Computational tools for studying viral antigenic evolution

Zorian T. Thornton

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington
2026

Reading Committee:
Frederick A. Matsen IV, Ph.D., Chair
Kelley Harris, Ph.D.
Jesse Bloom, Ph.D.

Program Authorized to Offer Degree:
Genome Sciences

University of Washington

Abstract

Computational tools for studying viral antigenic evolution

Zorian T. Thornton

Chair of the Supervisory Committee:

Professor Frederick A. Matsen IV, Ph.D.
Public Health Sciences Division, Fred Hutch Cancer Center

RNA viruses like influenza and SARS-CoV-2 evolve rapidly under selection from host immunity, accumulating mutations that enable escape from antibodies elicited by prior infection or vaccination. However, viral evolution is constrained by the requirement that surface proteins must remain functional: they must fold properly, express on the virion, and bind host cell receptors to mediate entry. Understanding how mutations affect these molecular phenotypes alongside immune evasion is essential for predicting viral evolution and designing effective vaccines.

This dissertation presents three computational tools for modeling genotype-phenotype relationships in viral proteins and for simulating viral evolution at the population level. These tools share a common emphasis on biological interpretability while leveraging modern statistical and machine learning methods to analyze high-throughput experimental data or benchmark viral surveillance methods.

The first tool, `torchdms`, suggests a framework to jointly model multiple phenotypes from deep mutational scanning experiments using biophysically motivated neural network architectures. We apply `torchdms` models to data from a deep mutational scan of the SARS-CoV-2 receptor binding domain, demonstrating accurate prediction of variant phenotypes and recovery of biologically meaningful mutational effects.

The second tool, `polyclonal`, models how mutations enable viral escape from polyclonal antibody responses. The model decomposes serum activity into epitope-specific components and quantifies mutation

effects at each epitope, capturing the nonlinear patterns of escape that arise when antibodies target multiple sites on a viral antigen.

The third tool, `antigen-prime`, is a forward-time epidemic simulator that models the coupled genetic and antigenic evolution of viruses under selection from host population immunity. We use `antigen-prime` to simulate 30 years of influenza-like evolution and benchmark methods for variant assignment and growth rate inference, revealing a previously undocumented failure mode in growth rate inference models.

Acknowledgements

First and foremost, thank you to God, my family, and my friends for your unwavering love and support. I know I'm a rather indecisive person, proudly boasting about how I was going to be one of the best pastry chefs in the world, to wanting to be a politician, to whatever it is that I am now. Y'all never judged or discouraged me, and I am so grateful for that even as I continue to try to figure things out. A special thanks to my friends from high school and college that I still talk to on a regular basis: your support and ability to pick me up while I'm down is something I will always cherish.

Thank you to the kindest person I know, who happens to be my advisor, Erick Matsen. The impact you have had on me as not only a scientist, but as a person is something I will be forever grateful for. I always felt like I had the freedom to explore whatever I wanted in the lab, and if I started to get a little *too* distracted with side quests that didn't directly relate to my thesis, you would gently steer me back on track. I know I'll likely never have an advisor as awesome as you, which makes the duration of this journey even more worthwhile.

Thank you to my current, former, and new committee members: Jesse Bloom, Kelley Harris, Sara Mostafavi, Tim Thornton, and Ivana Bozic. Thank you for everything from helpful insight and feedback on my many (failed) research ideas, to being a sounding board for my career goals outside of academia.

Thank you to Hugh Haddox and Will DeWitt for being two crucial mentors and friends during my time in the Matsen lab. Hugh, thank you for the wonderful journey that was `antigen-prime`, and for all of the fun chats about soccer, music, and life. Will, thank you for being such a great mentor and friend during my early years in the Matsen lab and Genome Sciences. There were times when I felt like my ideas were silly or baseless, and I'd message you personally on Slack just for a vibe check on them. Thank you for confirming the ideas that were good and bringing them up when I struggled to find my voice, and for telling

me when I was awfully wrong in the nicest way possible.

Thank you to Dr. Jane Robertson Evia for being a wonderful mentor and friend during my time at Virginia Tech and for introducing me to the world of academic research. From serving as my academic advisor in college, to encouraging me to give research a try, to checking in on me over the last six years as I've worked on this degree, I am forever grateful for your support, friendship, and guidance.

Thank you to some of the most impactful teachers I had in high school who played an instrumental role in me ending up here. First, thank you to my Algebra 2 teacher Mr. Jim Neufer. You were the first teacher to challenge me and push me to be the best I could be. I once called you a "math drill sergeant" and I'm glad we both knew I was right. Thank you for preparing me for the challenges to come.

Thank you to my calculus and statistics teacher, Mr. Barry Smith; I have always and will always refer to you as the best teacher I ever had. Thank you for being so passionate about teaching and challenging your students to strive for excellence. We spent three years and four courses together while I was in high school, and I can confidently say that if it weren't for you, I would likely be in Washington D.C. ripping my hair out right now instead of writing this thesis.

Finally, thank you to my US Government teacher, Dr. Erin Yearout-Patton. For a majority of high school, I was dead set on going into politics or international relations. The enthusiasm you had for teaching American Government and civics was so infectious that I carried around a copy of the United States Constitution in my backpack during my senior year of high school. Given the state of the world as I write this, I'm so grateful for your passion and ability to emphasize the importance of being an informed citizen and knowing how our government is supposed to work.

DEDICATION

To Katie. It may not be the cure for cancer, and it may be over a decade late, but hopefully this will still do some good. I'll always remember you, one of my first real friends. ♡

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Viruses like influenza and SARS-CoV-2 evolve rapidly under selective pressure from host immunity, accumulating mutations that enable escape from antibodies elicited by prior infection or vaccination. Understanding how mutations affect viral fitness and predicting which variants will spread requires connecting molecular phenotypes to immune recognition and population-level transmission dynamics.

This dissertation develops computational methods for studying how mutations to viral protein sequences affect phenotypes across these scales. Chapters 2 and 3 share a common modeling framework: interpretable models that represent mutational effects as additive contributions to latent phenotypes, transformed through nonlinear functions to capture the complexities of protein biophysics and immune recognition. Chapter 4 connects these molecular-level insights to population dynamics through simulation and uses the simulated data to benchmark computational methods for viral surveillance.

The following sections introduce the biological and computational background underlying this work, progressing from viral antigenic evolution and experimental methods for measuring mutational effects, through the fitness landscape and epistasis concepts that motivate our modeling approaches, to the immunological and epidemiological contexts in which viral evolution occurs. A detailed overview of each chapter's contributions appears in Section 1.6.

1.1 Viral antigenic evolution

RNA viruses evolve rapidly due to error-prone replication machinery that generates mutations at rates orders of magnitude higher than DNA-based organisms [1]. This high mutation rate, combined with large population sizes and short generation times, enables viral populations to explore sequence space and adapt to selective pressures on timescales relevant to individual infections and public health.

Antigenic drift is the gradual accumulation of mutations in viral surface proteins, which are the primary targets of host antibody responses [1]. As mutations arise that reduce antibody binding, viruses carrying these escape mutations gain a selective advantage in hosts with prior immunity. Over time, this process drives the viral population away from previously circulating strains, eroding the protection conferred by past infection or vaccination.

Influenza and SARS-CoV-2 exemplify the public health consequences of antigenic evolution. Seasonal influenza undergoes sufficient antigenic drift that vaccine strains must be updated annually, and vaccine effectiveness varies depending on the match between vaccine and circulating strains [2, 3]. SARS-CoV-2 has generated a succession of variants of concern: Alpha, Delta, Omicron and its sublineages, each with mutations in the spike (S) protein that enhance transmissibility, alter receptor binding, or enable escape from antibodies elicited by earlier variants or vaccines [4, 5, 6].

Understanding how mutations affect viral phenotypes such as protein expression levels, host receptor binding affinity, and immune escape is essential for predicting viral evolution and designing durable vaccines. The methods developed in this dissertation address different facets of this challenge: from quantifying mutational effects on molecular phenotypes (Chapter 2), to modeling escape from polyclonal antibody responses (Chapter 3), to simulating viral evolution under host population immunity (Chapter 4).

1.2 Deep mutational scanning

Deep mutational scanning (DMS) is an experimental technique that measures the functional effects of mutations across hundreds of thousands of protein variants in high throughput [7, 8, 9, 10]. A typical DMS experiment begins by creating a library of variants containing mutations to a gene of interest, expressing these variants, and selecting for a phenotype such as protein expression, ligand binding, or antibody es-

cape. By coupling selection with deep sequencing, researchers can quantify how each mutation affects the phenotype of interest.

Traditional DMS experiments focus on single-mutant libraries, where each variant contains exactly one amino acid substitution from the wildtype sequence [9, 11, 12, 13]. This design makes it straightforward to infer individual mutational effects on phenotype. However, single-mutant libraries provide limited information about how mutations interact when combined in the same protein. Recent advances in library construction have enabled DMS experiments that assay variants containing multiple mutations, providing deeper sampling of sequence space and enabling the study of epistatic interactions between mutations [5].

DMS has proven particularly valuable for studying viral proteins and immune proteins. When applied to viral surface proteins, DMS can reveal which mutations enhance receptor binding, enable antibody escape, or compromise protein stability [5, 14, 15]. This information is crucial for understanding viral evolution, predicting which variants may emerge in circulating populations, and designing vaccines that target conserved epitopes.

Despite its power, analyzing DMS data presents computational challenges. Variants with multiple mutations are sparsely sampled relative to the vast combinatorial space of possible sequences, and the nonlinear interactions between mutations, *epistasis*, make it difficult to predict the phenotype of unseen variants from observed data alone [16, 17, 18, 19, 20, 21]. These challenges have motivated the development of statistical and machine learning methods for inferring genotype-phenotype maps from DMS experiments, a topic we return to in Chapter 2.

1.3 Fitness landscapes and epistasis

A central goal of molecular biology is to understand how a protein's amino acid sequence determines its function. This relationship is formalized as a genotype-phenotype (G-P) map: a function that assigns each possible amino acid sequence a quantitative phenotype such as enzyme activity, binding affinity, or protein stability. The fitness landscape metaphor extends this concept by associating each sequence with a fitness value, where fitness often reflects the virus's ability to replicate and transmit. Understanding the shape of fitness landscapes is essential for predicting viral evolution, as antigenic drift drives populations toward fitness peaks and valleys.

If mutations combined additively or independently to phenotype, fitness landscapes would be smooth and predictable. However, mutations often interact: the phenotypic effect of a mutation depends on which other mutations are present in the sequence [22, 20, 21]. When these interactions are present, they can create rugged fitness landscapes with multiple peaks and valleys. Epistasis can be classified by its form, some examples include: magnitude epistasis, which occurs when the effect size of a mutation changes depending on genetic background, while sign epistasis occurs when a mutation is beneficial in one background but deleterious in another [18].

Several biophysical mechanisms can generate epistasis [16, 21]. Stability-mediated epistasis arises because a protein must fold before it can function [23, 17]. As a result, destabilizing mutations may be tolerated in a stable background but lethal when combined with other destabilizing mutations. Similarly, binding affinity may saturate at high levels, causing diminishing returns for affinity-enhancing mutations. These sources of nonlinearity create the rugged landscapes that shape evolutionary trajectories of proteins.

The *global epistasis* framework provides a tractable approach to modeling these nonlinearities [24]. Rather than attempting to fit pairwise or higher-order interaction terms, which is statistically intractable for the combinatorial space of possible sequences and can yield uninterpretable coefficients [19], global epistasis models assume that mutations contribute additively to a latent phenotype, which is then transformed through a nonlinear function $g()$ to produce the observed phenotype. This framework preserves interpretable per-mutation effect coefficients while capturing the nonlinear relationship between latent and observed phenotypes [24, 5]. This global epistasis framework underlies the methods developed in Chapters 2 and 3.

1.4 Antibodies and B cell immunity

The adaptive immune system defends against pathogens through both cellular and humoral responses. B cells are central to humoral immunity, generating a diverse repertoire of antigen-binding B cell receptors (BCRs) and their secreted antibody forms. Naive B cells, those that have not yet encountered their cognate antigen, become activated upon binding an antigen that fits their BCR, typically during initial exposure via infection or vaccination [25].

Following activation, B cells migrate to germinal centers in lymph nodes and undergo affinity maturation: repeated rounds of somatic hypermutation and selection that refine BCRs for improved antigen

binding [26, 27]. This process represents the host side of an evolutionary arms race: as pathogens evolve to escape existing antibodies, the immune system can generate new antibody variants with improved binding to the mutated antigens.

Antibodies recognize specific regions on antigens called *epitopes*. A single antigen typically presents multiple epitopes, and the antibody response to infection or vaccination is polyclonal, meaning the serum contains antibodies from many B cell lineages targeting different epitopes. This contrasts with monoclonal antibodies, which derive from a single B cell clone and bind a single epitope.

Not all antibodies that bind a viral antigen prevent infection. Neutralizing antibodies are those that block viral entry, typically by binding epitopes on surface proteins that mediate receptor attachment or membrane fusion [28, 29, 30].

Identifying which epitopes are targeted by the antibody response and how mutations affect antibody binding at each epitope is crucial for vaccine design. Some epitopes tolerate mutations readily and present pathways for immune escape; others are functionally constrained and represent potential vaccination targets. Technologies such as phage immunoprecipitation sequencing (PhIP-Seq) enable mapping of epitopes recognized by serum antibodies [31, 32], and when combined with deep mutational scanning, these methods can quantify how mutations affect binding at each epitope [33, 34].

Because polyclonal responses target multiple epitopes simultaneously, viral escape typically requires mutations at several sites. This creates complex patterns of antigenic epistasis, where the fitness benefit of an escape mutation depends on which other mutations are present, a theme we return to in Chapter 3.

1.5 Epidemiological modeling

Epidemiological models provide a mathematical framework for understanding how infectious diseases spread through populations [35]. The classic susceptible-infected-recovered (SIR) model partitions a population into compartments: susceptible individuals who can be infected, infected individuals who can transmit the pathogen, and recovered individuals who have cleared the infection and gained immunity. The dynamics of disease spread depend on transmission rates, recovery rates, and the fraction of the population in each compartment.

A key quantity derived from the SIR model is the effective reproduction number R_t, which reflects

the average number of secondary infections caused by a single infection [36, 37]. When $R_t > 1$, the number of infected individuals increases; when $R_t < 1$, the epidemic declines. In the special case where the population is fully susceptible, this quantity is called the basic reproduction number R_0: the average number of secondary infections caused by a single individual in an otherwise naive population.

For antigenically variable pathogens like influenza or SARS-CoV-2, R_t depends not only on intrinsic viral transmissibility but also on how well the virus can evade existing immunity in the population [38, 39]. A variant that is antigenically distinct from previously circulating strains effectively encounters a larger susceptible pool, increasing its R_t relative to other co-circulating variants. This creates a coupling between viral molecular evolution and population-level epidemiology: mutations that enable immune escape at the individual level translate to increased transmission at the population level.

Modeling this coupling requires integrating molecular-level information about genetic sequences and antigenic phenotypes with population-level models of transmission and immunity. In Chapter 4, we develop a simulator that combines these three types of information, enabling us to study how viral populations evolve under selection from host population immunity and to benchmark viral surveillance methods against ground-truth evolutionary dynamics.

1.6 Overview of chapters and contributions

This dissertation develops computational methods for modeling how mutations to viral protein sequences affect molecular phenotypes, and how these phenotypic effects shape viral evolution under immune selection. Chapters 2 and 3 share a common modeling framework: both represent mutational effects as additive contributions to latent phenotypes that are then transformed through nonlinear functions to produce observed phenotypes. Chapter 4 moves to the population level, modeling how viral populations evolve under selection from host immunity and benchmarking methods for analyzing viral surveillance data. These chapters progress through increasing levels of biological organization: from molecular phenotypes of individual proteins, to immune selection acting on single viral variants, to viral populations evolving under pressure from host population immunity.

Chapter 2: `torchdms`. This chapter operates at the level of molecular phenotypes, modeling how mutations affect protein folding and receptor binding. Deep mutational scanning experiments increasingly

measure multiple phenotypes simultaneously, such as protein expression and ligand binding. We introduce `torchdms`, a Python package that extends the global epistasis framework to jointly model multiple phenotypes. The Generalized Global Epistasis (GGE) architectures maintain interpretable mutational effect coefficients while capturing nonlinear and conditional dependencies between phenotypes, i.e., encoding the constraint that a protein must fold before it can bind. We apply these models to SARS-CoV-2 receptor binding domain data, demonstrating accurate prediction of held-out variants and recovery of biologically meaningful mutational effects.

Chapter 3: `polyclonal`. This chapter moves from static molecular phenotypes to the more complex setting of immune selection, where antibody pressure creates fitness landscapes that depend on host immune responses. Polyclonal antibody responses target multiple epitopes on viral antigens, creating complex patterns of escape where mutations at different epitopes combine synergistically. We present a biophysical model that decomposes polyclonal antibody activity into epitope-specific components and quantifies mutation effects at each epitope. The model captures antigenic epistasis. Mutations contribute additively to binding affinity at each epitope, but their effects on overall escape are nonlinear. This nonlinearity arises from the Hill curve relating affinity to fraction bound and the product across epitopes. We implement this model in the `polyclonal` Python package and validate it on simulated deep mutational scanning data.

Chapter 4: `antigen-prime`. This chapter scales up to the population level, examining how viral populations evolve to evade immunity across entire host populations over epidemiological timescales. Computational methods for influenza surveillance, including variant assignment and growth rate inference, are difficult to benchmark because natural viral populations lack known ground-truth values. We present `antigen-prime`, a forward-time epidemic simulator that jointly models the genetic and antigenic evolution of viruses under selection from host population immunity. We use `antigen-prime` to simulate 30 years of H3N2-like evolution and benchmark methods for assigning viral sequences to antigenic variants and inferring variant-specific growth rates. The benchmarks reveal that sequence-based clustering methods effectively recover antigenic groupings, and identify a previously undocumented failure mode in growth rate inference models.

Chapter 2

Inferring mutational effects from multi-mutant DMS data with `torchdms`

Deep mutational scanning (DMS) is a high-throughput experimental technique for measuring the effects of thousands of mutations on protein function. These experiments can provide measurements of multiple phenotypes for the library of variants, such as protein expression and ligand binding. However, it is difficult to attribute observed phenotypes to specific mutations when most variants contain multiple mutations, and nonadditive interactions between mutations make it challenging to predict phenotypes for unobserved variants. In this chapter, we describe `torchdms`, a Python package that extends the global epistasis framework to jointly model multiple phenotypes from DMS data. We introduce Generalized Global Epistasis (GGE) architectures that maintain interpretable mutational effect coefficients while capturing nonlinear and conditional dependencies between phenotypes. We apply `torchdms` to a published DMS dataset characterizing the effects of mutations in the SARS-CoV-2 receptor binding domain (RBD) on both protein expression and ACE2 binding affinity, demonstrating accurate predictions on held-out variants and recovery of biologically meaningful mutational effects.

2.1 Introduction

Understanding how mutations affect protein function is essential for interpreting disease-associated variants, predicting pathogen evolution, and engineering proteins with desired properties [9, 16]. Deep mutational scanning experiments now routinely measure the functional effects of thousands of protein variants, providing unprecedented views of sequence-function relationships. However, variants containing multiple mutations are sparsely sampled relative to the vast combinatorial space of possible sequences, making it difficult to predict phenotypes for unseen variants from observed data alone.

Computational models that infer genotype-phenotype (GP) maps from DMS data fall along a spectrum of interpretability and predictive power. At one extreme, linear models assume mutations act additively on phenotype, yielding directly interpretable coefficients for each possible amino acid substitution. These models capture the dominant effects of most mutations but fail to account for epistasis: the nonlinear interactions between mutations that arise from biophysical constraints such as protein stability thresholds and binding saturation [40, 41, 23]. At the other extreme, deep learning approaches, including protein language models, can achieve strong predictive performance by learning complex representations of sequence space [42, 43, 44]. However, these models function as "black boxes" whose internal parameters lack clear biophysical meaning, limiting their utility for understanding which mutations drive observed phenotypes.

The global epistasis (GE) framework provides a middle ground between these extremes [24]. GE models assume that mutations act additively on an unobserved latent phenotype, which is then transformed through a nonlinear function to produce the observed phenotype. This formulation captures key sources of nonlinearity, such as fitness thresholds and diminishing returns from beneficial mutations, while preserving interpretable coefficients for individual mutational effects. The signs and magnitudes of these coefficients indicate whether mutations are generally beneficial or deleterious across genetic backgrounds, and the shape of the nonlinear transformation reveals how the latent phenotype maps to observed fitness [24, 45].

Several software packages implement variants of the GE framework. The original GE model was demonstrated on single-phenotype DMS data, using Gaussian process regression for the nonlinear transformation [24]. MAVE-NN extends this approach by using neural networks and an information-theoretic training objective, enabling inference of more flexible nonlinearities [46]. The `dms-variants` package provides utilities for processing DMS data and fitting simple GE models to multiphenotype DMS experiments [5].

However, these implementations share a key limitation: they model each phenotype independently, preventing joint inference when DMS experiments measure multiple related phenotypes, such as protein expression and ligand binding.

Recent DMS experiments increasingly measure multiple phenotypes simultaneously [5], motivating models that can capture dependencies between phenotypes. For instance, a protein must fold properly before it can bind a ligand, so binding affinity measurements reflect both folding stability and intrinsic binding capacity. Modeling these phenotypes independently discards information about their relationship and may yield misleading estimates of mutational effects.

In this chapter, we introduce `torchdms`, a Python package that extends the GE framework to accommodate multiple phenotypes through what we term Generalized Global Epistasis (GGE) models. `torchdms` uses PyTorch [47] to implement flexible neural network architectures that can model dependencies between phenotypes while preserving interpretable mutational effect coefficients. We apply `torchdms` to published data on the SARS-CoV-2 receptor binding domain, demonstrating accurate predictions on held-out variants and recovery of biologically meaningful mutational effects.

2.2 Results

2.2.1 Generalized Global Epistasis models

We extend the global epistasis framework to accommodate multiple phenotypes through what we term Generalized Global Epistasis (GGE) models. The GGE framework computes one or more latent phenotypes as additive functions of mutational effects, then maps these through a flexible neural network to predict multiple observed phenotypes:

$$\hat{y} = g(\phi(x)) \tag{2.1}$$

$$\phi(x) = \beta_{wt} + \sum_{i}^{L} \beta_{i,a_i} \tag{2.2}$$

where β_{wt} is a wildtype offset, β_{i,a_i} is the effect of amino acid a_i at position i, and g is a neural network that maps the latent phenotype(s) to one or more observed phenotypes $\hat{y} \in \mathbb{R}^d$. This generalizes the standard GE

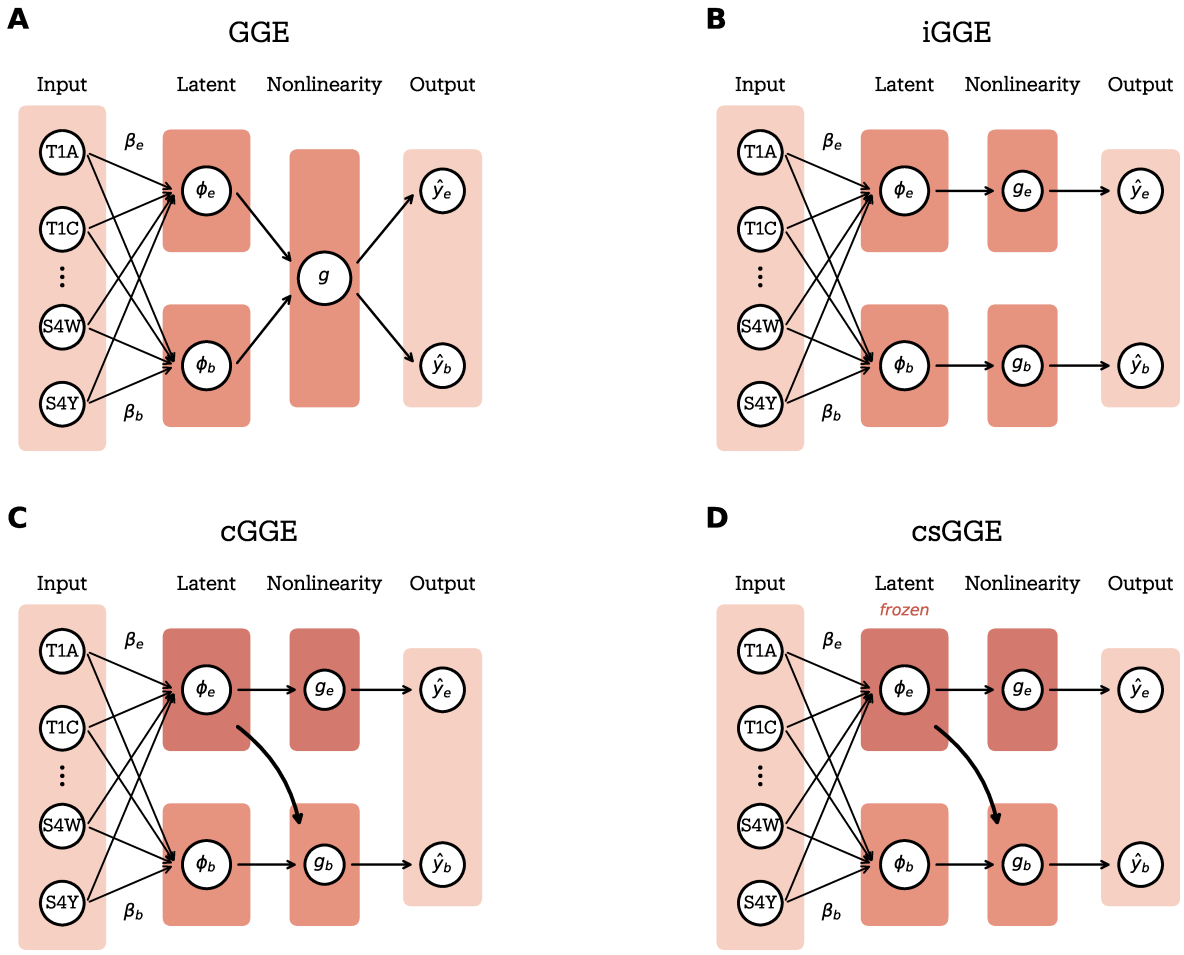

Figure 2.1: Global epistasis model architectures implemented in `torchdms`. Each architecture maps a binary-encoded input sequence through learned coefficient matrices (β) to latent phenotypes (ϕ), which are then transformed by nonlinear functions (g) to predict observed phenotypes ($\hat{y}$). **(A)** The Generalized Global Epistasis (GGE) model extends GE to multiple phenotypes (here, folding and binding) with separate coefficient matrices and a shared nonlinearity g. **(B)** The independent GGE (iGGE) model maintains fully separate pathways for each phenotype with no shared information. **(C)** The conditional GGE (cGGE) model incorporates a skip connection from the folding latent phenotype to the binding nonlinearity, encoding the biophysical prior that protein stability influences binding capacity. **(D)** The conditional sequential GGE (csGGE) model extends cGGE with a sequential training approach: the folding components are trained first and then the weights are frozen while the binding components are trained, preventing information from the binding phenotype from influencing folding predictions during optimization.

framework: when $d = 1$ and the latent space is one-dimensional, GGE reduces to the original GE model. By allowing multiple latent dimensions and multiple output phenotypes, GGE can capture shared structure across phenotypes through a common latent representation.

The independent GGE (iGGE) and conditional GGE (cGGE) models address specific experimental designs where phenotypes may be independent or conditionally dependent. The independent GGE (iGGE) model maintains separate latent spaces for each phenotype, with no information shared across phenotypes:

$$\hat{y}_p = g_p(\phi_p(x)) \tag{2.3}$$

$$\phi_p(x) = \beta_{wt,p} + \sum_{i}^{L} \beta_{i,a_i,p} \tag{2.4}$$

Each phenotype p has its own set of mutational effect coefficients $\beta_{i,a_i,p}$ and nonlinear transformation g_p. This architecture is equivalent to fitting separate GE models for each phenotype, which may be appropriate when phenotypes arise from independent biophysical processes.

The conditional GGE (cGGE) model captures dependencies between phenotypes by allowing one phenotype's latent space to inform predictions for another. This is particularly relevant for DMS experiments measuring both protein expression (or folding) and ligand binding, where binding requires proper folding:

$$\hat{y}_{\text{fold}} = g_{\text{fold}}(\phi_{\text{fold}}(x)) \tag{2.5}$$

$$\hat{y}_{\text{bind}} = g_{\text{bind}}(\phi_{\text{bind}}(x), \phi_{\text{fold}}(x)) \tag{2.6}$$

The binding prediction network receives input from both the binding latent space and the folding latent space, allowing it to learn that adequate expression is required for binding. We additionally define the conditional sequential GGE (csGGE) model, which extends cGGE with a sequential training approach: the folding components are trained first and then frozen while the binding components are trained. This prevents information from the binding phenotype from influencing folding predictions during optimization.

2.2.2 SARS-CoV-2 Receptor Binding Domain

To assess the ability of models implemented in `torchdms` to capture complex genotype-phenotype relationships in real DMS data, we applied our framework to a dataset measuring the effects of mutations in the SARS-CoV-2 receptor binding domain (RBD) on both protein expression and ACE2 binding affinity [5]. This dataset contains measurements for almost all possible single amino acid substitutions in the RBD, as well as a large number of variants with multiple mutations. Variants were assigned an expression score reflecting the amount of expressed RBD on the surface of yeast cells and a binding score reflecting the binding affinity of the expressed RBD to the ACE2 receptor, measured using Tite-Seq [48]. The full library contains 195,081 variants; after filtering for variants with measurements for both phenotypes and restricting to variants with at most six amino acid substitutions, the dataset contained 137,768 variants.

We divided the dataset into training, validation, and test sets using stratified splitting to ensure that each set contains an equal proportion of variants with different numbers of mutations. The training set contains 99,338 variants, the validation set contains 15,010 variants, and the test set contains 15,000 variants. Individual amino acid sequences are represented as one-hot encoded vectors $x \in \{0,1\}^{L\times|A|}$, where $L = 201$ is the length of the protein sequence and A is the amino acid alphabet of size 20.

We trained five model architectures on this dataset: a linear baseline, the shared-latent space GGE model, and the three two-track models. All models used a single hidden layer of 25 neurons and were trained for up to 500 epochs with early stopping (patience of 10 epochs), using L1 loss and three independent random initializations.

2.2.3 GGE models accurately predict held-out variant phenotypes

To compare how well each architecture captures genotype-phenotype relationships, we assessed predictive performance on the held-out test set by computing the coefficient of determination (R^2) between observed and predicted expression and ACE2 binding affinity (Figure 2.3). Linear models serve as a baseline for additive mutational effects, while the GGE and two-track models capture increasing levels of complexity.

The linear model achieved moderate performance, with R^2 values of 0.67 for expression and 0.53 for binding, indicating that additive effects capture a substantial portion of the variance but leave room for improvement. Notably, the linear model fails to capture the nonlinear relationship between mutational effects

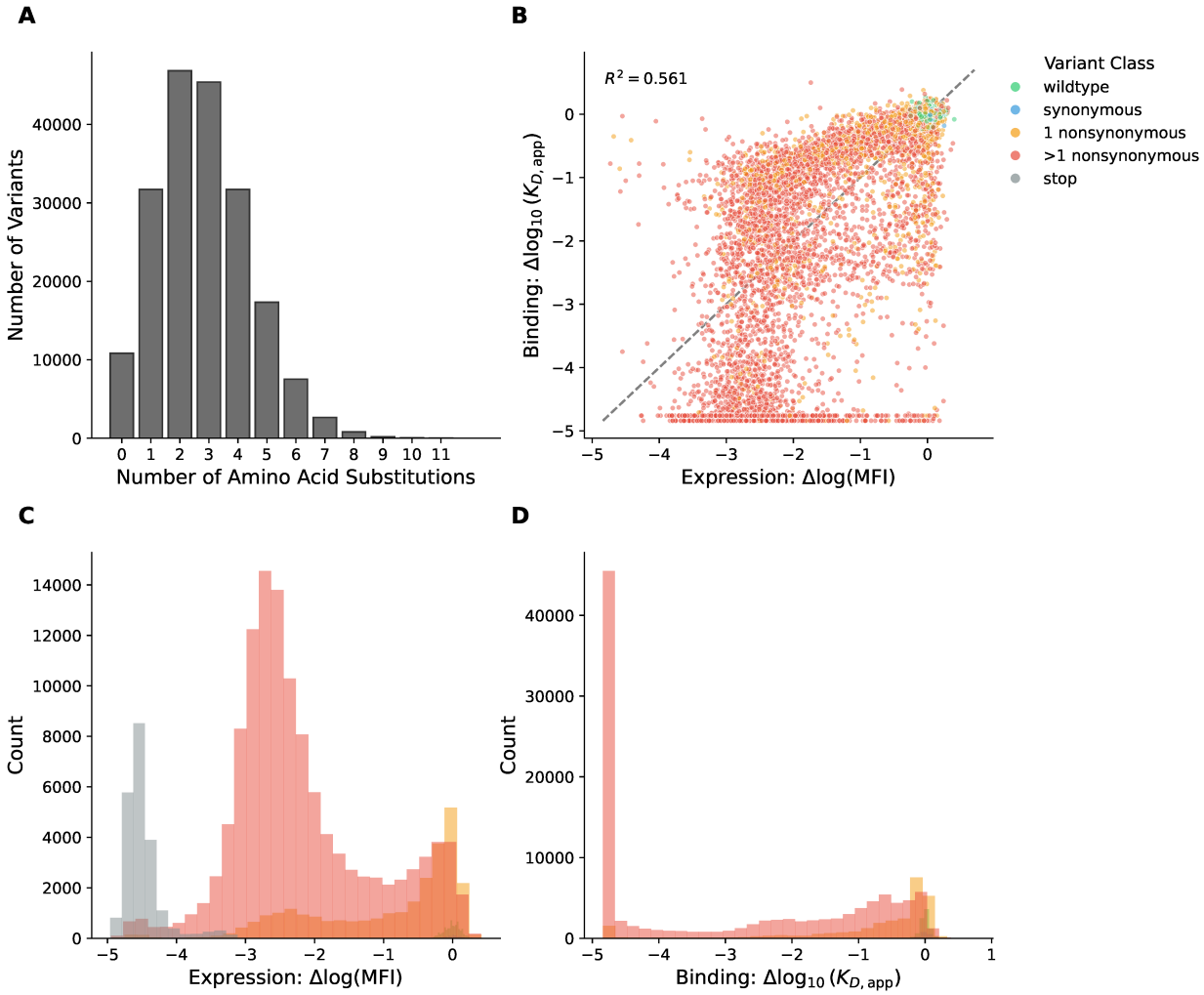

Figure 2.2: Overview of SARS-CoV-2 RBD DMS dataset structure and phenotype distributions. **(A)**: Distribution of the number of amino acid substitutions per variant in the dataset. **(B)**: Joint distribution of expression and binding scores for all variants in the dataset, colored by the number of amino acid substitutions. **(C)**: Distribution of expression scores colored by the type and number of amino acid substitutions. **(D)**: Distribution of binding scores colored by the type and number of amino acid substitutions.

and ACE2 binding, which serves as the main motivation for the GE and GGE frameworks (Figure 2.3).

The shared-latent GGE model improved performance substantially, achieving R^2 values of 0.86 for expression and 0.70 for binding, demonstrating the importance of modeling nonlinearities in the genotype-phenotype mapping. However, binding predictions remain less accurate than expression predictions, suggesting that a single shared latent dimension cannot fully capture both phenotypes.

The two-track models (iGGE, cGGE, csGGE) further improved performance, particularly for binding. By maintaining separate latent phenotypes for expression and binding, these architectures allow each phenotype to have its own additive representation while still modeling their interactions through the nonlinear network. The iGGE model achieved R^2 values of 0.94 for expression and 0.87 for binding, while the cGGE model achieved 0.93 and 0.85, respectively. The csGGE model achieved the best binding prediction ($R^2 = 0.89$) while maintaining strong expression performance ($R^2 = 0.94$), likely because its skip connection allows the model to capture the dependency of binding on expression.

Despite these improvements, all models struggle to predict expression scores for variants that fail to express, visible as a cluster of points with low observed expression but variable predicted values in Figure 2.3. This limitation reflects the inherent difficulty in modeling the sharp transition between functional and nonfunctional proteins.

2.2.4 Inferred mutational effects correlate with published estimates

To assess the biological interpretability of the learned β coefficients, we compared them against the single-mutant effects reported by Starr et al. [5] (Figure 2.4). The published effects were inferred using independent global epistasis models fit to each phenotype via the `dms-variants` package, providing an independent benchmark for our multiphenotype models.

The linear model achieved R^2 values of 0.79 for expression and 0.61 for binding (Figure 2.4, first column), reflecting that a purely additive model recovers coefficients similar to those from independent GE fits.

The shared-latent GGE model showed reduced agreement, with R^2 values of 0.68 for expression and 0.44 for binding (Figure 2.4, second column). This drop likely reflects the constraint of mapping both phenotypes from a shared two-dimensional latent space: the model must find β coefficients that simultaneously

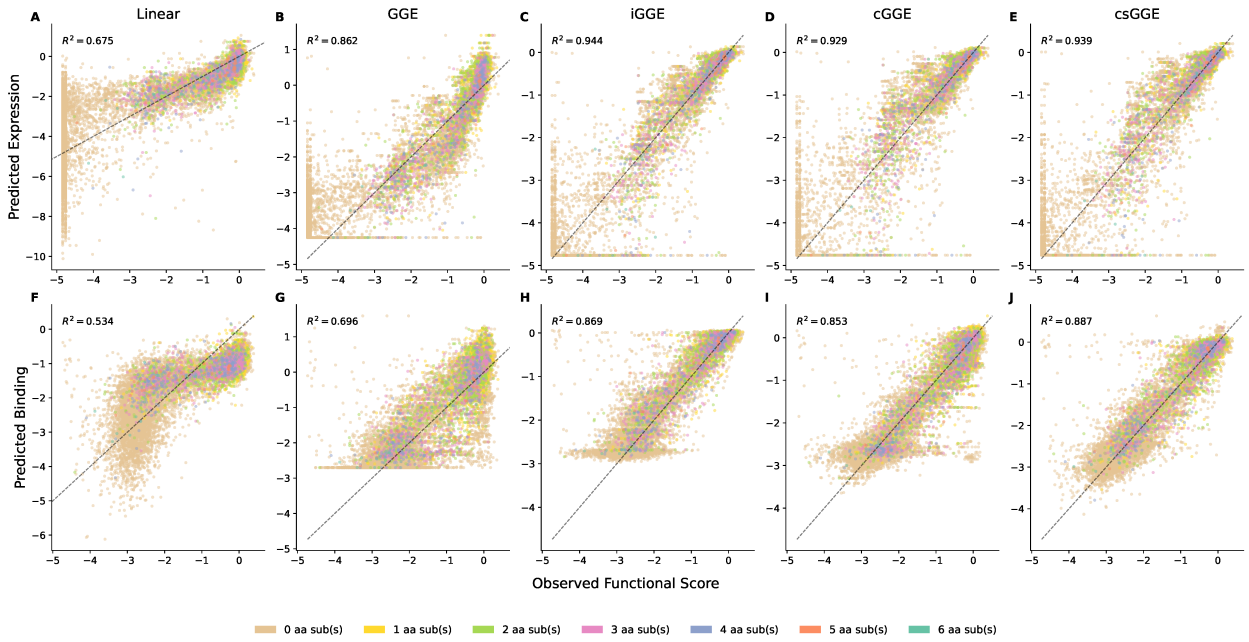

Figure 2.3: Model performance on RBD DMS data. Observed vs. predicted functional scores for expression (top row) and binding (bottom row) across five model architectures. Points are colored by number of amino acid substitutions. Dashed lines indicate perfect prediction ($y = x$).

predict both phenotypes, potentially causing interference between mutational effects for different phenotypes. Furthermore, it is unclear which latent dimension corresponds to which phenotype, complicating interpretation of the learned coefficients.

The two-track models recovered more accurate single-mutant effects (Figure 2.4, columns 3–5). The iGGE model achieved R^2 values of 0.82 for expression and 0.66 for binding, while the cGGE model achieved 0.80 and 0.64, respectively. The csGGE model showed the strongest agreement for binding ($R^2 = 0.69$) while maintaining high expression agreement ($R^2 = 0.82$). By maintaining separate β coefficient matrices for each phenotype, these architectures can learn phenotype-specific additive effects without the compromises required by the shared-latent GGE model.

Several factors may contribute to the remaining discrepancies between `torchdms` β values and the published estimates. First, gauge-fixing conventions differ between packages: while both set wildtype coefficients to zero, normalization of observed mutations may differ. Second, we use L1 loss for robustness to outliers, whereas `dms-variants` infers coefficients using maximum likelihood estimation under a Cauchy noise model. Finally, `torchdms` optimizes both phenotypes jointly, which may produce slightly

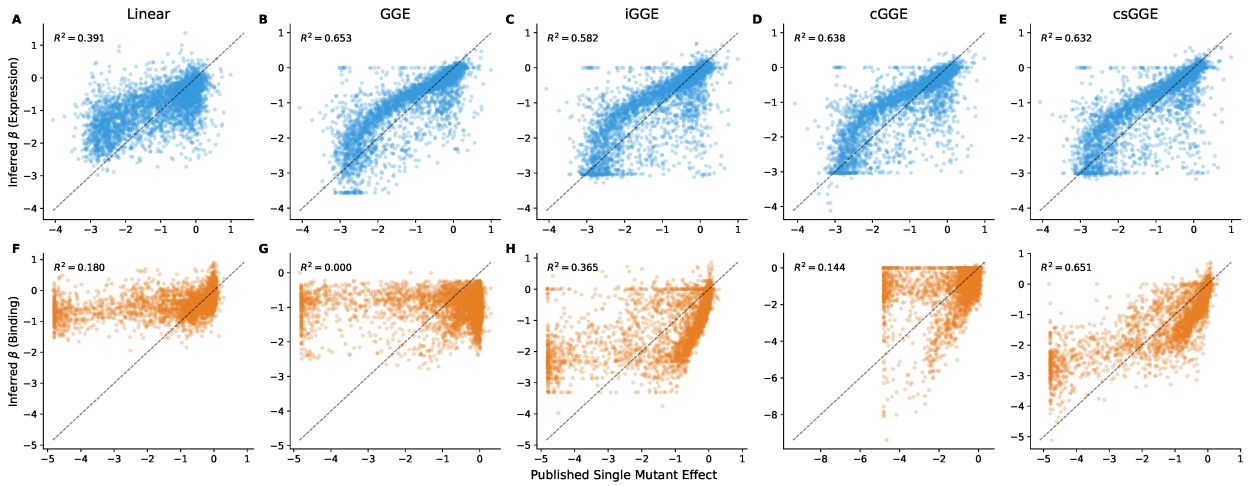

Figure 2.4: Comparison of mutational effect coefficients (β) inferred by `torchdms` models with single-mutant effects reported by Starr et al. [5] for expression (top row) and binding (bottom row). The published effects were inferred using independent global epistasis models fit to each phenotype via `dms-variants`. Each point represents one of 3,797 single amino acid substitutions. Dashed lines indicate perfect agreement.

different coefficient estimates than fitting each phenotype independently. The iGGE model is a notable exception: with fully independent phenotype tracks (separate β coefficients and separate nonlinearities), it is structurally equivalent to fitting two independent GE models yet still shows imperfect agreement with the published estimates.

2.2.5 Learned nonlinearities reveal latent-to-observed phenotype mappings

Understanding how models transform latent phenotypes into predictions can reveal whether they capture expected biophysical relationships, such as the requirement that a protein must fold before it can bind. To visualize these learned transformations, we plotted model predictions across a grid of latent phenotype values (Figures 2.5 and 2.6). For the two-track models, the x-axis represents the binding latent phenotype and the y-axis represents the expression latent phenotype. Background color indicates model predictions, while overlaid points show test set variants colored by their observed phenotype values.

For expression predictions (Figure 2.5), all models show a horizontal gradient: predicted expression depends primarily on the expression latent phenotype (y-axis) and is largely independent of the binding latent (x-axis). This is consistent with the biological expectation that expression is determined by protein stability and folding, independent of binding affinity.

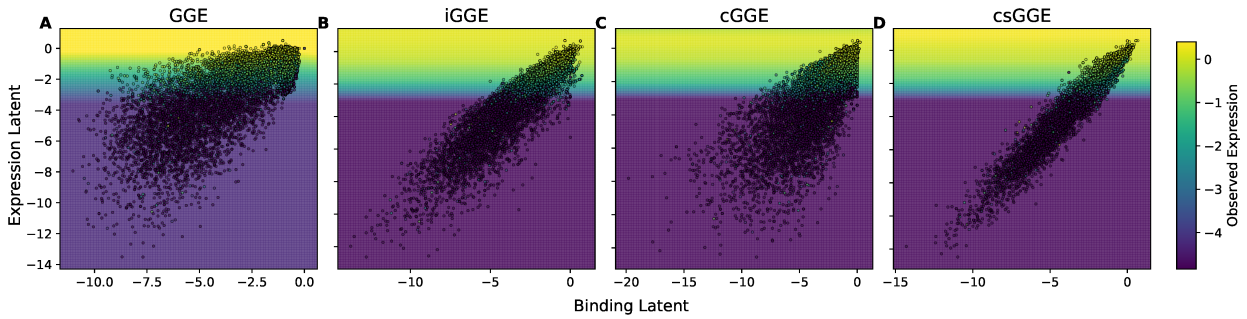

Figure 2.5: Learned nonlinear mapping from latent phenotype space to observed expression functional scores. Background color shows model predictions across the two-dimensional latent space; points show test set variants colored by observed expression score.

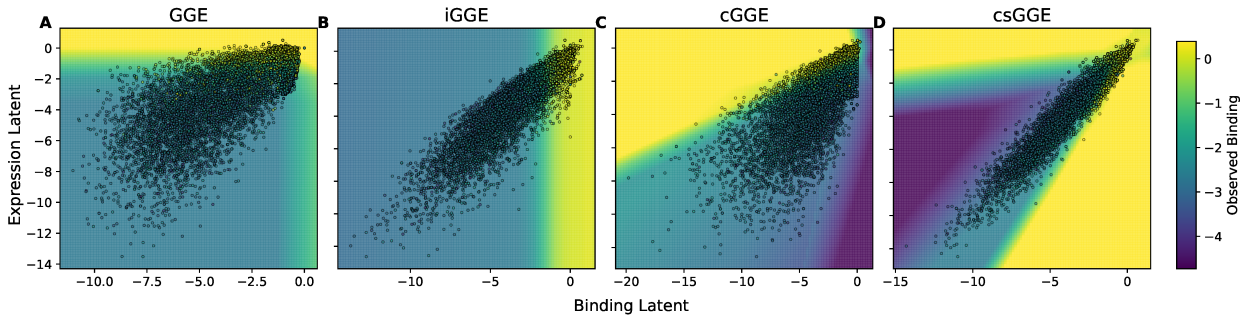

Figure 2.6: Learned nonlinear mapping from latent phenotype space to observed binding functional scores. Background color shows model predictions across the two-dimensional latent space; points show test set variants colored by observed binding score.

For binding predictions (Figure 2.6), the two-track models reveal a more complex dependency structure. Predicted binding depends on both latent phenotypes, with a diagonal or curved gradient indicating that high binding requires both high expression latent (the protein must fold) and high binding latent (the protein must bind ACE2). This captures the biological constraint that a protein must be expressed and properly folded before it can bind its target. The cGGE and sGGE models show examples of this behavior, with binding predictions increasing along the diagonal from low expression/low binding to high expression/high binding.

The shared-latent GGE model (leftmost column in both figures) shows less interpretable structure, as both phenotypes must be predicted from the same two-dimensional latent space without designated dimensions for each phenotype.

2.3 Methods

2.3.1 Data preparation

Individual amino acid sequences are represented as one-hot encoded vectors $x \in \{0,1\}^{L\times|A|}$, where L is the length of the protein sequence and A is the amino acid alphabet. We distinguish variants with n amino acid substitutions from wildtype as n-mutants. For each sequence x_i, we have measured DMS phenotype(s) $y_i \in \mathbb{R}^d$, where d is the number of measured phenotypes (e.g., ligand binding, protein expression). The dataset is defined as $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, where N is the number of sequences in the DMS library.

We split DMS data into training, validation, and test sets for hyperparameter tuning and model evaluation. To ensure that identical amino acid sequences are not present in multiple sets, we group all identical sequences together before assignment. We define a *stratum* as the set of all n-mutants for a given n and perform stratified splitting to ensure each subset contains an equal proportion of n-mutants. Users can optionally group all single mutants into the training set, though this is not recommended for libraries with many higher-order mutants.

2.3.2 Model architectures

The GGE framework and its variants (iGGE, cGGE, csGGE) are described in Section 2.2.1. All architectures share a common structure: a linear embedding layer that computes latent phenotypes $\phi(x)$ from one-hot encoded sequences via learned β coefficients, followed by a nonlinear network g that maps latent phenotypes to observed phenotypes. The nonlinear network consists of one or more fully connected hidden layers with ReLU activations. For the RBD experiments, we used a single hidden layer of 25 neurons for all models.

2.3.3 Model training and loss functions

We train models using the Adam optimizer with a batch size of 500 variants. A learning rate scheduler reduces the learning rate by a factor of 0.5 if the validation loss does not decrease for a specified number of epochs. Early stopping halts training when validation loss fails to improve, preventing overfitting. Models are trained with multiple independent random initializations to ensure results are not sensitive to initial conditions.

`torchdms` supports several loss functions: mean squared error (MSE), mean absolute error (MAE/L1), Huber loss, and root mean squared error (RMSE). MAE is preferred for robustness to outliers, which are common in DMS datasets due to measurement noise.

2.3.4 Regularization and constraints

L1 regularization L1 regularization prevents overfitting by penalizing large coefficient values. `torchdms` supports L1 (lasso) regularization on both the β coefficients and the weights of the nonlinear network g. The regularization penalty is added to the loss function:

$$\mathcal{L}_{\text{reg}} = \lambda_\beta \|\beta\|_1 + \lambda_g \|W_g\|_1 \tag{2.7}$$

where λ_β controls the regularization strength on the latent phenotype coefficients and λ_g controls the regularization on the nonlinear network weights W_g.

Low-rank approximation When many mutations are unobserved in the training data, the β coefficient matrix may overfit to noise. `torchdms` offers low-rank regularization via truncated singular value decomposition (SVD) applied after each gradient step. The β coefficients are reshaped into a matrix $B \in \mathbb{R}^{|A| \times L}$ where rows correspond to amino acids and columns to positions. We compute the SVD $B = U\Sigma V^\top$ and reconstruct a rank-r approximation:

$$B_r = U_r \Sigma_r V_r^\top \tag{2.8}$$

where U_r, Σ_r, and V_r contain only the top r singular vectors and singular values. This low-rank constraint encourages the model to learn smooth patterns across positions and amino acids, improving generalization to unobserved mutations.

Monotonicity constraints The original GE framework assumed a monotonic relationship between latent and observed phenotypes. `torchdms` supports this constraint by restricting all weights in the nonlinear network g (excluding bias terms) to be nonnegative. After each gradient step, negative weights are clamped to zero:

$$w \leftarrow \max(w, 0) \tag{2.9}$$

Combined with monotonic activation functions (e.g., ReLU), this ensures the mapping from latent to observed phenotype is monotonically increasing (or decreasing, depending on a sign parameter). Monotonicity constraints are optional and may be relaxed when the data suggest non-monotonic relationships.

2.3.5 Gauge-fixing

Gauge-fixing constraints ensure interpretable coefficients by removing degrees of freedom in the latent space. Without these constraints, the β coefficients in GGE models are not identifiable: any constant added to all coefficients at a position can be absorbed by the nonlinearity g. We apply gauge-fixing constraints after each gradient step. Let $\mathcal{M}$ denote the set of mutations observed in the training data. For each latent dimension, we set coefficients for wildtype amino acids and unobserved mutations to zero:

$$\beta_{i,a_i} = 0 \quad \text{if } a_i = a_i^{wt} \text{ or } (i, a_i) \notin \mathcal{M} \tag{2.10}$$

We then project the remaining coefficients to have mean zero at each position:

$$\beta_{i,a_i} \leftarrow \beta_{i,a_i} - \frac{1}{|\mathcal{M}_i|} \sum_{a \in \mathcal{M}_i} \beta_{i,a} \tag{2.11}$$

where $\mathcal{M}_i$ is the set of observed mutations at position i. This ensures that coefficient values reflect deviations from the average mutational effect at each position, with wildtype serving as the reference.

2.3.6 Evaluation metrics

Model predictive performance is evaluated using the coefficient of determination (R^2) on held-out test data:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}, \tag{2.12}$$

where y_i are observed phenotypes, $\hat{y}_i$ are model predictions, and $\bar{y}$ is the mean observed phenotype.

To assess the biological interpretability of inferred mutational effects, we compare the learned β coefficients to published single-mutant effect estimates when available. For the RBD dataset, we compare against the single-mutant effects reported by Starr et al. [5], which were inferred using independent global epista-

sis models fit to each phenotype via the `dms-variants` package. We compute the Pearson correlation between the `torchdms`-inferred β values and the published estimates.

2.4 Discussion

This chapter presented `torchdms`, a software package that extends the global epistasis framework to jointly model multiple phenotypes from deep mutational scanning data. A key feature of the Generalized Global Epistasis models is that they maintain interpretable mutational effect coefficients (β) while allowing flexible nonlinear mappings to observed phenotypes. The two-track architectures (iGGE, cGGE, csGGE) enable modeling of phenotype dependencies—such as the requirement that a protein must fold before it can bind a ligand—while preserving phenotype-specific additive representations.

We demonstrated `torchdms` on the SARS-CoV-2 RBD dataset from Starr et al. [5], where two-track models achieved strong predictive performance ($R^2 > 0.85$ for both expression and binding) and recovered mutational effects that correlate well with independently published estimates. The visualization of learned nonlinearities revealed interpretable structure: expression predictions depend primarily on the expression latent phenotype, while binding predictions depend on both latents, capturing the biological constraint that proteins must fold before they can bind.

The models make several simplifying assumptions that warrant discussion. First, the additive latent phenotype representation assumes that mutational effects combine linearly before nonlinear transformation, which may not hold for proteins with strong pairwise or higher-order epistasis. While `torchdms` supports prelatent interaction layers to capture pairwise effects, these substantially increase model complexity. Second, the gauge-fixing procedure that centers coefficients at each position makes the β values dependent on which mutations were observed in the training data; coefficients are not directly comparable across datasets with different mutation coverage. Third, predictions may become less accurate for variants with many mutations relative to the training distribution, as the model has limited opportunity to learn higher-order effects.

The GGE framework relates closely to other global epistasis implementations. The original global epistasis model [24] used I-spline basis functions to guarantee monotonic nonlinearities, a more principled approach than `torchdms`'s weight clamping. MAVE-NN [46] and `dms-variants` include explicit noise models and information-theoretic metrics that quantify how much sequence variation is captured;

`torchdms` does not model measurement noise and relies on standard loss and R^2 metrics. The multiphenotype extensions in `torchdms` resemble fuzzy logic models that encode multiple biophysical traits [49] and connect conceptually to the multi-epitope polyclonal model described in Chapter 3. In both cases, an underlying additive genotype-phenotype map is transformed through nonlinear functions that capture how phenotypes combine.

Several limitations of `torchdms` could be addressed in future work. We hypothesized that the cGGE and csGGE architectures would outperform iGGE by capturing the interdependence between expression and binding, but this did not pan out for prediction: iGGE achieved the best predictive performance. However, csGGE did recover mutational effects most accurately, suggesting that modeling conditional dependencies may be more important for coefficient interpretation than for prediction. The lack of an explicit noise model prevents decomposition of variance into measurement error versus biological variability; incorporating measurement models similar to MAVE-NN would enable prediction intervals and better uncertainty quantification. `torchdms` is also not suitable for inferring mutational effects across DMS libraries created with different wildtype backgrounds, as the gauge-fixing procedure relies on a single wildtype reference. `multidms`, a similar software package, has since been developed to address this challenge by inferring shifts in mutational effects across different genetic backgrounds [50]. Finally, while the cGGE and csGGE architectures are designed to model conditional dependencies between expression and binding phenotypes, validating these models would require datasets where phenotypes are measured independently: a constraint not satisfied by the RBD data, where binding is only measured for variants that pass an expression threshold.

2.4.1 Code and data availability

The `torchdms` source code is archived at `https://github.com/matsengrp/torchdms`. This project is no longer actively maintained.

Chapter 3

A biophysical model of viral escape from polyclonal antibodies

Understanding how mutations combine to escape polyclonal antibodies is critical for predicting viral evolution, as polyclonal responses can target multiple distinct epitopes on a viral antigen. This chapter presents a biophysical model that decomposes polyclonal antibody activity into epitope-specific components and quantifies mutation effects at each epitope. We implement this model in `polyclonal`, a software package that infers these parameters from deep mutational scanning data. Validation on simulated deep mutational scanning experiments demonstrates that the software accurately recovers underlying biophysical parameters and predicts escape by arbitrary mutation combinations. The software is available at `https://jbloomlab.github.io/polyclonal`.

3.1 Introduction

Many viruses evolve antigenically to escape polyclonal antibody responses elicited from vaccination or prior infection [51, 52, 53, 54]. Understanding how viral mutations affect antibody escape is critical for interpreting viral evolution and predicting vaccine efficacy. Neutralization assays remain the gold standard for assessing how viral variants escape immune responses, but they are limited to retrospective analyses of a relatively small number of viral variants of interest [55, 56]. Genomic epidemiology enables real-

time tracking of viral evolution through public sequence databases, but cannot directly assess the antigenic phenotypes of emerging variants [57, 58, 59, 4].

Deep mutational scanning (DMS) experiments bridge this gap by systematically measuring how large libraries, often hundreds of thousands, of viral protein variants affect antibody binding and neutralization [12, 13, 60, 6, 61, 62]. A critical insight from DMS studies is that polyclonal antibody responses typically target multiple distinct epitopes on a viral protein, with antibodies targeting different epitopes having differing neutralization potencies [61]. This epitope heterogeneity means that escape from polyclonal sera requires mutations at multiple sites, and the phenotypic effect of combining mutations depends on whether they target the same or different epitopes.

However, the expansive sequence space of viral proteins makes it infeasible to experimentally measure all possible mutation combinations. This motivates the development of computational models that can predict how combinations of mutations affect antibody escape. Existing approaches include basic transformations of DMS data [63], models integrating phylogenetic and antigenic data [64], sequence-based prediction methods [65, 66], and neural networks trained on either DMS [44] or sequence data [42, 43]. While these methods have proven useful, many lack biophysically interpretable parameters that connect to the underlying biology of antibody-antigen interactions.

In this chapter, we present a biophysical model of viral escape from polyclonal antibodies. This model decomposes polyclonal sera into distinct epitope classes, each with interpretable parameters describing antibody activity and mutation escape effects. The model can be fit to DMS data and used to predict how combinations of mutations would be neutralized by the sera used to generate the experimental data [67]. We also describe `polyclonal`, a software package implementing this model, and validate it using both simulated and experimental data.

3.2 Results

3.2.1 Conceptualizing antibody epitopes

Viral antigens can be partitioned into distinct epitopes based on patterns of antibody escape. Classic experiments on influenza tested viral mutants against large panels of monoclonal antibodies [68, 69, 70]. Viral

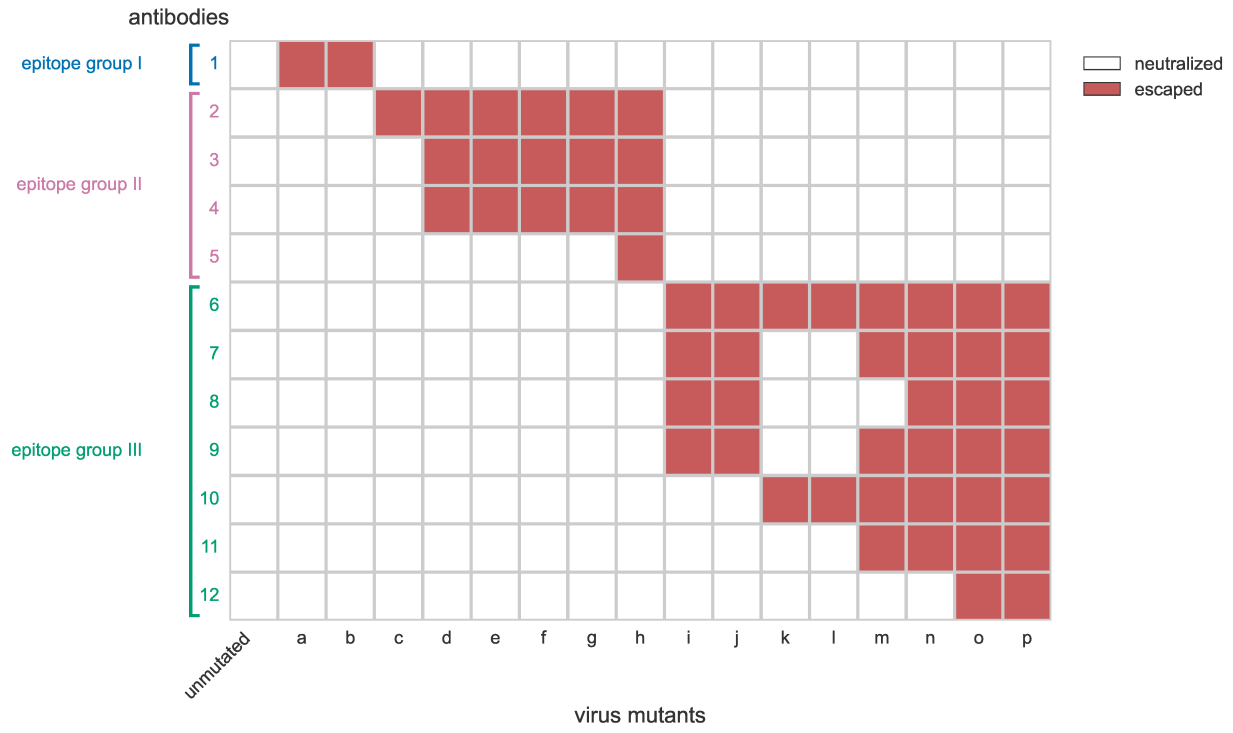

Figure 3.1: Heatmap displaying classical experimental data adapted from Webster and Laver (1980) [70]. Antibodies can be grouped into epitopes based on whether they share viral escape mutants. Influenza virus mutants (a–p) were selected using individual antibodies from a large panel of monoclonal antibodies. Each viral mutant was then tested to see if it was neutralized or escaped by the other antibodies in the panel. Antibodies were then grouped into epitopes based on their shared escape mutants.

escape mutants were first selected using individual antibodies and then tested against other antibodies in the panel. Groups of antibodies were escaped by similar viral mutants (Figure 3.1), and antibodies sharing common escape mutants were inferred to recognize a common epitope on the viral protein. In Figure 3.1, Antibodies 2–5 recognize a similar epitope since they are escaped by many of the same viral mutants. Based on these studies, the H3 influenza hemagglutinin was divided into five distinct epitopes [71, 72].

This epitope framework provides a straightforward way to conceptualize viral escape from polyclonal serum. Polyclonal serum contains multiple antibodies recognizing discrete epitopes, so any variant that fully escapes must possess mutations that allow escape at multiple epitopes. This framework has been applied to SARS-CoV-2, where the receptor-binding domain has been divided into multiple epitope classes [29, 73, 30].

Consider viral escape from a hypothetical polyclonal mixture of two antibodies that bind distinct epi-

topes, where Antibody 1 has higher functional activity than Antibody 2. A single mutation escaping Antibody 1 causes marked escape from the overall mixture since the more active antibody no longer contributes (Figure 3.2A). A single mutation escaping Antibody 2 but leaving Antibody 1 intact has little effect on the overall mixture, since the more active Antibody 1 can still bind (Figure 3.2B).

This framework makes distinct predictions about the effects of multiple mutations depending on whether they are in the same or different epitopes. If a variant has two mutations in Antibody 1's epitope and one mutation already escapes binding by Antibody 1, the second mutation has a largely redundant effect (Figure 3.2C). Multiple escape mutations in the same epitope have diminishing returns;even when all Antibody 1 molecules are unbound, Antibody 2 can still bind. However, if a variant gains mutations in the epitopes of both antibodies, both antibodies are escaped and the polyclonal mixture has no activity (Figure 3.2D).

Grouping antibodies by epitopes is a simplifying approximation. Two antibodies targeting a similar epitope may be escaped by slightly different mutations, and some antibodies bind idiosyncratic regions outside or between major epitopes [28, 73, 74, 61, 62]. In Figure 3.1, Antibodies 6 and 12 are grouped into the same epitope, even though some viral mutants only escape one of these two antibodies. Nonetheless, the concept of epitopes provides a valuable framework for interpreting polyclonal antibody escape and forms the basis of the quantitative model we present here.

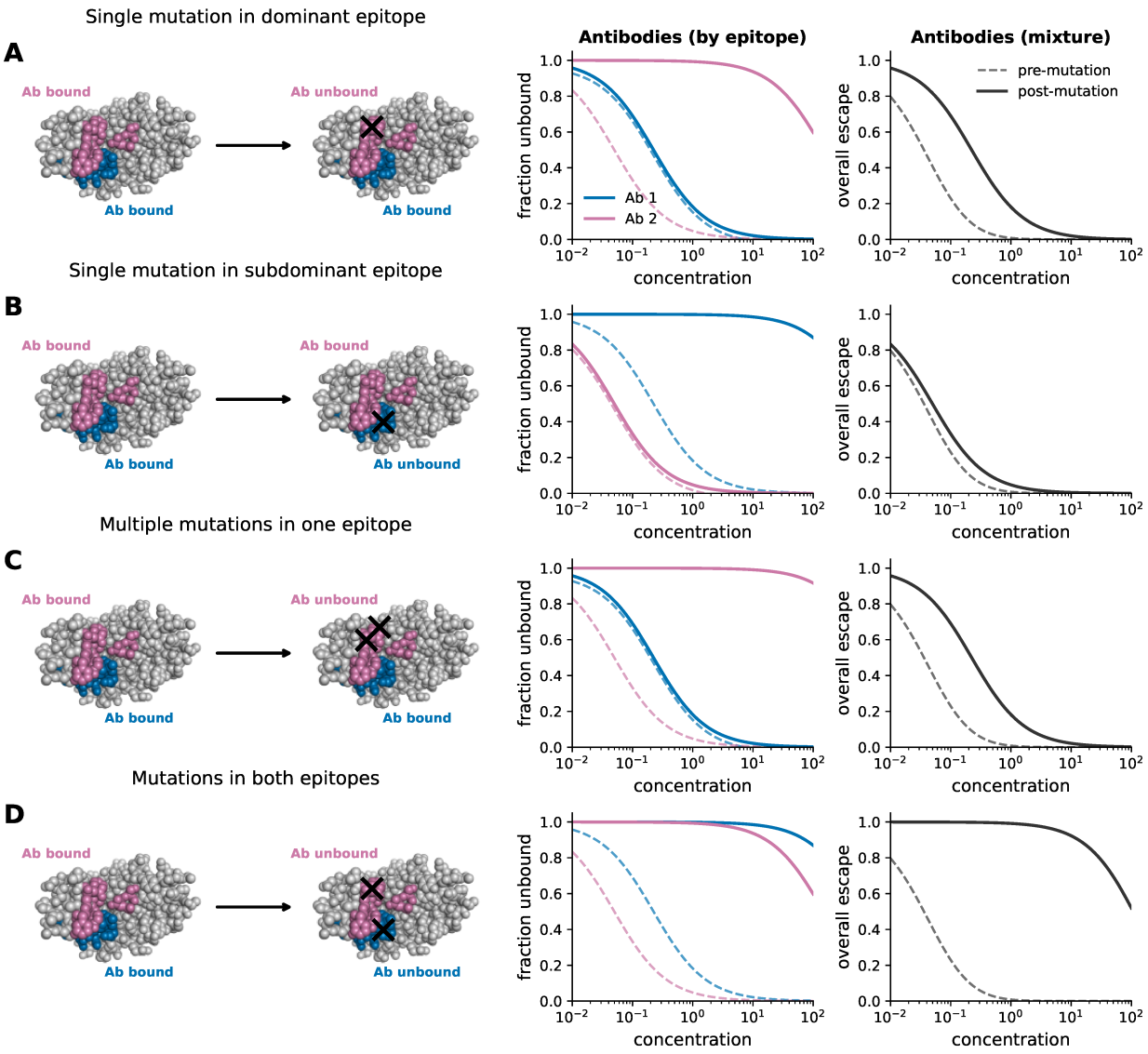

Figure 3.2: Viral escape from a hypothetical mixture of antibodies targeting two different epitopes. Each row illustrates a different mutation scenario and its impact on antibody binding. **(A)** A single mutation in the dominant epitope eliminates binding by the dominant antibody (Ab 2, pink) but not the subdominant antibody (Ab 1, blue), resulting in partial escape from the mixture. **(B)** A single mutation in the subdominant epitope eliminates binding by the subdominant antibody but not the dominant antibody, resulting in minimal change in overall escape. **(C)** Multiple mutations in the dominant epitope still only eliminate binding by antibodies targeting that epitope; the subdominant antibody continues to neutralize. **(D)** Mutations in both epitopes are required to fully escape the polyclonal mixture, as each mutation eliminates binding by antibodies targeting its respective epitope. Dashed lines indicate pre-mutation binding curves; solid lines indicate post-mutation curves.

3.2.2 A biophysical model of polyclonal antibody escape

Here we formalize the epitope-based model of polyclonal antibody escape in terms of experimental measurables and biophysical quantities. Let $p(v, c)$ be the fraction of viral variant v that escapes a mixture of polyclonal antibodies at concentration c, a quantity measurable from neutralization assays. Suppose the antibodies bind to one of E epitopes on the viral antigen, and let $U_e(v, c)$ be the fraction of variant v with an unbound epitope e at concentration c. Assuming antibodies bind independently without competition, the overall escape fraction is the product of unbound fractions across epitopes:

$$p(v, c) = \prod_{e=1}^{E} U_e(v, c). \tag{3.1}$$

The fraction unbound at each epitope follows a Hill equation [75] with coefficient one:

$$U_e(v, c) = \frac{1}{1 + c \exp(-\phi_e(v))}, \tag{3.2}$$

where $-\phi_e(v)$ represents the functional activity of antibodies targeting epitope e against variant v. More positive values of $-\phi_e(v)$ indicate higher functional activity. The activity $\phi_e(v)$ depends on both antibody-intrinsic factors such as binding affinity and extrinsic factors such as antibody concentration in the mixture.

We express $\phi_e(v)$ in terms of mutation effects by assuming mutations have additive effects on free energy of binding for antibodies at each epitope [24, 45]:

$$\phi_e(v) = -a_{wt,e} + \sum_{m=1}^{M} \beta_{m,e} b(v)_m, \tag{3.3}$$

where $a_{wt,e}$ is the functional activity of antibodies binding epitope e on the unmutated antigen, $\beta_{m,e}$ is the extent to which mutation m reduces antibody binding at epitope e, $b(v)_m$ is one if variant v has mutation m and zero otherwise, and M is the total number of possible mutations. Larger values of $a_{wt,e}$ indicate stronger antibody activity targeting the epitope, while larger values of $\beta_{m,e}$ correspond to greater contribution to antibody escape.

This model captures a key feature of antigenic epistasis: mutations have additive effects on binding affinity at each epitope but manifest non-linear effects on overall escape due to the Hill curve (Equation 3.2)

and the product across epitopes (Equation 3.1). Multiple mutations in the same epitope therefore show diminishing returns, while mutations across different epitopes combine synergistically(Figure 3.2).

3.2.3 Validation of `polyclonal` with a simulated deep mutational scanning dataset

We validated our software by simulating a hypothetical polyclonal antibody mixture targeting three major neutralizing epitopes (class 1–3) on the SARS-CoV-2 RBD [62] (Figure 3.3A). We assigned each epitope a pre-mutation functional activity ($a_{wt,e}$) based on the typical hierarchy of neutralizing activities in human polyclonal serum (class 2 $>$ class 3 $>$ class 1) [62] (Figure 3.3B). We assigned $\beta_{m,e}$ values using prior deep mutational scanning studies that measured effects of all single RBD mutations on escape from prototypical monoclonal antibodies targeting each epitope (Figure 3.3C). The site of greatest total escape is 417 for class 1, 484 for class 2, and 444 for class 3.

We then computationally simulated a realistic deep mutational scanning dataset containing 30,000 RBD variants with an average of two amino acid mutations per variant. We calculated the true $p(v, c)$ for each variant using our biophysical model at three concentrations representing IC97.5, IC99.9, and IC99.998 against the unmutated RBD. We added Gaussian noise ($\sigma = 0.05$) to the $p(v, c)$ measurements, truncated values to be between 0 and 1 to reflect experimental errors, and built in a 1% probability of adding or subtracting a mutation from a variant to simulate sequencing errors [67].

We found that `polyclonal` successfully inferred the underlying properties of our hypothetical polyclonal mixture when fit to the noisy, simulated data. The predicted $a_{wt,e}$ values closely matched the true values (Figure 3.4A), indicating that we can deconvolve the dominance hierarchy of epitopes targeted by a polyclonal mixture. The predicted $\beta_{m,e}$ values strongly correlated with the true values (Figure 3.4B; $R^2 = 0.60$ for class 1, $R^2 = 0.96$ for class 2, and $R^2 = 0.95$ for class 3), indicating that we can learn the effects of mutations on antibody escape at each epitope. The lower correlation for class 1 can be attributed to its relative subdominance: mutations to a subdominant epitope manifest little effect on the functional activity of the overall mixture when immunodominant epitopes remain unaffected (Figure 3.2B), and so only have measurable effects in the subset of mutants with mutations in more dominant epitopes.

We used the fit model to predict escape by variants not seen in our experiments. We simulated an independent deep mutational scanning dataset with a higher number of mutations (three) on average per

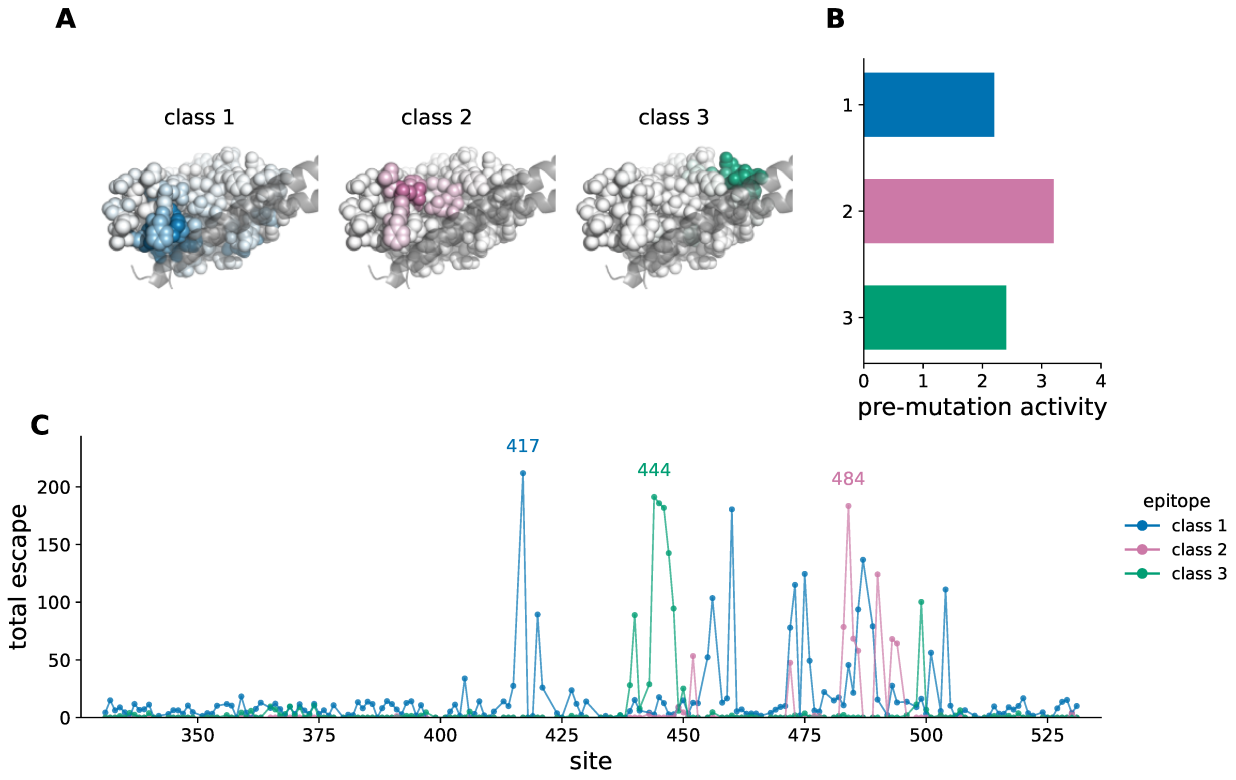

Figure 3.3: A hypothetical polyclonal antibody mixture targeting SARS-CoV-2 RBD. **(A)** Structures of the SARS-CoV-2 RBD in complex with key helices of ACE2. RBD sites are shaded to indicate the extent to which they mediate escape from antibodies targeting each epitope in the hypothetical polyclonal mixture. **(B)** Pre-mutation functional activities of antibodies ($a_{wt,e}$) for each epitope in the mixture. Note that the class 2 epitope is immunodominant, whereas the class 1 epitope is subdominant. **(C)** Sum of positive mutation escape effects ($\beta_{m,e}$) at each RBD site for each epitope. Key sites with high escape are labeled: site 417 (class 1), site 444 (class 3), and site 484 (class 2).

variant. We found that our fit model could predict the IC90 (the concentration required for $p(v, c) = 0.1$) of each variant in the independent dataset with high accuracy (Figure 3.4C; $R^2 = 0.97$). Thus, `polyclonal` can make predictions over all variants with mutations observed in the deep mutational scanning library.

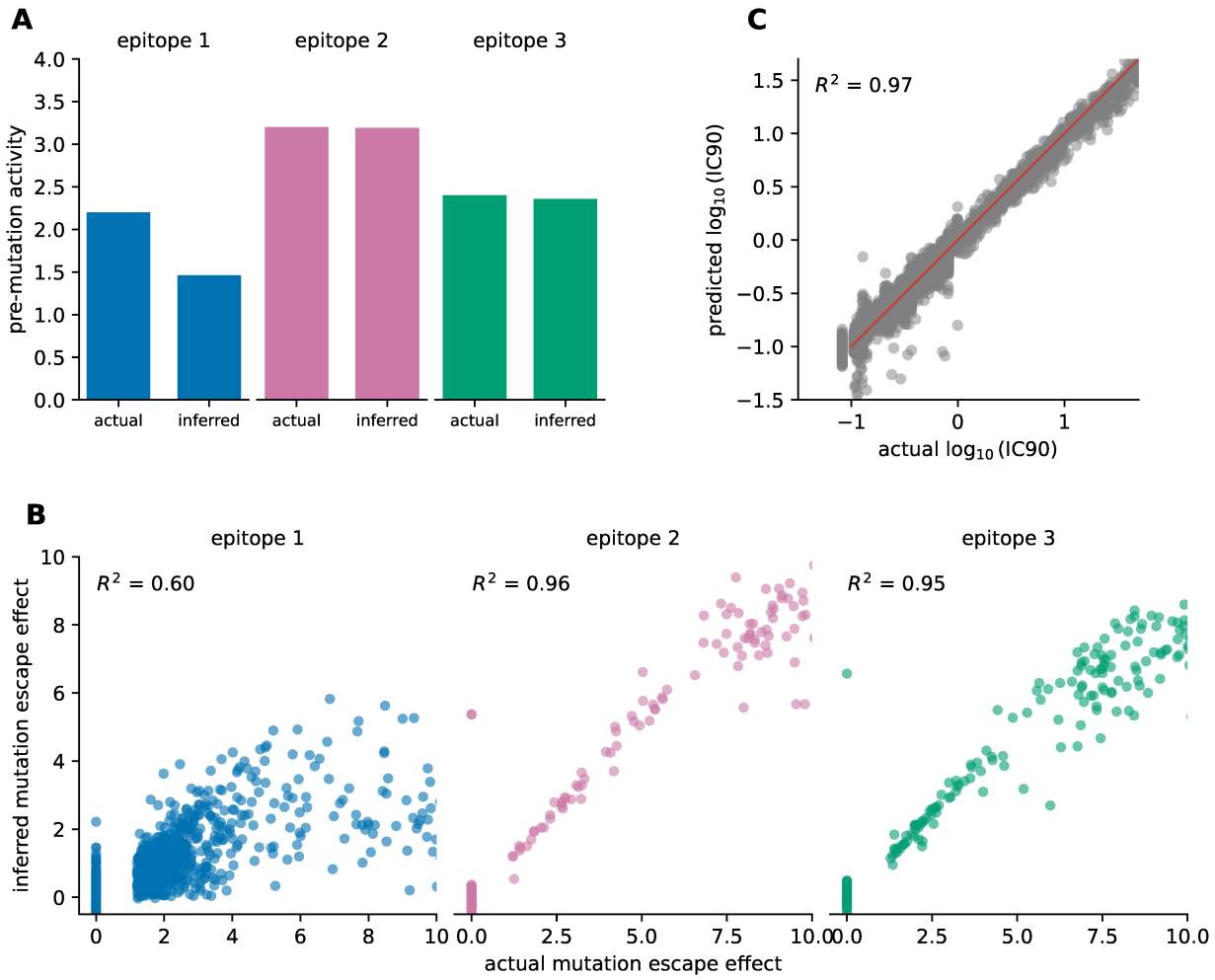

Figure 3.4: Model validation using a computationally simulated deep mutational scanning dataset. **(A)** Pre-mutation functional activities of antibodies ($a_{wt,e}$) inferred by the fit model match the actual $a_{wt,e}$ for each epitope. **(B)** Mutation escape effects ($\beta_{m,e}$) inferred by the fit model strongly correlate with the actual $\beta_{m,e}$ for each epitope used in the simulation. Mutation escape effects for the immunodominant epitope 2 are most correlated, while escape effects for the subdominant epitope 1 are least correlated. **(C)** The IC90s predicted by the fit model strongly correlate with the actual IC90s of an independently simulated dataset with a higher mutation rate (three mutations on average per variant).

3.3 Discussion

This chapter presented a biophysical model that captures how mutations combine to escape polyclonal antibodies. This model is implemented in `polyclonal`, a software package that infers epitope-specific parameters from deep mutational scanning data. A key feature of the model is that its formulation allows for natural modeling of antigenic epistasis: mutations have additive effects on binding affinity at each epitope, but the Hill curve relating affinity to fraction bound and the product across epitopes produce non-linear effects on overall escape. This results in mutations in the same epitope showing diminishing returns while mutations across different epitopes can combine synergistically to escape polyclonal serum.

The model makes several simplifying assumptions that warrant discussion. First, grouping antibodies by discrete epitopes approximates the continuous landscape of antibody binding; in reality, each antibody is unique and two antibodies targeting a similar epitope may be escaped by slightly different mutations. Despite this simplification, decades of experimental work have demonstrated that epitope-based representations provide a useful framework for interpreting viral escape [71, 72, 29]. Second, the model does not account for the fact that multiple antigens decorate each virion, instead modeling escape as the state where no antibodies bind a single idealized antigen. Third, the model assumes antibody binding at one epitope does not influence affinity at other epitopes. These assumptions could be relaxed in future versions of the model, though we expect them to hold reasonably well for variants with modest numbers of mutations. Predictions may become less accurate for variants with many mutations relative to the parental strain used in the deep mutational scanning experiment.

The single-epitope version of our model relates closely to global epistasis models used for interpreting other types of deep mutational scanning data [24, 46] (see also Chapter 2). In both cases, an underlying additive genotype-phenotype map is transformed through a non-linear function to produce the observed phenotype. The multi-epitope version extends this framework and resembles fuzzy logic models that encode multiple biophysical traits [49].

Following publication of this work, the `polyclonal` software has been applied to deep mutational scanning studies of both SARS-CoV-2 and influenza, enabling predictions of viral escape that correlate well with traditional neutralization assays [76, 77, 78, 79, 80]. These applications have contributed to understanding which mutations drive immune evasion and helped predict the evolutionary success of circulating

viral clades. More broadly, this work demonstrates how biophysical models can translate high-throughput experimental data into actionable predictions for viral surveillance and vaccine design.

3.4 Methods

3.4.1 Model fitting

The goal is to estimate the biophysical model parameters that best predict the deep mutational scanning escape measurements under biologically motivated constraints. We define a loss function robust to outliers:

$$L = \sum_{v,c} h_{\delta_{\text{loss}}}(p(v,c) - \hat{p}(v,c)), \tag{3.4}$$

where $p(v,c)$ is the actual escape fraction for variant v at concentration c, $\hat{p}(v,c)$ is the predicted escape fraction, and $h_\delta(x)$ is a scaled Pseudo-Huber function:

$$h_\delta(x) = \frac{\delta^2 \left(\sqrt{1 + (x/\delta)^2} - 1\right)}{\delta}. \tag{3.5}$$

The parameter δ controls when the loss transitions from quadratic (L2-like) to linear (L1-like) behavior. When the residual x is small, the loss is approximately L2; when x is large, the loss becomes L1-like, making it robust to outliers. A default $\delta_{\text{loss}} = 0.01$ was used in all fitted models. To minimize this loss function, we use the gradient-based L-BFGS-B method implemented in `scipy.optimize.minimize` [81].

3.4.2 Regularization

We implement penalty terms to regularize parameters under biologically motivated constraints. We regularize the escape values ($\beta_{m,e}$) using a Pseudo-Huber function, based on the assumption that most mutations should not mediate escape:

$$R_{\text{escape}} = \lambda_{\text{escape}} \sum_{m,e} h_{\delta_{\text{escape}}}(\beta_{m,e}), \tag{3.6}$$

where λ_{escape} is the regularization strength and δ_{escape} is the Pseudo-Huber delta parameter. Default parameters $\lambda_{\text{escape}} = 0.02$ and $\delta_{\text{escape}} = 0.1$ were used.

We also regularize the variance of escape values at each site, based on the assumption that mutations at a site involved in escape will exhibit similar effects:

$$R_{\text{spread}} = \lambda_{\text{spread}} \sum_{e,i} \frac{1}{M_i} \sum_{m \in i} \left(\beta_{m,e} - \frac{1}{M_i} \sum_{m' \in i} \beta_{m',e} \right)^2, \tag{3.7}$$

where λ_{spread} is the regularization strength, i ranges over all sites, M_i is the number of mutations at site i, and $m \in i$ indicates all mutations at site i. The default parameter $\lambda_{\text{spread}} = 0.25$ was used.

We regularize the pre-mutation functional activities ($a_{wt,e}$) using a Pseudo-Huber function, as these should be less positive (or even negative) unless there is clear evidence to the contrary:

$$R_{\text{activity}} = \lambda_{\text{activity}} \sum_{e} h_{\delta_{\text{activity}}}(a_{wt,e}) \quad \text{if } a_{wt,e} \geq 0, \tag{3.8}$$

where $\lambda_{\text{activity}}$ is the regularization strength, δ_{activity} is the Pseudo-Huber delta parameter, and $R_{\text{activity}} = 0$ when $a_{wt,e} < 0$. Default parameters $\lambda_{\text{activity}} = 1.0$ and $\delta_{\text{activity}} = 0.1$ were used.

3.4.3 Determining the number of epitopes

Fitting the model requires specifying the number of epitopes a priori, but the true number of epitopes targeted by antibodies in polyclonal serum is unknown in practice. Our approach is similar to the "elbow method" commonly used to determine the optimal number of clusters in k-means clustering. We start by fitting a model with one epitope and iteratively fit models with an increasing number of epitopes. At some point, the Nth epitope becomes redundant, evidenced by a highly negative $a_{wt,e}$ value (indicating that antibodies against this epitope would never be bound) and all near-zero $\beta_{m,e}$ values. This indicates that the previous fit model, containing $N - 1$ epitopes, best describes the polyclonal mixture. An example of this approach applied to the simulated RBD data is available at `https://jbloomlab.github.io/polyclonal/specify_epitopes.html`.

3.4.4 Simulating deep mutational scanning libraries

We computationally simulated deep mutational scanning libraries based on the hypothetical polyclonal antibody mixture shown in Figure 3.3. All libraries contained 30,000 variants but differed by their mutation rate,

with variants containing an average of one, two, three, or four mutations following a Poisson distribution. Variant escape was simulated under six concentrations representing IC91.443, IC97.488, IC99.441, IC99.9, IC99.985, and IC99.998 against the unmutated RBD antigen. For more details on how these experiments were simulated, see `https://jbloomlab.github.io/polyclonal/simulate_RBD.html`.

To compare model fitting across experimental conditions, models were fit with identical parameters except for the variable of interest: library mutation rate, library size, or antibody concentration. To determine the impact of library mutation rate, models were fit to simulated datasets containing one, two, three, or four mutations on average, all containing 30,000 variants measured at three concentrations. To determine the impact of library size, models were fit to different-sized subsets of variants randomly sampled from a library with three mutations on average per variant. To determine the impact of antibody concentration, models were fit to a library measured at one or multiple antibody concentrations. For more information on these experimental design simulations, see `https://jbloomlab.github.io/polyclonal/expt_design.html`.

3.4.5 Publication and acknowledgments

The material in this chapter is available as a published paper [67]. Timothy Yu and Jesse Bloom contributed to conceptualizing the biophysical model and implementing the software. William DeWitt provided helpful discussions regarding regularization approaches.

3.4.6 Code and data availability

Software source code is available at: `https://github.com/jbloomlab/polyclonal`

Code to reproduce analyses and figures is available at: `https://github.com/zorian15/phd-thesis-analysis/`

Chapter 4

Simulating influenza sequences under a model of host cross-immunity with `antigen-prime`

This chapter presents `antigen-prime`, a forward-time epidemic simulator that jointly models the genetic and antigenic evolution of viruses under selection from host population immunity. While previous chapters focused on modeling mutational effects from experimental data, this chapter addresses the challenge of simulating realistic viral evolution to benchmark computational methods used for influenza surveillance and vaccine strain selection. The simulator enables generation of ground-truth data for evaluating methods that assign viral sequences into variants and estimate variant-specific growth rates—tasks that are difficult to benchmark using natural data where true values are unknown.

4.1 Introduction

Influenza virus infects about one billion and kills hundreds of thousands of people globally each year [2]. Vaccination provides the primary method of prevention, but the virus undergoes rapid antigenic evolution, necessitating annual vaccine updates [1]. To track the virus's evolution, laboratories worldwide sequence thousands of influenza genomes each year, sharing data through public databases [58, 82]. Public health

agencies also monitor influenza spread through case counts and hospitalizations.

It is of interest to develop computational methods to use the above sources of data to help guide vaccine updates [83]. One goal of such methods is to partition observed viral sequences into groups (i.e. "variants") with similar antigenic phenotypes. Another goal is to estimate variant-specific growth rates, helping to quantify variant fitness in the present and forecast future variant abundance.

For seasonal influenza virus, two main computational methods exist for variant assignment. One method involves building a phylogenetic tree of observed sequences and then partitioning the tree into clades that correspond to unique variants [84]. This method can incorporate prior knowledge about which mutations are most likely to impact the virus's antigenicity. Another method involves computing pairwise distances between observed sequences, and then using dimensionality reduction to embed sequences in a low-dimensional space. Clustering of sequences in this space can then be used to identify unique variants [85]. Both methods have proven useful for interpreting real-time influenza surveillance data [85, 86].

A variety of methods exist for estimating variant-specific growth advantages from viral surveillance data. Some only consider variant-specific frequencies over time, such as the fitness model by Luksza and Lässig [87] or multinomial logistic regression approaches [88, 89, 90]. Others also consider observed numbers of case counts over time, such as the fixed growth advantage (FGA) and growth advantage random walk (GARW) models from the `evofr` framework [91].

Although these methods are widely used, it is challenging to benchmark their accuracy. That is because natural populations lack known ground-truth variant assignments and growth advantages. Experiments such as neutralization assays or hemagglutination inhibition assays can be used to quantify antigenic phenotypes of influenza viruses in a laboratory setting. However, it can be difficult for experiments to fully capture antigenic selection in nature due to the high level of heterogeneity in human immune responses to influenza [92].

A complementary approach is to simulate viral evolution under defined selective pressures and then use the simulated data, and associated ground-truth quantities, to benchmark the above methods. Doing so would require a simulator that models three key aspects of influenza evolution: (1) genetic sequence evolution through a biologically realistic mutation processes, (2) coupled antigenic evolution where mutations in epitope regions alter a virus's antigenic phenotype, allowing escape from host immunity, and (3) sustained viral transmission in a large host population, with case counts tracked over time. While numer-

ous pathogen-evolution simulators exist [93, 94, 95, 96], none fully integrate all three of these components. `SANTA-SIM` [93] models genetic sequence evolution, but not antigenic phenotypes or transmission dynamics. Conversely, `antigen` [94] models transmission dynamics and antigenic evolution, but does not model genetic sequence evolution. `FAVITES` [95] models contact network transmission without antigenic evolution, while `virolution` [96] models within-host evolution without population-level antigenic drift.

4.2 Results

4.2.1 Summary of `antigen-prime`

We sought to build a simulator that integrates all three components of influenza evolution listed above. We achieved this by extending the `antigen` simulator [94], which already models transmission dynamics and antigenic evolution. We updated `antigen` to include explicit genetic sequences that evolve through mutation and drive antigenic change. We call this updated simulator `antigen-prime`.

The original `antigen` implements a deme-structured SIR model to simulate viral transmission dynamics and antigenic evolution under selective pressure from host immunity. Each infected host carries a single virus object that is associated with coordinates that represent the virus's location in a two-dimensional antigenic space. The Euclidean distance between two viruses in this space approximates their antigenic similarity, that is, the degree of immune cross-reactivity between them, with smaller distances corresponding to greater cross-reactivity. Mutations to the virus move it in this space. Individual hosts maintain immune memory as a collection of antigenic coordinates from previous infections. When an infected host comes in contact with a susceptible host, the minimum Euclidean distance in antigenic space between the virus from the infected host and viruses from the susceptible host's immune memory determines the infection risk, with smaller distances conferring greater protection. Selection to escape immune memories in the host population drives viral antigenic evolution. However, viruses in `antigen` are not associated with genetic sequences as they exist only as points in antigenic space.

In `antigen-prime`, we updated the virus object to include not only the virus's coordinates in antigenic space, but also a nucleotide sequence of a protein-coding gene (Figure 4.1A). Simulations initialize with a single virus with a defined sequence, and this sequence evolves as the virus replicates and spreads in

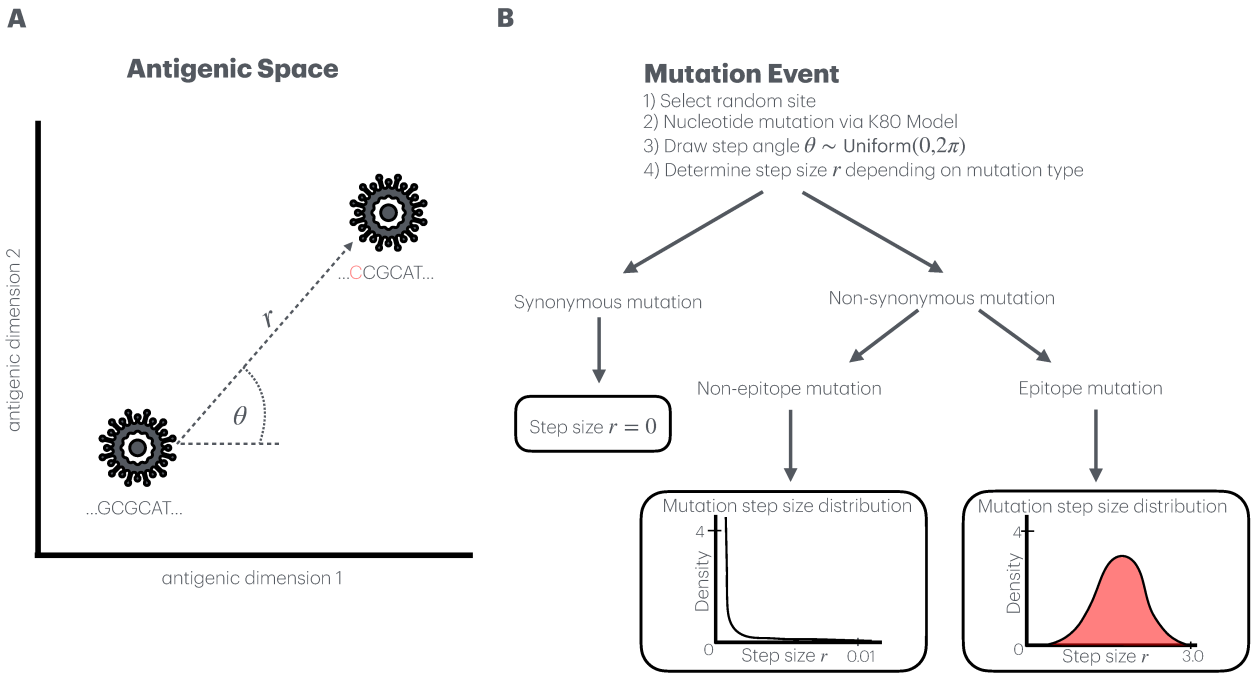

Figure 4.1: Schematic of how `antigen-prime` simulates coupled genetic and antigenic evolution of viruses. **A**: Each virus is represented by both a nucleotide sequence and coordinates in a d-dimensional antigenic space. Mutation events simultaneously change the nucleotide sequence and move the virus in antigenic space. **B**: During mutation events, a nucleotide site is randomly selected and mutated according to the K80 model, while antigenic coordinates are updated based on a sampled step direction θ and step size r that depends on whether the mutation is synonymous or non-synonymous and whether it occurs in an epitope or non-epitope site.

the host population. The user must define a subset of amino-acid sites in the protein to be "epitope sites", where mutations have large antigenic effects (all remaining sites are considered to be "non-epitope sites"). In this chapter, we used an influenza hemagglutinin (HA) nucleotide sequence that codes for a protein of length 566 amino acids. We classified 49 of these amino-acid sites to be epitope sites, using the set of epitope sites defined in Luksza and Lässig [87].

The nucleotide sequence of a virus object mutates at a user-defined mutation rate. Since there is only one virus object per infected host, this object's sequence is like a consensus sequence within the host, and the rate at which it mutates represents the rate at which mutations reach sufficient frequency within the host to transmit to new hosts upon contact events. Thus, we model this rate as a function of both neutral mutation and within-host purifying selection. We model the neutral mutation process using the K80 mutation model [97], parameterized by a transition/transversion ratio $\kappa = 5.0$ (user-configurable parameter

set to match empirical observations in influenza [98, 99]). Specifically, mutation events randomly select a nucleotide site and assign a new nucleotide by drawing from the K80-model probability distribution. We model purifying selection by rejecting a subset of mutations based on their properties. Mutations that introduce stop codons are rejected. Nonsynonymous mutations are probabilistically either accepted or rejected according to a user-defined "acceptance rate", with separate acceptance rates for epitope and non-epitope sites. In this chapter, we use acceptance rates of 0.75 and 0.2 at epitope and non-epitope sites, respectively, which we tuned to reproduce empirical H3N2 mutation statistics [100]. The higher acceptance rate at epitope sites reflects their greater tolerance of amino-acid change, consistent with positive selection for immune escape, whereas the lower rate at non-epitope sites reflects stronger purifying selection, as mutations there more often disrupt hemagglutinin function.

The antigenic effect of a nucleotide mutation depends on if and how it changes the protein sequence (Figure 4.1B). Synonymous mutations do not alter the virus's antigenic coordinates. Nonsynonymous mutations at epitope sites cause large movements in antigenic space with step sizes drawn from a gamma distribution with mean 0.6 antigenic units (analogous to the original `antigen`). Nonsynonymous mutations at non-epitope sites cause very small movements with step sizes drawn from a gamma distribution with mean 1×10^{-5} antigenic units. The direction of movement (θ) is random for both types of mutations. Thus, in `antigen-prime`, antigenic evolution is mostly driven by nonsynonymous mutations at epitope sites, while other mutations accumulate with little-to-no antigenic impact. Because the step direction is drawn independently for each mutation event, the same epitope substitution arising in different sequence backgrounds generally produces different antigenic displacements rather than a single, convergent antigenic effect, a design choice inherited from the original `antigen` [94]. `antigen-prime` also optionally supports fixed, pre-defined antigenic vectors for specific amino-acid changes, which would instead make a given substitution produce a consistent antigenic displacement. However, constraining mutations to fixed vectors reduces the diversity and variance of antigenic coordinates across the viral population, so we disabled this feature for our long-term simulation and note that it may be better suited to shorter-term simulations. Consequently, data generated under the random-direction model are best suited for benchmarking methods that do not rely on convergent antigenic evolution.

In summary, the main difference between `antigen` and `antigen-prime` is that virus objects in

`antigen-prime` have sequences. In `antigen`, mutation events directly result in changes to the virus object's location in antigenic space. In `antigen-prime`, mutation events first act on the virus's sequence, which in turn can result in changes to the virus's location in antigenic space. Other aspects of the simulator are the same, including antigenic selection on viruses to escape immune memory in the host population.

Related to antigenic selection, we have also updated `antigen-prime` to record population immunity over time. This feature enables benchmarking of variant-assignment methods by providing explicit fitness values for viruses. Every t days, the simulator samples n hosts and records the centroid of the most recent entry in each host's immune history, representing the average antigenic profile of recent infections. These immunity centroids enable fitness calculation for sampled viruses by computing infection risk based on the host centroid values at that time.

4.2.2 Simulating long-term influenza evolution using `antigen-prime`

We used `antigen-prime` to simulate 30 years of H3N2 seasonal influenza evolution. We initialized simulations with the full-length HA sequence from A/Beijing/32/1992. Each simulation comprised three geographical demes: the Northern hemisphere, Southern hemisphere, and tropics, each with 30 million hosts. We ran 30 parallel replicate simulations to quantify variability in outcomes.

As with the original `antigen`, `antigen-prime` simulations reproduced influenza-like phylogenetic structure and rates of antigenic evolution. In natural H3N2 evolution, the average time between two contemporaneous sequences and their most recent common ancestor is expected to be $\sim$3.22 years [101], reflecting influenza's spindly phylogenetic structure. The average rate of antigenic change is expected to be $\sim$1.6 antigenic units per year, as quantified using experimental data from hemagglutination-inhibition assays [51, 102]. Many `antigen-prime` simulations resulted in values close to these numbers (Figure B.1).

`antigen-prime` simulations also reproduced influenza-like mutational patterns. As an empirical reference, we considered a phylogenetic tree that Huddleston et al. made using seasonal H3N2 influenza HA sequences collected over 25 years [100]. For this tree, and for each of our simulated trees, we counted the number of epitope and non-epitope mutations observed along the branches of the tree, separately doing so for trunk branches and side branches, and using the set of epitope sites from Luksza and Lässig [87]. Many of the simulated trees had counts similar to the empirical tree (Figure B.2). Table 4.1 shows counts

Mutation Characteristic	Empirical Value (25 years)	Scaled Value (30 years)	Simulation Value (30 years)
Trunk Mutations			
Epitope mutations	50	60	66
Non-epitope mutations	32	38	49
Ratio (epitope:non-epitope)	1.56	1.56	1.35
Side Branch Mutations			
Epitope mutations	485	582	1254
Non-epitope mutations	1177	1412	3294
Ratio (epitope:non-epitope)	0.41	0.41	0.38

Table 4.1: Comparison of mutation patterns between an empirical phylogeny and the simulated phylogeny used for downstream analysis. The table shows counts of mutations occurring in epitope and non-epitope regions along the trunk vs. side branches. Empirical values are derived from 25 years of H3N2 HA sequences from Huddleston et al. [100], with trunk and side branch annotations following the methodology of Bedford et al. [103]. Scaled values extrapolate empirical counts to 30 years ($\times 1.2$) for comparison with simulation values. Simulation values are from a 30 year simulation.

for a single simulation that reproduced realistic influenza-like dynamics in terms of mutational patterns, phylogenetic structure, and antigenic movement, and which we describe below in more detail. On the trunk, the simulated and empirical trees have similar mutation counts, and similar ratios in counts between epitope and non-epitope mutations. On side branches, the mutation counts are substantially higher for the simulated tree because this tree has substantially more sequences (this high sampling density was useful for downstream benchmarking purposes). However, the ratio of counts between epitope and non-epitope mutations, which is not sensitive to sampling density, is similar between the simulated and empirical trees. Given the complexity of natural influenza evolution and the simplifying assumptions in our model, we expect approximate rather than precise agreement with empirical values.

In the following sections, we use the simulation characterized in Table 4.1 to benchmark methods for analyzing influenza sequences. This simulation reproduced various aspects of influenza evolution. A phylogenetic tree of viruses sampled from this simulation exhibits a characteristic ladder-like structure (Figure 4.2A). Case counts across the three demes display realistic seasonal patterns, with Northern and Southern demes exhibiting opposite seasonal peaks and the tropics showing more constant transmission (Figure 4.2B). The distribution of sampled viruses in antigenic space reveals distinct antigenic clusters that show sustained antigenic evolution over time (Figure 4.2C). Epitope mutations accumulate approximately linearly over time, beginning from the roughly 10 mutations carried over from the discarded 10-year burn-in period (Figure 4.2D). Global variant frequency dynamics show sequential variant emergence-and-replacement patterns

observed in the original `antigen` [94] (Figure 4.2E). Panels A, B, C, and E confirm that `antigen-prime` recapitulates antigenic and epidemiological dynamics already demonstrated by the original `antigen` [94], whereas panel D reflects a capability unique to `antigen-prime`: epitope mutations accumulating at a measurable rate that is tied to the underlying genetic sequence evolution.

4.2.3 Benchmarking variant-assignment methods

We sought to use the above `antigen-prime` simulation to benchmark methods for grouping viral genetic sequences into variants with similar antigenicity, with the simulation providing ground-truth coordinates of each virus in antigenic space. We evaluated three methods. The first method uses k-means clustering of viruses by their ground-truth antigenic coordinates, providing a baseline for comparison. The second method uses Neher's clade suggestion algorithm [84] to cluster viruses based on a phylogenetic tree built from their sequences. This approach defines variants by considering tree topology, overall genetic divergence, and amino acid changes at epitope sites, and has been applied in efforts to guide seasonal influenza vaccine composition [86]. We parameterized this method using the same set of 49 epitope sites that we used in the simulation. The third method uses `pathogen-embed` [85] to compute pairwise Hamming distances between viral sequences, then use those data to project sequences into a low-dimensional space using t-SNE, and cluster sequences in that space using k-means. We configured the antigenic and sequence-based methods to produce 30 variants, and the phylogenetic method to produce approximately 30 variants (resulting in 31) for comparison. When visualized in antigenic space, all three methods create largely distinct clusters (Figure 4.3A-C), even though the latter two approaches use only genetic data. As expected, the first approach results in strictly non-overlapping clusters. The latter two approaches result in clusters that are mostly separated, but have some overlap, reflecting the inherent difficulty in cleanly resolving antigenic boundaries from genetic data alone.

We quantified method performance using the following three metrics. First, we quantified the number of circulating variants per year. We tuned each method to produce approximately three per year to reflect real-world influenza dynamics [86]. Each method resulted in similar traces of this quantity over time (Figure 4.3D).

Second, we quantified the variance in antigenic fitness among viruses assigned to a given variant (Fig-

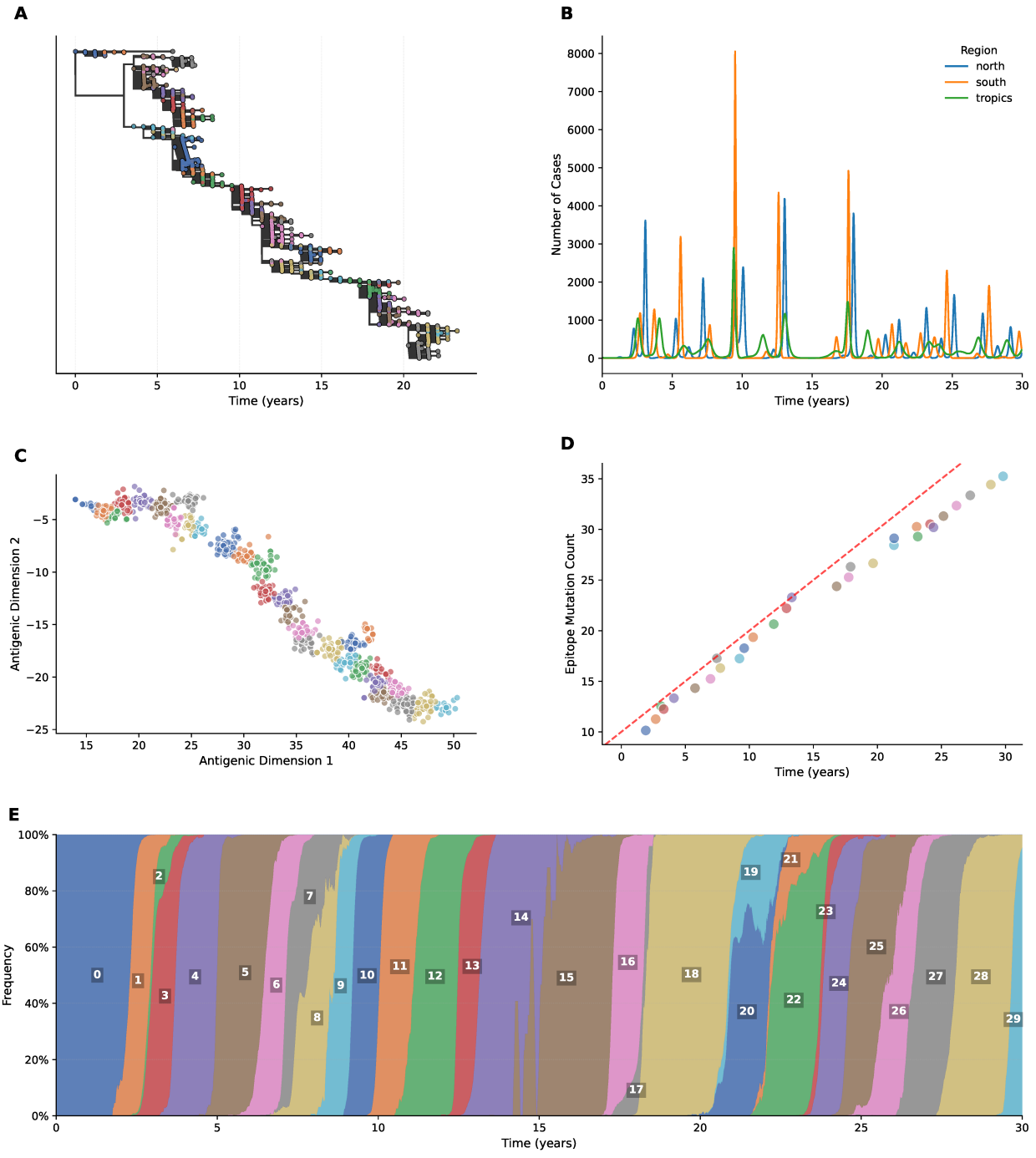

Figure 4.2: Genetic, antigenic, and epidemiological dynamics from a 30 year `antigen-prime` simulation. **(A)** Phylogenetic tree of 150,000 viruses sampled from the simulation, with tips colored by antigenic variant assignment. **(B)** Case counts across three geographic demes (North, Tropics, South), demonstrating realistic seasonal epidemic patterns with hemispheric differences. **(C)** Antigenic space of sampled viruses colored by variant, with variants assigned using k-means clustering on antigenic coordinates. The data show distinct antigenic clusters. **(D)** Epitope mutations accumulate approximately linearly over time (red dashed line shows 1 mutation/year reference), indicating consistent antigenic drift throughout the simulation. The plotted 30-year window follows a discarded 10-year burn-in (see Methods), so approximately 10 epitope mutations have already accumulated at the start of the window. **(E)** Global variant frequency dynamics aggregated across all demes, showing sequential variant emergence and replacement patterns. Panels A, B, C, and E reproduce dynamics first shown with the original `antigen` [94], while panel D is unique to `antigen-prime`.

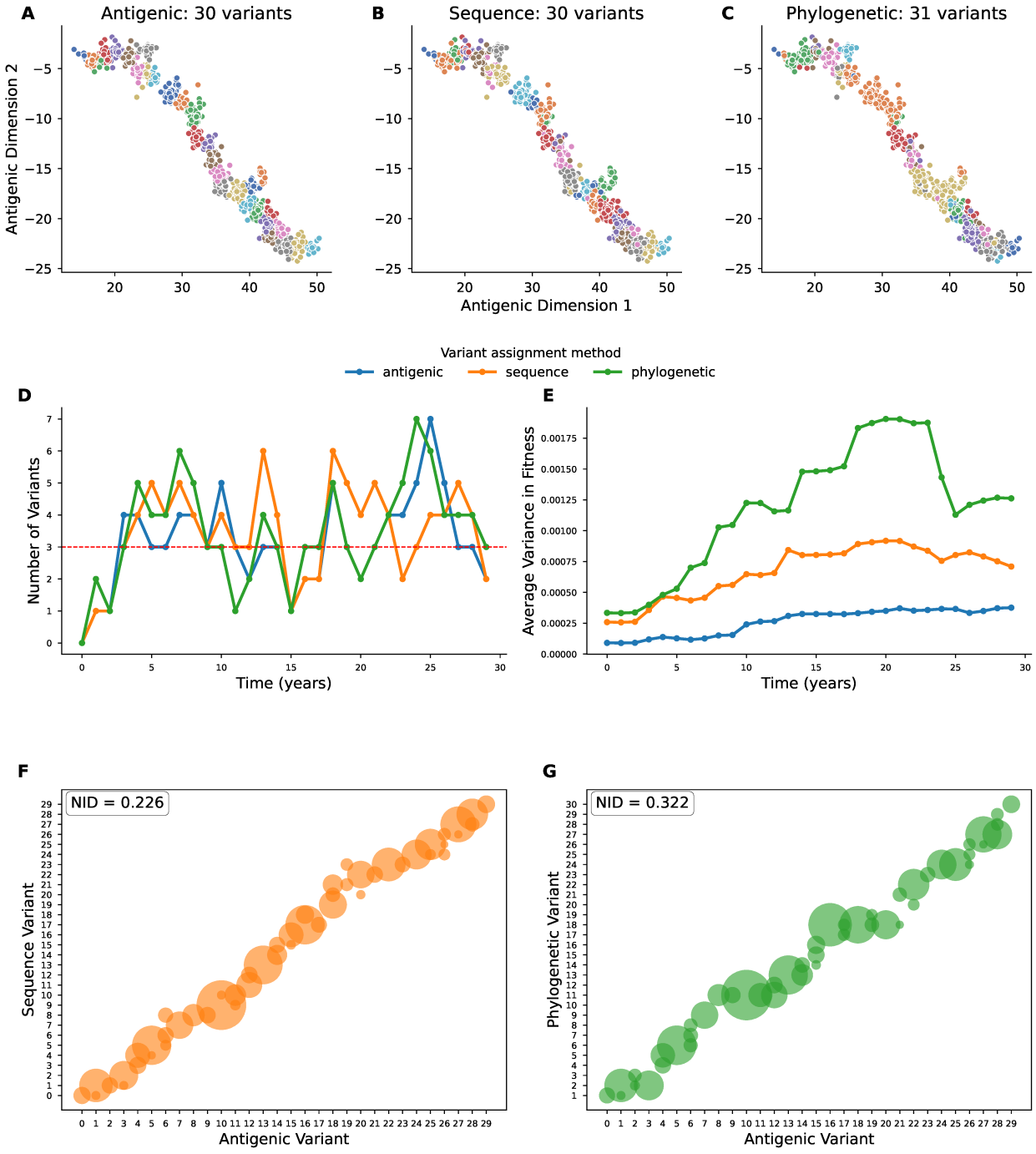

Figure 4.3: Benchmarking variant-assignment methods. **(A-C)** Variant assignments visualized in simulated antigenic space. Each panel shows the same viral sequences plotted by their antigenic coordinates, with colors representing different variants. **(D)** Number of variants circulating over time. The red dashed line marks three variants per year, reflecting typical real-world dynamics. **(E)** Average within-variant fitness variance over time. Lower values indicate better grouping of viruses with similar fitness. **(F)** Variant-assignment agreement between the sequence-based method and ground-truth antigenic variants. Bubble sizes represent the number of sequences shared between compared variants. The overlap is high (NID = 0.226). **(G)** Same as panel F, but for the phylogenetics-based method. The overlap is intermediate (NID = 0.322).

ure 4.3E). Specifically, for each year, we used the annual host population immunity centroid from that year to calculate the fitness of each sampled virus from the entire simulation, where fitness represents infection risk and is a function of the distance between a virus and the immunity centroid in antigenic space. For each variant, we then computed the variance in these fitness values among viruses assigned to that variant, and then averaged these variances across all variants. The blue line from Figure 4.3E shows the performance of the first variant-assignment method, which clusters viruses by their ground-truth coordinates in antigenic space, and which we use as a baseline for evaluating the other two methods. The blue line is close to zero for all years, indicating low variance in fitness values and thus high performance. The other two methods also result in low variance (see the orange and green lines), as expected from their ability to effectively cluster viruses in antigenic space (Figure 4.3B/C). However, the variance of these methods is $\sim$3-10-fold higher than the blue baseline, reflecting the observation that there is some overlap between clusters. The sequence-based method shows consistently lower variance, and thus higher performance, than the phylogenetic method.

Third, we quantified the extent that the sequence- and phylogenetics-based assignments agreed with the ground-truth assignments from k-means clustering in antigenic space (Figure 4.3F,G). We quantified agreement using the normalized information distance (NID) [104]; a distance metric ranging from 0 to 1 where lower values indicate more agreement between two variant assignments and higher values indicate less agreement. Both the sequence-based and phylogenetic-based methods resulted in assignments with high overlap with ground-truth assignments, with the former method showing better overlap (NID = 0.226) than the latter (NID = 0.322).

Overall, this benchmark indicates that both the sequence-based and phylogenetic-based variant-assignment methods are effective at grouping viruses into antigenically similar variants, with the former method showing higher performance. The success of these approaches also helps validate that `antigen-prime` successfully implements the fundamental coupling between genetic sequences and antigenic phenotypes that drives influenza evolution.

4.2.4 Benchmarking methods for inferring variant-specific growth rates

Next, we sought to use the simulated data to benchmark the ability of models to infer variant-specific growth rates. Natural data on SARS-CoV-2 has been used to benchmark the ability of MLR models to perform this task [105]. But, previous studies have not benchmarked the more sophisticated FGA and GARW models.

To derive ground-truth growth rates from the simulated data, we divided the 30 years of data into overlapping one-year windows staggered every six months, capturing influenza seasons in both Northern and Southern demes. We further divided the simulated data by North, South, and Tropics demes. Then, for each window from each deme, we computed variant-specific frequencies and growth rates in weekly time bins, using variant assignments from k-means clustering of viruses in antigenic space. Specifically, we defined the empirical growth rate of variant v as the per-bin change in its log incidence,

$$r_v(t, d) = \frac{\log I_v(t, d) - \log I_v(t - \Delta, d)}{\Delta}, \tag{4.1}$$

where $I_v(t, d) = f_v(t, d)\, C(t, d)$ is the incidence of variant v in deme d, given by the product of its frequency $f_v(t, d)$ and the total case counts $C(t, d)$, and Δ is the spacing between time bins in days. Because this growth rate reflects the change in *absolute* variant incidence rather than relative frequency, a variant can have a negative growth rate while still increasing in relative frequency, as occurs when overall incidence declines during a seasonal trough but the variant declines more slowly than its competitors. We omitted some variants at some time points due to insufficient case or sequence count data to accurately derive growth rates.

We then tested the ability of the FGA and GARW models from `evofr` to recover these growth rates. We separately fit each model to each window of data from each deme. As input, the models take the total number of reported cases amongst the host population and the observed counts for each variant in the sampled viral population. They then use a Bayesian approach to estimate probability distributions of variant-specific growth rates and frequencies over time. To analyze model predictions, we sampled 500 times from the inferred posterior distributions for variant growth rates and report the median values and 95 % highest posterior density (HPD) intervals. For each analysis window, we then calculated the mean absolute error (MAE) between predicted and ground-truth growth rates.

Below, we mostly focus on results from the GARW model as it allows growth advantages to vary smoothly over time, accommodating situations where the fixed-advantage assumption made in the FGA model may break down: for instance, due to shifting population immunity or cross-immunity between variants [91]. Results for the simpler FGA model show comparable performance and are available in the supplement (Figure B.3, Figure B.4).

The performance of the GARW model varied across analysis windows. For many windows, the MAE values are low near zero, indicating accurate growth-rate predictions (Figure 4.4). Figure 4.4B provides a detailed view of one such window, which shows good agreement between variant-specific frequencies and growth rates inferred by the model (see the lines) and those directly derived from the simulated data (see the circles).

However, for several other windows, the MAE values are substantially higher, pointing to inaccurate growth-rate predictions. Examining these windows revealed a failure mode that has not been documented in previous studies. Figure 4.4C-E provides three examples of this failure mode. In each case, there is one variant that initially predominates and then begins to decline in frequency as one or two low-frequency variants start increasing in frequency. The model accurately predicts the frequency trajectory of each variant. However, in many weekly time bins, the model does not accurately predict variant-specific growth rates, especially in bins near the middle or end of the window when the initial predominant variant begins to substantially decline in frequency. Often, the prediction error is larger than the estimated model uncertainty (observed growth rates fall outside the HPD interval). Strikingly, there are multiple examples where the predicted and observed growth rates have opposing signs, like in Figure 4.4D where the pink variant is predicted to have a positive growth rate in several time bins where it actually has a negative growth rate, or in Figure 4.4C where the opposite is true for the purple and red variants. In general, in time bins with enough data to quantify growth rates for multiple variants, when the predicted growth rate is inaccurate for one variant, it tends to be inaccurate for the other variants by roughly the same amount and in roughly the same direction, indicating systematic bias. This bias may stem from an underlying assumption in both the GARW and FGA models that variant-specific growth rates tend to change in a concerted manner.

In all, this benchmark indicates that the GARW and FGA models are largely effective at predicting variant-specific growth rates from the simulated data. However, it also revealed potential problems with

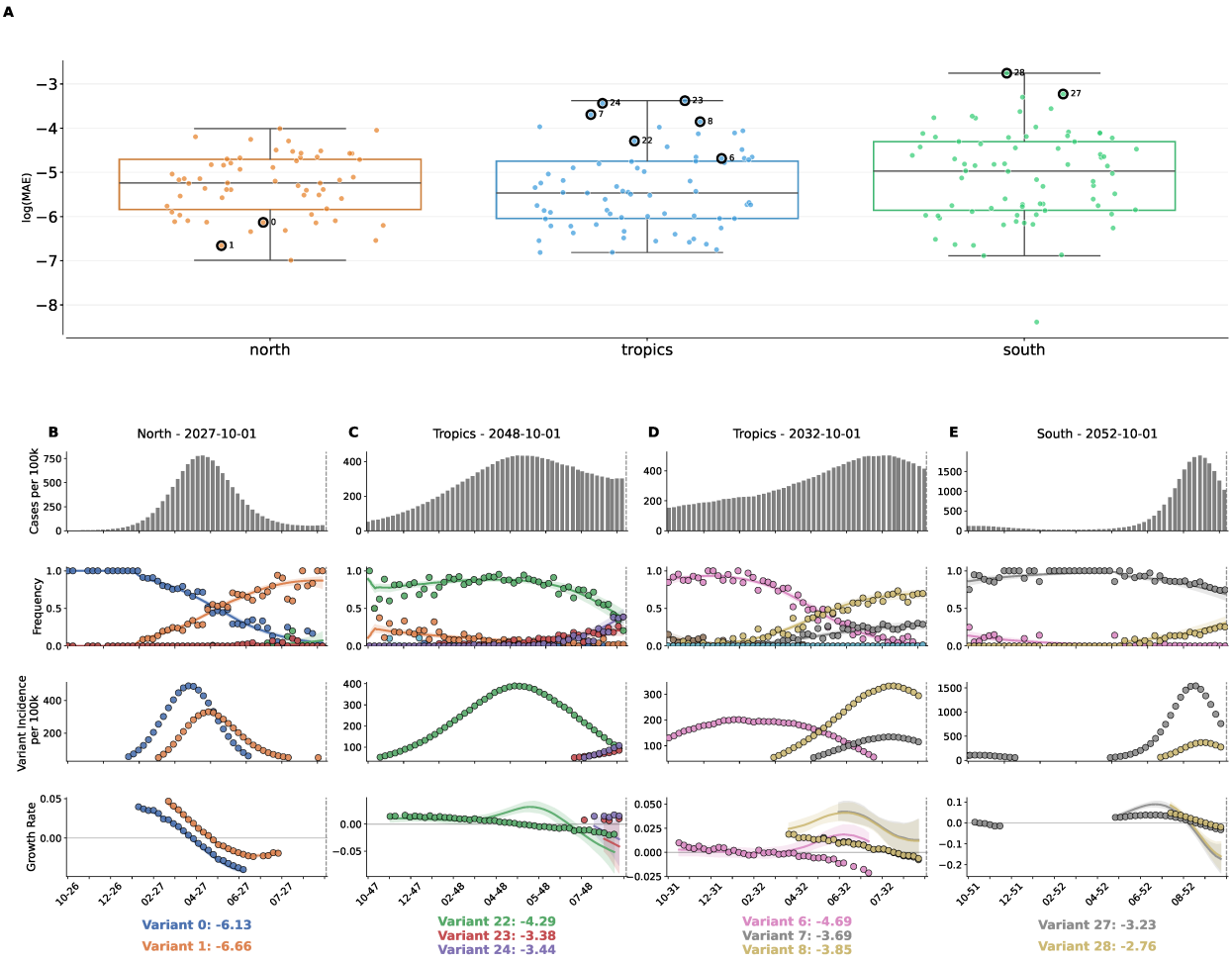

Figure 4.4: GARW model performance on variant growth-rate inference. **(A)** Distribution of log MAE values of inferred variant-specific growth rates across geographic demes. Each data point represents a single variant from a specific training window. Variants selected for panels B-E are circled. **(B-E)** Example analysis windows. Each panel shows case counts per 100,000 hosts over time (top), variant frequencies (second), case counts by variant (third), and inferred growth rates (bottom) for a given window. Growth rates reflect the rate of change of *absolute* variant incidence (variant frequency $\times$ total case counts; Equation 4.1), so both variants can show negative growth rates during a seasonal decline in overall incidence even as the fitter variant rises in relative frequency. Points show values derived directly from the simulated data. Not all variants are shown at all time points due to data filtering on observed count thresholds for a series of timepoints (see Methods). Solid lines show median of model inferences and shaded regions indicate 95 % HPD intervals. Average $\log$ MAE values for each variant are reported below the growth-rate plots. **(B)** Successful inference of frequencies and growth rates (North, 2027-10-01). **(C)** Growth rates underestimated near end of the analysis window (Tropics, 2048-10-01). **(D)** Growth rates overestimated for multiple variants (Tropics, 2032-10-01). **(E)** Growth rates inaccurately inferred for both variants 27 and 28 (South, 2052-10-01).

these models that motivate additional investigation.

4.3 Discussion

We have presented `antigen-prime`, a simulator that jointly models the genetic and antigenic evolution of viruses under selection from host population immunity. We showed that `antigen-prime` can be used to simulate H3N2 seasonal influenza-like dynamics over long timescales. The dynamics are influenza-like in terms of their phylogenetic structure, antigenic evolution, and sequence-level mutation patterns, with mutations at epitope sites driving antigenic change. We then used the simulated data to benchmark computational methods that are currently used to interpret influenza surveillance data, and have relevance for public-health decision-making.

Our work helps address a fundamental challenge in evaluating these computational methods: the lack of ground-truth data in natural viral populations makes it difficult to assess whether the methods accurately capture the biological processes they aim to model. Simulated data with known ground-truth values can be used to benchmark methods. But, the field lacks simulators of pathogen evolution that model both genetic and antigenic evolution under selection from host population immunity. `antigen-prime` fills the gap, enabling benchmarking of the methods we analyzed in this chapter.

The benchmarking results update our understanding of the efficacy of these methods. The benchmark on variant assignment showed that both the sequence-based and phylogenetic-based methods were effective at grouping viruses with similar antigenic properties even without explicit access to fitness or antigenic information. Interestingly, the sequence-based approach performed better than the phylogenetic one. The reason for this is not immediately evident to us, and could warrant future exploration. Future work could also benchmark variant assignment over shorter evolutionary time scales relevant for interpreting real-time influenza evolution. We note that performance on this benchmark is probably higher than expected for natural viral populations, since predicting phenotype from genotype is more challenging in nature due to factors not captured in `antigen-prime`, such epistasis between mutations and variable immunity in the host population.

The benchmark on growth-rate inference showed that the GARW and FGA methods performed well in most time windows, but also revealed windows where model predictions dramatically differed from ground-

truth. While the GARW method is conceptually more flexible, this additional flexibility did not provide substantial advantages over the simpler FGA method in our benchmark. Investigating low-performance windows identified a previously undocumented failure mode. Thus, these results helped identify limitations to these methods, and could be used to help guide future method development. We focused this benchmark on the FGA and GARW models because they infer growth rates from variant frequencies together with case counts, the same quantities from which we derive our ground-truth growth rates, and because they reflect approaches currently used in influenza surveillance. We did not directly benchmark the fitness model of Luksza and Lässig [87], despite its conceptual influence on `antigen-prime`, because that model predicts clade frequencies from a sequence-based fitness function rather than consuming case counts; adapting it to recover the case-count-based growth rates evaluated here would require re-deriving the model, which we leave to future work. In the future, `antigen-prime` simulated data could also be used to benchmark the ability of methods to forecast influenza evolution, which is also highly relevant to developing effective vaccines.

Despite capturing many features of viral evolution under immune selection, `antigen-prime` makes several simplifying assumptions that limit its biological realism. The fitness of simulated viruses is determined solely by their antigenic phenotype. However, other models of influenza evolution also model potential fitness costs of mutations at non-epitope sites [87], where mutations can disrupt HA's ability to mediate viral entry. Additionally, the model maintains static epitope sites throughout the simulation, and employs a simple mutation model without epistatic interactions. The antigenic space itself is also a simplification. We represent antigenic phenotype in two dimensions, following the antigenic cartography of Smith et al. [51], which showed that influenza antigenic relationships are largely captured in two dimensions; higher-dimensional representations that might capture finer antigenic structure are a direction for future work. Furthermore, because antigenic relationships are measured as Euclidean distances, cross-reactivity in our model is necessarily symmetric, so that immunity to one virus protects against a second exactly as much as the converse; cross-reactivity measured in nature, for example by hemagglutination-inhibition assays, is often asymmetric. Relaxing this symmetry assumption is another avenue for future work.

Our representation of host immunity is similarly simplified. Each host's immunity is a set of discrete antigenic coordinates recorded from its prior infections, and susceptibility to a new virus is governed by the

single nearest prior infection (the minimum antigenic distance; Equation 4.2), rather than by the cumulative, broadly cross-reactive antibody landscapes that characterize real influenza immunity [106, 107]. Hosts retain all prior infections indefinitely, so the model does not represent waning of immunity, and it likewise omits immune imprinting, whereby a host's first childhood exposure shapes subsequent responses and creates systematic differences across birth cohorts [108, 109, 110, 111], as well as vaccination, which is a dominant driver of population immunity in the settings the benchmarked methods target. These simplifications are inherited from the `antigen` framework [94] as a deliberate tradeoff toward tractability. Together with the antigenic-space assumptions above, they likely make the simulated immunity landscape smoother and more predictable than its natural counterpart, so the benchmarking results reported here are best interpreted as performance bounds under a simplified immunity landscape rather than as estimates of performance on real surveillance data. Incorporating waning, vaccination, and imprinting is a goal of future work.

In all, `antigen-prime` provides a powerful framework for simulating seasonal influenza evolution, enabling researchers to benchmark and guide development of methods for interpreting influenza surveillance data. In the future, `antigen-prime` could be tuned to simulate evolutionary dynamics of other viruses, helping to benchmark methods for a variety of pathogens related to human health.

4.4 Methods

4.4.1 Data and code availability

The `antigen-prime` simulator source code is available at `https://github.com/matsengrp/antigen-prime`. Analysis scripts, simulation outputs, and code to generate all figures are available at `https://github.com/matsengrp/antigen-forecasting`.

4.4.2 Publication and acknowledgments

The material in this chapter is pending review at *Virus Evolution* [112]. Thien Tran contributed to updating the Java software. John Huddleston, Marlin Figgins, and Hugh Haddox provided useful discussions of forecasting models and variant assignment methods.

4.4.3 Implementation of `antigen-prime`

`antigen-prime` is implemented as a Java program forked from the original `antigen` simulator [94]. The software is compiled using Maven and requires Java 11 or higher, with dependencies specified in the project's `pom.xml` file. The complete source code is available at `https://github.com/matsengrp/antigen-prime`.

Two major extensions distinguish `antigen-prime` from the original simulator: (1) explicit genetic sequence evolution with site-specific mutation effects on antigenic movement, and (2) periodic sampling of host population immunity to enable fitness calculations for downstream analysis.

The core simulation algorithm follows the discrete-event SIR (Susceptible-Infected-Recovered) framework described in Bedford et al. [94], with key extensions to couple genetic and antigenic evolution. The simulation proceeds through discrete time steps, modeling viral transmission dynamics across structured host populations (demes) while tracking both genetic sequences and antigenic coordinates for each virus. At each time step, the algorithm processes infection events based on host susceptibility and viral fitness, applies stochastic mutations to viral genomes according to the K80 model, updates antigenic coordinates based on mutation type and location, and samples host immunity profiles periodically to calculate population-level immunity centroids.

The risk that host h is infected upon contact with virus v is a piecewise-linear function of $d^{\star}_{v,h}$, the minimum Euclidean antigenic distance between v and the antigenic coordinates stored in h's immune memory:

$$\mathrm{ROI}_{h,v}(t, d) = \min\big(1,\ \max\big(1-\omega,\ s\, d^{\star}_{v,h}\big)\big)\,, \tag{4.2}$$

where s is the slope converting antigenic distance into infection risk (the `smithConversion` parameter) and ω is the immunity conferred by an antigenically identical prior infection (the `homologousImmunity` parameter). A host with no relevant prior immunity (large $d^{\star}_{v,h}$) is therefore fully susceptible ($\mathrm{ROI}_{h,v} = 1$), whereas a host previously infected by an antigenically identical virus retains a residual infection risk of $1-\omega$. We used $s = 0.07$ and $\omega = 0.95$, matching the original `antigen` model [94].

In this chapter, we use an overall viral mutation rate of $\mu = 10^{-3}$ mutations per virus per day. Epitope sites comprise $\sim$9% of the sequence (49/566 sites) and have an acceptance rate of 0.75, resulting in an

effective mutation rate of $\mu \times 0.065$ mutations per virus per day at epitope sites. Non-epitope sites comprise the remaining $\sim$91% of the sequence and have a lower acceptance rate of 0.2, reflecting greater purifying selection at these sites, resulting in an effective rate of $\mu \times 0.183$ mutations per virus per day. We tuned these acceptance rates to reproduce mutational patterns observed in seasonal influenza in nature (Figure B.2).

4.4.4 Simulation parameterization and host population immunity sampling

We simulated 40 years of influenza evolution across three demes, and discarded the first 10 years as burn-in. We parameterized mutation step-size distributions to reflect the differential antigenic impact of epitope versus non-epitope mutations. Non-epitope sites used $r_{ne} \sim \text{Gamma}(\alpha = 1, \beta = 0.0001)$ while epitope sites used $r_e \sim \text{Gamma}(\alpha = 2.25, \beta = 0.267)$. All mutations received a step direction $\theta \sim \text{Uniform}(0, 2\pi)$.

We sampled 150,000 viruses over the course of the simulation, proportionally by prevalence across demes and time, yielding approximately 4,400 unique nucleotide sequences to provide adequate count data for downstream growth-rate inference. To track population-immunity dynamics, we calculated and saved the antigenic centroid from the most recent infection stored in the immune memories of 10,000 hosts from each deme every 365 days. The host population immunity centroid at time t is calculated as:

$$H_t = \frac{1}{|H|} \sum_{h \in H} (\text{ag}_1^*, \text{ag}_2^*)_h \tag{4.3}$$

where H is the set of all sampled hosts across demes, $|H|$ is the total number of hosts sampled, and $(\text{ag}_1^*, \text{ag}_2^*)_h$ represents the two-dimensional antigenic coordinates of the most recent infection for host h. This centroid represents the average antigenic position of the host population's immunity at time t and is used to calculate virus fitness values in the downstream variant assignment benchmark.

4.4.5 Simulation selection for benchmarking applications

We applied three criteria for selecting simulations suitable for benchmarking: (1) realistic epidemiological dynamics with seasonal epidemic patterns in the temperate demes and year-round transmission in the tropics, (2) summary statistics matching empirical A/H3N2 HA genetic and antigenic evolution, and (3) sufficient viral diversity for robust benchmarking of methods for variant assignment and growth-rate inference.

To achieve the first criterion, we maintained the same antigenic and epidemiological parameter values from the original `antigen` paper by Bedford *et al.* [94]. We also set the overall mutation rate to $\mu = 10^{-3}$ mutation events per individual per day. For the second criterion, we ran 120 total simulations with four different parameter configurations for epitope and non-epitope mutation acceptance rates (epitope: 0.75 or 1.0; non-epitope: 0.1 or 0.2), using 30 replicates for each configuration. All 120 simulations ran to completion without viral population extinction (Figure B.1, Figure B.2).

We define the mean pairwise genealogical diversity π_G as the average total branch length between randomly sampled virus pairs:

$$\pi_G = \frac{1}{n} \sum_{i=1}^{n} \left[(t_{v_A^i} - t_{\text{MRCA}^i}) + (t_{v_B^i} - t_{\text{MRCA}^i}) \right] \tag{4.4}$$

where n is the number of sampled pairs, t_v is the birth time of virus v, and t_{MRCA} is the birth time of the most recent common ancestor for each pair. We then applied a filtering criterion to focus on "flu-like" simulations: $\pi_G \leq 9.0$ years. This filtering reduced the dataset from 120 to 83 qualifying simulations. For downstream variant assignment and growth rate inference benchmarking, we selected a single simulation with epitope acceptance rate of 0.75 and non-epitope acceptance rate of 0.2. This simulation had a simulated π_G of 3.6 years and TMRCA of 3.4 years, and average antigenic movement of 1.6 units per year, closely matching empirical observations. The mutation summary statistics reported in Table 4.1 represent this selected simulation. Final selection involved confirming that phylogenetic trees and antigenic space distributions exhibited realistic influenza-like dynamics by visual inspection. Phylogenetic trees were inspected to ensure they displayed ladder-like structures, and case count dynamics were examined to confirm seasonal epidemic patterns in temperate demes and year-round transmission in the tropics.

4.4.6 Variant assignment for `antigen-prime` simulations

Three variant-assignment methods were applied to the 5,900 unique sequences: antigenic clustering (ground truth), sequence-based clustering, and phylogenetic variant assignment. Antigenic variants were defined by k-means clustering ($k = 30$) on two-dimensional antigenic coordinates. Sequence-based variants were assigned using `pathogen-embed` [85]: (1) sequence alignment with `MAFFT` via `augur align`, (2) pairwise Hamming distance calculation, (3) t-SNE embedding in 2D space, and (4) k-means clustering

($k = 30$) on embeddings. The $k = 30$ parameter for both methods reflects empirical influenza dynamics of approximately three variants per year over the 30 year simulation.

Phylogenetic variants were assigned using the Neher clade assignment algorithm [84] following phylogeny reconstruction and ancestral inference. The algorithm was configured to use the same epitope sites defined by Luksza and Lässig [87] that are used in the `antigen-prime` simulation. Phylogeny inference used `IQ-TREE` via `augur tree` with `augur refine` refinement. Ancestral reconstruction applied `augur ancestral` and `augur translate` with default parameters. Clade assignment used parameters: bushiness branch scale 1.0, divergence scale 2.0, branch length scale 2.0, minimum clade size 22 sequences, targeting approximately 30 variants (resulting in 31). All variant assignments were re-labeled chronologically by average birth date of constituent viruses.

Within-variant fitness variance was calculated to evaluate how well each method grouped viruses with similar fitness. We computed fitness for all viruses annually using the host population immunity centroid (Equation 4.3), then calculated average within-variant fitness variance for each assignment method.

$$\mathrm{WSS}(t) = \frac{1}{|V|} \sum_{v \in V} \mathrm{Var}(\omega_{t,v}) \tag{4.5}$$

where V is the set of all variants assigned by a method, $|V|$ is the total number of variants, and $\omega_{t,v}$ represents the fitness values of all viruses in variant v at time t.

Agreement between variant assignments was quantified using the normalized information distance (NID) [104]:

$$\mathrm{NID}(X, Y) = 1 - \frac{I(X, Y)}{H(X, Y)} \tag{4.6}$$

where $I(X, Y)$ is the mutual information between assignments X and Y, and $H(X, Y)$ is their joint entropy. NID ranges from 0 (identical assignments) to 1 (completely independent assignments).

4.4.7 Inferring variant growth rates with `evofr` forecasting models

We implemented two forecasting models from the `evofr` [91] framework to infer variant growth rates from simulated surveillance data. The Fixed Growth Advantage (FGA) model implements a renewal equation approach where each variant has a fixed multiplicative growth factor, while the Growth Advantage Random

Walk (GARW) model allows variant growth advantages to vary smoothly over time using a random walk prior.

Data preparation involved extracting weekly variant-specific sequence counts and case counts from simulation outputs for each analysis timepoint. Data were separated by deme with analysis dates representing both in-season and out-of-season periods across different epidemic contexts. The retrospective observation window was limited to 365 days from each analysis date.

Model fitting used the `evofr` software package for both FGA and GARW models. Model-specific hyperparameters were initialized with spline basis functions of order 4 with 10 knots to model time-varying parameters. Generation time and reporting delay distributions were defined for the renewal equation models, with generation times parameterized as gamma distributions: $g(\tau) \sim \text{Gamma}(\text{mean} = 3.0, \text{std} = 1.2)$. Sequence count data used Dirichlet-Multinomial likelihood with concentration parameter 100, while case count data used Negative Binomial likelihood with dispersion parameter 0.05. Variational inference approximated posterior distributions of model parameters using full-rank variational inference with 50,000 iterations and learning rate of 0.01, generating 500 posterior samples per model. We focused on the inferred variant growth rates in this work.

4.4.8 Benchmarking growth rate inference performance

Empirical growth rates ($r_{\text{data},v}$) were calculated from simulated surveillance data using spline-based smoothing to reduce noise. For each variant v and location d, sequence data processing used univariate spline interpolation by log-transforming sequence counts to handle data skew and stabilize variance, applying a cubic univariate spline (degree $k = 3$) with smoothing factor $s = 1.0$ to the log-transformed data, then transforming the smoothed values back to the original scale.

After obtaining smoothed sequence counts, we calculated variant-specific incidence by multiplying total case counts by variant frequencies, then computed empirical growth rates as the change in log-transformed variant incidence over time:

$$r_{\text{data},v}(t_i) = \frac{\ln(C_d(t_i) \cdot f_{v,d}(t_i)) - \ln(C_d(t_{i-1}) \cdot f_{v,d}(t_{i-1}))}{\Delta t} \tag{4.7}$$

where $C_d(t_i)$ represents the total case counts in location d at time t_i, $f_{v,d}(t_i)$ represents the smoothed

frequency of variant v in location d at time t_i, and Δt is the time difference between observations. This approach scales the total disease burden by variant-specific frequencies, providing a representation of variant-specific growth rates.

Data filtering ensured reliable growth-rate estimates by excluding time points with smoothed sequence counts below 10 sequences (due to high sampling variance), variant frequencies below 1% of the total population (to avoid stochastic effects in rare variants), and variant incidence below 50 cases on any given day (to ensure sufficient case data). We required a minimum of 3 consecutive valid time points for a variant to be included in the growth-rate benchmarking analysis.

Performance on growth-rate inference was evaluated using mean absolute error (MAE) between the medians of the inferred growth rate posteriors ($r_{\text{model},v}$) and empirical ($r_{\text{data},v}$) growth rates for each variant:

$$\text{MAE}_v = \frac{1}{n} \sum_{i=1}^{n} |r_{\text{data},v}(t_i) - r_{\text{model},v}(t_i)| \tag{4.8}$$

where n is the number of time points for variant v. We report $\log(\text{MAE})$ for each variant to facilitate comparison across the small error ranges typically observed. The identification of analysis windows with unforeseen pathologies was done by tediously looking at many analysis windows with exceptionally high errors. Complete results, including detailed performance metrics for all analysis windows, are available at `https://github.com/matsengrp/antigen-forecasting/notebooks/`.

Chapter 5

Conclusion

This dissertation presents three computational tools for studying the molecular and antigenic evolution of rapidly evolving viruses like influenza and SARS-CoV-2. A primary motivation for this work was to develop methods for analyzing data produced by relatively new experimental techniques such as deep mutational scanning of viral proteins and for benchmarking computational methods used in viral surveillance and forecasting. These tools leverage domain knowledge from protein biophysics, immunology, and epidemiology alongside modern machine learning methods to model complex genotype-phenotype relationships while maintaining biological interpretability. They enable flexible yet rigorous analysis of high-throughput experimental data and provide a framework for evaluating methods that inform public health decisions.

All three tools are implemented as open-source software packages with documentation and tutorials to facilitate adoption by the research community. Beyond the specific applications presented here, each tool is designed to be flexible enough for broader use. `torchdms` provides a general framework for inferring genotype-phenotype maps from deep mutational scanning data and can be applied to any protein system where multi-mutant DMS data are available, not only viral proteins. `polyclonal` can model antibody escape for any viral surface protein given appropriate deep mutational scanning measurements, and has already been applied to study immune escape in both SARS-CoV-2 and influenza. `antigen-prime` is parameterizable for different pathogens and could be tuned to simulate the evolutionary dynamics of other antigenically variable viruses such as RSV or seasonal coronaviruses, enabling benchmarking of surveillance methods across diverse epidemiological contexts.

The global epistasis framework underlying `torchdms` assumes that mutations combine additively before nonlinear transformation, which may not hold for proteins with strong pairwise or higher-order epistatic interactions. Prediction accuracy also degrades for variants with many more mutations than those in the training data, reflecting the inherent challenge of extrapolating beyond the observed sequence space. Additionally, all models struggle to capture sharp transitions between functional and non-functional proteins and the boundary between folded and unfolded states proves difficult to predict on the RBD data used. The current implementation also lacks an explicit noise model, forgoing decomposition of variance into measurement error versus biological variability and preventing proper uncertainty quantification. Future extensions could incorporate measurement error models to enable prediction intervals, as in `MAVE-NN`. The related `multidms` package already addresses one limitation by enabling inference of mutational effects across DMS libraries constructed on different wildtype backgrounds.

The `polyclonal` model makes several simplifying assumptions about antibody-antigen interactions. It groups antibodies into discrete epitopes, but in reality two antibodies targeting similar regions may be escaped by slightly different mutations, and some antibodies bind idiosyncratic regions outside or between major epitopes. The model also assumes that antibody binding at one epitope does not influence binding at other epitopes, and that escape can be modeled for a single idealized antigen rather than accounting for the multiple antigens decorating each virion. Furthermore, the number of epitopes must be specified *a priori*, though the true number is often unknown in practice. These assumptions are expected to hold reasonably well for variants with modest numbers of mutations, but predictions may become less accurate for highly mutated variants. Mutations to subdominant epitopes also proved difficult to characterize: their effects on overall escape are masked when immunodominant epitopes remain intact. Despite these limitations, `polyclonal` has already been applied to study immune escape in both SARS-CoV-2 and influenza, demonstrating how biophysically motivated models can translate high-throughput deep mutational scanning data into predictions relevant for viral surveillance and vaccine design.

`antigen-prime` makes several simplifying assumptions to enable tractable simulation of viral evolution over epidemiological timescales. The fitness of simulated viruses is determined solely by their antigenic phenotype, which is primarily driven by mutations made at epitope sites, omitting potential fitness costs of mutations at non-epitope sites that may constrain evolutionary trajectories in nature. The model also main-

tains static epitope sites throughout the simulation and employs a simple mutation model without epistatic interactions between sites. As a result, model and method performance on our benchmarks are likely higher than would be expected for a natural viral population, where predicting phenotype from genotype is complicated by things like epistasis and variable immunity in the host population. However, our benchmarks also revealed a previously undocumented failure mode in growth rate inference models: when one variant initially predominates then declines as low-frequency variants rise, the models accurately predict frequency trajectories but fail to predict growth rates, often producing estimates with the opposite sign of the true values. This systematic bias may stem from an underlying assumption that variant-specific growth rates change in a concerted manner. Future work could use `antigen-prime` to benchmark methods over shorter timescales more relevant to real-time surveillance, or to evaluate forecasting methods that predict which variants will dominate in the future.

Together, these tools illustrate how computational methods that integrate domain knowledge with flexible statistical and machine learning frameworks can address questions about viral evolution spanning molecular to population scales. The progression from modeling mutational effects on individual proteins, to quantifying escape from polyclonal antibodies, to simulating viral evolution under population immunity reflects the multi-scale nature of viral antigenic evolution itself. By making these tools openly available and documenting their limitations, we hope to contribute to the broader infrastructure for understanding and responding to the evolution of viruses that threaten human health.

Bibliography

[1] Velislava N Petrova and Colin A Russell. The evolution of seasonal influenza viruses. *Nat. Rev. Microbiol.*, 16(1):47–60, January 2018.

[2] Florian Krammer, Gavin J D Smith, Ron A M Fouchier, Malik Peiris, Katherine Kedzierska, Peter C Doherty, et al. Influenza. *Nat. Rev. Dis. Primers*, 4(1):3, June 2018.

[3] Amanda C Perofsky and Martha I Nelson. The challenges of vaccine strain selection. *Elife*, 9:e62955, October 2020.

[4] 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(7902):679–686, March 2022.

[5] Tyler N Starr, Allison J Greaney, Sarah K Hilton, Daniel Ellis, Katharine H D Crawford, Adam S Dingens, et al. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell*, 182(5):1295–1310.e20, September 2020.

[6] Allison J Greaney, Tyler N Starr, Pavlo Gilchuk, Seth J Zost, Elad Binshtein, Andrea N Loes, et al. Complete mapping of mutations to the SARS-CoV-2 spike receptor-binding domain that escape antibody recognition. *Cell Host Microbe*, 29(1):44–57.e9, January 2021.

[7] Douglas M Fowler, Carlos L Araya, Sarel J Fleishman, Elizabeth H Kellogg, Jason J Stephany, David Baker, et al. High-resolution mapping of protein sequence-function relationships. *Nat. Methods*, 7 (9):741–746, September 2010.

[8] Carlos L Araya and Douglas M Fowler. Deep mutational scanning: assessing protein function on a massive scale. *Trends Biotechnol.*, 29(9):435–442, September 2011.

[9] Douglas M Fowler and Stanley Fields. Deep mutational scanning: a new style of protein science. *Nature Methods*, 11(8):801–807, aug 2014.

[10] Matteo Cagiada, Kristoffer E Johansson, Audrone Valanciute, Sofie V Nielsen, Rasmus Hartmann-Petersen, Jun J Yang, et al. Understanding the origins of loss of protein function by analyzing the effects of thousands of variants on activity and abundance. *Mol. Biol. Evol.*, March 2021.

[11] Hugh K Haddox, Adam S Dingens, and Jesse D Bloom. Experimental estimation of the effects of all amino-acid mutations to HIV's envelope protein on viral replication in cell culture. *PLoS Pathog.*, 12 (12):e1006114, December 2016.

[12] Adam S Dingens, Hugh K Haddox, Julie Overbaugh, and Jesse D Bloom. Comprehensive mapping of HIV-1 escape from a broadly neutralizing antibody. *Cell Host Microbe*, 21(6):777–787.e4, June 2017.

[13] Juhye M Lee, Rachel Eguia, Seth J Zost, Saket Choudhary, Patrick C Wilson, Trevor Bedford, et al. Mapping person-to-person variation in viral mutations that escape polyclonal serum targeting influenza hemagglutinin. *Elife*, 8, August 2019.

[14] Adam S Dingens, Dana Arenz, Haidyn Weight, Julie Overbaugh, and Jesse D Bloom. An antigenic atlas of HIV-1 escape from broadly neutralizing antibodies distinguishes functional and structural epitopes. *Immunity*, 50(2):520–532.e3, February 2019.

[15] Michael B Doud and Jesse D Bloom. Accurate measurement of the effects of all amino-acid mutations on influenza hemagglutinin. *Viruses*, 8(6):155, June 2016.

[16] Michael J Harms and Joseph W Thornton. Evolutionary biochemistry: revealing the historical and physical causes of protein properties. *Nat. Rev. Genet.*, 14(8):559–571, August 2013.

[17] Nicholas C Wu, C Anders Olson, and Ren Sun. High-throughput identification of protein mutant

stability computed from a double mutant fitness landscape. *Protein Sci.*, 25(2):530–539, February 2016.

[18] Zachary R Sailer and Michael J Harms. Molecular ensembles make evolution unpredictable. *Proc. Natl. Acad. Sci. U. S. A.*, 114(45):11938–11943, November 2017.

[19] Zachary R Sailer and Michael J Harms. Uninterpretable interactions: epistasis as uncertainty. *bioRxiv*, page 378489, July 2018.

[20] Frank J Poelwijk, Michael Socolich, and Rama Ranganathan. Learning the pattern of epistasis linking genotype and phenotype in a protein. *Nat. Commun.*, 10(1):4213, September 2019.

[21] Charlotte M Miton, Karol Buda, and Nobuhiko Tokuriki. Epistasis and intramolecular networks in protein evolution. *Curr. Opin. Struct. Biol.*, 69:160–168, August 2021.

[22] C Anders Olson, Nicholas C Wu, and Ren Sun. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Curr. Biol.*, 24(22):2643–2651, November 2014.

[23] Lizhi Ian Gong, Marc A Suchard, and Jesse D Bloom. Stability-mediated epistasis constrains the evolution of an influenza protein. *Elife*, 2:e00631, May 2013.

[24] Jakub Otwinowski, David M McCandlish, and Joshua B Plotkin. Inferring the shape of global epistasis. *Proc. Natl. Acad. Sci. U. S. A.*, 115(32):E7550–E7558, August 2018.

[25] Naomi E Harwood and Facundo D Batista. Early events in B cell activation. *Annu. Rev. Immunol.*, 28:185–210, 2010.

[26] Gabriel D Victora and Luka Mesin. Clonal and cellular dynamics in germinal centers. *Curr. Opin. Immunol.*, 28:90–96, June 2014.

[27] Luka Mesin, Jonatan Ersching, and Gabriel D Victora. Germinal center B cell dynamics. *Immunity*, 45(3):471–482, September 2016.

[28] Michael B Doud, Scott E Hensley, and Jesse D Bloom. Complete mapping of viral escape from neutralizing antibodies. *PLoS pathogens*, 13(3):e1006271, 2017.

[29] Christopher O Barnes, Claudia A Jette, Morgan E Abernathy, Kim-Marie A Dam, Shannon R Esswein, Harry B Gristick, et al. Sars-cov-2 neutralizing antibody structures inform therapeutic strategies. *Nature*, 588(7839):682–687, 2020.

[30] 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(7898): 657–663, 2022.

[31] H Benjamin Larman, Zhenming Zhao, Uri Laserson, Mamie Z Li, Alberto Ciccia, M Angelica Martinez Gakidis, et al. Autoantigen discovery with a synthetic human peptidome. *Nat. Biotechnol.*, 29 (6):535–541, May 2011.

[32] Divya Mohan, Daniel L Wansley, Brandon M Sie, Muhammad S Noon, Alan N Baer, Uri Laserson, et al. PhIP-Seq characterization of serum antibodies using oligonucleotide-encoded peptidomes. *Nat. Protoc.*, 13(9):1958–1978, September 2018.

[33] Meghan E Garrett, Hannah L Itell, Katharine H D Crawford, Ryan Basom, Jesse D Bloom, and Julie Overbaugh. Phage-DMS: A comprehensive method for fine mapping of antibody epitopes. *iScience*, 23(10):101622, October 2020.

[34] Meghan E Garrett, Jared Galloway, Helen Y Chu, Hannah L Itell, Caitlin I Stoddard, Caitlin R Wolf, et al. High-resolution profiling of pathways of escape for SARS-CoV-2 spike-binding antibodies. *Cell*, May 2021.

[35] William Ogilvy Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927.

[36] Anne Cori, Neil M Ferguson, Christophe Fraser, and Simon Cauchemez. A new framework and software to estimate time-varying reproduction numbers during epidemics. *American journal of epidemiology*, 178(9):1505–1512, 2013.

[37] Katelyn M Gostic, Lauren McGough, Edward B Baskerville, Sam Abbott, Keya Joshi, Christine

Tedijanto, et al. Practical considerations for measuring the effective reproductive number, r t. *PLoS computational biology*, 16(12):e1008409, 2020.

[38] Julia R Gog and Bryan T Grenfell. Dynamics and selection of many-strain pathogens. *Proc. Natl. Acad. Sci. U. S. A.*, 99(26):17209–17214, December 2002.

[39] 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 Research*, 5(112):112, 2020.

[40] Jesse D Bloom, Sy T Labthavikul, Christopher R Otey, and Frances H Arnold. Protein stability promotes evolvability. *Proc. Natl. Acad. Sci. U. S. A.*, 103(15):5869–5874, April 2006.

[41] Jesse D Bloom and Frances H Arnold. In the light of directed evolution: Pathways of adaptive protein evolution. *Proc. Natl. Acad. Sci. U. S. A.*, 106(Supplement 1):9995–10000, June 2009.

[42] Brian Hie, Ellen D Zhong, Bonnie Berger, and Bryan Bryson. Learning the language of viral evolution and escape. *Science*, 371(6526):284–288, January 2021.

[43] Nicole N Thadani, Sarah Gurev, Pascal Notin, Noor Youssef, Nathan J Rollins, Daniel Ritter, et al. Learning from prepandemic data to forecast viral escape. *Nature*, 622(7984):818–825, October 2023.

[44] Joseph M Taft, Cédric R Weber, Beichen Gao, Roy A Ehling, Jiami Han, Lester Frei, et al. Deep mutational learning predicts ACE2 binding and antibody escape to combinatorial mutations in the SARS-CoV-2 receptor-binding domain. *Cell*, 185(21):4008–4022.e14, October 2022.

[45] Jakub Otwinowski. Biophysical inference of epistasis and the effects of mutations on protein stability and function. *Mol. Biol. Evol.*, 35(10):2345–2354, October 2018.

[46] Ammar Tareen, Mahdi Kooshkbaghi, Anna Posfai, William T Ireland, David M McCandlish, and Justin B Kinney. MAVE-NN: learning genotype-phenotype maps from multiplex assays of variant effect. *Genome Biol.*, 23(1):98, April 2022.

[47] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, et al. PyTorch: An imperative style, High-Performance deep learning library. December 2019.

[48] Rhys M Adams, Thierry Mora, Aleksandra M Walczak, and Justin B Kinney. Measuring the sequence-affinity landscape of antibodies with massively parallel titration curves. *Elife*, 5, December 2016.

[49] Shira Warszawski, Ravit Netzer, Dan S Tawfik, and Sarel J Fleishman. A "fuzzy"-logic language for encoding multiple physical traits in biomolecules. *Journal of molecular biology*, 426(24):4125–4138, 2014.

[50] Hugh K Haddox, Jared G Galloway, Bernadeta Dadonaite, Jesse D Bloom, Frederick A Matsen, and William S DeWitt. Jointly modeling deep mutational scans identifies shifted mutational effects among SARS-CoV-2 spike homologs. *bioRxivorg*, page 2023.07.31.551037, August 2023.

[51] Derek J Smith, Alan S Lapedes, Jan C de Jong, Theo M Bestebroer, Guus F Rimmelzwaan, Albert D M E Osterhaus, et al. Mapping the antigenic and genetic evolution of influenza virus. *Science*, 305 (5682):371–376, July 2004.

[52] Scott E Hensley, Suman R Das, Adam L Bailey, Loren M Schmidt, Heather D Hickman, Akila Jayaraman, et al. Hemagglutinin receptor binding avidity drives influenza a virus antigenic drift. *Science*, 326(5953):734–736, October 2009.

[53] 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, February 2014.

[54] Rachel T Eguia, Katharine H D Crawford, Terry Stevens-Ayers, Laurel Kelnhofer-Millevolte, Alexander L Greninger, Janet A Englund, et al. A human coronavirus evolves antigenically to escape antibody immunity. *PLoS Pathog.*, 17(4):e1009453, April 2021.

[55] G K Hirst. Studies of antigenic differences among strains of influenza a by means of red cell agglutination and its inhibition. *J. Exp. Med.*, 78(5):407–423, November 1943.

[56] Marciela M DeGrace, Elodie Ghedin, Matthew B Frieman, Florian Krammer, Alba Grifoni, Arghavan Alisoltani, et al. Defining the risk of SARS-CoV-2 variants on immune protection. *Nature*, 605(7911): 640–652, May 2022.

[57] Stefan Elbe and Gemma Buckland-Merrett. Data, disease and diplomacy: GISAID's innovative contribution to global health: Data, disease and diplomacy. *Global Chall.*, 1(1):33–46, January 2017.

[58] Yuelong Shu and John McCauley. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.*, 22(13):30494, March 2017.

[59] Andrew Rambaut, Edward C Holmes, Áine O'Toole, Verity Hill, John T McCrone, Christopher Ruis, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.*, 5(11):1403–1407, November 2020.

[60] Nicholas C Wu, Andrew J Thompson, Juhye M Lee, Wen Su, Britni M Arlian, Jia Xie, et al. Different genetic barriers for resistance to HA stem antibodies in influenza H3 and H1 viruses. *Science*, 368 (6497):1335–1340, June 2020.

[61] Allison J Greaney, Andrea N Loes, Katharine H D Crawford, Tyler N Starr, Keara D Malone, Helen Y Chu, et al. Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. *Cell Host Microbe*, 29(3):463–476.e6, March 2021.

[62] Allison J Greaney, Tyler N Starr, Christopher O Barnes, Yiska Weisblum, Fabian Schmidt, Marina Caskey, et al. Mapping mutations to the SARS-CoV-2 RBD that escape binding by different classes of antibodies. *Nat. Commun.*, 12(1):4196, July 2021.

[63] 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(1):veac021, May 2022.

[64] Richard A Neher, Trevor Bedford, Rodney S Daniels, Colin A Russell, and Boris I Shraiman. Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses. *Proc. Natl. Acad. Sci. U. S. A.*, 113(12):E1701–9, March 2016.

[65] Hailiang Sun, Jialiang Yang, Tong Zhang, Li-Ping Long, Kun Jia, Guohua Yang, et al. Using sequence data to infer the antigenicity of influenza virus. *mBio*, 4(4), August 2013.

[66] William T Harvey, Donald J Benton, Victoria Gregory, James P J Hall, Rodney S Daniels, Trevor Bedford, et al. Identification of low- and high-impact hemagglutinin amino acid substitutions that drive antigenic drift of influenza a(H1N1) viruses. *PLoS Pathog.*, 12(4):e1005526, April 2016.

[67] Timothy C Yu, Zorian T Thornton, William W Hannon, William S DeWitt, Caelan E Radford, Frederick A Matsen, 4th, et al. A biophysical model of viral escape from polyclonal antibodies. *Virus Evol*, 8(2):veac110, December 2022.

[68] W G Laver, G M Air, R G Webster, W Gerhard, C W Ward, and T A Dopheide. Antigenic drift in type a influenza virus: sequence differences in the hemagglutinin of hong kong (H3N2) variants selected with monoclonal hybridoma antibodies. *Virology*, 98(1):226–237, October 1979.

[69] J W Yewdell, R G Webster, and W U Gerhard. Antigenic variation in three distinct determinants of an influenza type a haemagglutinin molecule. *Nature*, 279(5710):246–248, May 1979.

[70] R G Webster and W G Laver. Determination of the number of nonoverlapping antigenic areas on hong kong (H3N2) influenza virus hemagglutinin with monoclonal antibodies and the selection of variants with potential epidemiological significance. *Virology*, 104(1):139–148, July 1980.

[71] DC Wiley, IA Wilson, and JJ Skehel. Structural identification of the antibody-binding sites of hong kong influenza haemagglutinin and their involvement in antigenic variation. *Nature*, 289(5796):373–378, 1981.

[72] JJ Skehel, DJ Stevens, RS Daniels, AR Douglas, M Knossow, IA Wilson, et al. A carbohydrate side chain on hemagglutinins of hong kong influenza viruses inhibits recognition by a monoclonal antibody. *Proceedings of the National Academy of Sciences*, 81(6):1779–1783, 1984.

[73] Luca Piccoli, Young-Jun Park, M Alejandra Tortorici, Nadine Czudnochowski, Alexandra C Walls, Martina Beltramello, et al. Mapping neutralizing and immunodominant sites on the sars-cov-2 spike receptor-binding domain by structure-guided high-resolution serology. *Cell*, 183(4):1024–1042, 2020.

[74] Tyler N Starr, Nadine Czudnochowski, Zhuoming Liu, Fabrizia Zatta, Young-Jun Park, Amin Ad-

detia, et al. Sars-cov-2 rbd antibodies that maximize breadth and resistance to escape. *Nature*, 597 (7874):97–102, 2021.

[75] Tal Einav and Jesse D Bloom. When two are better than one: Modeling the mechanisms of antibody mixtures. *PLoS Comput. Biol.*, 16(5):e1007830, May 2020.

[76] Bernadeta Dadonaite, Katharine H. D. Crawford, Caelan E. Radford, Ariana G. Farrell, Timothy C. Yu, William W. Hannon, et al. A pseudovirus system enables deep mutational scanning of the full SARS-CoV-2 spike. *Cell*, 186(6):1263–1278.e20, 2023. doi: 10.1016/j.cell.2023.02.001.

[77] 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(8021):617–626, 2024. doi: 10.1038/s41586-024-07636-1.

[78] Bernadeta Dadonaite, Jenny J. Ahn, Jordan T. Ort, Jin Yu, Colleen Furey, Annie Dosey, et al. Deep mutational scanning of H5 hemagglutinin to inform influenza virus surveillance. *PLOS Biology*, 22 (11):e3002916, 2024. doi: 10.1371/journal.pbio.3002916.

[79] Bernadeta Dadonaite, Sheri Harari, Brendan B. Larsen, Lucas Kampman, Alex Harteloo, Anna Elias-Warren, et al. Spike mutations that affect the function and antigenicity of recent KP.3.1.1-like SARS-CoV-2 variants. *Journal of Virology*, page e0142325, 2025. doi: 10.1128/jvi.01423-25.

[80] Timothy C Yu, Caroline Kikawa, Bernadeta Dadonaite, Andrea N Loes, Janet A Englund, and Jesse D Bloom. Pleiotropic mutational effects on function and stability constrain the antigenic evolution of influenza haemagglutinin. *Nat. Ecol. Evol.*, pages 1–15, December 2025.

[81] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17: 261–272, 2020. doi: 10.1038/s41592-019-0686-2.

[82] James Hadfield, Colin Megill, Sidney M Bell, John Huddleston, Barney Potter, Charlton Callender, et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, 34(23):4121–4123, December 2018.

[83] Dylan H Morris, Katelyn M Gostic, Simone Pompei, Trevor Bedford, Marta Łuksza, Richard A Neher, et al. Predictive modeling of influenza shows the promise of applied evolutionary biology. *Trends Microbiol.*, 26(2):102–118, February 2018.

[84] Richard A Neher, John Huddleston, Trevor Bedford, Nicola S Lewis, Ruth Harvey, Monica Galiano, et al. Nomenclature for tracking of genetic variation of seasonal influenza viruses. *medRxiv*, page 2025.12.06.25341755, December 2025.

[85] Sravani Nanduri, Allison Black, Trevor Bedford, and John Huddleston. Dimensionality reduction distills complex evolutionary relationships in seasonal influenza and SARS-CoV-2. *Virus Evolution*, 10(1):veae087, 11 2024.

[86] John Huddleston, Trevor Bedford, Jennifer Chang, Jover Lee, and Richard A Neher. Seasonal influenza circulation patterns and projections for february 2024 to february 2025. 2024.

[87] Marta Luksza and Michael Lässig. A predictive fitness model for influenza. *Nature*, 507(7490): 57–61, March 2014.

[88] Kimihito Ito, Chayada Piantham, and Hiroshi Nishiura. Predicted dominance of variant Delta of SARS-CoV-2 before Tokyo Olympic Games, Japan, July 2021. *Euro Surveill.*, 26(27), July 2021.

[89] 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. *bioRxiv*, page 2021.03.17.21253775, March 2021.

[90] 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(6599):1327–1332, June 2022.

[91] Marlin D Figgins and Trevor Bedford. Inferring variant-specific effective reproduction numbers from combined case and sequencing data. September 2025. doi: 10.7554/elife.104802.1. URL `http://dx.doi.org/10.7554/eLife.104802.1`.

[92] Caroline Kikawa, Andrea N Loes, John Huddleston, Marlin D Figgins, Philippa Steinberg, Tachianna Griffiths, et al. High-throughput neutralization measurements correlate strongly with evolutionary success of human influenza strains. *eLife*, November 2025. doi: 10.7554/elife.106811.2. URL http://dx.doi.org/10.7554/eLife.106811.2.

[93] Abbas Jariani, Christopher Warth, Koen Deforche, Pieter Libin, Alexei J Drummond, Andrew Rambaut, et al. SANTA-SIM: simulating viral sequence evolution dynamics under selection and recombination. *Virus Evol*, 5(1):vez003, January 2019.

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

[95] Niema Moshiri, Manon Ragonnet-Cronin, Joel O Wertheim, and Siavash Mirarab. FAVITES: simultaneous simulation of transmission networks, phylogenetic trees and sequences. *Bioinformatics*, 35 (11):1852–1861, June 2019.

[96] Nicolas Ochsner, Judith Bouman, Timothy G Vaughan, Tanja Stadler, Sebastian Bonhoeffer, and Roland R Regoes. Viral simulation reveals overestimation bias in within-host phylodynamic migration rate estimates under selection. *bioRxiv*, page 2025.01.06.631458, January 2025.

[97] M Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.*, 16(2):111–120, December 1980.

[98] Raul Rabadan, Arnold J Levine, and Harlan Robins. Comparison of avian and human influenza a viruses reveals a mutational bias on the viral genomes. *J. Virol.*, 80(23):11887–11891, December 2006.

[99] Jesse D Bloom and Matthew J Glassman. Inferring stabilizing mutations from protein phylogenies: application to influenza hemagglutinin. *PLoS Comput. Biol.*, 5(4):e1000349, April 2009.

[100] 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, September 2020.

[101] Matthew Scotch, Temitope O C Faleye, Jillian M Wright, Sarah Finnerty, Rolf U Halden, and Arvind Varsani. Campus-based genomic surveillance uncovers early emergence of a future dominant a(H3N2) influenza clade. *iScience*, 28(12):113941, December 2025.

[102] Björn F Koel, David F Burke, Theo M Bestebroer, Stefan van der Vliet, Gerben C M Zondag, Gaby Vervaet, et al. Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. *Science*, 342(6161):976–979, November 2013.

[103] Trevor Bedford, Steven Riley, Ian G Barr, Shobha Broor, Mandeep Chadha, Nancy J Cox, et al. Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature*, 523(7559): 217–220, July 2015.

[104] M Li, X Chen, X Li, B Ma, and P M B Vitanyi. The similarity metric. *IEEE Trans. Inf. Theory*, 50 (12):3250–3264, December 2004.

[105] Eslam Abousamra, Marlin Figgins, and Trevor Bedford. Fitness models provide accurate short-term forecasts of SARS-CoV-2 variant frequency. *PLOS Computational Biology*, 20(9):1–20, 09 2024. doi: 10.1371/journal.pcbi.1012443.

[106] Judith M Fonville, Samuel H Wilks, Sarah L James, Annette Fox, Maria Ventresca, Mathilde Aban, et al. Antibody landscapes after influenza virus infection or vaccination. *Science*, 346(6212):996–1000, 2014. doi: 10.1126/science.1256427.

[107] Carole Henry, Nai-Ying Zheng, Min Huang, Alexandra Cabanov, Kenneth T Rojas, Kaval Kaur, et al. Preexisting immunity shapes distinct antibody landscapes after influenza virus infection and vaccination in humans. *Science Translational Medicine*, 13(573):eabd3601, 2021. doi: 10.1126/scitranslmed.abd3601.

[108] Katelyn M Gostic, Monique Ambrose, Michael Worobey, and James O Lloyd-Smith. Potent protection against H5N1 and H7N9 influenza via childhood hemagglutinin imprinting. *Science*, 354(6313): 722–726, 2016. doi: 10.1126/science.aag1322.

[109] Sarah Cobey and Scott E Hensley. Immune history and influenza virus susceptibility. *Current Opinion in Virology*, 22:105–111, 2017. doi: 10.1016/j.coviro.2016.12.004.

[110] Philip Arevalo, Huong Q McLean, Edward A Belongia, and Sarah Cobey. Earliest infections predict the age distribution of seasonal influenza A cases. *eLife*, 9:e50060, 2020. doi: 10.7554/eLife.50060.

[111] Marcos C Vieira, Celeste M Donato, Philip Arevalo, Guus F Rimmelzwaan, Thomas Wood, Liza Lopez, et al. Lineage-specific protection and immune imprinting shape the age distributions of influenza B cases. *Nature Communications*, 12(1):4313, 2021. doi: 10.1038/s41467-021-24566-y.

[112] Z T Thornton, T Tran, M D Figgins, J Huddleston, T Bedford, and F A D Matsen, IV. antigen-prime: Simulating coupled genetic and antigenic evolution of influenza virus. *bioRxiv*, 2026. Preprint.

Chapter A

Supplementary Material for `torchdms`

A.1 Conditional Dependencies and Parameter Interpretability

The cGGE and csGGE architectures both incorporate a skip connection from the fold latent phenotype (ϕ_f) to the binding nonlinearity (g_b), encoding the biophysical prior that protein folding influences binding capacity. However, these architectures differ in their training procedures, with important consequences for parameter interpretability.

A.1.1 Skip Connections in cGGE

The cGGE architecture's skip connection allows binding predictions to depend on folding, but joint training may compromise interpretability. The binding nonlinearity receives two inputs:

$$\hat{y}_b = g_b(\phi_b(v), \phi_f(v)), \tag{A.1}$$

where $\phi_b(v) = \beta_b^\top x_v$ and $\phi_f(v) = \beta_f^\top x_v$ are the latent phenotypes computed from the binary encoding x_v of variant v. This skip connection allows the model to capture the dependency between folding and binding: variants that fail to fold properly cannot bind effectively, regardless of mutations in the binding interface.

During joint training on both phenotypes, gradients from the binding loss $\mathcal{L}_b$ propagate through two paths:

- **Binding pathway**: $\mathcal{L}_b \to g_b \to \phi_b \to \beta_b$

- **Skip connection**: $\mathcal{L}_b \rightarrow g_b \rightarrow \phi_f \rightarrow \beta_f$

The second path means that β_f is optimized not only to predict folding phenotypes but also to provide information useful for binding prediction. This gradient interference can compromise the interpretability of β_f as a pure representation of mutational effects on folding.

A.1.2 Sequential Training in csGGE

The csGGE architecture addresses this interpretability concern through sequential training:

- **Phase 1**: Train the fold pathway (β_f, g_f) on folding data alone.
- **Phase 2**: Freeze the fold pathway and train only the binding pathway (β_b, g_b) on binding data.

By freezing β_f before introducing binding gradients, the fold coefficient matrix remains a pure representation of mutational effects on protein stability. The skip connection still provides ϕ_f as input to g_b, preserving the biophysical dependency, but gradients cannot flow back to modify β_f.

Chapter B

Supplementary Material for `antigen-prime`

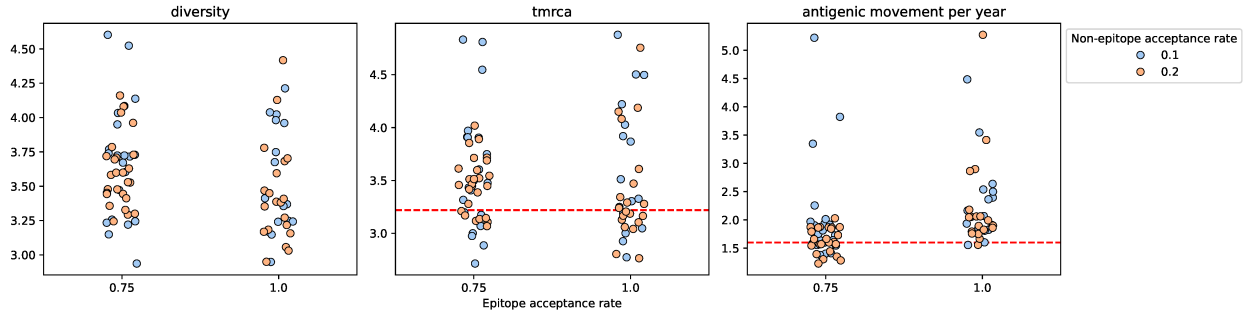

Figure B.1: Genealogical and antigenic summary statistics for 120 simulations of 30 years of H3N2-like evolution. Each point represents a single simulation. Red dashed horizontal lines represent empirical values reported in previous studies. The 9.0 diversity cutoff was used in Bedford et al. [94], and the antigenic movement per year metric was chosen to reflect results observed in Smith et al. [51] and Koel et al. [102]. **A**: Genealogical diversity (π_G). **B**: Time to most recent common ancestor (TMRCA), red dashed line used as a target value based on data from nature [101]. **C**: Antigenic movement per year, red dashed line used as a target value based on data from nature.

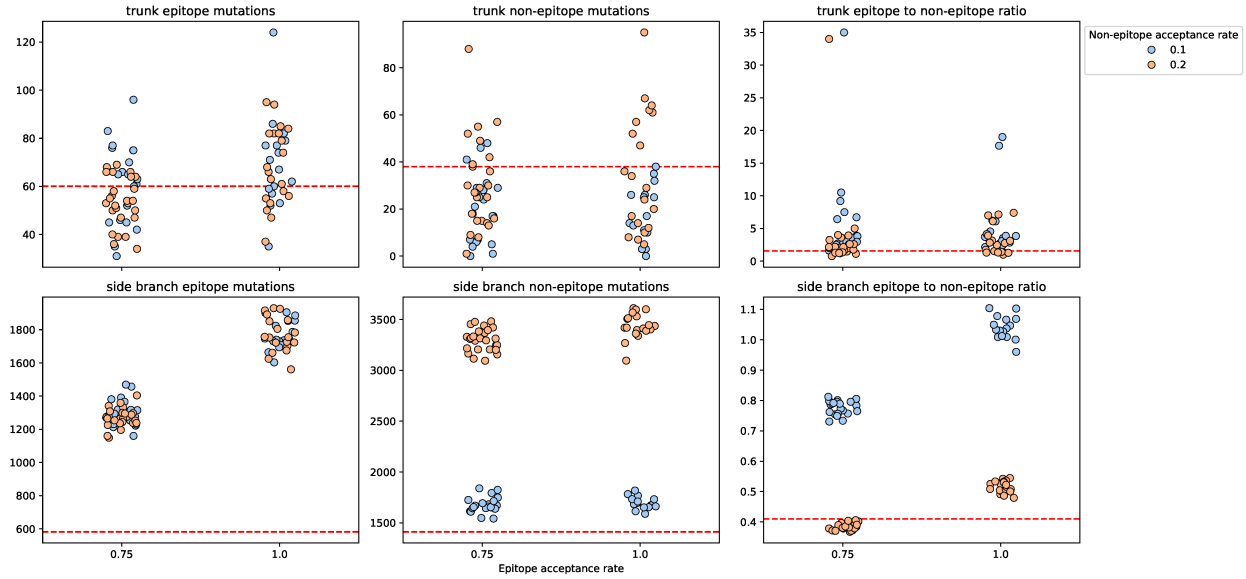

Figure B.2: Summary of `antigen-prime` mutation statistics for 120 simulations of 30 years across various epitope/non-epitope acceptance rate configurations. Each point represents results from a single simulation. Red dashed horizontal lines represent empirical values reported in Table 4.1. **A**: Total number of epitope mutations observed on the trunk of the phylogeny. **B**: Total number of non-epitope mutations observed on the trunk of the phylogeny. **C**: Ratio of epitope to non-epitope mutations observed on the trunk of the phylogeny. **D**: Total number of epitope mutations observed on side branches of the phylogeny. **E**: Total number of non-epitope mutations observed on side branches of the phylogeny. **F**: Ratio of epitope to non-epitope mutations observed on side branches of the phylogeny.

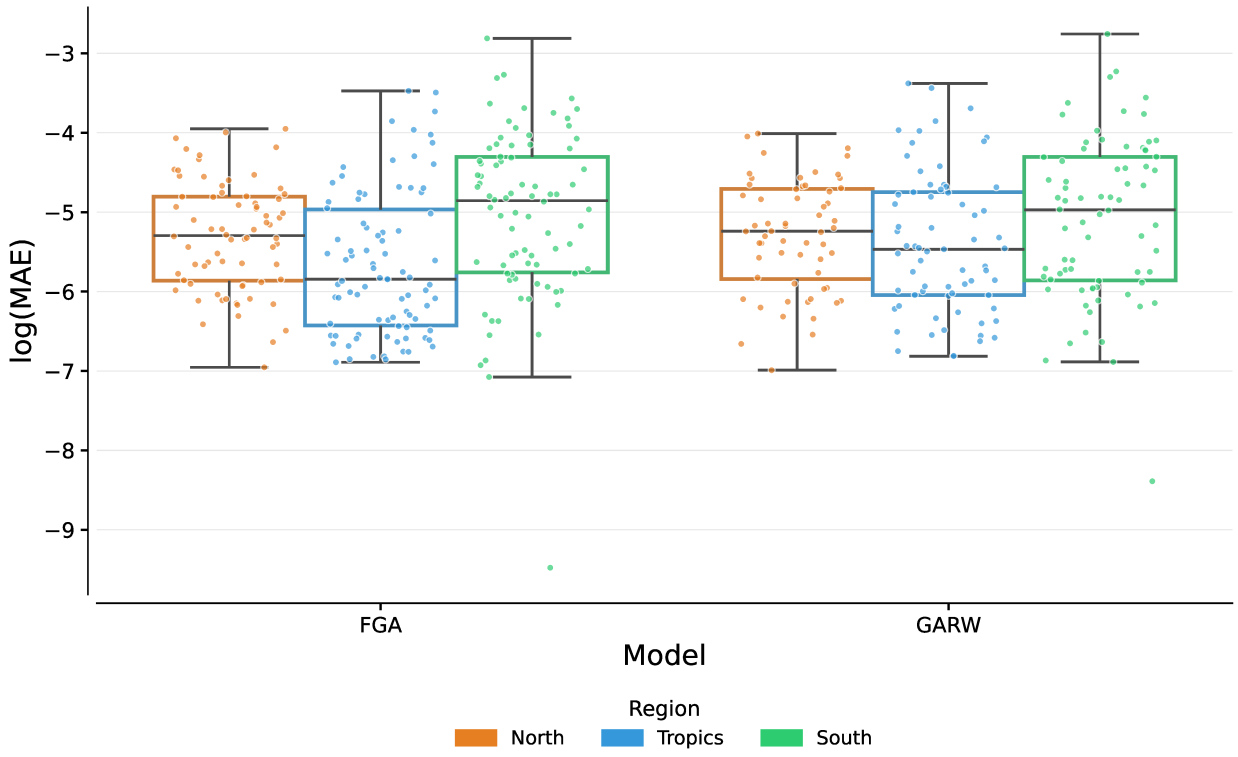

Figure B.3: Distribution of variant-specific growth rate inference errors comparing FGA and GARW models across geographic demes. Log-mean absolute error ($\log$ MAE) distributions show the performance of both model types in inferring exponential growth rates (r_{model}) compared to empirical growth rates (r_{data}) calculated from variant frequency dynamics. Each data point represents a single variant within a training window, with boxplots showing the distribution of errors and individual points overlaid. Both FGA and GARW models demonstrate comparable performance with similar median errors and error variance across all geographic demes.

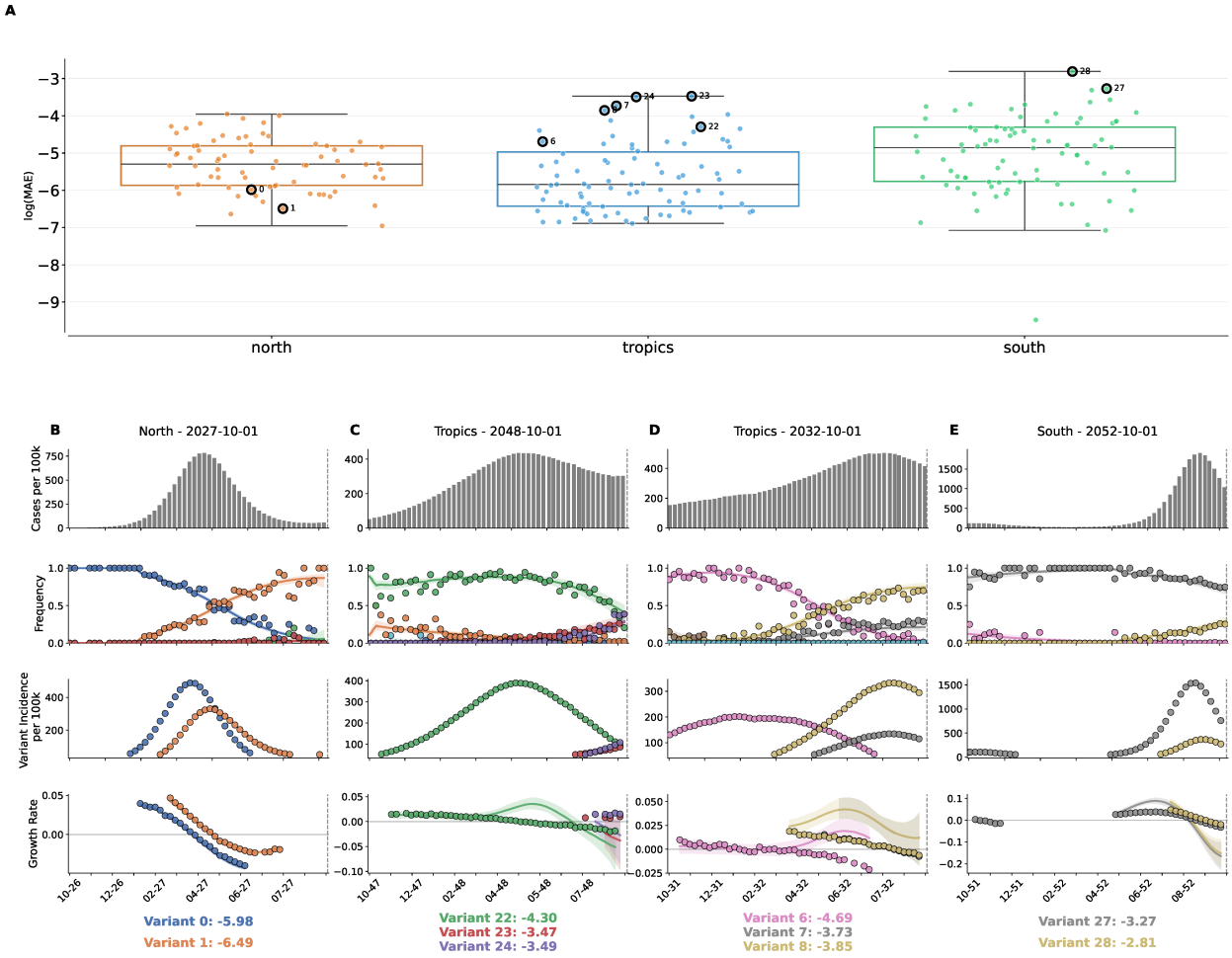

Figure B.4: FGA model performance for variant growth-rate inference. **(A)** Distribution of log MAE values of inferred variant-specific growth rates across geographic demes. Each data point represents a single variant from a specific training window. Variants selected for panels B-E are circled. **(B-E)** Examples of FGA model performance across different training windows. Average log MAE values for each variant are reported below the growth-rate plots. **(B)** Successful inference of frequencies and growth rates (North, 2027-10-01). **(C)** Growth rates underestimated near end of the analysis window (Tropics, 2048-10-01). **(D)** Growth rates overestimated for multiple variants (Tropics, 2032-10-01). **(E)** Growth rates inaccurately inferred for both variants 27 and 28 (South, 2052-10-01).