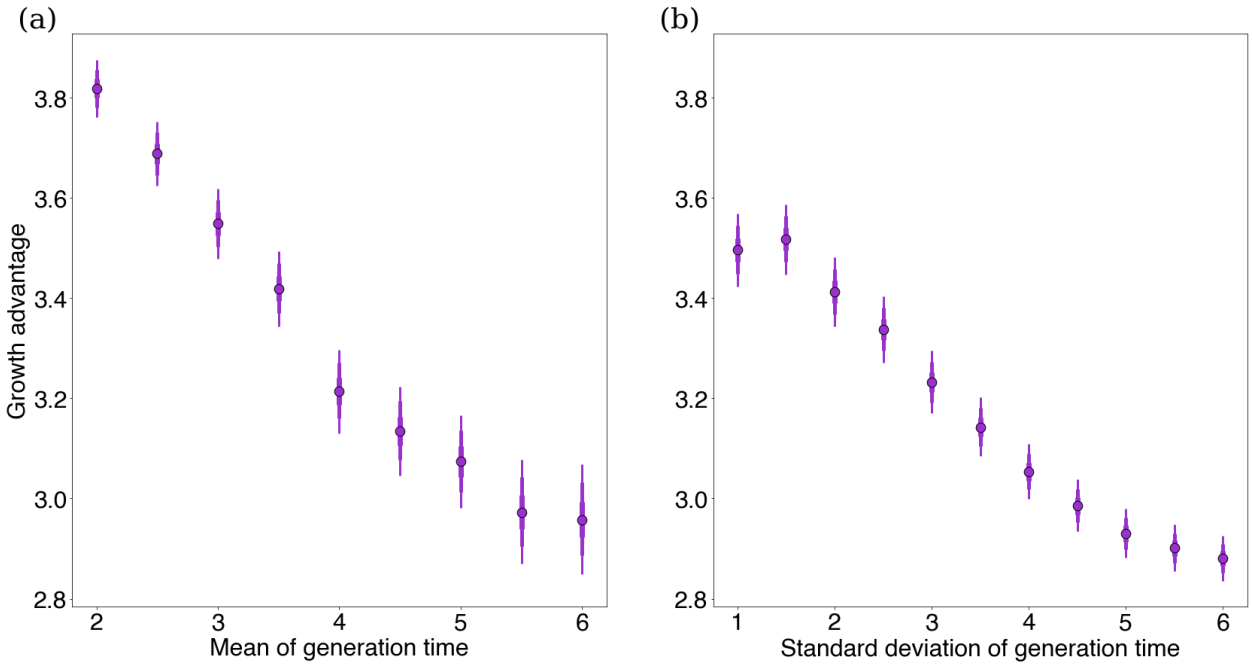

Figure A.4: **Sensitivity of growth advantages to changes in generation time.** (a) We vary the mean of Omicron generation time keeping a constant standard deviation 1.2 and plot against exponential growth rates for Delta in Washington state on July 1st, 2021 using our fixed growth model. (b) The same as (a), but we instead vary the standard deviation of Omicron generation time keeping a constant mean 3.2.

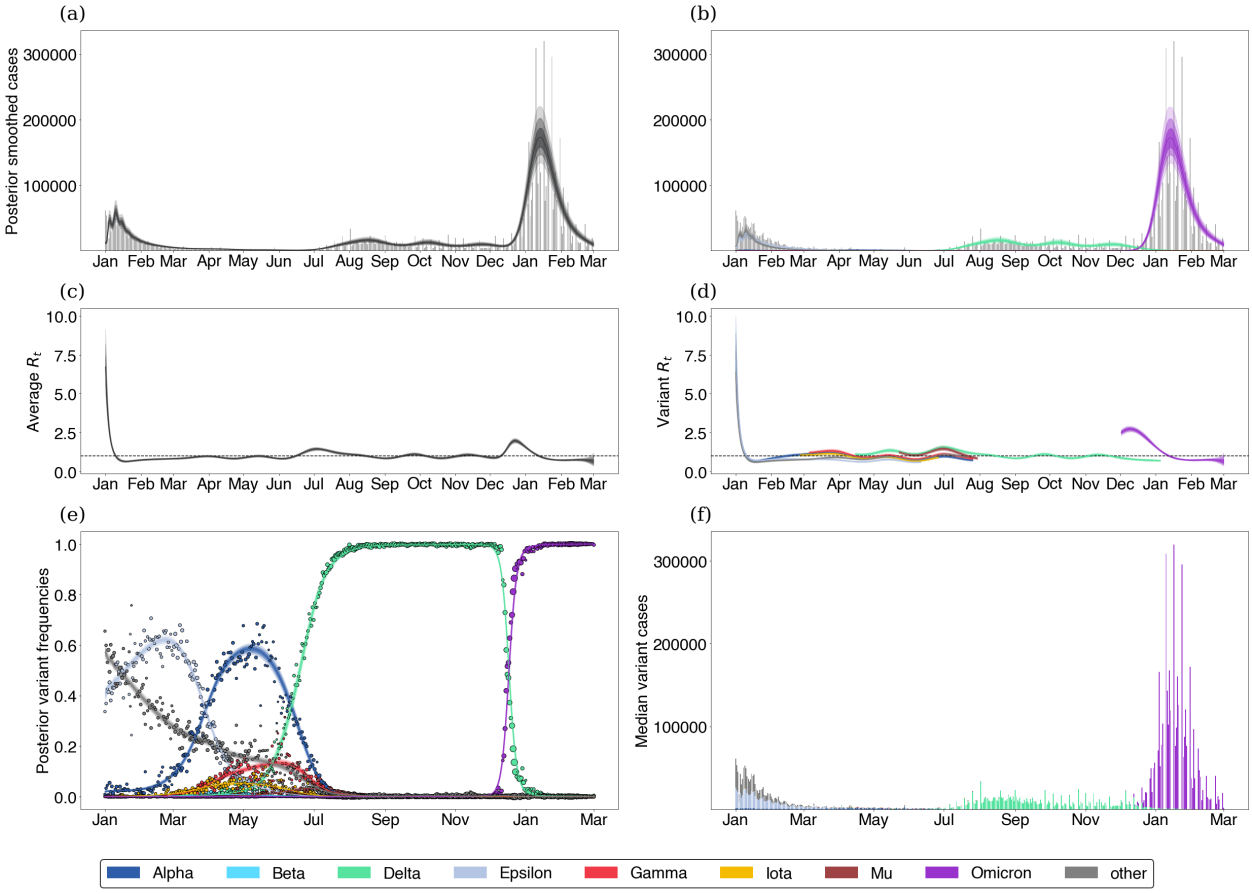

Figure A.5: **Fitting the GARW model to California data.**

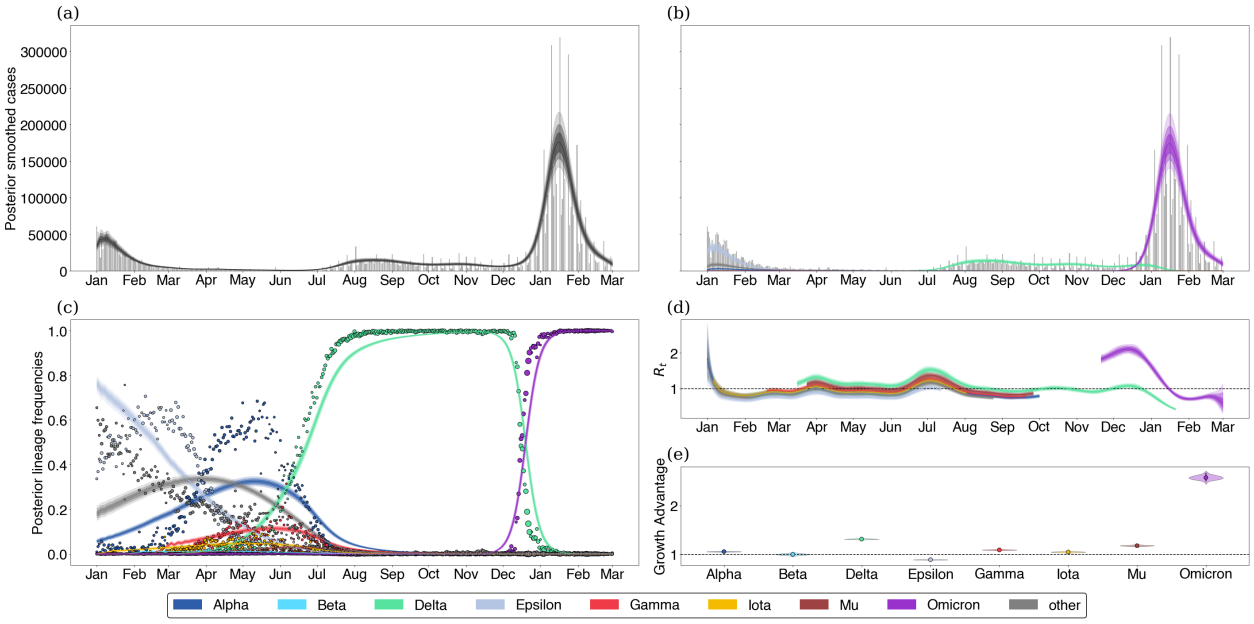

Figure A.6: **Fitting the fixed growth advantage model to California data.**

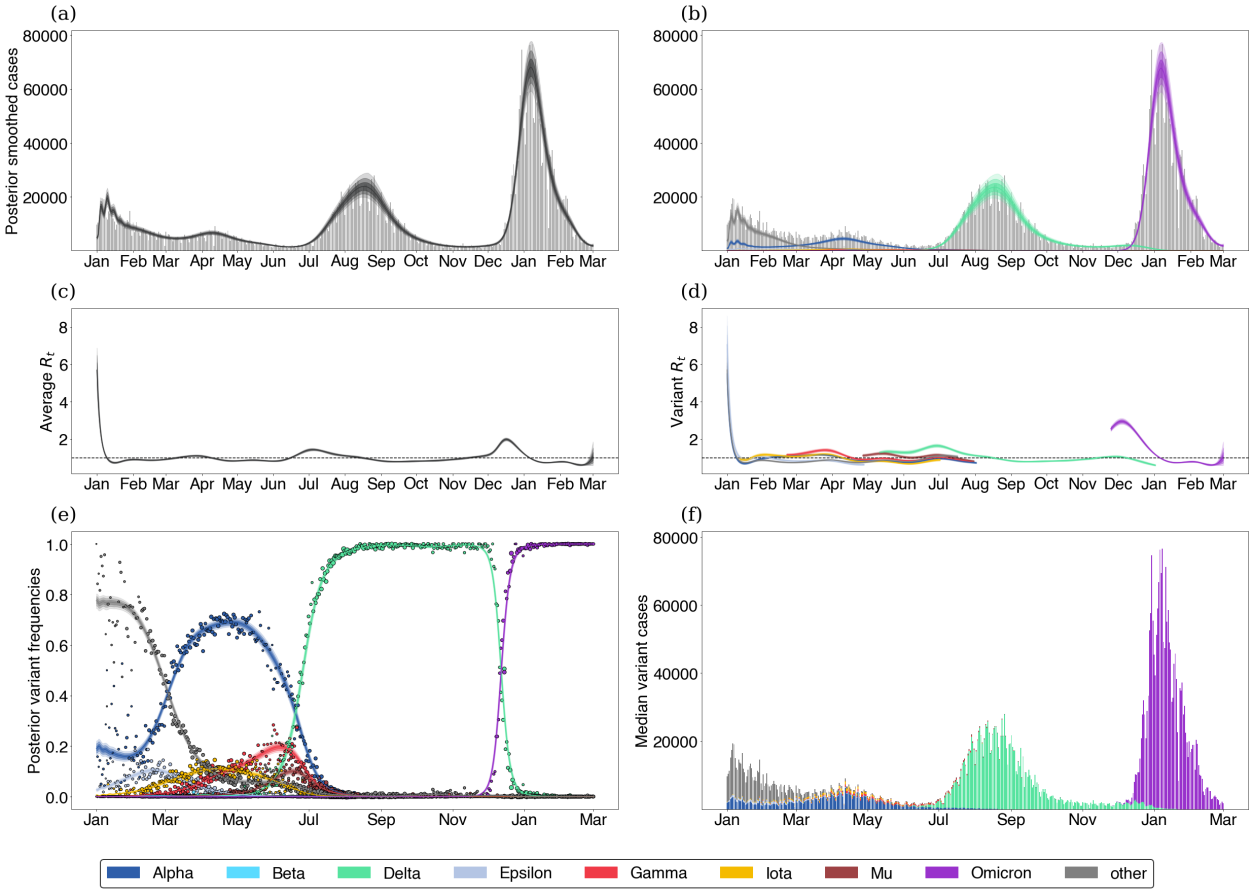

Figure A.7: **Fitting the GARW model to Florida data.**

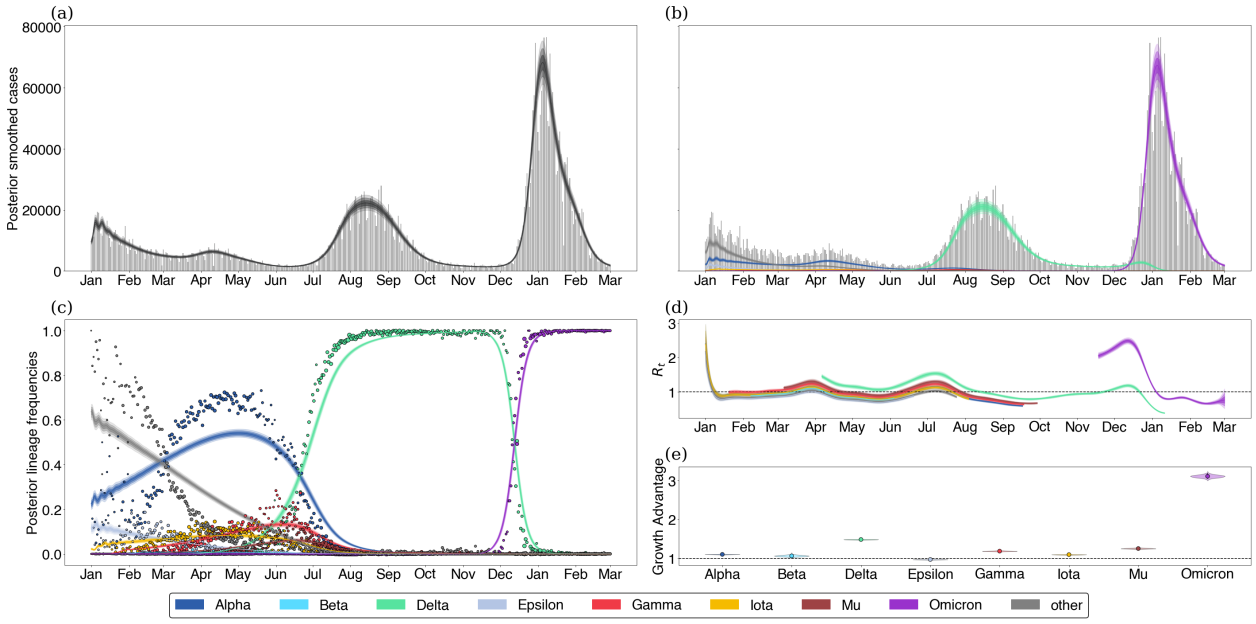

Figure A.8: **Fitting the fixed growth advantage model to Florida data.**

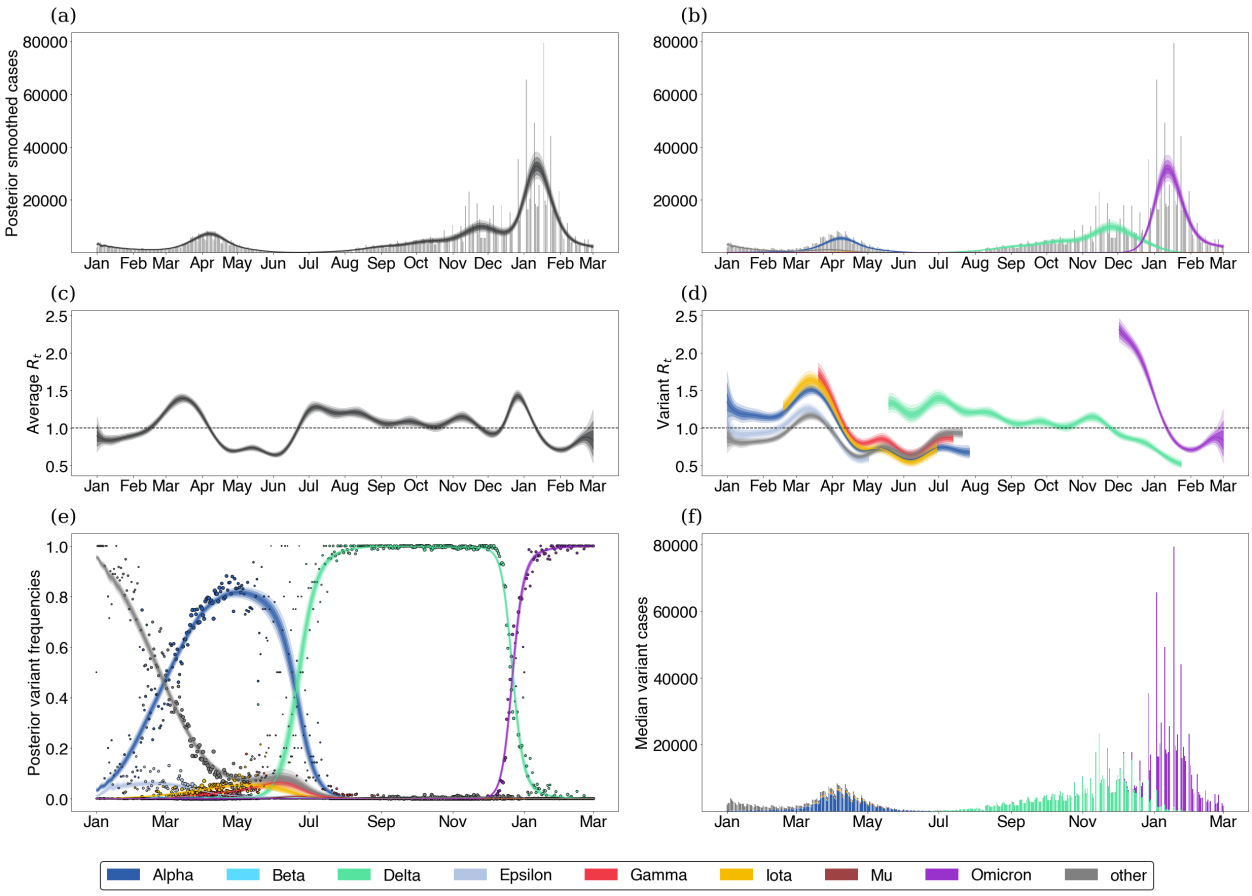

Figure A.9: **Fitting the GARW model to Michigan data.**

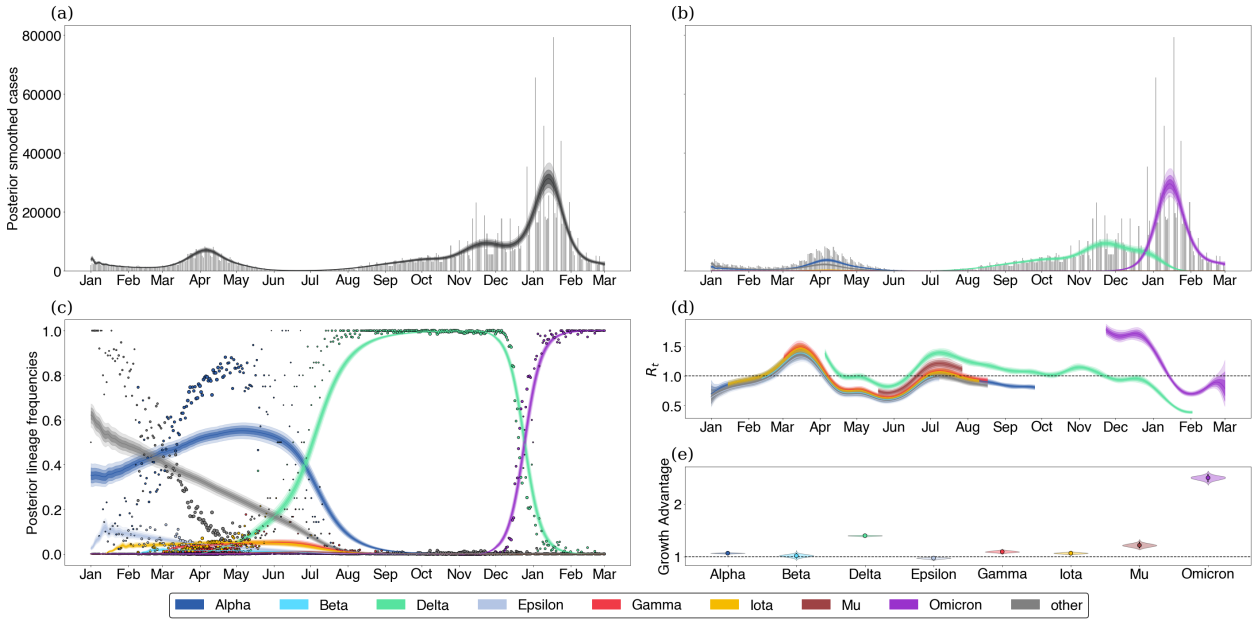

Figure A.10: **Fitting the fixed growth advantage model to Michigan data.**

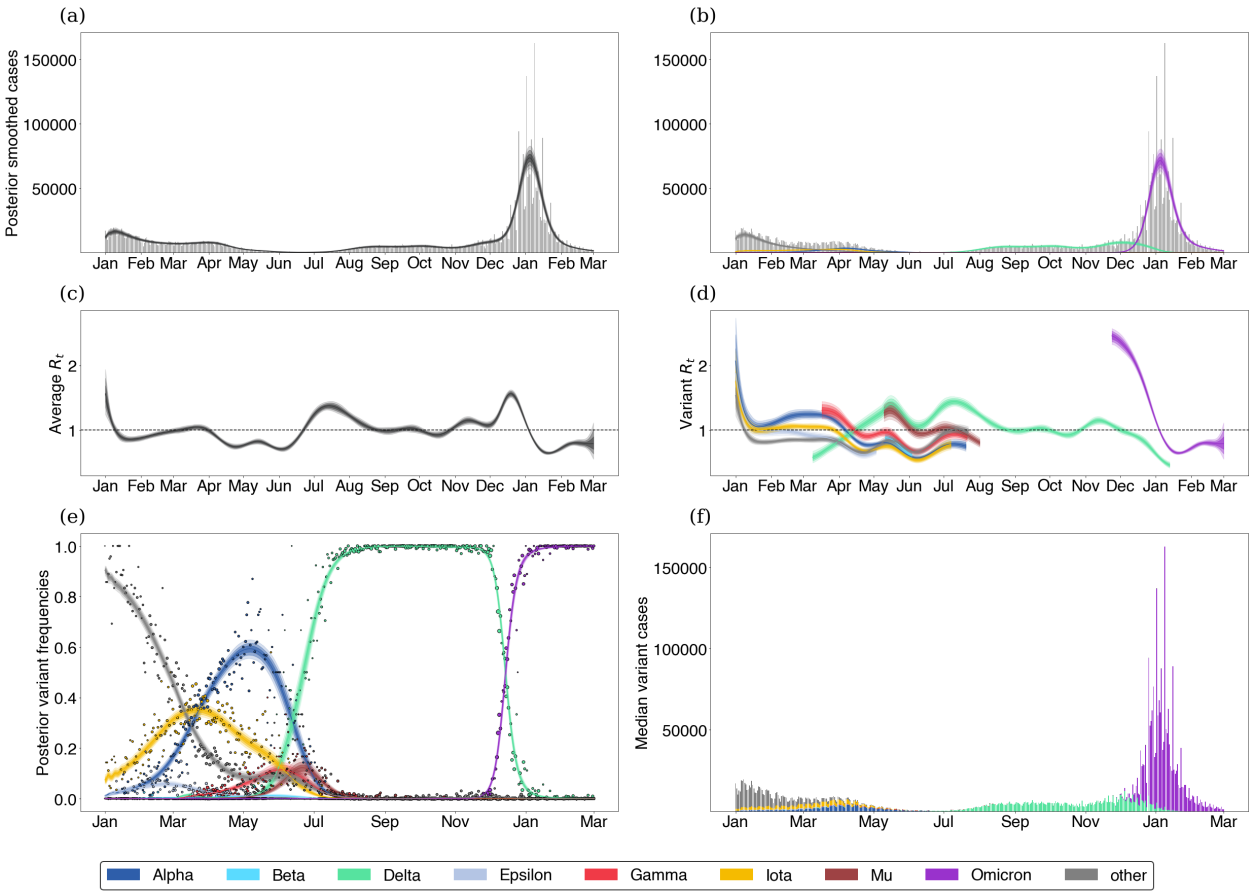

Figure A.11: **Fitting the GARW model to New York state data.**

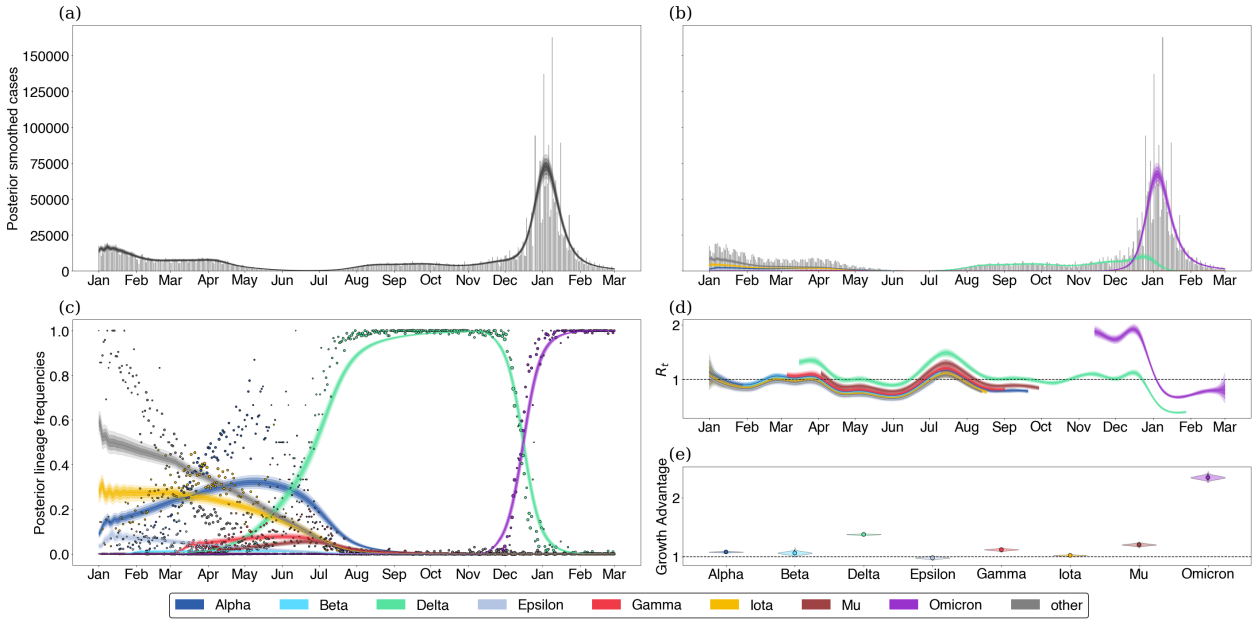

Figure A.12: **Fitting the fixed growth advantage model to New York state data.**

Appendix B

SUPPLEMENTARY MATERIALS FOR CHAPTER 3

B.1 Supplementary Figures

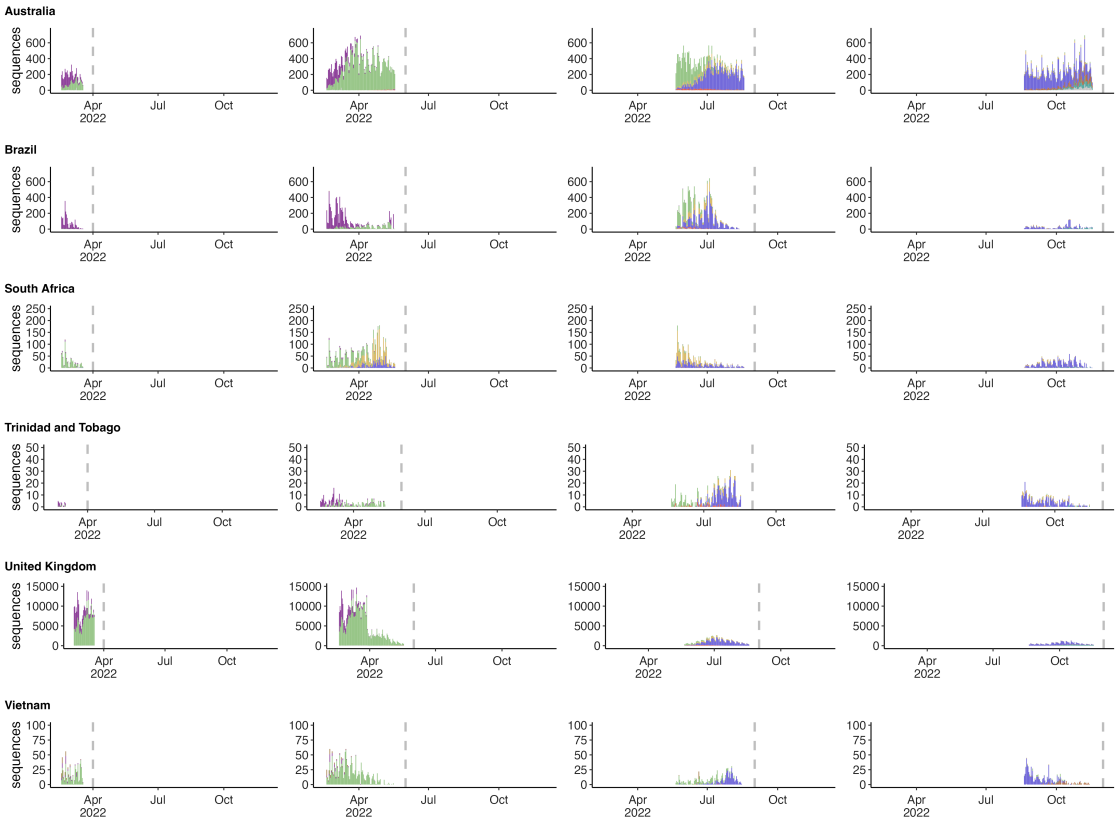

Figure B.1: **Reconstructing available data sets for Australia, Brazil, South Africa, Trinidad and Tobago, the United Kingdom, and Vietnam.** (A) Variant sequence counts categorized by Nextstrain clade at 4 different analysis dates.

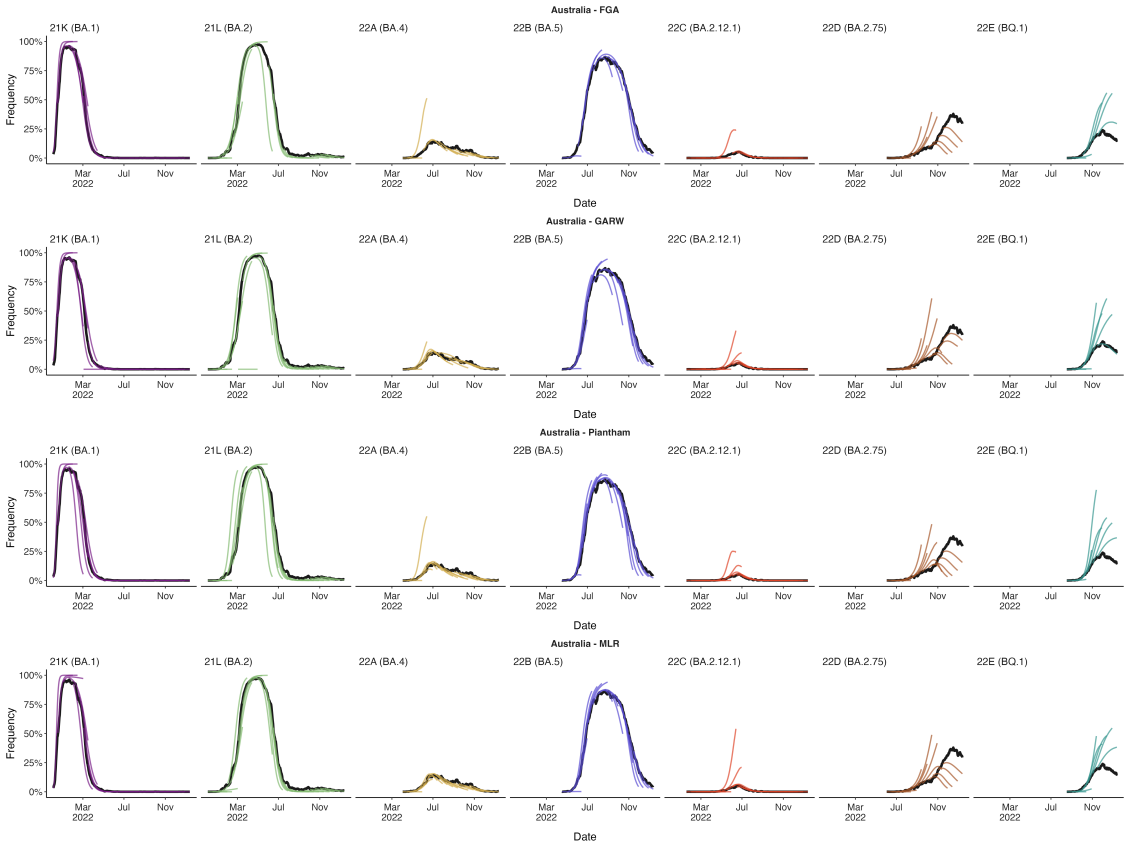

Figure B.2: **Reconstructing predictions for Australia** (A) +30 day frequency forecasts for variants in bimonthly intervals using the MLR model for Australia. Each forecast trajectory is shown as a different colored line. Retrospective smoothed frequency is shown as a thick black line.

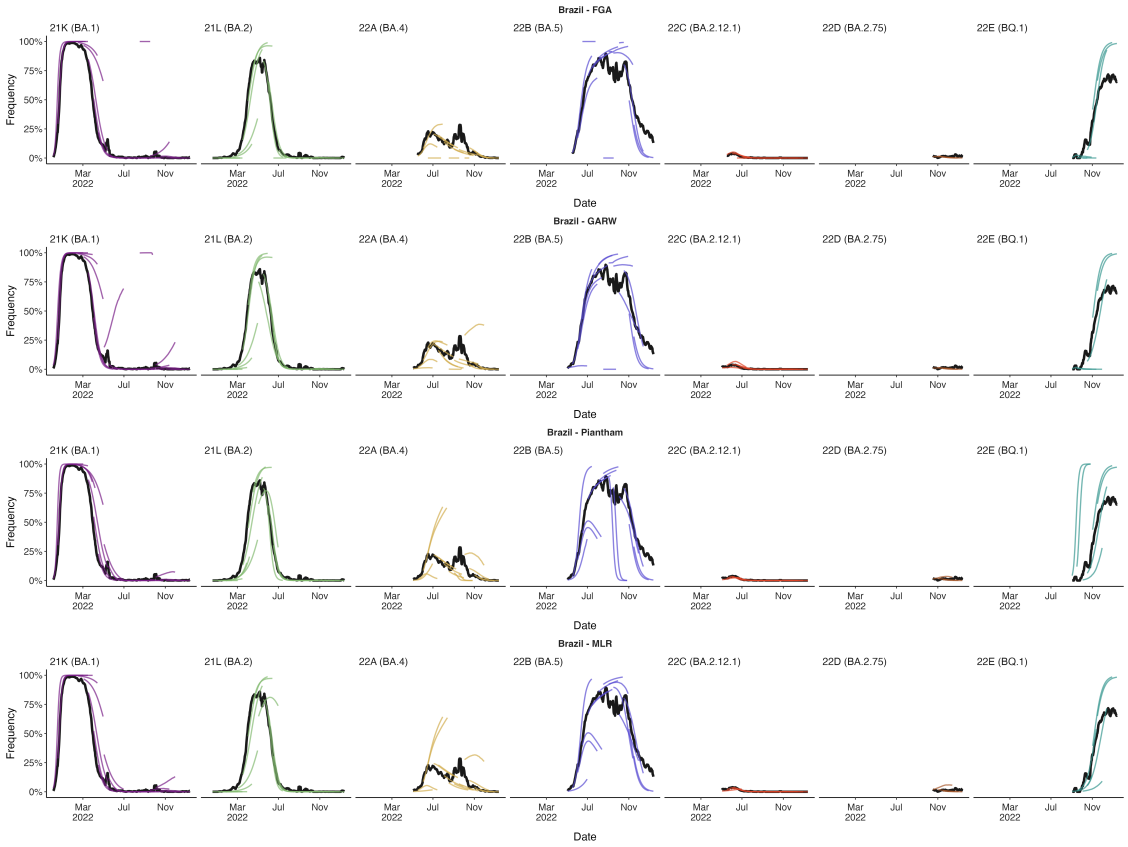

Figure B.3: **Reconstructing predictions for Brazil** (A) +30 day frequency forecasts for variants in bimonthly intervals using the MLR model for Brazil. Each forecast trajectory is shown as a different colored line. Retrospective smoothed frequency is shown as a thick black line.

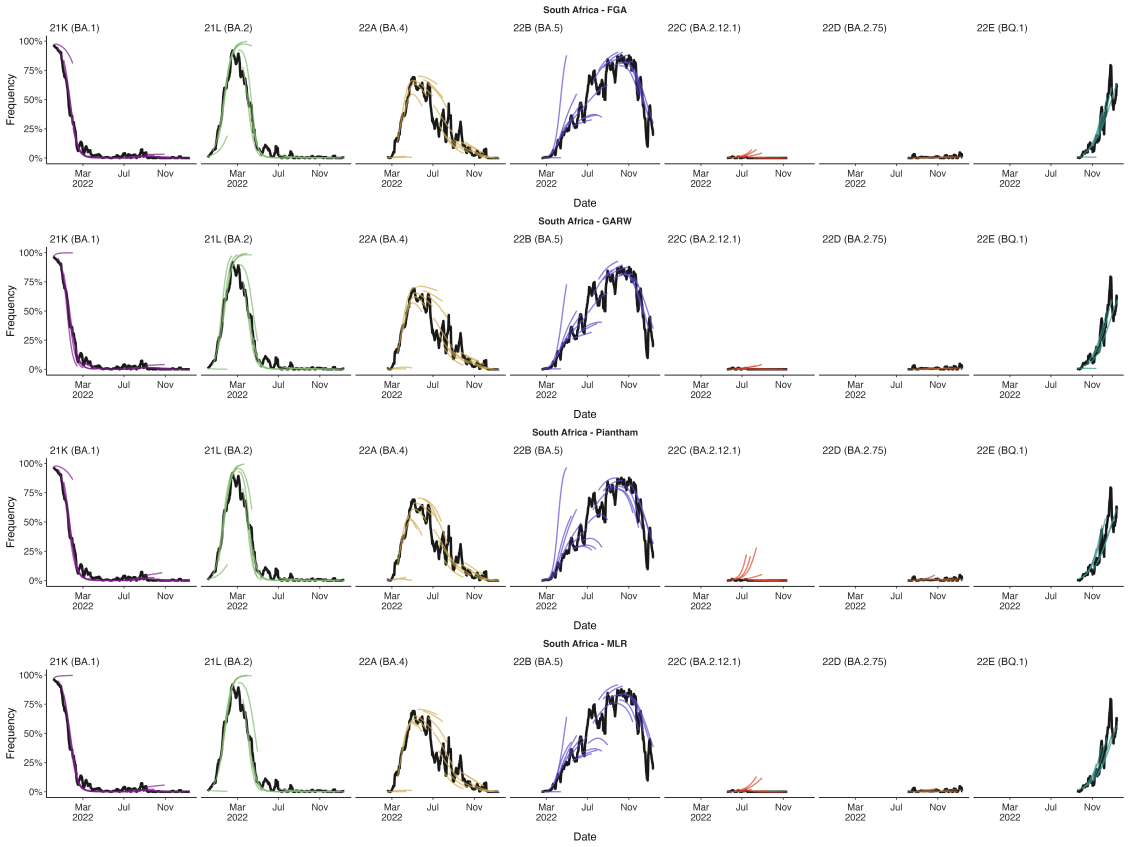

Figure B.4: **Reconstructing predictions for South Africa** (A) +30 day frequency forecasts for variants in bimonthly intervals using the MLR model for South Africa. Each forecast trajectory is shown as a different colored line. Retrospective smoothed frequency is shown as a thick black line.

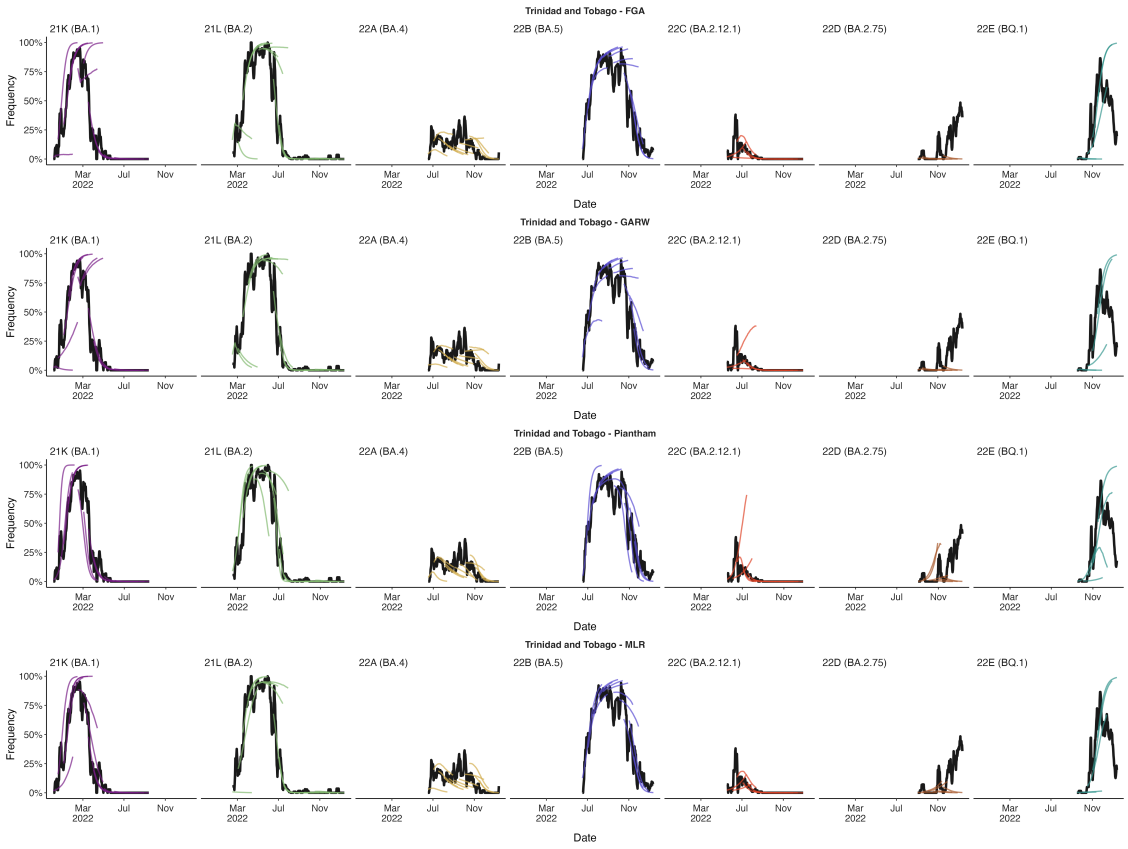

Figure B.5: **Reconstructing predictions for Trinidad and Tobago** (A) +30 day frequency forecasts for variants in bimonthly intervals using the MLR model for Trinidad and Tobago. Each forecast trajectory is shown as a different colored line. Retrospective smoothed frequency is shown as a thick black line.

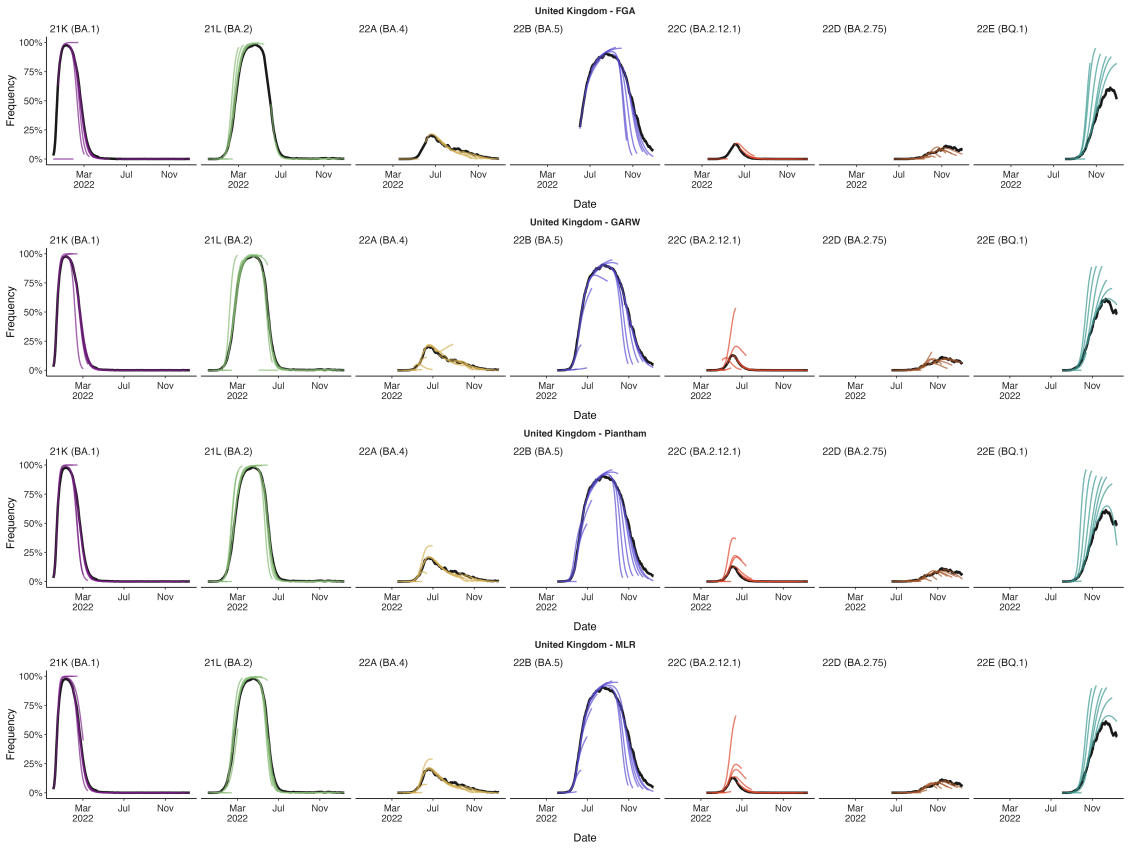

Figure B.6: **Reconstructing predictions for the United Kingdom** (A) +30 day frequency forecasts for variants in bimonthly intervals using the MLR model for the United Kingdom. Each forecast trajectory is shown as a different colored line. Retrospective smoothed frequency is shown as a thick black line.

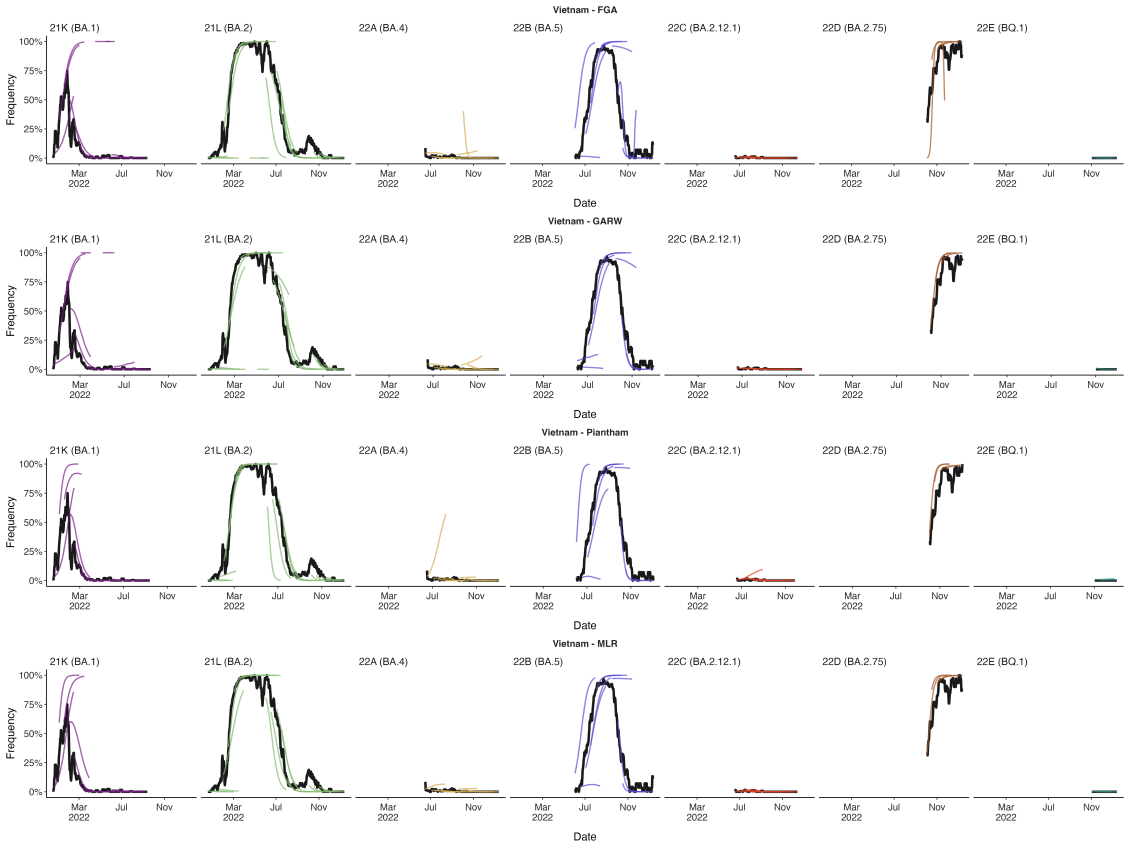

Figure B.7: **Reconstructing predictions for Vietnam** (A) +30 day frequency forecasts for variants in bimonthly intervals using the MLR model for Vietnam. Each forecast trajectory is shown as a different colored line. Retrospective smoothed frequency is shown as a thick black line.

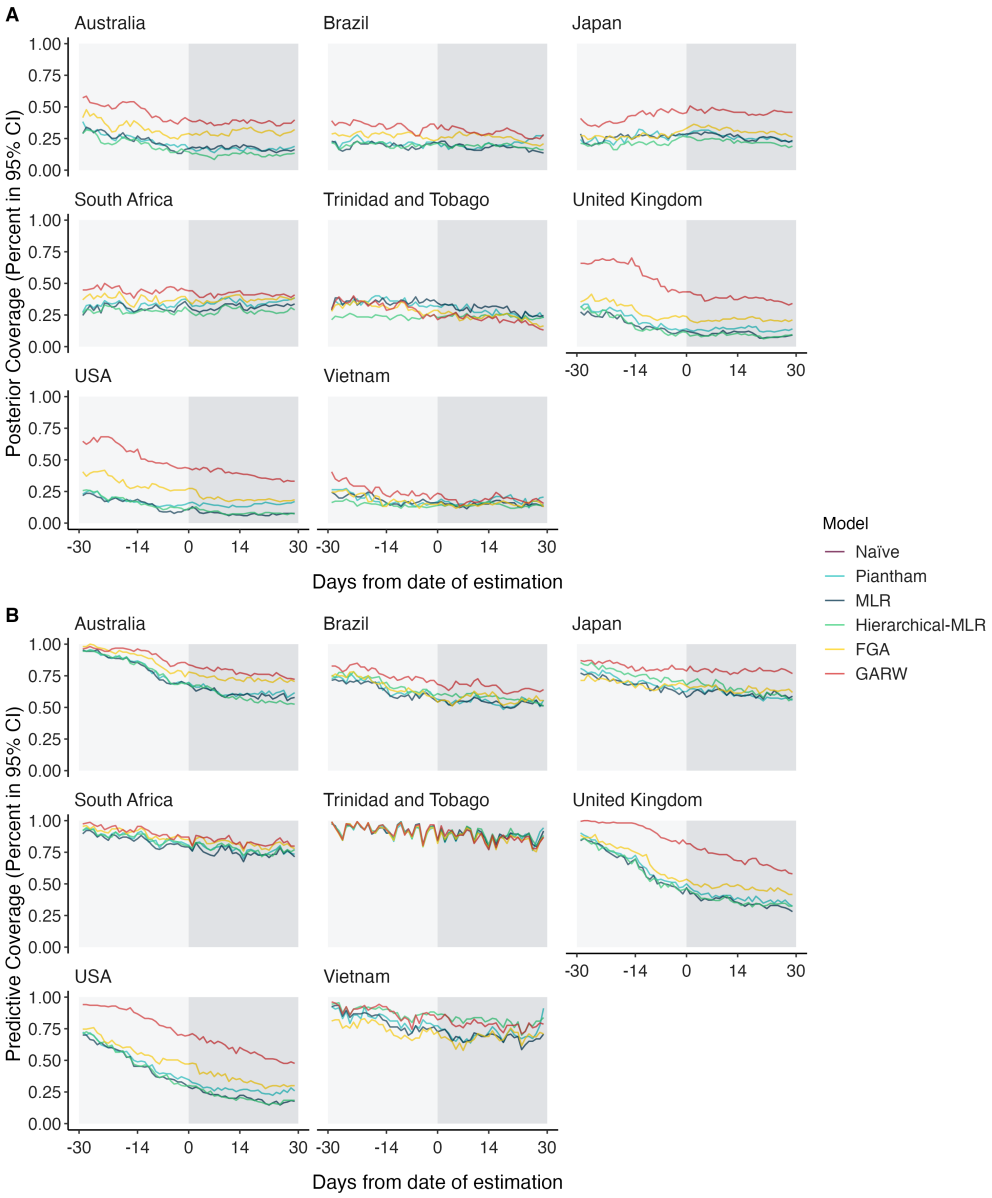

Figure B.8: **Posterior and predictive coverage for estimates across countries and models** (A) The proportion of estimates lying within the 95% confidence intervals (CIs) of posterior latent frequencies across lag times (-30,-30). (B) The proportion of estimates lying within the 95% confidence intervals (CIs) of posterior predictive sample frequencies across lag times (-30,-30). We generate the posterior predictive sample frequencies by sampling random counts for each variant using their posterior latent frequencies conditioning on the total sequences being those observed retrospectively.

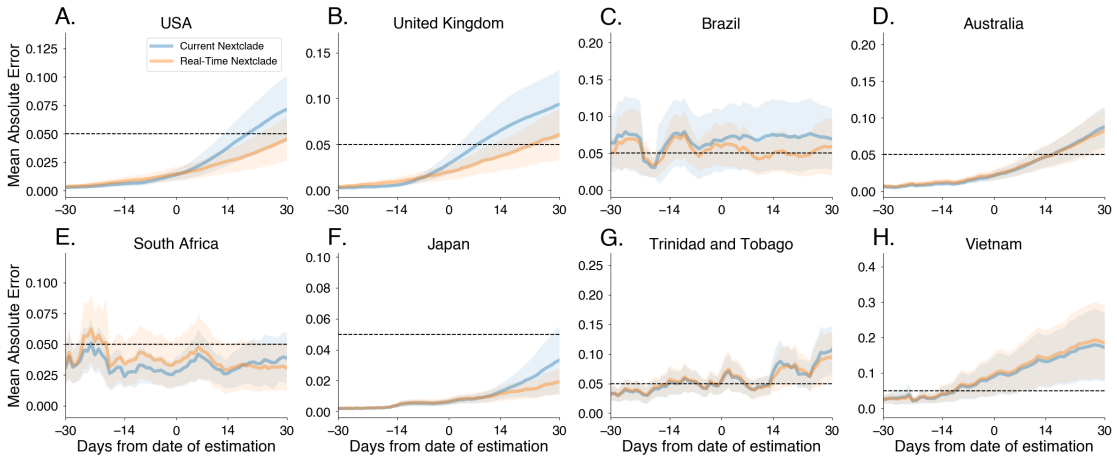

Figure B.9: **Comparing the accuracy of short-term forecast models under retrospective vs real-time clade assignments.** (A-H) Mean absolute error for MLR as a function of days since date of estimation, starting from 30 day hindcasts to 30 days forecasts. Intervals shown have width of two standard errors of the mean. We compare retrospective Nextstrain clade assignments made today ('Current Nextclade') to Nextstrain clade assignments available in Oct 2022 ('Real-time Nextclade'). We find that errors are qualitatively similar regardless of Nextclade version with errors being potentially higher for the current Nextclade version.

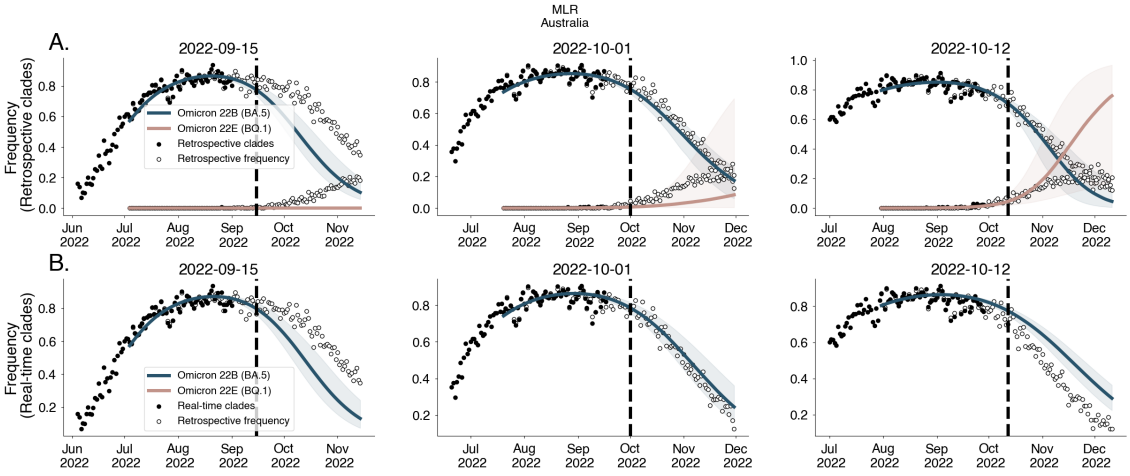

Figure B.10: **Forecasts for Australia using clade designations under retrospective vs real-time clade assignments** Forecasts from MLR fit to data generated using retrospective Nextstrain clade designations ('Current Nextclade') (A) and Nextstrain clade assignments available in Oct 2022 ('Real-time Nextclade') (B).

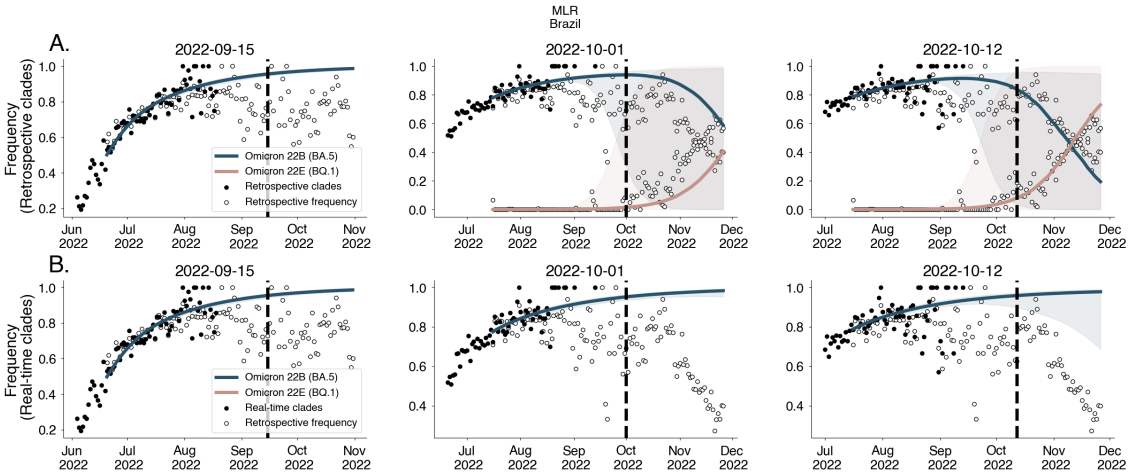

Figure B.11: **Forecasts for Brazil using clade designations under retrospective vs real-time clade assignments** Forecasts from MLR fit to data generated using retrospective Nextstrain clade designations ('Current Nextclade') (A) and Nextstrain clade assignments available in Oct 2022 ('Real-time Nextclade') (B).

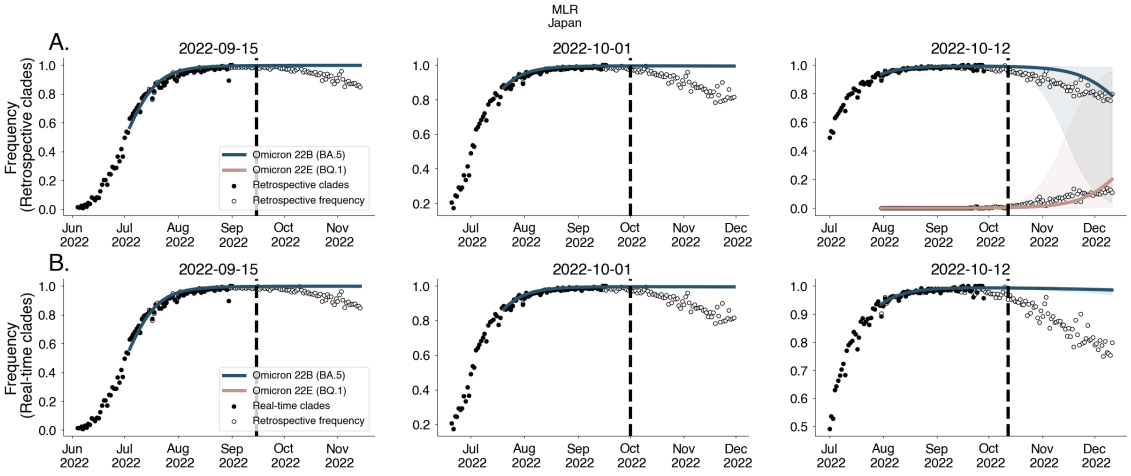

Figure B.12: **Forecasts for Japan using clade designations under retrospective vs real-time clade assignments** Forecasts from MLR fit to data generated using retrospective Nextstrain clade designations ('Current Nextclade') (A) and Nextstrain clade assignments available in Oct 2022 ('Real-time Nextclade') (B).

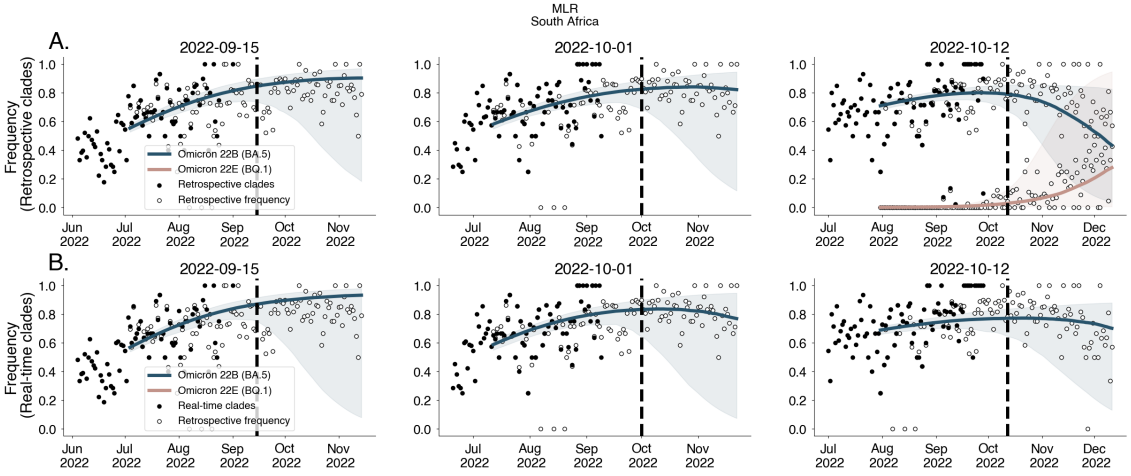

Figure B.13: **Forecasts for South Africa using clade designations under retrospective vs real-time clade assignments** Forecasts from MLR fit to data generated using retrospective Nextstrain clade designations ('Current Nextclade') (A) and Nextstrain clade assignments available in Oct 2022 ('Real-time Nextclade') (B).

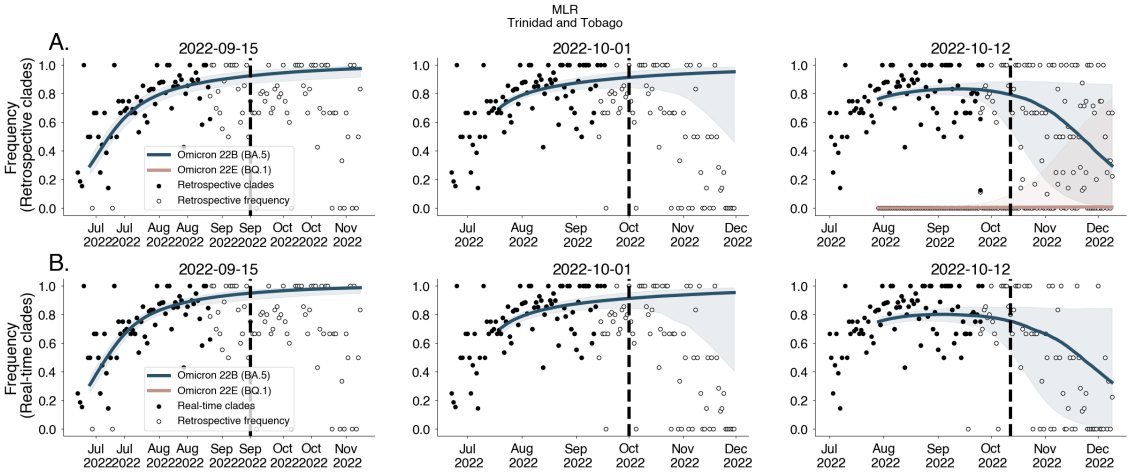

Figure B.14: **Forecasts for Trinidad and Tobago using clade designations under retrospective vs real-time clade assignments** Forecasts from MLR fit to data generated using retrospective Nextstrain clade designations ('Current Nextclade') (A) and Nextstrain clade assignments available in Oct 2022 ('Real-time Nextclade') (B).

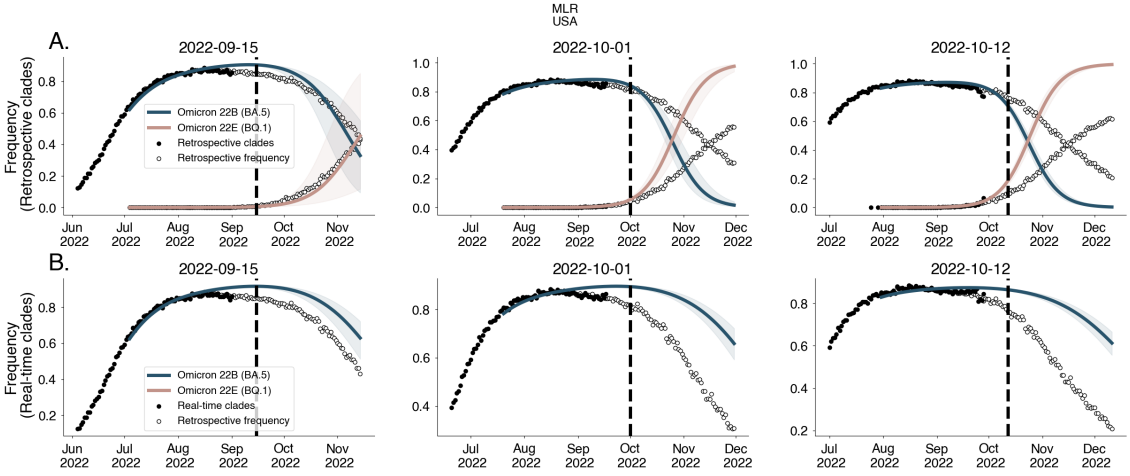

Figure B.15: **Forecasts for United States using clade designations under retrospective vs real-time clade assignments** Forecasts from MLR fit to data generated using retrospective Nextstrain clade designations ('Current Nextclade') (A) and Nextstrain clade assignments available in Oct 2022 ('Real-time Nextclade') (B).

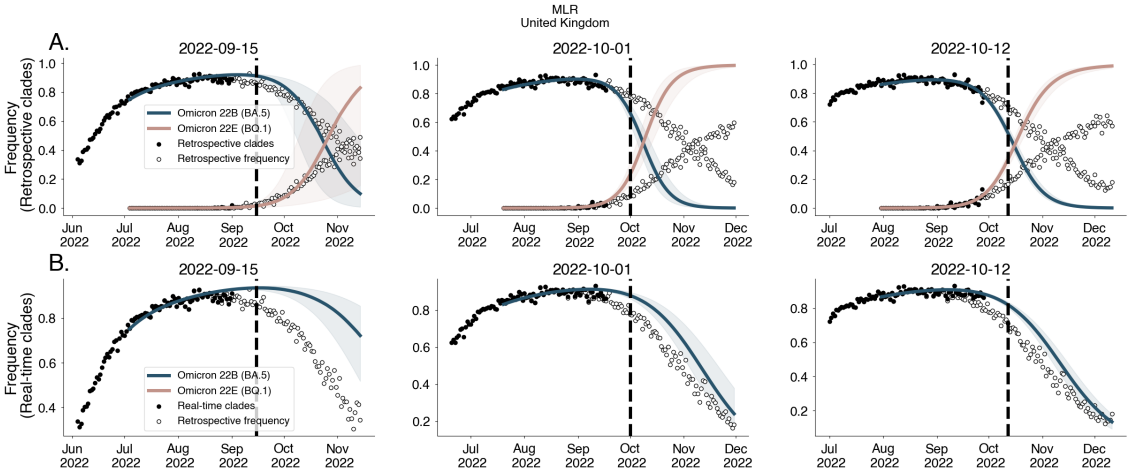

Figure B.16: **Forecasts for United Kingdom using clade designations under retrospective vs real-time clade assignments** Forecasts from MLR fit to data generated using retrospective Nextstrain clade designations ('Current Nextclade') (A) and Nextstrain clade assignments available in Oct 2022 ('Real-time Nextclade') (B).

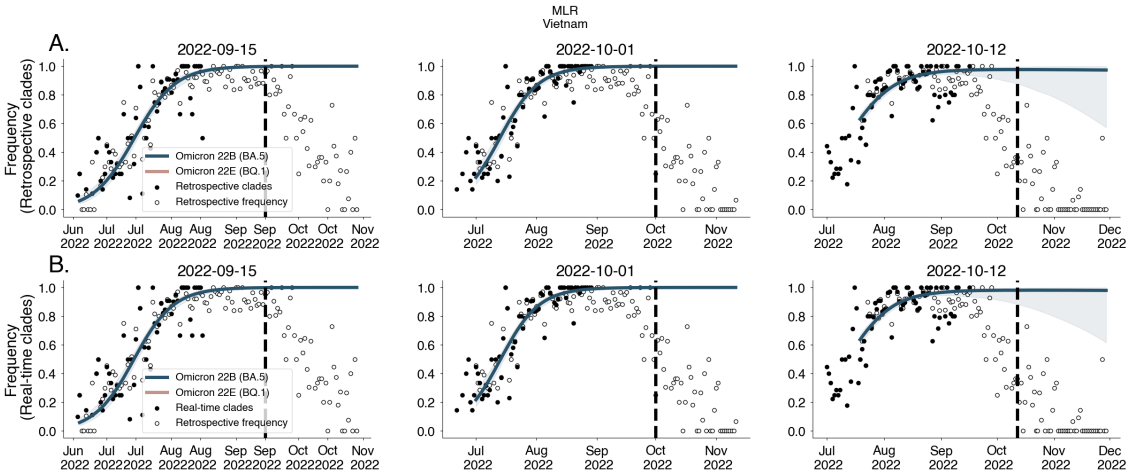

Figure B.17: **Forecasts for Vietnam using clade designations under retrospective vs real-time clade assignments** Forecasts from MLR fit to data generated using retrospective Nextstrain clade designations ('Current Nextclade') (A) and Nextstrain clade assignments available in Oct 2022 ('Real-time Nextclade') (B).

Appendix C

SUPPLEMENTARY MATERIALS FOR CHAPTER 4

C.1 Materials and Methods

Generating sequence counts We prepared sequence count data sets using the Nextstrain-curated SARS-CoV-2 sequence metadata [40] which is created using the GISAID EpiCoV database [48]. These sequences were tallied according to either their annotated Nextstrain clade or Pango lineage [4] depending on the data set to produce sequence count for each variant, for each day over the period of interest, and in each country analyzed.

Likelihood of sequence counts given frequencies The models discussed in this paper use observed counts of variant sequences to inform the underlying variant frequency in the population. This is accomplished using a multinomial likelihood, so that given count of sequences $S_v(t)$ of variant v at time t and total sequences $N(t)$ collected at time t, we have that

$$S_v(t) \sim \text{Multinomial}(N(t), f_v(t)), \tag{C.1}$$

where $f_v(t)$ is the frequency of variant v at time t. This is a simple model of sequence counts to frequencies and does not account for over-dispersion of sequence counts relative to a multinomial. However, all models can be extended to estimate and account for over-dispersion by replacing the above likelihood with a Dirichlet-Multinomial likelihood.

Approximate Gaussian processes for relative fitness estimation To generate smooth non-parametric estimates of variant growth rates, we develop a Gaussian process based model for relative fitnesses. That is, we model the relative fitness for each variant over time $\lambda_v(t)$ as a multivariate normal distribution:

$$\boldsymbol{\lambda}_v \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{C.2}$$

$$\boldsymbol{\Sigma}_{s,t} = K_\theta(s,t), \tag{C.3}$$

where K_θ is a potentially parameterized kernel function. This induces a structure on the covariance of the relative fitness values over time points s and t.

For computational efficiency, we implement a Hilbert Space Gaussian Process (HSGP) approximation instead of fitting V independent Gaussian processes. This approximation allow us to share basis functions between variants [70]. Under this approximation, the relative fitnesses are computed as

$$\lambda_v(t) \approx \sum_{j=1}^{m} S_\theta(\sqrt{\mu_j})^{1/2} \cdot \varphi_j(t) \cdot \beta_j, \tag{C.4}$$

where S_θ is the spectral density of the kernel K_θ, μ_j and φ_j are the eigenvalues and eigenfunctions of the Laplacian, and $\beta_j \sim \text{Normal}(0,1)$ [70]. Since the eigenvalues and eigenfunctions are shared across variants, this allows us to re-use values across variants, simplifying the computation to a matrix multiplication as

$$\boldsymbol{\lambda}_t = \boldsymbol{\Phi}_t \sqrt{\mathbf{S}_\theta} \boldsymbol{\beta}. \tag{C.5}$$

For the analyses in this paper, we use this approximate Gaussian process with a Matérn 5/2 kernel and shared hyperparameters across variants. We demonstrate this model for simulated data from Fig. 4.1 and show resulting relative fitnesses through time in Fig. C.1.

Correlations are insufficient for mechanism identification To assess how vaccination uptake affects the growth advantage of a variant with increased transmissibility, we simulate the spread of a more transmissible variant across populations with different initial past exposure and vaccination levels. This enables us to isolate the effects of transmissibility within different immunity landscapes, examining how relative fitness and growth advantage shift based on population vaccination coverage alone in the absence of immune escape. We

begin with the 2-variant SIR model described in Supplementary Text C.2.1. We simulate this model for 100 days with generation time $\tau = 1/\gamma = 3.0$ days, $R_{0,W} = 1.4$, $I_W(0) = 100$ individuals, $I_v(0) = 1$ individual, a 50% transmissibility increase $\rho = 0.5$, and no immune escape $\eta = 0.0$. We divide the period into early and late epidemic with the breakpoint being $t = 50$. In Fig. 4.3D-E, we estimate the log growth advantage for the variant in the early and full periods using a logit-linear model

$$\log\left(\frac{f_v(t)}{1 - f_v(t)}\right) = \beta t/\tau + \alpha, \tag{C.6}$$

where we take the model slope β to be our log growth advantage.

We repeat these simulations for a range of vaccination levels starting from 0% and ending at 65%.

Predicting epidemic growth rate from selective pressure The derivation of the selective pressure metric shows that the selective pressure can be a useful tool in predicting the epidemic growth rate. To develop a predictive model of epidemic growth rate using selective pressure, we begin by generating estimates of selective pressure and epidemic growth rate from a period with high sequencing and case surveillance.

We take sequence count and case count data from all states in the United States between January 2021 and November 2022. State-level daily case counts were obtained from USAFacts downloaded on August 7, 2024 at https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/.

Using the sequence counts, we compute selective pressure estimates from relative fitness and frequencies estimated with our approximate Gaussian process relative fitness model. From the case data, we derive the empirical growth rate using a 14-day moving average on case counts $\hat{C}_t$ and computing the empirical growth rate as $\hat{r}_t = \log(\hat{C}_t) - \log(\hat{C}_{t-1})$. We then use the past 28 days of selective pressure to predict the empirical growth rate.

We use a gradient boosting regressor model which is fit using a mean absolute error loss function. This model was selected as it achieved the minimal error via time series

cross-validation averaged across 10 splits among candidate models (Fig. C.9). The candidate models include linear regression, ridge regression, Lasso regression, random forests, and gradient-boosted trees as implemented in scikit-learn [63]. We additionally tune the hyperparameters of this model using grid search cross-validation.

We validate our model by comparing our predicted epidemic growth rates to held-out case data for US states, and additionally to estimates of the epidemic growth rates in England derived from data from the Office for National Statistics (ONS) Coronavirus Infection Survey [67]. Estimates of prevalence from the ONS Infection Survey were obtained for January 2022 to September 2022 from www.ons.gov.uk/peoplepopulationandcommunity/ healthandsocialcare/conditionsanddiseases/datasets/coronaviruscovid19infectionsurveydata. Epidemic growth rates are computed on this data in the same way as the state-level analysis.

Latent immune factor model We show that relative fitness dynamics can be explained by low-dimensional immunity when transmission dynamics are described with compartmental models (Supplementary Text C.2.1). This motivates a model to learn this low-dimensional structure that is inspired by latent-factor models. We start by assuming that the relative fitness of variant v at time t and in geographic location g can be described by D latent factors so that

$$\lambda_v^g(t) = \sum_{d=1}^{D} \eta_{v,d}\, \varphi_d^g(t). \tag{C.7}$$

As the structure here resembles Equation C.30, we call $\eta_{v,d}$ "pseudo-escape" of variant v from group d and φ_d^g "pseudo-immunity" group d in geographic location g. To make this more consistent with our intuition here, we model φ_d^g to be in $[0, 1]$ and model it as smoothly varying in time. We model $\text{logit}(\varphi_d^g)$ using 4th order splines with 6 knots placed uniformly over the time period modeled. Though we choose to model these latent factors with splines, other models would work here. For example, one alternative would be the approximate Gaussian processes described above. Additionally, in order to ensure identifiability of the parameter estimates, we fix some base variant v^* which fitness is defined relative to, so that

$\eta_{v^*,d} = 0$ for all $1 \leq d \leq D$. For the same reason, we fix the order of components, so that the components are numbered in decreasing order by their share in the arbitrarily defined base geography.

We apply this model to SARS-CoV-2 sequence counts in the period between March 2023 to March 2024 for 14 countries. To access the necessary number of immune dimensions, we vary the number of immune dimensions between $D = 2$ to $D = 12$. Looking at the loss for the latent factor model for increasing D, we choose $D = 10$ for our primary analysis by noting the point at which the loss function seems to stagnate with increasing D i.e. the "elbow" method (Fig. C.12).

We compare the distances between variant pairs in our estimated pseudo-escape space to distances in log2 titer. Using human titer data from Jian et al [45], we compute neutralization titer distances as the average of differences in log2 neutralization titers between pairs of variants for a cohort of individuals. This analysis is repeated among 1,000 bootstrapped samples to create a distribution of R^2 values (Fig. C.10). Additionally, we subset this by exposure history and repeat this analysis to find which exposure groups best explain distances in pseudo-escape space (Fig. C.11).

Data and code accessibility

Source code used to generate figures, model implementations, and sequence count data are available at github.com/blab/relative-fitness-mechanisms.

C.2 Supplementary Text

C.2.1 Exponentially growing populations to frequency dynamics

We consider a viral population consisting of V exponentially-growing variant viruses each with prevalence I_v. Defining the time-varying growth rate for the prevalence of variant v as $r_v(t)$, we can model the prevalence using an ordinary differential equation

$$\frac{dI_v}{dt} = r_v(t)I_v(t), \quad v = 1, 2, \ldots, V. \tag{C.8}$$

The above differential equation has a known solution in terms of the integral of the time-varying growth rate and initial prevalence,

$$I_v(t) = I_v(0) \exp\left(\int_0^t r_v(s)ds\right), \tag{C.9}$$

where $I_v(0)$ is the initial prevalence of variant v.

Now turning to the frequency dynamics of the population, we write the frequency of variant v in the population as $f_v(t) = I_v(t)/\sum_{u=1}^{V} I_u(t)$. This allows us to derive an ODE for variant frequency in terms of the variant growth rates using the quotient rule for differentiation

$$\frac{df_v}{dt} = f_v\left(\sum_{u=1}^{V}[r_v(t) - r_u(t)]f_u\right) \tag{C.10}$$

$$= f_v\left(r_v(t) - \sum_{u=1}^{V} r_u(t)f_u\right). \tag{C.11}$$

This system of differential equations resembles a logistic growth equation and can be shown to have the following solution in terms of the initial frequencies $f_v(0)$ and the variant growth rates

$$f_v(t) = \frac{f_v(0)\exp(\int_0^t r_v(s)ds)}{\sum_{u=1}^{V} f_u(0)\exp(\int_0^t r_u(s)ds)}. \tag{C.12}$$

The above representation of the variant frequency will serve as a centerpiece for many of the arguments to follow. We see that by tracking the rate at which variant viruses are spreading, we can construct the corresponding frequency dynamics without knowing the absolute prevalence of any variant.

Relative frequency and relative fitness Using the above equation for the variant frequencies, we can write the relative frequency of variant v over u as $x_{v,u}(t) = f_v(t)/f_u(t)$ to see

$$x_{v,u}(t) = \frac{f_v(t)}{f_u(t)} = \frac{f_v(0)}{f_u(0)} \exp\left(\int_0^t [r_v(s) - r_u(s)]ds\right) \tag{C.13}$$

$$= x_{v,u}(0) \exp\left(\int_0^t \lambda_{v,u}(s)ds\right). \tag{C.14}$$

Notice this relative frequency change depends on the initial relative frequencies and the *relative fitness* $\lambda_{v,u}(t) = r_v(t) - r_u(t)$ of v over u. This relative fitness has the same units as the exponential growth rate (e.g. per day). Using the definition of relative fitness, we can notice that

$$\lambda_{v,u}(t) = r_v(t) - r_u(t) = \frac{d}{dt}\left[\log\left(x_{v,u}(t)\right)\right] = -\lambda_{u,v}(t). \tag{C.15}$$

We can see that there is a symmetry in the relative fitnesses and that the associated frequency dynamics depend on the differences between relative fitnesses. This suggests that absolute fitness (in terms of the growth of infections) may not be inferable from frequencies alone. This definition of relative fitness becomes essential in describing various existing modeling approaches for frequency dynamic data and motivates possible extensions since we can represent these models as having the form:

$$f_v(t) = \frac{f_v(0) \exp(\int_0^t \lambda_{v,v^*}(s)ds)}{\sum_{u=1}^V f_u(0) \exp(\int_0^t \lambda_{u,v^*}(s)ds)}, \tag{C.16}$$

where the growth rate of v is expressed as relative to an arbitrary pivot variant v^*.

Cumulative relative-fitness and frequency change Above we saw that within our framework frequency change over time intervals depends only on the cumulative relative fitness over time intervals $\Lambda_{v,u}(0,t) = \int_0^t \lambda_{v,u}(s)ds$. We can then characterize approaches for modeling frequency change in terms of how they represent, estimate, and forecast these relative fitnesses. This framework includes various existing methods for analyzing frequency data such as the seasonal influenza forecasting models of Lässig and Łuksza [55] and Huddleston et al [42], multinomial logistic regression for frequency estimation [5] and the SARS-CoV-2 mutational fitness model of Obermeyer et al [60].

Though this framework can be used to describe existing statistical methods for frequency modeling, it is also applicable to traditional compartmental models of epidemics. In fact, applying these ideas to compartmental models enables to see how mechanistic assumptions on the transmission process determine relative fitness of variant viruses.

Two-strain SIR For simplicity, we will begin by analyzing a two-strain SIR model in which a variant virus v can differ from wildtype virus wt by increased intrinsic transmissibility (via η_T) and immune escape against wild-type immunity (via η_E). This gives a system of 5 ordinary differential equations

$$\frac{dS}{dt} = -\beta S I_W - \beta\rho S I_v, \tag{C.17}$$

$$\frac{dI_W}{dt} = \beta S I_W - \gamma I_W, \tag{C.18}$$

$$\frac{dI_v}{dt} = \beta(1+\rho) S I_v + \beta(1+\rho)\eta\varphi_W I_v - \gamma I_v, \tag{C.19}$$

$$\frac{d\varphi_W}{dt} = \gamma I_W - \beta(1+\rho)\eta\varphi_W I_v, \tag{C.20}$$

$$\frac{d\varphi_v}{dt} = \gamma I_v, \tag{C.21}$$

where I_W denotes wild-type prevalence, I_v denotes variant prevalence and φ_W denotes immunity derived from wild-type infection. In this model, the variant virus can infect both susceptible individuals S and individuals with immunity to wild-type virus φ_W. Increased

intrinsic transmissibility increases the baseline transmission rate from β in wild-type to $\beta\rho$ in the variant virus and immune escape increases the transmission rate against those with wildtype immunity, so that the at-risk population is $\eta\varphi_W$.

Writing that $r_W(t) = \beta S - \gamma$ and $r_v(t) = \rho\beta S + \beta(1+\rho)\eta\varphi_W - \gamma$, we can then write the relative fitnesses as:

$$\lambda_{v,W}(t) = \rho\beta S(t) + (1+\rho)\eta\beta\varphi_W(t). \tag{C.22}$$

From this representation of relative fitness, we can see that given fixed increases to overall transmission ($\rho > 0$) or immune escape ($\eta > 0$), the observed fitness boost at the level of variant relative fitness still depends on the proportion of the population at risk for infection.

n-strain SIR This model can also be extended to an n-strain SIR model where each variant strain v_i with $2 \leq i \leq n$ is described by its own advantage parameters $\theta_i = (\rho^{(i)}, \eta^{(i)})$ relative to the wildtype ($\theta_W = \theta_1 = (0,0)$)

$$\frac{dS}{dt} = -\beta S I_W - \beta(1+\rho) S I_{v_i} \tag{C.23}$$
$$\frac{dI_W}{dt} = \beta S I_W - \gamma I_W \tag{C.24}$$
$$\frac{dI_{v_i}}{dt} = \beta(1+\rho_i) S I_{v_i} + \beta(1+\rho_i)\eta_i\varphi_W I_{v_i} - \gamma I_{v_i} \tag{C.25}$$
$$\frac{d\varphi_W}{dt} = \gamma I_W - \beta(1+\rho_i)\eta_i\varphi_W I_{v_i} \tag{C.26}$$
$$\frac{d\varphi_{v_i}}{dt} = \gamma I_{v_i}, \quad i \in \{2, \ldots, n\}. \tag{C.27}$$

In this formulation, the variant viruses compete only for susceptible population and those with previous wild-type infection. This formulation can be generalized to allow for competition between all variants for any exposure history and will be discussed in the following sections. In Fig. 4.1, we implement and simulate a 3-strain model with wildtype as above, an escape variant E with $\theta_2 = (0, \eta)$, and a transmissibility increase variant T with $\theta_3 = (\rho, 0)$.

Models of immune escape against heterogeneous backgrounds We'll now consider a model where all hosts are assumed to fall into one of B immune backgrounds φ_b for $b = 1, \ldots, B$. We assume that infection by each variant v then leaves recovered hosts in the corresponding immune background of the most recent infection b_v. Variant transmission then occurs via immune escape against a background leading to a matrix of escape rates $\boldsymbol{\eta} = \eta_{v,b}$ for variants v and background b.

We can then write the system of ordinary differential equations as

$$\frac{dI_v}{dt} = \beta \sum_{1 \leq b \leq B} \eta_{v,b} \varphi_b I_v - \gamma I_v, \quad v = 1, \ldots, V \tag{C.28}$$

$$\frac{d\varphi_b}{dt} = -\beta \sum_{1 \leq v \leq V} \eta_{v,b} \varphi_b I_v + \sum_{v:\ b_v = b} \gamma I_v. \tag{C.29}$$

With this model, susceptible and recovered compartments in the standard SIR model can be thought of as immune backgrounds. This allows us to represent the standard SIR model as $S = \varphi_S$, $I = I_W$, $R = \varphi_W$ and $\eta_{W,S} = 1, \eta_{W,W} = 0$ and $b_W = W$. We can also think of the two-strain SIR with $\rho = 1$ as a special case of this model where we set $S = \varphi_S, \eta_{W,S} = 1, \eta_{W,W} = 0, \eta_{v,S} = 1, \eta_{v,W} = \eta$ and keep all other parameters the same.

With this formulation of immune escape, we can then write the relative fitnesses in terms of the escape rates $\eta_{v,b}$ and the immune background proportions φ_b as

$$\lambda_{v,u}(t) = \beta \sum_{1 \leq b \leq B} (\eta_{v,b} - \eta_{u,b}) \varphi_b(t). \tag{C.30}$$

Under this model of immune escape, we can see relative fitness among variants can be decomposed into differences in immune escape among immune backgrounds within a population. Due to the dependence here on the proportion of each immune background in determining fitness, this suggests that the overall distribution of susceptibility to strains is potentially an important consideration when translating individual-level measures of immune escape to population-level estimates of variant fitness. Understanding the size and complexity of this immune space may therefore be useful for parameterization and forecasting of variant frequencies. However, the extent to which modeling this complexity affects estimates of

relative fitness also depends on how quickly the distribution of immune backgrounds change i.e. $\frac{d\varphi_b}{dt}$.

Though the derivation above uses a simplified model of using most recent infection to sort individuals into an immune group, we show that a more complicated model that accounts for the entire exposure history of the host also gives a similar decomposition to relative fitness in Supplementary Text C.2.3.

C.2.2 Revisiting existing models for frequency growth

Using the theory developed for exponentially-growing variant populations, we now re-visit existing methods for modeling viral frequency dynamics.

Multinomial Logistic Regression We begin with multinomial logistic regression (MLR) with fixed relative fitness. This model can be written as

$$f_v(t) = \frac{f_v(0)\exp(\lambda_v t)}{\sum_u f_u(0)\exp(\lambda_u t)}, \tag{C.31}$$

where $f_v(t)$ is the frequency of variant v at time t and λ_v is the relative fitness of variant v. This provides estimates of the relative fitness compared to some reference strain u^* for which $\lambda_{u^*} = 0$. In this model, initial frequencies $f_v(0)$ and relative fitness λ_v are estimated from frequency dynamics. Converting this estimate to an estimate of transmission advantage (relative effective reproduction number) requires assuming a delta distribution of the generation time [87].

Comparing this to equation C.22, we can see this model of fixed relative fitness results from assuming that the at-risk populations are constant over-time. This assumption is useful since it requires no outside knowledge of the at-risk population and relative infection rates, though this may be less useful for longer forecasts or when there is large turnover in at-risk populations due to infection.

Fitness models of seasonal influenza Motivated by the observed antigenic evolution of seasonal influenza, Lässig and Łuksza [55] and Huddleston et al [42] approximate the

cumulative relative fitness between influenza seasons on the level of individual strains as

$$\Lambda_{v,u}(t+\Delta t,t) = (\beta_1 x_{v,1} + \cdots + \beta_p x_{v,p})\Delta t = (\boldsymbol{\beta}\cdot\mathbf{x}_v)\Delta t, \tag{C.32}$$

where the relative fitness is determined by strain-specific predictors $\mathbf{x}_v$ and the regression parameter $\boldsymbol{\beta}_v$ are estimated.

This formulation fits neatly into the framework we've developed as the cumulative fitness here can be written as the integral of a relative fitness $\lambda_{v,u} = \boldsymbol{\beta}\cdot\mathbf{x}_v$ over the time period of interest:

$$\Lambda_{v,u}(t+\Delta t,t) = \int_t^{t+\Delta t} \lambda_{v,u}(s)ds \ = \int_t^{t+\Delta t} (\boldsymbol{\beta}\cdot\mathbf{x}_v)ds. \tag{C.33}$$

Therefore, these models can be thought as regression-based predictors of relative fitness where frequency and external covariates contribute to estimated relative fitness.

C.2.3 Relative fitness for full immune history models

We show that the simple background model is consistent with an expanded immune history model. Beginning with the model from Lazebnik and Bunimovich-Mendrazitsky 2022 [53], we consider the differential equation for the individuals with strain infection history J and current infecting strain i $R_J I_i$

$$\frac{dR_J I_i}{dt} = -\gamma_{J,i} R_J I_i + \beta_{J,i} R_J \sum_{K\in P(M), i\notin K} R_K I_i. \tag{C.34}$$

Here, infection can occur from any individual infected with strain i assuming their past immune history does not include i and the infected are any recovered individual with immune history J R_J. To compute the strain growth rate, we can sum over all possible immune

histories for individuals infected with strain i, so that

$$\frac{dI_i}{dt} = \sum_{J \in P(M), i \notin J} \frac{dR_J I_i}{dt} \tag{C.35}$$

$$= -\gamma_i I_i + \sum_{J \in P(M), i \notin J} \beta_{i,J} R_J \sum_{K \in P(M), i \notin K} R_K I_i \tag{C.36}$$

$$= -\gamma_i I_i + \sum_{J \in P(M), i \notin J} \beta_{i,J} R_J I_i \tag{C.37}$$

$$= \left(-\gamma_i + \sum_{J \in P(M), i \notin J} \beta_{i,J} R_J \right) I_i \tag{C.38}$$

$$= \left(-\gamma + \beta \sum_{J \in P(M), i \notin J} \eta_{i,J} R_J \right) I_i. \tag{C.39}$$

$$\tag{C.40}$$

Assuming that the transmission rate can be decomposed as a base transmission rate β and a strain i and immune history J specific escape rate $\eta_{i,J}$ and that the recovery rate is constant, we notice this is identical to our previous immune background model. Therefore, our relative fitnesses simplify to

$$\lambda_{i,j} = \beta \sum_{B \in P(M)} (\eta_{i,B} - \eta_{j,B}) R_B, \tag{C.41}$$

where for simplicity we define $\eta_{v,B} = 0$ if $v \in B$.

C.2.4 Selective pressure and contribution to epidemic growth rates

In this section, we derive our selective pressure metric $\psi(t)$ and show how it contributes to the overall epidemic growth rate in the population.

Beginning again from our assumption of inhomogeneous exponential growth, we can write a differential equation for the total prevalence $I(t) = \sum_v I_v(t)$,

$$\frac{dI}{dt} = \sum_v \frac{dI_v}{dt} = \sum_v r_v(t) I_v(t) \tag{C.42}$$

$$= \left(\sum_v r_v(t) f_v(t) \right) I(t), \tag{C.43}$$

where we've used that $I_v(t) = f_v(t)I(t)$. This allows us to see that $\overline{r}(t) = \sum_v r_v(t) f_v(t)$ is the average growth rate of the prevalence. Re-writing the average in terms of some base exponential growth rate and the relative fitnesses so that $r_v(t) = \lambda_v(t) + r_W(t)$, we get that $\overline{r}(t) = \sum_v \lambda_v(t) f_v(t) + r_W(t)$. We can simplify this by writing $\overline{r}(t) = \overline{\lambda}(t) + r_W(t)$ where $\overline{\lambda}(t) = \sum_v \lambda_v(t) f_v(t)$ is the mean fitness of the population. We can now look at the rate of change in the average growth rate by taking its derivative

$$\frac{d\overline{r}}{dt} = \frac{dr_W}{dt} + \frac{d\overline{\lambda}}{dt} \tag{C.44}$$

$$= \frac{dr_W}{dt} + \sum_v \left[\frac{d\lambda_v}{dt} f_v(t) + \lambda_v(t) \frac{df_v}{dt} \right] \tag{C.45}$$

$$= \frac{dr_W}{dt} + \sum_v \left[\frac{d\lambda_v}{dt} f_v + \lambda_v f_v (\lambda_v - \overline{\lambda}) \right] \tag{C.46}$$

$$= \frac{dr_W}{dt} + \sum_v \frac{d\lambda_v}{dt} f_v(t) + \sum_v \lambda_v (\lambda_v - \overline{\lambda}) f_v \tag{C.47}$$

$$= \frac{dr_W}{dt} + \sum_v \frac{d\lambda_v}{dt} f_v(t) + \sum_v \lambda_v^2 f_v - \overline{\lambda} \sum_v \lambda_v f_v \tag{C.48}$$

$$= \frac{dr_W}{dt} + \mathbb{E}_{f(t)} \left[\frac{d\lambda_v}{dt} \right] + \mathrm{Var}_{f(t)}[\lambda_v]. \tag{C.49}$$

Here, we've written the last line in terms of expectations relative to sampling according to the frequency distribution. This shows us that the change in the average growth rate of the epidemic can be written in terms of the growth rate of the pivot category, the mean rate of change in the relative fitness, and the variance of the relative fitnesses. We will call terms which can be computed in terms of quantities derived from frequencies alone the selective pressure

$$\psi(t) = \mathbb{E}_{f(t)}\left[\frac{d\lambda_v}{dt}\right] + \mathrm{Var}_{f(t)}[\lambda_v]. \tag{C.50}$$

We can use this idea to directly write the prevalence in terms of the selective pressure and the base growth rate. First, we define a cumulative selective pressure

$$\Psi(t) = \int_0^t \psi(s)ds. \tag{C.51}$$

We can then use this to reconstruct the relative incidence

$$\frac{I(t)}{I(0)} = \exp\left(\int_0^t \overline{r}(s)ds\right) \tag{C.52}$$

$$= \exp\left(\int_0^t [r_W(s) + \Psi(s)]ds\right). \tag{C.53}$$

C.3 Supplementary Figures

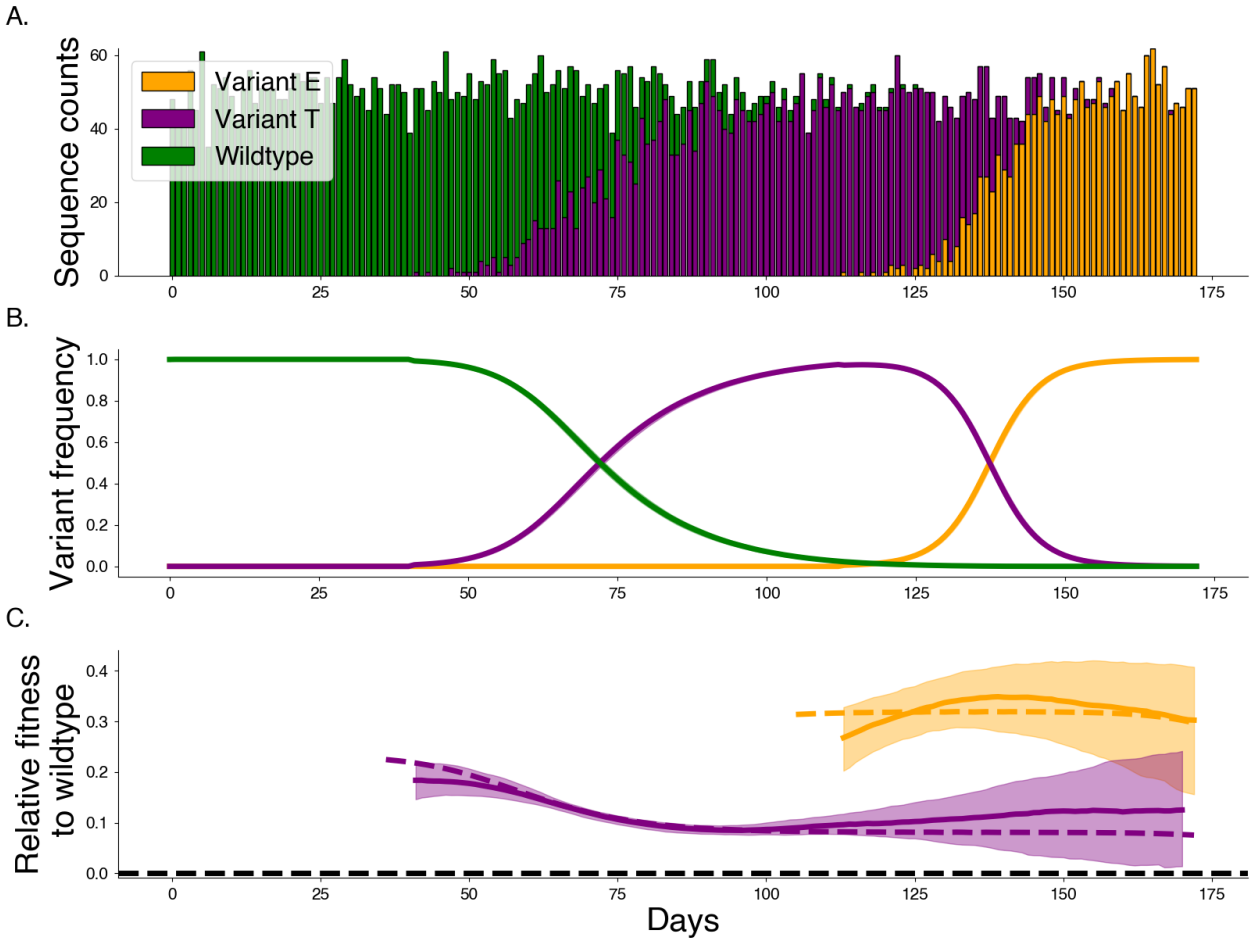

Figure C.1: **Estimating relative fitness with Gaussian processes.** Gaussian processes allow us a non-parametric estimate of the relative fitness for variants through time. This figure uses Gaussian processes to model the 3 variant example shown in Fig. 4.1. A. Synthetic sequence counts generated using a multinomial distribution with frequencies from Fig. 4.1C. B. Frequencies and posterior frequencies according to Gaussian process model. Intervals show the 80% credible interval. C. Relative fitnesses. Dashed line shows true relative fitnesses from underlying mechanistic model.

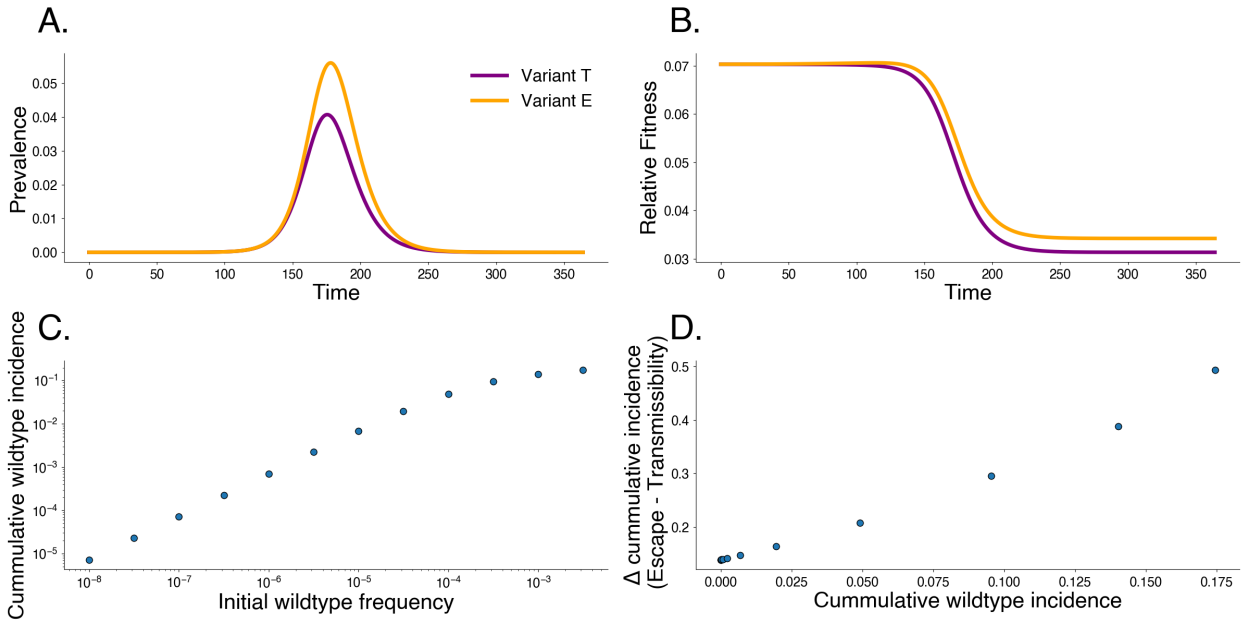

Figure C.2: **Differences in fitness mechanisms impact frequency and prevalence in the short-term.** Comparing simulations from two independent two-variant systems with either an escape variant E (orange) or a transmissibility variant T (purple). We fix the initial relative fitness for the two variants using Equation 4.4 and simulate dynamics for 365 days. A. The prevalence for the variants. B. The relative fitness from the variants. C. The cumulative wildtype incidence as a function of the initial wildtype frequency. D. The difference between the cumulative incidence between the escape variant and the transmissibility variant as a function of wildtype incidence.

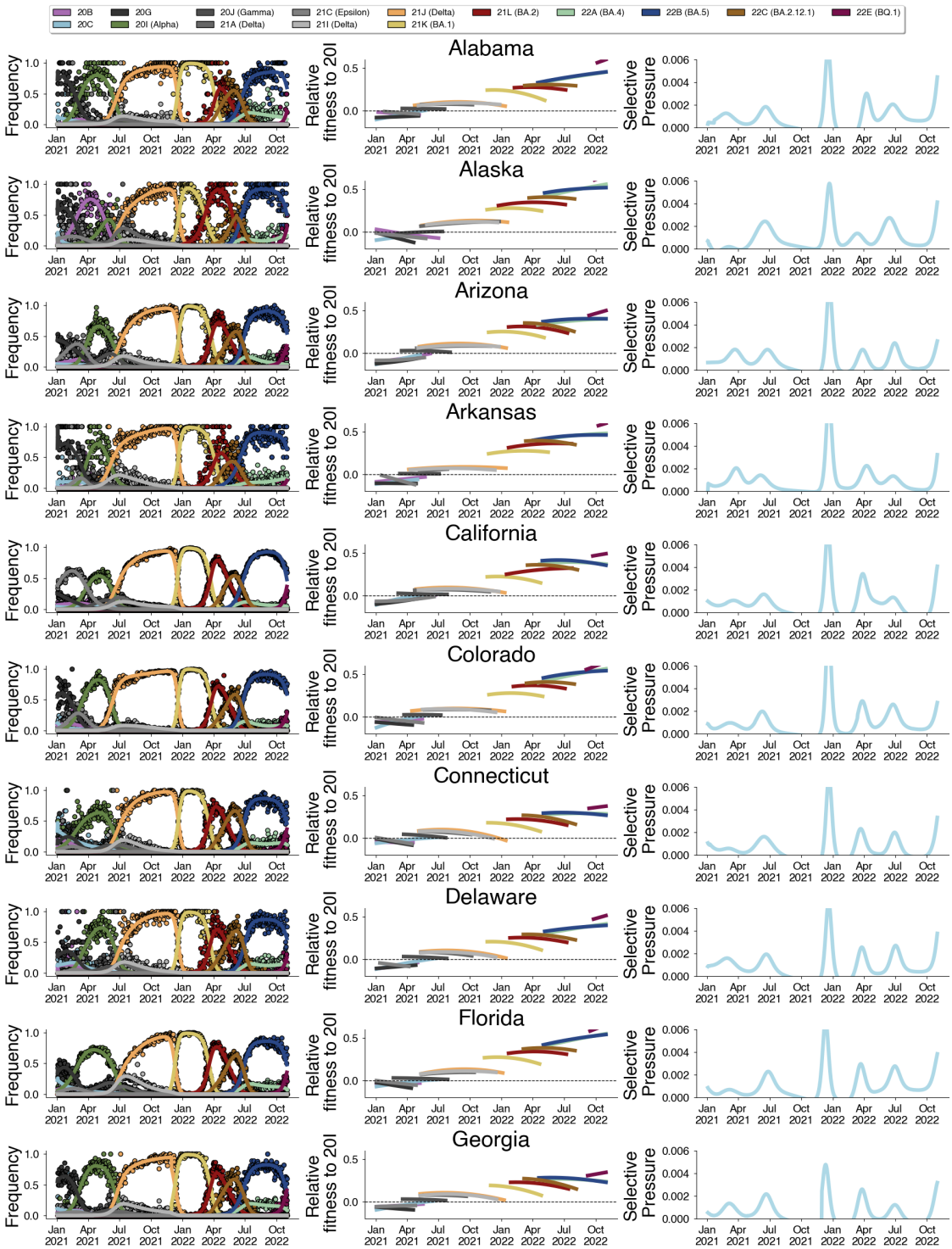

Figure C.3: **Estimated variant frequencies, relative fitnesses, and selective pressure. Alabama through Georgia.**

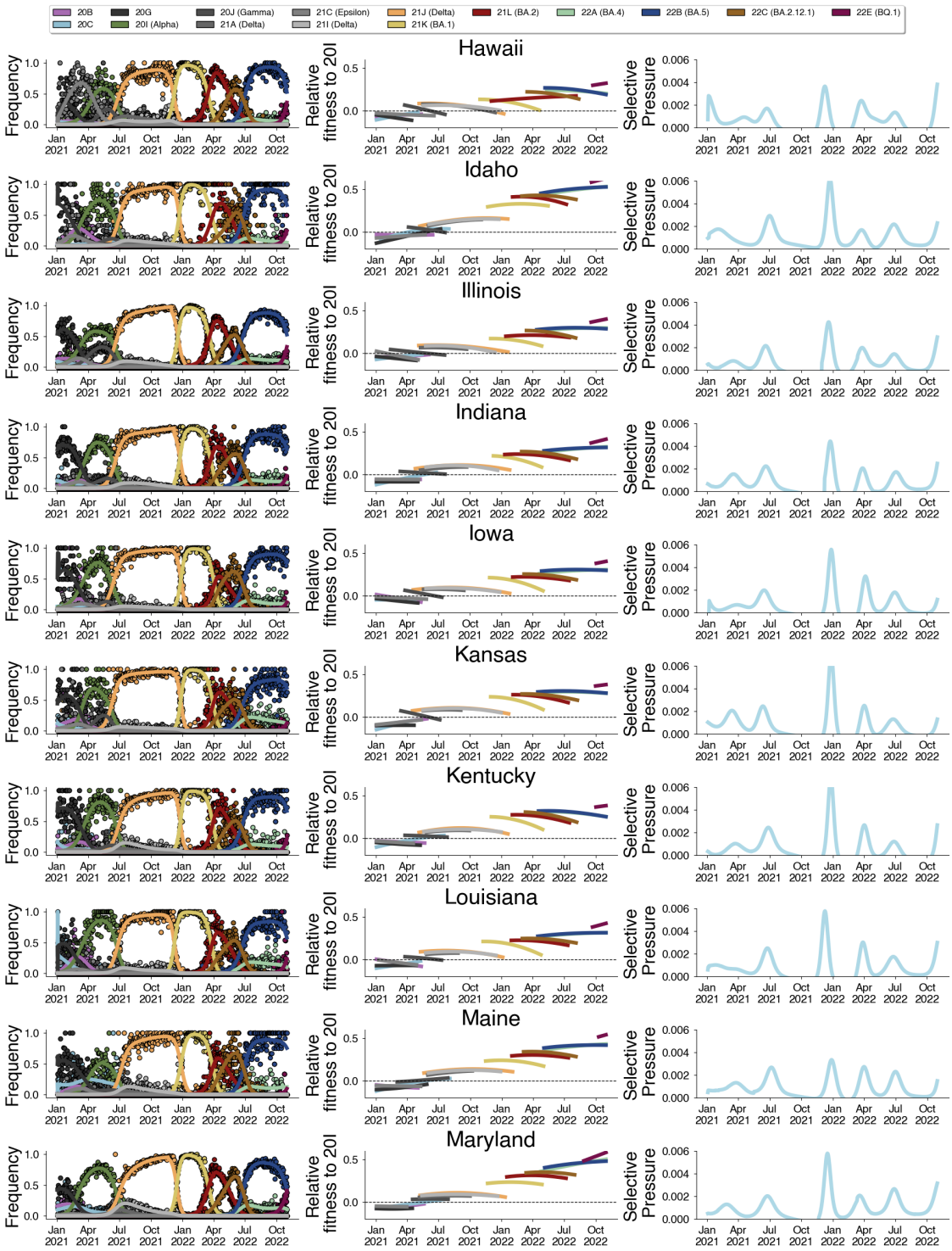

Figure C.4: **Estimated variant frequencies, relative fitnesses, and selective pressure. Hawaii to Maryland.**

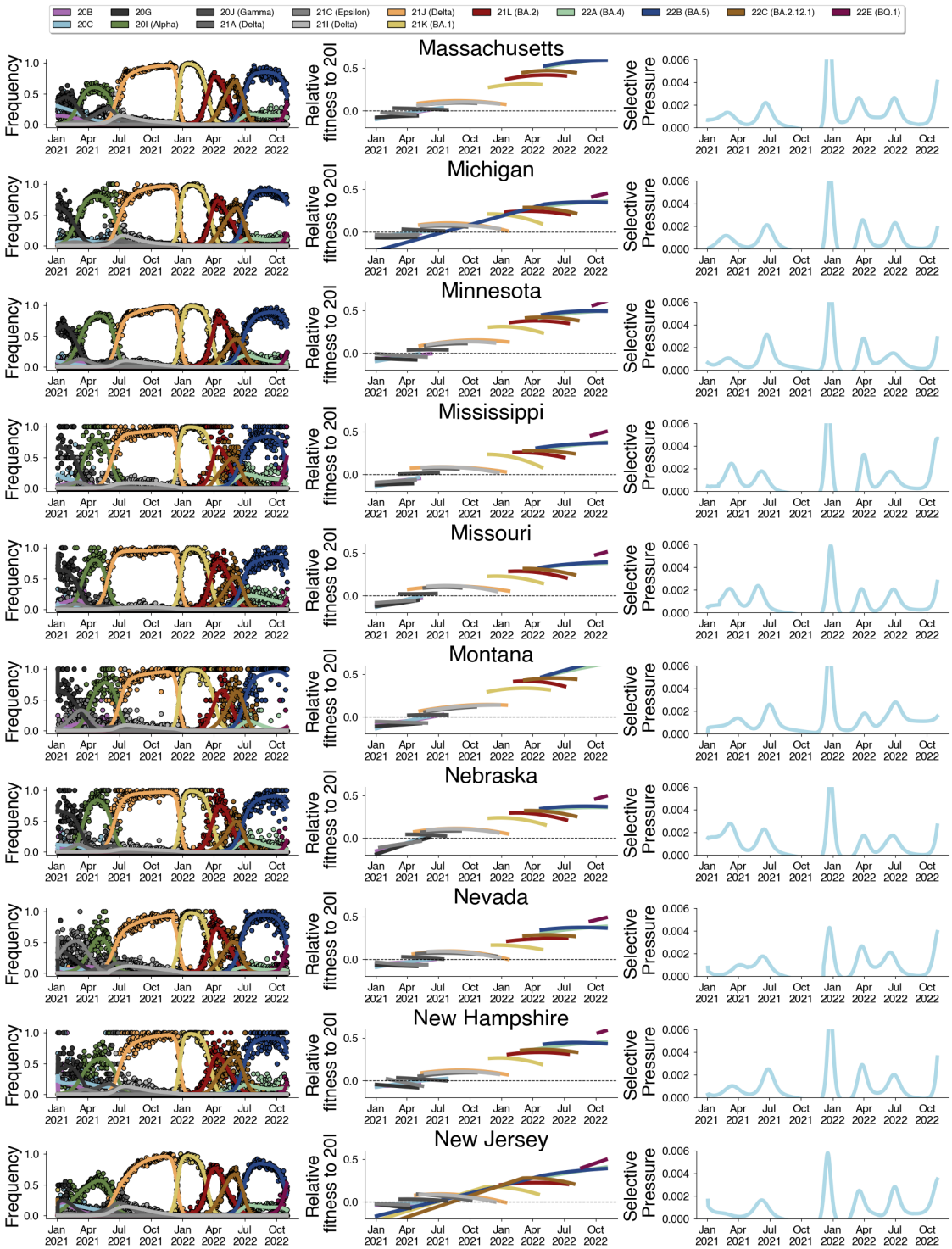

Figure C.5: **Estimated variant frequencies, relative fitnesses, and selective pressure. Massachusetts to New Jersey.**

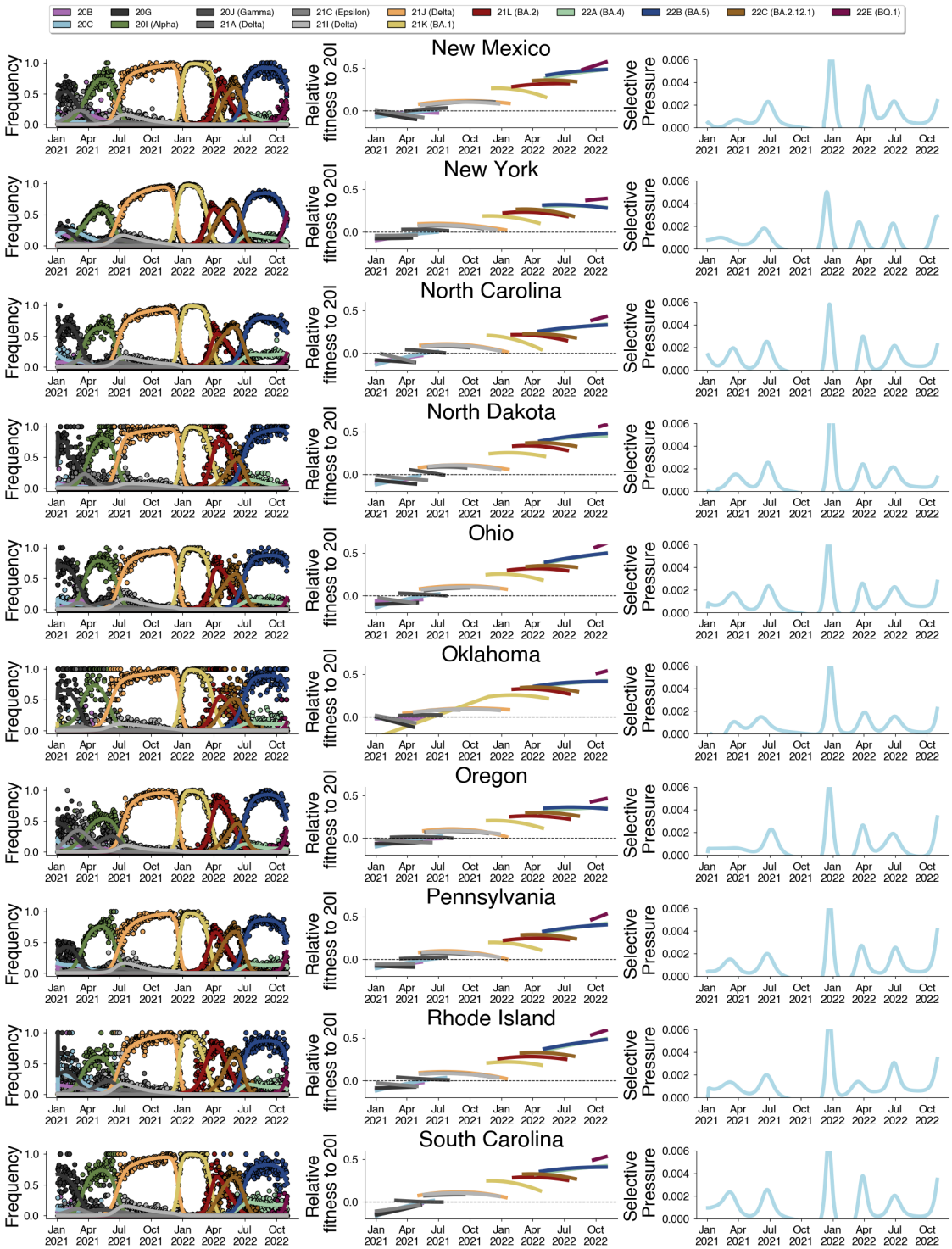

Figure C.6: **Estimated variant frequencies, relative fitnesses, and selective pressure. New Mexico to South Carolina.**

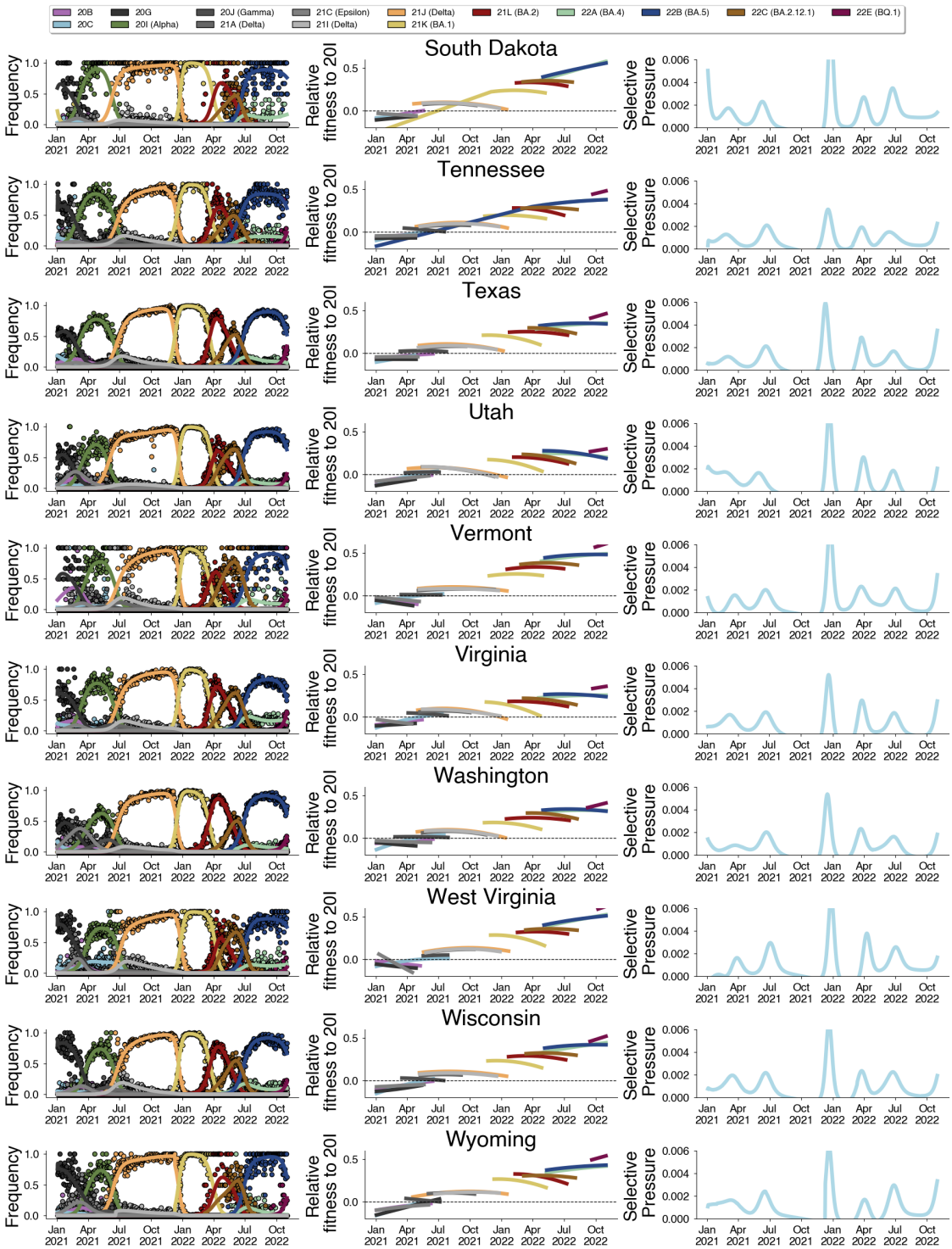

Figure C.7: **Estimated variant frequencies, relative fitnesses, and selective pressure. South Dakota to Wyoming.**

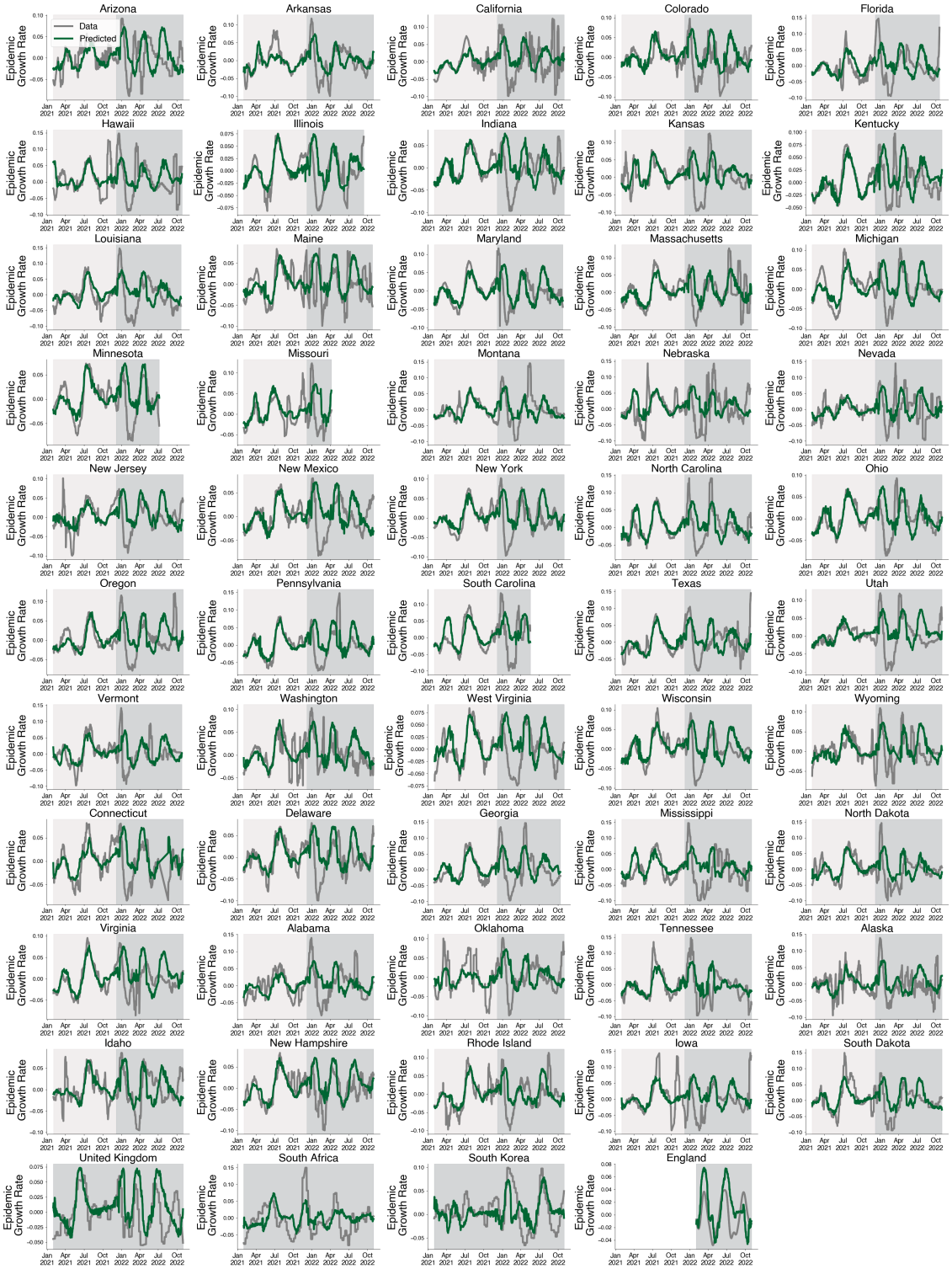

Figure C.8: **Predictions for empirical growth rate using selective pressure for all locations.**

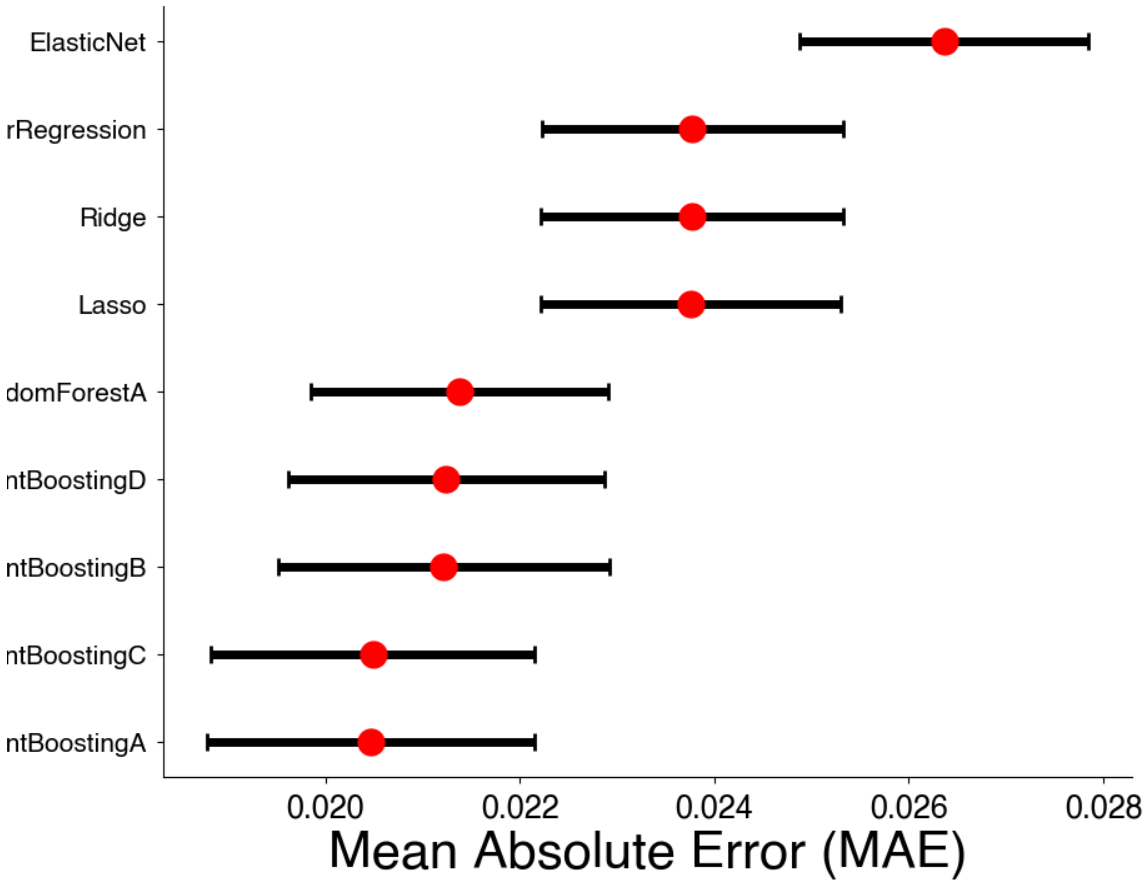

Figure C.9: **Cross-validation error by model** We compare the errors between models fit on 10 time series cross-validation splits.

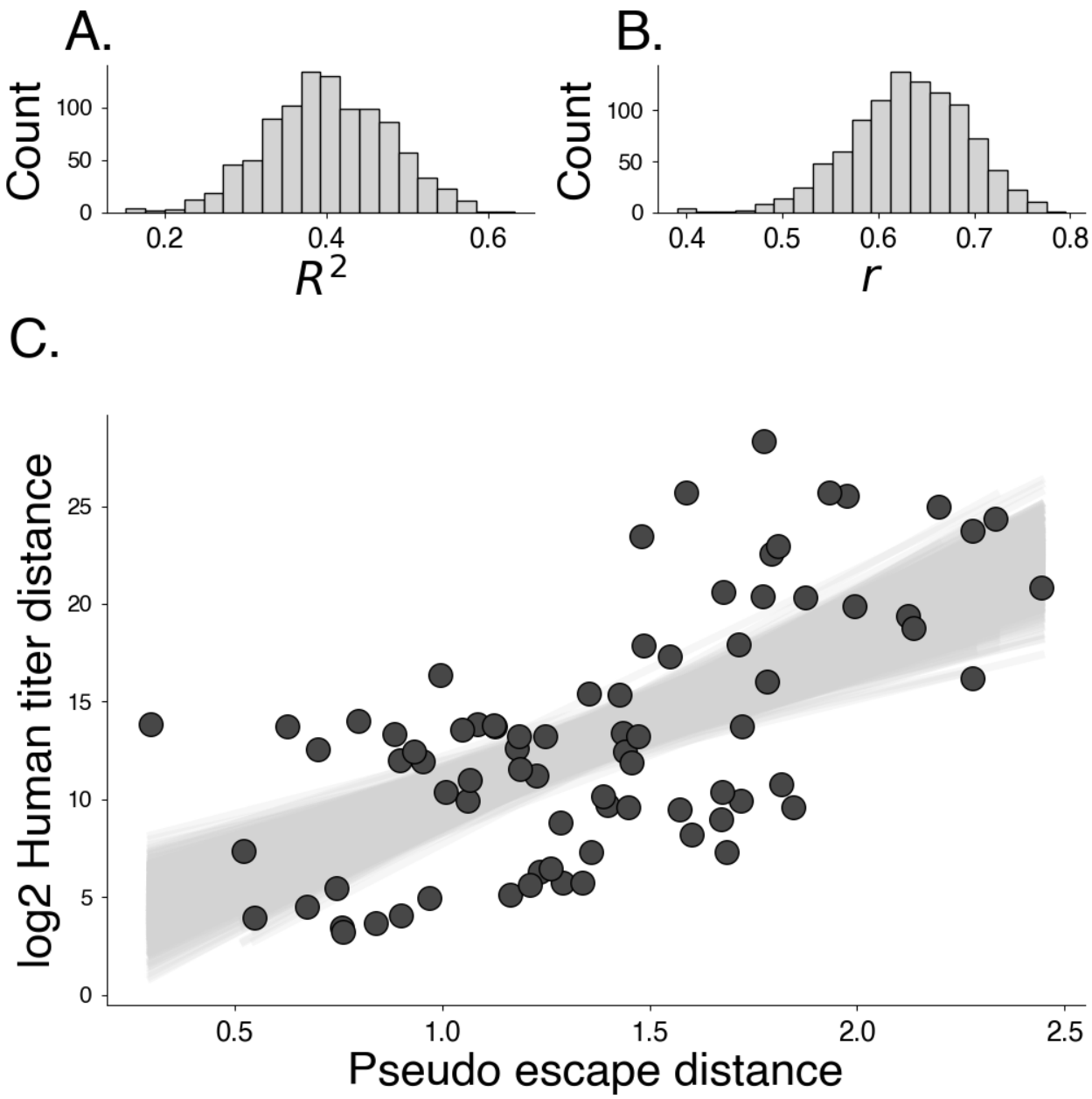

Figure C.10: **Bootstrapping pseudo-immune distance and human titer distance analysis** ($N_{\mathbf{replicate}} = 1,000$).

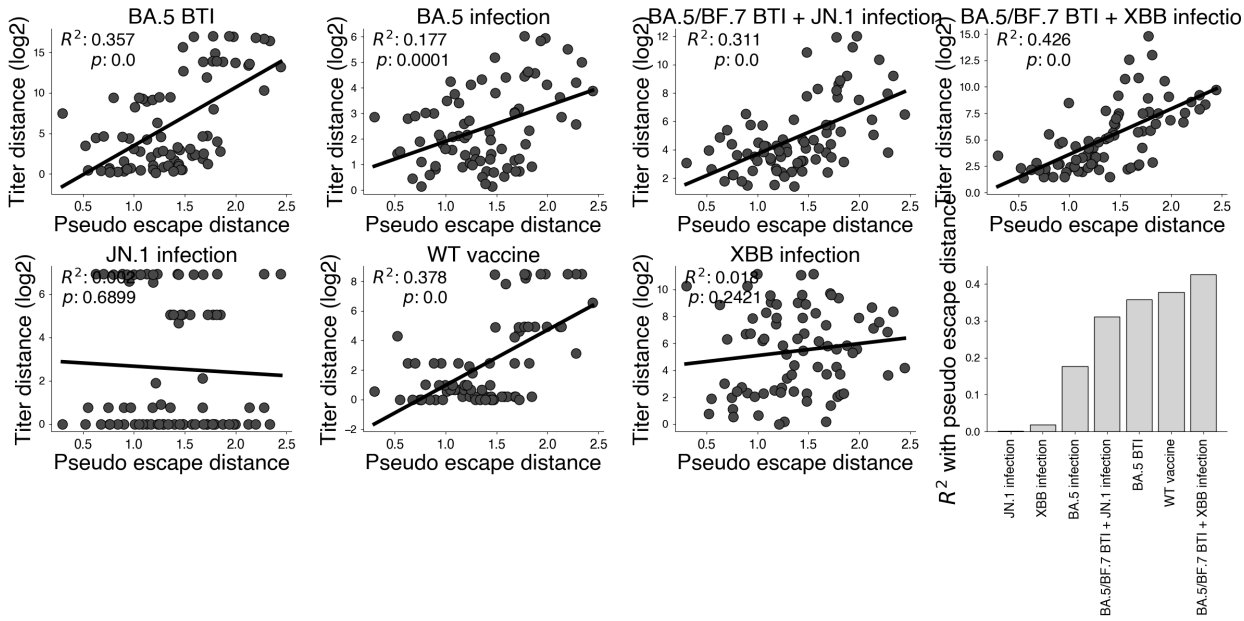

Figure C.11: **Comparing pseudo-escape distance and titer distance between exposure groups.**

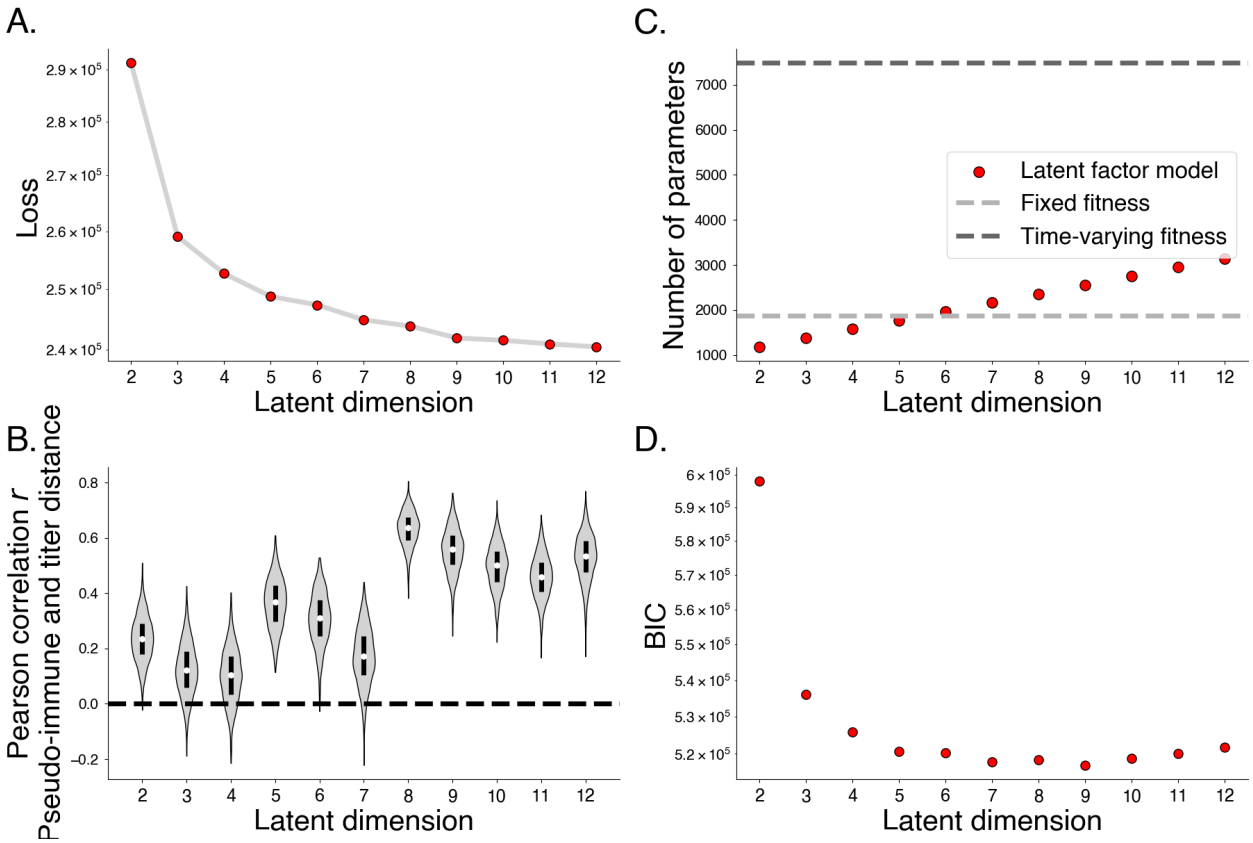

Figure C.12: **Comparing latent factor model by number of latent immune dimensions.** A. Maximum a posteriori loss by number of latent immune dimensions. B. Spearman correlation between pseudo-immune and titer distance by number of latent immune dimensions. C. Number of parameters by number of latent immune dimensions. D. Bayesian Information Criterion (BIC) by number of latent immune dimensions.

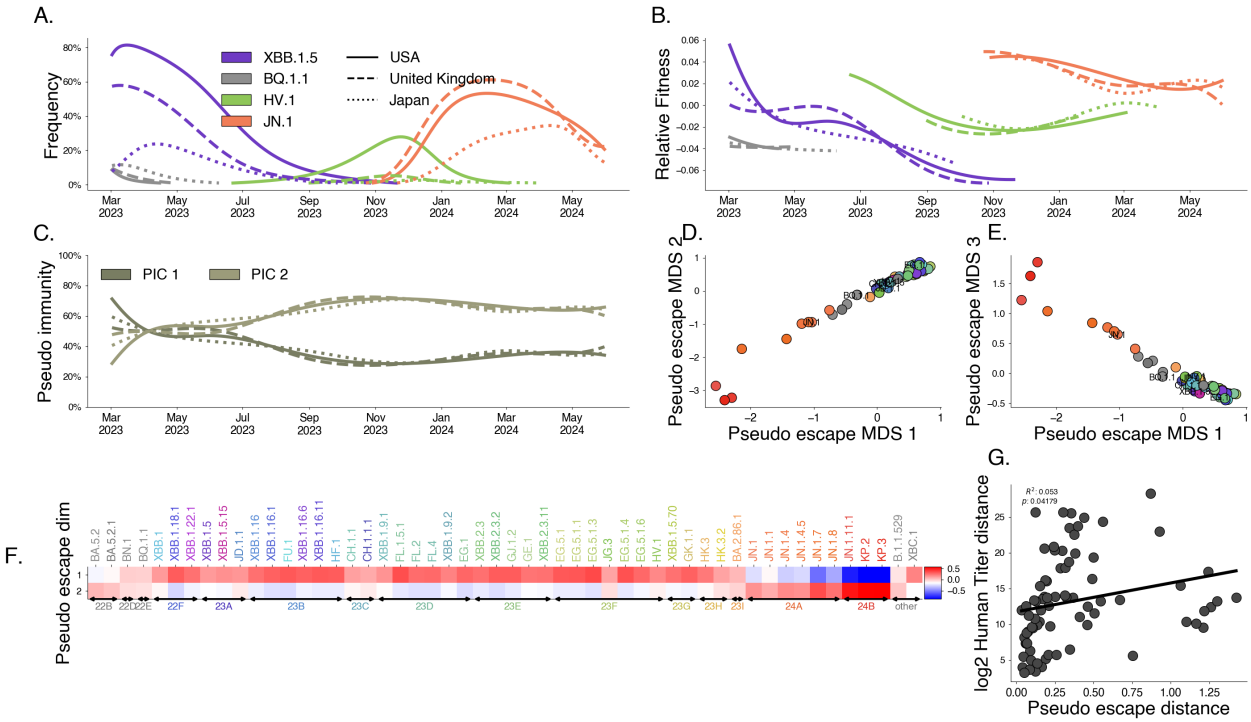

Figure C.13: **Latent factor model with $D = 2$ pseudo immune dimensions.**

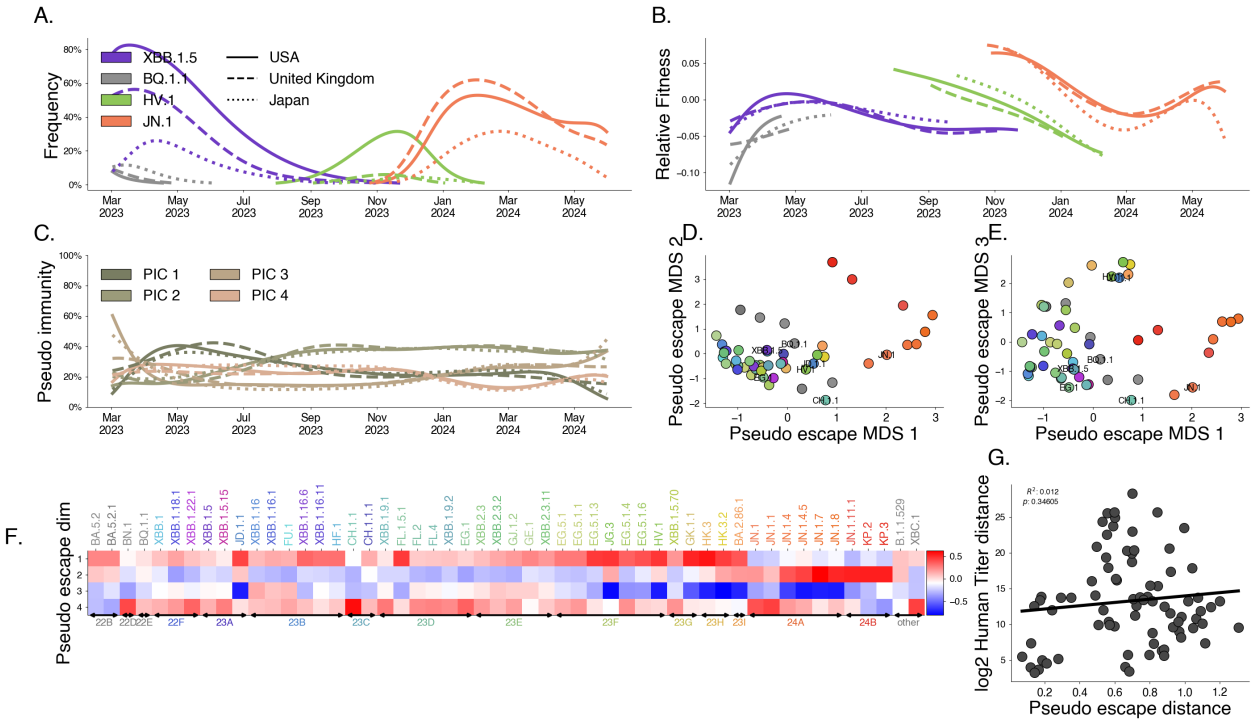

Figure C.14: **Latent factor model with $D = 4$ pseudo immune dimensions.**

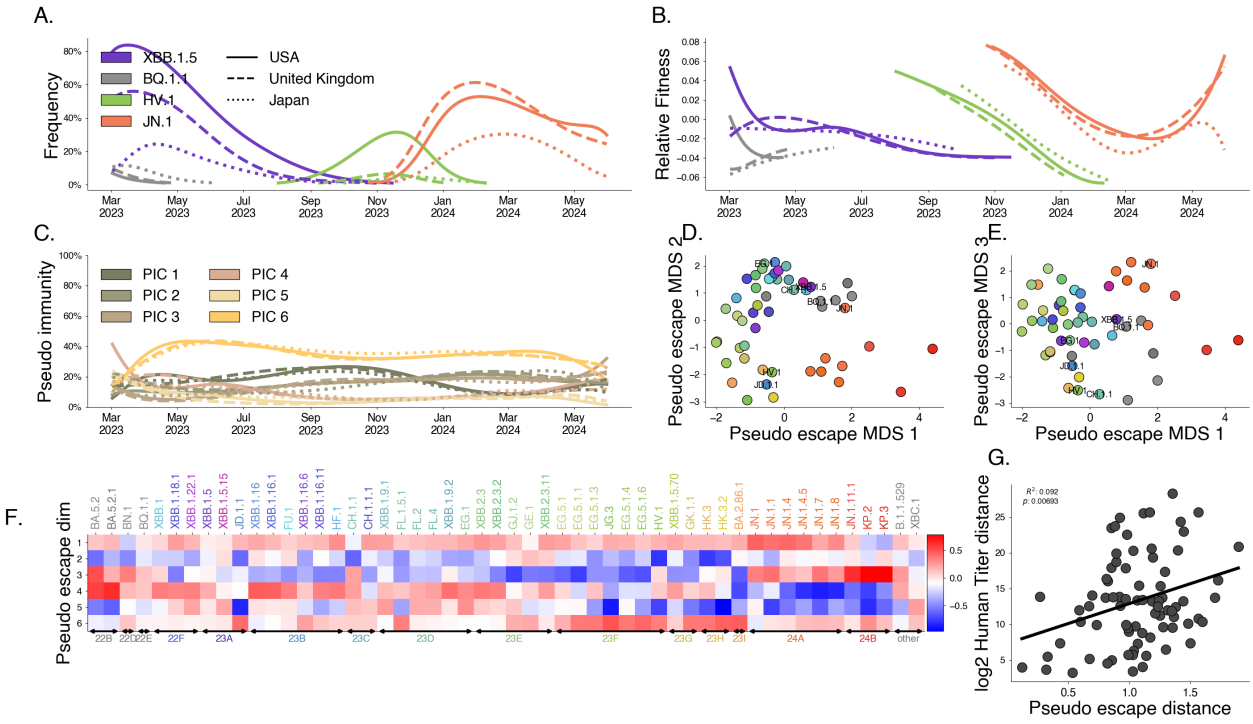

Figure C.15: **Latent factor model with $D = 6$ pseudo immune dimensions.**

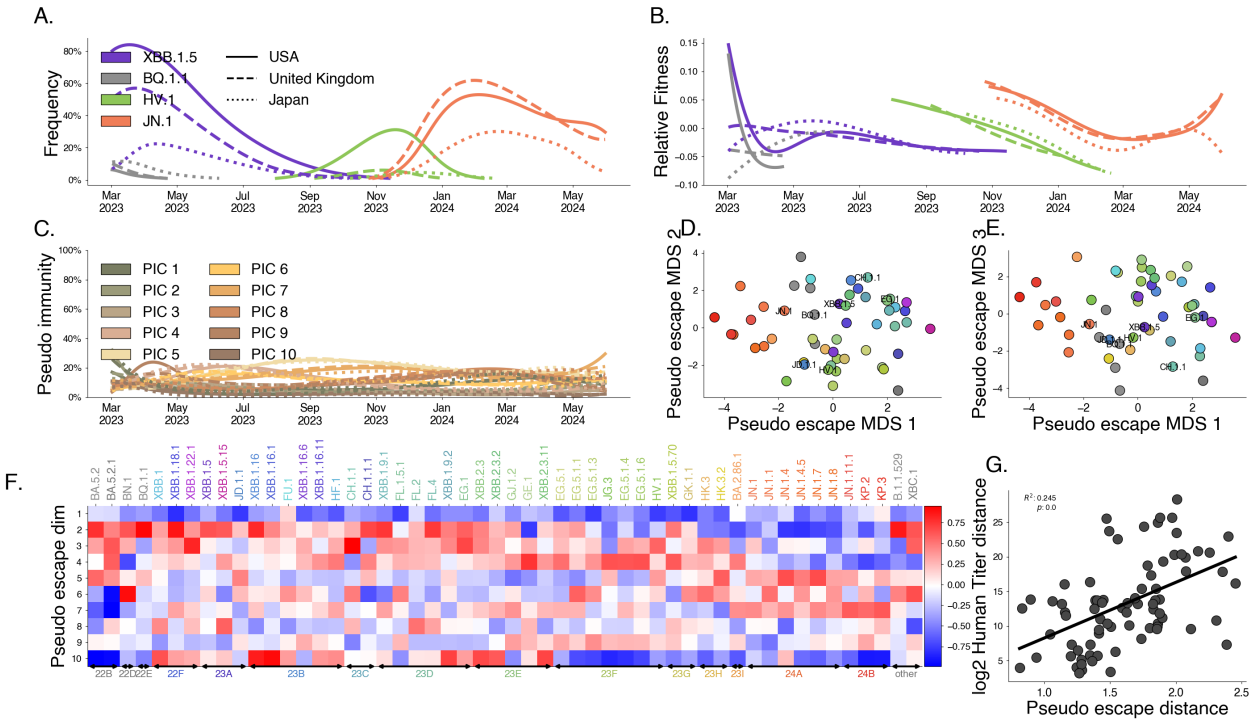

Figure C.16: **Latent factor model with $D = 10$ pseudo immune dimensions.**

Appendix D

SUPPLEMENTARY MATERIALS FOR CHAPTER 5

D.1 Supplementary Text

Connection to phylogenetic generalized least squares

Due to shared evolutionary history among a population, the independence assumptions of regression analyses are often violated. To address this, phylogenetic least squares (PGLS) has emerged as a way of accounting for covariance in the traits between samples from a population. In PGLS, we assume trait values x are assumed to have distribution:

$$x \sim \text{Normal}(X\beta, \Sigma), \tag{D.1}$$

where X is matrix of features, β is the regression coefficients, and Σ is a covariance matrix derived from a phylogenetic tree, capturing the covariance due to shared evolutionary history.

We show that our approach of modeling innovations between parent-child lineage pairs similarly captures this evolutionary structure.

When we directly model trait changes along branches using parent-child relationships, we focus on the changes $\Delta x_i = x_{\text{child},i} - x_{\text{parent},i}$. We assume that

$$\Delta x_i = Z_i\gamma + \varepsilon_i, \tag{D.2}$$

where Z_i is a vector of features for branch i e.g. branch length, γ is a vector of coefficients, and ε_i is a random error term with mean 0 and variance σ_i^2.

The trait x_i at node i can then be written as the sum of changes from the root to that node:

$$x_i = x_{\text{root}} + \sum_{j \in \text{path to } i} (Z_j \gamma + \varepsilon_j) \tag{D.3}$$

$$= x_{\text{root}} + \left(\sum_{j \in \text{path to } i} Z_j \right) \gamma + \sum_{j \in \text{path to } i} \varepsilon_j. \tag{D.4}$$

Simplifying it in this way allows us to easily analyze the covariance between the traits for two variants

$$\Sigma_{ik} = \text{Cov}(x_i, x_k) = \text{Cov} \left(\sum_{j \in \text{path to } i} \varepsilon_j, \sum_{l \in \text{path to } k} \varepsilon_l \right) \tag{D.5}$$

$$= \sum_{j \in \text{shared}(i,k)} \sigma_j^2. \tag{D.6}$$

We can see that this covariance Σ_{ik} is the sum of variances along the branches shared by nodes i and k.

Using our model of growth advantage innovation, we derive an expression for x_i that naturally suggests an equivalent covariance matrix Σ under PGLS. This covariance arises directly from the cumulative variances of changes along shared evolutionary paths, providing a concrete connection between the two methods, reflect the shared evolutionary paths.

D.2 Supplementary Figures

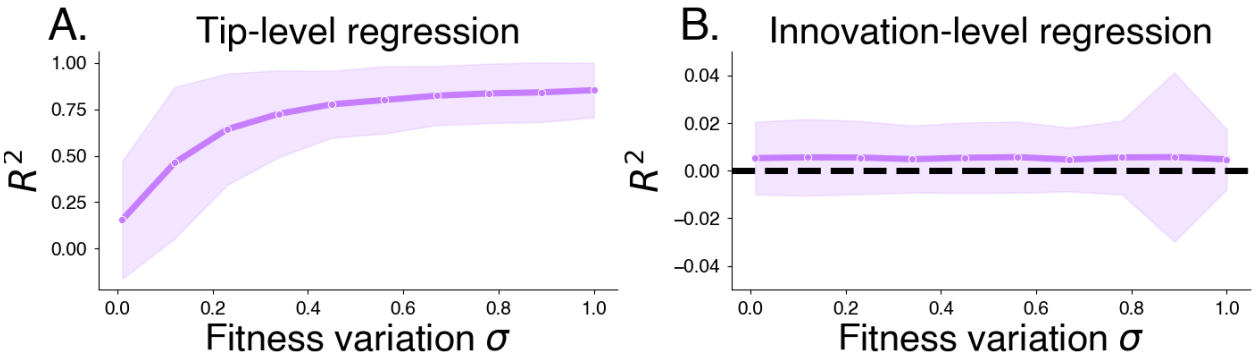

Figure D.1: **Variance explained as a function of fitness variation.** Relationship between fitness variation and the strength of regression (R^2) for tip-level (A) and innovation-level (B) analyses. Points represent the mean R^2 for each fitness variation, while the shaded regions show two standard deviations of the mean across 500 replicate simulations.

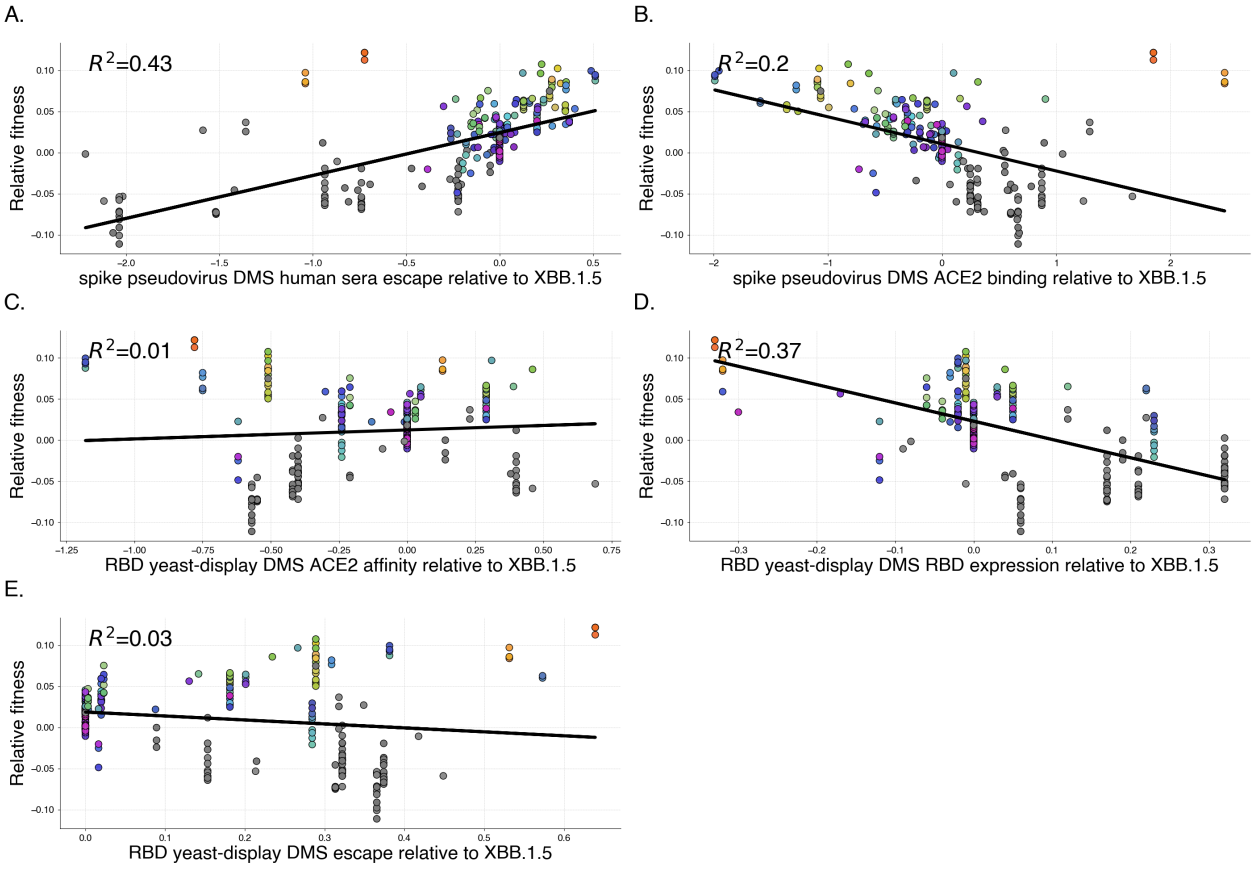

Figure D.2: **Naive regressions between molecular phenotypes and relative fitness**

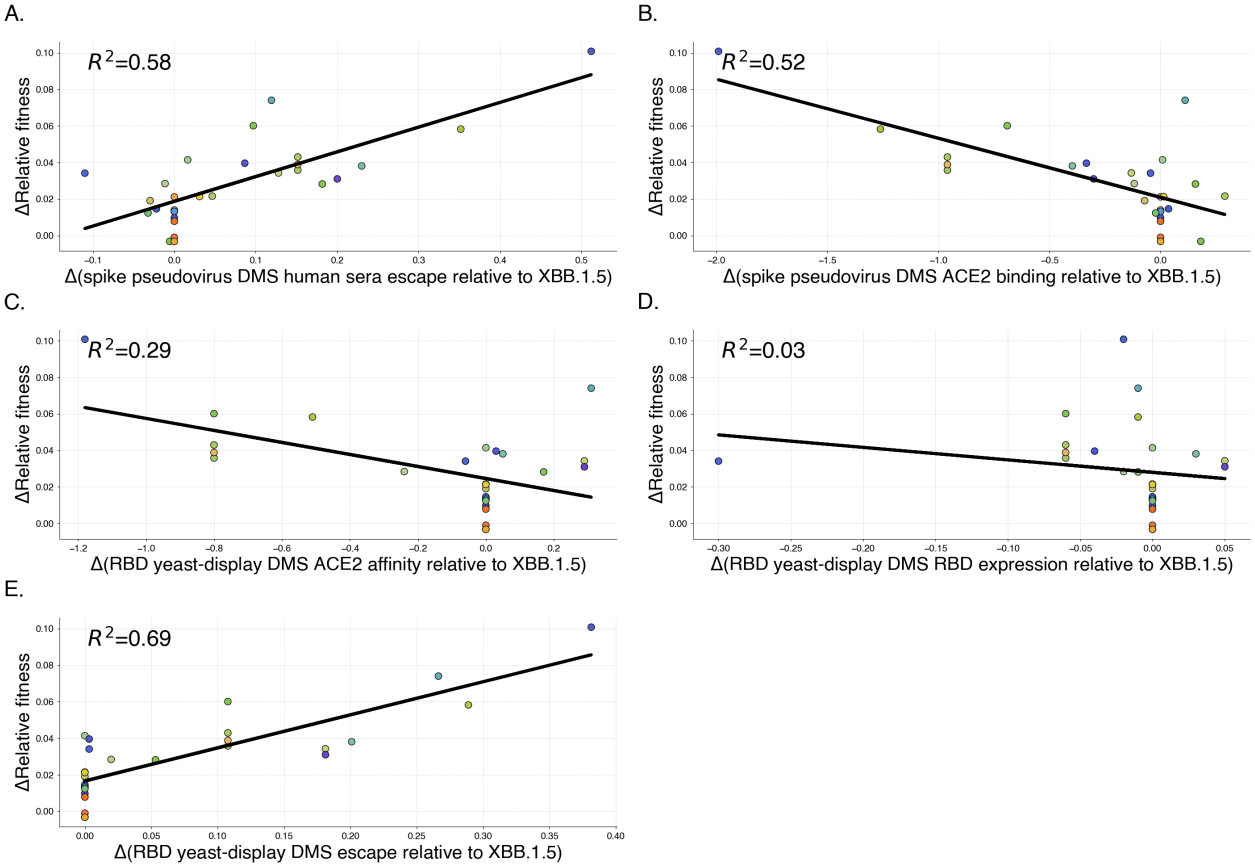

Figure D.3: **Innovation regressions between molecular phenotypes and relative fitness**