Integrating Complex Biomedical Datasets into Differential Equation Models

Dissertation
zur
Erlangung des Doktorgrades (Dr. rer. nat.)
der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von
Simon Merkt
aus
Mainz

Bonn 2024

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

Gutachter/Betreuer: Prof. Dr. Jan Hasenauer
Gutachter: Prof. Dr. Kevin Thurley

Tag der Promotion: 01. Juli 2025
Erscheinungsjahr: 2025

Abstract

This thesis focuses on developing mathematical models utilizing complex biomedical datasets to understand epidemiological and cell-biological processes. The thesis is divided into three primary research articles and addresses how differential equation models can be applied to enhance insights into disease dynamics and cellular communication processes.

The research is motivated by the need to better understand complex biological processes, particularly epidemiological patterns in populations and cellular dynamics. The COVID-19 pandemic highlighted the importance of life sciences in global health and underscored the role of computational methods and mathematical modeling in analyzing biomedical data. Mechanistic models incorporating biological knowledge are chosen over purely data-driven models due to their ability to integrate prior information and predict beyond the available data. The thesis focuses on ordinary differential equations (ODEs) and partial differential equations (PDEs) as tools to model time-dependent biological processes.

The thesis addresses two main research questions:

I. How can mechanistic models enhance classical cohort studies to understand disease dynamics better?

II. How can mechanistic models be utilized to capture communication processes derived from single-cell data?

The research involves integrating data into differential equation models through the following steps:

1. Data Preprocessing: Reducing the complexity of high-dimensional datasets using methods like Principal Component Analysis, diffusion maps, and clustering.
2. Model Construction: Developing ODE and PDE models tailored to specific biomedical phenomena.
3. Parameter Estimation: Using statistical methods and gradient-based optimization techniques to estimate unknown parameters in the models.
4. Uncertainty Analysis: Employing techniques like Markov Chain Monte Carlo to assess the reliability of model predictions.

The thesis consists of three key studies:

A. SEIR Modeling for COVID-19 in Ethiopia: The first study focuses on integrating antibody data into Susceptible-Exposed-Infectious-Recovered (SEIR) models to predict the spread of SARS-CoV-2 in Ethiopia. Three models are developed: a basic SEIR model, an extended SEIR model, and one including a variant model for a new virus strain. Key insights include the impact of healthcare workers' exposure and the underreporting of infections in Ethiopia.

B. Multivariant and Antibody-Level Modeling: This second study extends the SEIR modeling to a model including multiple virus variants and one focusing on antibody levels. They provide insights into how variant-specific immunity and vaccination affect the spread of COVID-19. The model predicts early and widespread vaccination could have significantly mitigated infection waves in Ethiopia.

C. PDE Modeling of Cell-to-Cell Communication: The final study develops a PDE-ODE model to describe immune cell activation via ligand-mediated communication. Applied to single-cell RNA sequencing data, the model highlights the role of cell-to-cell communication in immune responses and demonstrates the impact of ligand concentration on cellular dynamics.

The thesis concludes that mechanistic models, particularly ODE and PDE systems, are powerful tools for integrating biomedical datasets and providing valuable insights into disease dynamics and cellular processes. Future research could explore incorporating more detailed genetic data to enhance model accuracy or multi-scale modeling, linking within-host viral dynamics to population-level models.

Contents

List of Figures

List of Algorithms

I am very grateful for the supervision and guidance I obtained from Jan Hasenauer and all the opportunities he provided me by granting me a position as a PhD student in his lab and during that time.
Moreover, I want to thank Kevin Thurley for agreeing to be my thesis' second reviewer and Thomas Schultz and Felix Meißner for agreeing to be part of my thesis committee.
Additionally, I would like to thank my cooperation partners, without whom many of my projects would not have been possible.
Last but not least, I want to thank all members and former members of the Hasenauer Lab for the fruitful exchanges during the years and for the great time I had as a PhD student. In particular, I would like to thank Clemens, Dilan, Jakob, Jonas, Lea, Manu, Polina and Vincent for their valuable feedback.

Simon Merkt, Bonn 2024

1 Introduction

This thesis is structured in a cumulative format, comprising three research articles that present the integration of complex biomedical datasets into differential equation models:

A. *Seroepidemiology and model-based prediction of SARS-CoV-2 in Ethiopia: longitudinal cohort study among front-line hospital workers and communities* [1].

B. *Long-term monitoring of SARS-CoV-2 seroprevalence and variants in Ethiopia provides prediction for immunity and cross-immunity* [2].

C. *A dynamic model for Waddington's landscape accounting for cell-to-cell communication* [3].

Each article was (co-)authored by the author of this thesis, with contributions from various co-authors who assisted with data collection, analyses beyond the scope of this thesis, and supervision. They are reprinted as part of this thesis in Appendices A, B and C and the specific contributions the author of this thesis made to each paper will be detailed in the respective publication summary sections.
In addition to the primary research of this thesis' author in the area mentioned above, the author also contributed to the following projects:

- Development of the parameter estimation standard PEtab [4].
- COVID-19 modeling using early Wuhan public data [5].
- Mini-batching method for large-scale cancer cell line optimization [6].
- Development of the parameter estimation toolbox pyPESTO [7].
- Cost-effectiveness analysis of vaccination strategies in Ethiopia [8].

This thesis is structured in six sections. In the remainder of Section 1, we provide background, motivation, and goals for the research presented in this thesis and state the research question. In Section 2, we introduce the mathematical framework of the methods employed in addressing these questions. Next, in Sections 3, 4, and 5, we provide a summary of each publication focusing on the contributions of this thesis' author. Finally, in Section 6, we conclude by recapitulating the achieved scientific results, the advancements in knowledge gained regarding the scientific questions, and an outlook on potential future research.

1.1 Background, Motivation, and Goals

Life science is an umbrella term for several sciences investigating different aspects of living entities and processes at various scales, such as biology, genetics, ecology, and medicine [9]. Their goals range from a deep understanding of systems to improving health outcomes. Recently, the COVID-19 pandemic has highlighted the crucial role of life sciences in ensuring global public health. On a macro scale, most countries tracked the disease spread and monitored specific variants within their populations, while the World Health Organization (WHO) tracked global dynamics. This led to the development of non-clinical intervention strategies like lockdowns, isolation of infected individuals, and facial mask mandates [10]. Simultaneously, researchers worldwide focused on understanding the virus

at a micro level and leveraging this knowledge to combat its spread. The discovery of the virus's spike protein as a cell entry mechanism and its use in traditional vector-based and novel mRNA vaccination methods stands out as a critical achievement [11]. Even though the threat of SARS-CoV-2 has decreased due to widespread immunity from infections and vaccinations, other health challenges remain. On the one hand, there are more structural challenges, like reproductive rights and malnutrition, that are best addressed in politics and logistics. However, there are also the threats posed by particular diseases like mutated viruses with pandemic potential, diabetes, and various forms of cancer [12]. The recent pandemic also underscored the crucial role of computational methods for life sciences [13]. Statistics on past disease trends and population model predictions became a staple in the daily news, while sophisticated computational techniques significantly aided vaccine development. This period has particularly highlighted the necessity of mathematical models for systematically analyzing complex biological phenomena and predicting future outcomes. Echoing George Box's famous assertion that "all models are wrong, but some are useful," one can always choose from various modeling approaches, each with different simplifying assumptions and core ideas to approximate reality most helpfully.

Modeling approaches can be categorized as mainly data-driven approaches or approaches driven by already-known mechanisms. Often, researchers use terms like machine learning, empirical modeling, and artificial intelligence to refer to data-driven approaches [14, 15]. As there is no consensus definition, particularly on machine learning and artificial intelligence, we will use empirical models as an umbrella term for the data-driven approaches. Where empirical models deduce patterns solely from observed data, their counterpart, the mechanistic models, incorporate known mechanics into the model before calibrating it with data. Recently, empirical models, particularly large-scale deep neural networks such as Google's AlphaFold—which predicts protein structures—have garnered significant attention [16]. With the rise of big data, e.g., in cell biology driven by next-gen sequencing [17], one might be inclined to conclude that empirical modeling is the sole future of modeling in life sciences. However, it is crucial to recognize that these two approaches differ fundamentally in their core assumptions, leading to distinct advantages and disadvantages depending on the research goals, data complexity and availability, as well as prior knowledge. Baker et al. [15] identified three key differences:

1. Mechanistic models inherently struggle to incorporate data of different scales, a task more straightforward in machine learning.

2. The mechanistic framework allows for integrating prior information, enabling work with smaller datasets, whereas machine learning always requires large datasets.

3. Mechanistic models can be employed for predictions on questions outside the scope of the integrated data after validation. In contrast, machine learning predictions are limited to within the scope of the training data.

Here, we focus on mechanistic models, in particular systems of differential equations, which describe changes in a state with respect to a function of the state itself. We employ two subclasses of differential equations: ordinary differential equations (ODEs) and partial differential equations (PDEs). ODEs, which describe mean effects over time, are a very commonly used modeling framework in systems biology for integrating complex biological data to understand the underlying processes [18]. Specifically in epidemics, they have a long history of being employed to model the spread of infectious diseases through

populations via compartmental modeling. They are still one of the main tools used nowadays [19–21]. PDEs, on the other hand, also account for spatially non-homogeneous effects and can be very useful for describing cell populations across states [22–24].
Since the complexity, which is required to model biomedical phenomena, often prohibits purely analytical analysis of such models, numerical methods have been developed tailored explicitly to solving ODEs and PDEs [25, 26] and will be employed by us. Moreover, models usually have unknown quantities, which must be inferred from measurement data. Such inference problems are also unlikely to be solvable analytically; hence, numerical inference strategies have been developed to address this issue [27].
The inclusion of mechanistic knowledge keeps computational and data demands, in general, lower than in purely empirical models. However, one still has to leverage which simplifications are reasonable to effectively integrate data into these models and use them to gain insights. Or, in the language of George Box, one has to find out how wrong the model can be while still being functional.

1.2 Research Question

In this thesis, we will focus on the question of *how to integrate complex biomedical datasets into differential equation models to enhance our understanding of epidemiological patterns and cellular communication processes*. This question will be addressed by answering the following two subquestions, where the first aims to gain insights on the macro-level of a population of people and the second at the micro-scale of single cells:

I. How can mechanistic models enhance classical cohort studies to understand disease dynamics better?

II. How can mechanistic models be utilized to capture communication processes derived from single-cell data?

These questions are addressed in three publications, which are summarized in Sections 3, 4 and 5: **Question I** is tackled in [1] and [2], which are summarized in Subsections 3 and 4, respectively. The former presents three ODE models capturing the development of SARS-CoV-2 antibody levels in community members and health care workers in two Ethiopian cities. The latter is based on the medical follow-up study, where the original survey of antibody prevalence was extended and enriched by variant information obtained from sequenced positive PCR tests. These data sets are integrated into two ODE models, where one captures the variant dynamic by describing infection pathways of up to four infections, and the other makes full use of the antibody data to predict immunity levels and retrospective vaccination effects. **Question II** is addressed in [3] where a PDE-ODE system is introduced, which captures the developmental potential landscape of communicating cells by describing cell densities and ligand concentrations simultaneously. It is summarized in Subsection 5. However, before presenting the results, we must first introduce the core mathematical concepts used to address these research questions.

2 Mathematical Framework

For data integration into biomedical models, particularly for the three publications presented in this thesis, the basic workflow usually consists of the following steps:

1. *Data Preprocessing*: As data collected from experiments might not be directly usable within a computational model, preprocessing of the data becomes necessary. This process can range from simple tasks like aggregating antibody data over time to more complex procedures, such as trajectory inference on single-cell sequencing data.

2. *Model Construction*: Formulating an appropriate model is crucial. If the model is incorrectly specified or simulated, statistical assumptions might be violated, and conclusions drawn from the model might be invalid. We rely on established frameworks in differential equations, but these frameworks must be adapted and enhanced for each case.

3. *Parameter Estimation*: After constructing the model, the unknown parameters must be estimated from the collected data to conclude dynamics and dependencies of interest. While we use established frameworks for this purpose, selecting suitable, statistically motivated objective functions and efficient optimization algorithms is essential to ensure feasible and accurate parameter estimation.

4. *Uncertainty Analysis*: As parameter point estimates depend on the collected data and only a subsample of the population can be observed, the parameter estimates are inherently associated with uncertainty. Uncertainty analysis is used to quantify the uncertainty amount and understand the parameter estimates' reliability.

In the following, we will discuss the mathematical backgrounds of methodologies for these steps. We aim to balance generalizability and a clear focus on the specific applications relevant to our research.

2.1 Data Preprocessing

In life sciences, complex datasets often exhibit multimodal characteristics and significant variability in statistical power, resolution, and detail across different observations, time points, and conditions. This inherent variability requires a careful balance between maintaining granularity, preserving statistical power, and managing computational demands. Additionally, the high-dimensional nature of the data presents challenges, as it is not always immediately apparent which underlying phenomena should be modeled. As a result, preliminary data analysis may be necessary to supplement standard aggregation techniques. While the field of data processing and analysis is vast and could warrant multiple theses, this section will specifically focus on the dimension reduction, clustering, and smoothing methods employed in this work.

2.1.1 Principal Component Analysis

Datasets are often inherently high-dimensional, especially in single-cell transcriptomic sequencing. Visualizing and analyzing such datasets directly can be computationally challenging and may obscure meaningful patterns. To address this, dimensionality reduction techniques are commonly applied for visualization and before proceeding with

further analysis. One widely used linear dimensionality reduction method is *Principal Component Analysis* (PCA) [28].

Consider a high-dimensional dataset $\mathcal{D} \subset \mathbb{R}^{n_\mathcal{D}}$, where $n_\mathcal{D}$ represents the number of dimensions. Suppose this dataset has been normalized to have zero mean, and we seek a lower-dimensional representation. The objective of PCA is to find orthogonal axes, denoted as $(v_i)_{i=1}^{n_\mathcal{D}}$, that capture the greatest variance in the data, with the first axis capturing the greatest variance, the second capturing the next greatest variance orthogonal to the first, and so on. These axes, known as *principal components*, serve as a new coordinate system. Dimensionality reduction is then achieved by projecting data points $d \in \mathcal{D}$ onto the first n_r principal components $(v_i)_{i=1}^{n_r}$, where $n_r \ll n_\mathcal{D}$, thus preserving as much variance as possible.

To ensure no single dimension dominates the analysis, it is typical to standardize the data before applying PCA. Standardization shifts non-zero mean data to have zero mean and scales each feature to have unit variance, preventing dimensions with disproportionately large variance from skewing the results.

The principal components are derived by solving a series of maximization problems. The first principal component v_1 is the direction along which the variance of the data is maximized:

$$v_1 = \arg\max_{\|v\|=1} \sum_{d \in \mathcal{D}} (d^T v)^2.$$

Subsequent components v_k are found by removing the projection of the data onto the previously computed components and then maximizing the variance in the remaining subspace:

$$\mathcal{D}_k = \left\{ d - \sum_{i=1}^{k-1} d^T v_i v_i^T \,\middle|\, d \in \mathcal{D} \right\},$$
$$v_k = \arg\max_{\|v\|=1} \sum_{d \in \mathcal{D}_k} (d^T v)^2.$$

The components are usually derived by performing an eigenvalue decomposition on the covariance matrix of the data or through singular value decomposition (SVD) of the data matrix. The eigenvectors represent the directions in which the data varies the most, while the eigenvalues measure the amount of variance captured along each direction. By ordering the eigenvectors based on their associated eigenvalues, the desired sequence of principal components, $(v_i)_{i=1}^{n_\mathcal{D}}$.

Each data point $d \in \mathcal{D}$ can then be represented in the reduced space spanned by the first n_r components, using the coordinates $(d^T v_i)_{i=1}^{n_r}$. The transformation from the original coordinate system $\mathbb{R}^{n_\mathcal{D}}$ to the principal components can be interpreted as an orthogonal transformation (involving rotations and reflections) of the data around the origin, aligning the principal components with the axes of maximum variance.

The dimensionality of the reduced dataset, n_r, is often chosen based on the percentage of variance explained by the first n_r components. This percentage is computed as the ratio of the sum of the n_r largest eigenvalues to the sum of all eigenvalues. Commonly, n_r is selected for visualization purposes (i.e., $n_r = 2$ or $n_r = 3$) or to ensure that the explained variance exceeds a predetermined threshold.

While PCA is a powerful tool for dimensionality reduction, it has limitations when applied to nonlinear datasets. For instance, PCA struggles to capture structure in datasets where

samples of one class encircle samples of another. To address such cases, kernel PCA [29] can be employed. Kernel PCA maps the data into a higher-dimensional feature space before performing PCA, though this requires careful tuning of hyperparameters to ensure effective results.

2.1.2 Diffusion Maps

To overcome the limitations of linear methods in the case of nonlinear data, one can employ *diffusion maps*. This nonlinear technique assumes the data lies on a low-dimensional submanifold within the sample space $\mathbb{R}^{n_\mathcal{D}}$ [30]. Learning this manifold is done by performing a random walk through the graph of data points. For computing the transition probabilities, we need a kernel k, which is required to be symmetric, i.e., $k(x,y) = k(y,x)$, and positivity preserving, i.e., $k(x,y) \geq 0$. Through normalization by $d(x) = \sum_{y\in\mathcal{D}} k(x,y)$ we obtain a transitionkernel of a Markov chain on $\mathcal{D}$:

$$p(x,y) = \frac{k(x,y)}{d(x)}$$

with the property $\sum_{y\in\mathcal{D}} p(x,y) = 1$. To construct the diffusion map, we compute the eigenvalues λ_i and eigenvectors ψ_i of the $|\mathcal{D}| \times |\mathcal{D}|$-dimensional transition matrix P, where $P_{xy} = p(x,y)$. If we assume that the graph is connected, finite and that $k(x,x) > 0$, we obtain a sequence of eigenvalues $1 = \lambda_0 > |\lambda_1| \geq |\lambda_2| \geq \ldots$, where the first eigenvector ψ_0 corresponds to the unique stationary distribution

$$\pi(y) = \frac{d(y)}{\sum_{z\in\mathcal{D}} d(z)}.$$

Hence, the first eigenvalue-eigenvector pair will often be omitted in the following steps. To capture the intrinsic geometry of the data, one defines the t-step *diffusion distance* between two points as

$$D_t(x,y) = \left(\sum_{z\in\mathcal{D}} \frac{(p(x,z) - p(y,z))^2}{\pi(z)} \right)^{\frac{1}{2}}.$$

This distance measures how similar two points are based on their connectivity in the data manifold, and it can be shown that

$$D_t(x,y) = \left(\sum_{j\geq 1} \lambda_j^{2t} (\psi_j(x) - \psi_j(y))^2 \right)^{\frac{1}{2}}, \tag{1}$$

where $\psi_j(x)$ denotes the x-th entry of ψ_j. Now, one can embed the data into a low-dimensional space using the first n_r informative diffusion maps $(\lambda_j^t \psi_j)_{j=1}^{n_r}$. Informative here means that one is excluding the first eigenvalue and eigenvector, corresponding to the stationary distribution, Since (1) ensures that the Euclidean distance in the reduced space approximates the t-step diffusion distance in the original data space, we obtain a meaningful representation of the data's intrinsic structure.
In the context of single-cell analysis, Gaussian kernels are a reasonable choice [31]. Hence, we define

$$k(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2} \right)$$

as the kernel. Alternatively, one can use data-dependent kernel bandwidths σ, e.g., by incorporating information about the neighborhood of the data points and obtaining

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2(\sigma_x^2 + \sigma_y^2)}\right).$$

Usually, one wants to disregard local cell densities since varying densities might be an artifact of the measurement process without any underlying biological relationship influencing those variations. Therefore, we divide the kernel by the densities around x and y to obtain transition probabilities:

$$p(x, y) = C(x)\frac{k(x, y)}{\sum_{x' \in \mathcal{D}} k(x', y) \sum_{y' \in \mathcal{D}} k(x, y')},$$

where $C(x)$ is the normalizing constant, such that $\sum_y p(x, y) = 1$ for all $x \in \mathcal{D}$. Moreover, implementations like [32] set $t = 0$ such that the diffusion maps simplify to $(\psi_j)_{j=1}^{n_r}$, the eigenvectors of P.

2.1.3 Diffusion Pseudotime

In the case of the data $\mathcal{D}$ representing single-cell measurements, in addition to dimensionality reduction, researchers are often interested in obtaining developmental trajectories. These so-called *trajectory inference* methods can be viewed as reduction methods to a one-dimensional space that includes an ordering representing the developmental stage. In this section, we introduce *diffusion pseudotime* [33], which builds directly upon the theory of diffusion maps.
Diffusion pseudotime is defined as the distance:

$$\mathrm{dpt}(x, y) = \|M(x, \cdot) - M(y, \cdot)\|,$$

where $M = \sum_{t=1}^{\infty} \bar{P}^t$. Here, $\bar{P}$ is the transition matrix from the previous section, where the eigenvector ψ_0 corresponding to the largest eigenvalue 1 and the steady state distribution is subtracted from P, resulting in $\bar{P} = P - \psi_0 \psi_0^T$. Since all remaining eigenvalues of $\bar{P}$ have absolute values smaller than 1, we can use the identity from spectral theory $M = \sum_{t=1}^{\infty} \bar{P}^t = (I - \bar{P})^{-1} - I$ for the computation of M. The above-defined diffusion pseudotime distance represents the probability of reaching cell y starting at cell x over infinite time. To infer a trajectory, we select a fixed root cell $r \in \mathcal{D}$, e.g., a known stem cell, and assign diffusion pseudotime values to all $d \in \mathcal{D}$ using $\mathrm{dpt}(r, d)$. We obtain the developmental trajectory by sorting all cells by their assigned values in ascending order.

2.1.4 Uniform Manifold Approximation and Projection

Here, we introduce another nonlinear dimension reduction method mainly employed for visualization: *Uniform Manifold Approximation and Projection* (UMAP) [34]. Like diffusion maps, it tries to capture the manifold on which the data lies and operates on a graph constructed between data points. It has solid theoretical backgrounds in algebraic topology and fuzzy set theory, motivating some choices of computational derivation, which we will focus on in this section. The core idea is to capture the local connectivity of the data manifold by a collection of fuzzy 1-simplicial sets and to represent those in a low-dimensional space. A 1-simplex is a line segment connecting two points, and "fuzziness"

implies that the connection of points is not binary on or off. Still, there is a certain ambiguity or strength of connection, e.g., the connection for two close points could be 0.9 while for two far away points, it could be 0.1. UMAP starts by constructing the k-nearest neighbour graph in the original data space. Then, it computes the fuzzy simplicial set, which captures the local connectivity of the data, by assigning probabilities to the edges of the graph based on the distances between points, thereby creating a fuzzy topological representation. To be precise: For each pair of data points x_i and x_j, it calculates a membership strength p_{ij} that represents the likelihood of x_j being a neighbor of x_i. This is done using a smooth, exponential decay function of the distance between x_i and x_j:

$$p_{ij} = \exp\left(-\frac{d(x_i, x_j) - \rho_i}{\sigma_i}\right)$$

Here, $d(i, j)$ is the distance between points x_i and x_j, usually Euclidean distance, ρ_i is the distance to the closest neighbor of x_i (to account for local density variations), and σ_i is a scaling factor ensuring that each point has approximately the same number of significant neighbors n_{neighbor}. Since p_{ij} is not symmetric, it represents a directed graph with weighted edges, and we obtain an undirected graph by fuzzy set union, i.e. $\bar{p}_{ij} = p_{ij} + p_{ji} - p_{ij}p_{ji}$. In the probabilistic sense, this represents the probability of x_i being in the neighbor set of x_j or x_j being in the set of x_i. UMAP then constructs a low-dimensional representation by optimizing the embedding that minimizes the cross-entropy function aligning the fuzzy simplicial sets of the high-dimensional and low-dimensional data:

$$\sum_{e \in E} w_h(e) \log\left(\frac{w_h(e)}{w_l(e)}\right) + (1 - w_h(e)) \log\left(\frac{1 - w_h(e)}{1 - w_l(e)}\right), \tag{2}$$

where E is the set of possible 1-simplices, i.e., in the graph sense, edges between nodes, and $w_h(e)$ and $w_l(e)$ are the weights of each edge e in the high, respectively low, dimensional space. Here, a second hyper-parameter comes into play: the minimal distance allowed between points in the low-dimensional representation. The choice of this minimal distance and the number of neighbors n_{neighbor} is crucial for obtaining meaningful low-dimensional representations. Generally speaking, higher n_{neighbor} will preserve more global structure and be more computationally demanding. In contrast, minimal distance controls the spread of the data points in the low-dimensional representation, and a higher minimal distance leads to a higher spread. In practice, the UMAP algorithm starts at random initializations of low-dimensional representations. It uses a differentiable approximation of (2) and the gradient descent method to obtain the embedding.

2.1.5 k-Means

A common question in data analysis is whether clusters can be deduced from the data alone without prior knowledge. While dimension reduction methods aim at finding structure in the variable space, clustering approaches search for structure in the observations and, therefore, can be employed complementary to or independent of dimension reduction. One approach for clustering datasets is the method of *k-means*, which aims to find a partition of the dataset $\mathcal{D}$ such that the variance in each cluster is minimized [35]. This minimization equates to the minimization of the sum of squared distances of each data point to its cluster center, i.e., mean. Formally, the objective is denoted as

$$\underset{(\mathcal{D}_1,\ldots,\mathcal{D}_k)\in\mathcal{C}}{\arg\min} \sum_{i=1}^{k} \sum_{d \in \mathcal{D}_i} \|d - \mu_i\|^2, \tag{3}$$

where

$$\mathcal{C} = \left\{ (\mathcal{D}_1, \ldots, \mathcal{D}_k) \text{ subsets of } \mathcal{D} \,\middle|\, \dot{\bigcup}_{i=1}^{k} \mathcal{D}_i = \mathcal{D} \right\}$$

and μ_i is the mean of $\mathcal{D}_i$. Algorithmically, this is usually implemented in an iterative approach similar to expectation-maximization: The process begins with a random selection of k initial means, $\mu_1, \ldots, \mu_k$. Each $\mathcal{D}_i$ is then defined as the set of data points with the minimal distance to μ_i, i.e.,

$$\mathcal{D}_i = \{d \in \mathcal{D} : \|d - \mu_i\|^2 \leq \|d - \mu_j\|^2 \quad \forall j = 1, \ldots, k\},$$

and points assigned to multiple clusters are removed from all but one, e.g., randomly or by a rule on their index. Subsequently, μ_i is redefined as the mean of $\mathcal{D}_i$, and this process is repeated until convergence is achieved. To ensure that the clustering globally minimizes the total variance described in (3), the algorithm is typically run multiple times, accounting for the random initialization of the means. As for previously discussed methods, choosing the right hyperparameter, k, is crucial since it is rarely the case that one knows how many clusters to expect. Clustering quality can be evaluated using quality measures like *silhouette scores*, where mean inner cluster distances are compared to minimum in-between cluster distances [36]. Hence, in practice, one can run the clustering algorithm for multiple values of k, and a-posteriori chose the best scoring k.

2.1.6 Kernel Density Estimates

Data obtained from biological experiments or medical surveys rarely cover time intervals sufficiently dense to provide a complete picture of the distributions that underlie the biological processes or disease progression being studied. If one wants to integrate certain data directly into the model, smoothing techniques can be required. The smoothing method we will present here is the *kernel density estimation*, where one tries to capture the underlying density function f of n independent and identically distributed samples $d^1, \ldots, d^n \in \mathcal{D}$ [37, 38]. The kernel density estimator of the samples underlying density f is defined as

$$\hat{f}_h(d) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{d - d^i}{h}\right),$$

where $K : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is the kernel and $h > 0$ is called the *bandwidth*. The choice of K is crucial and can have various forms. However, for $\hat{f}_h$ to be a density, i.e., integrate to 1, also K has to be a density. In our cases, we will resort to the standard normal density and exponential density functions. The bandwidth h controls the smoothness of the resulting density estimate, with larger values of h producing smoother estimates and smaller values capturing more detail, and there are rules of thumb to choose it [39].

2.2 Differential Equation Models

Not all biomedical phenomena are in homeostasis, and sometimes, one wants to capture the evolution over time, e.g., of a spreading disease or a developing cell population. Time-dependent mathematical models $x(t)$ can express such biological or epidemiological processes. The model is called deterministic if it is not dependent on any randomness. Since biological processes usually involve randomness, a deterministic model represents mean effects. Assuming the process is smooth enough, it is often easier to formulate

a relation between the changes of the process over time and the process itself than to explicitly state the process itself, $\frac{d}{dt}x(t) = f(t, x)$. This is called a differential equation: An ordinary differential equation (ODE) in case only the time derivative is involved and a partial differential equation (PDE) if there are also derivatives with respect to other variables involved, i.e., $\frac{d}{dt}x(t, s) = f(t, s, x, \nabla x, \nabla^2 x, \ldots)$ for some state variable s. Both types of differential equations employ distinct analytical and numerical solution methods, which we will introduce in the following.

2.2.1 Ordinary Differential Equations

Since higher-order ODE systems can be rewritten as first-order systems by treating derivatives of x as additional states, we will not mention them explicitly in this section. We start by defining the general problem setup:

Definition 1. For $x_0 \in \mathbb{R}^n$, $t_0, T \in \mathbb{R}_{\geq 0}$, such that $t_0 < T$, and a mapping $f : [t_0, T] \times \mathbb{R}^n \to \mathbb{R}^n$ the equation system

$$\begin{aligned} \frac{d}{dt}x(t) &= f(t, x) \qquad \forall t \in [t_0, T] \\ x(t_0) &= x_0 \end{aligned}$$

is called an *initial value problem* (IVP).

It is a well-known result in the theory of ODEs that if f is Lipschitz, the IVP has a unique local solution in the neighborhood of t_0 (Picard-Lindelöff, e.g., [40]). Moreover, in the biomedical context, the systems usually behave well enough that their solution extends to the whole time interval $[t_0, T]$.
Unless f is linear, finding the analytical solution of the IVP is not straightforward. Hence, iterative numerical approximation algorithms like the Euler method or more advanced Runge-Kutta methods are employed [25]. The former can be briefly described by

$$x_{i+1} = x_i + hf(t_0 + ih, x_i)$$

for step size h. The latter is a whole family of methods achieving higher accuracy by adding a weighted average of multiple function evaluations in each approximation step, leading to explicit or implicit equations. A prominent example is the Runge-Kutta method of fourth-order (RK4):

$$\begin{aligned} x_{n+1} &= x_n + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4), \\ t_{n+1} &= t_n + h, \end{aligned}$$

where

$$\begin{aligned} k_1 &= f(t_n, x_n), \\ k_2 &= f\left(t_n + \frac{h}{2}, x_n + h\frac{k_1}{2}\right), \\ k_3 &= f\left(t_n + \frac{h}{2}, x_n + h\frac{k_2}{2}\right), \\ k_4 &= f\left(t_n + h, x_n + hk_3\right). \end{aligned}$$

Since the Euler method only uses one function evaluation at each step, it belongs to the family of *single-step* solvers. Runge-Kutta methods use intermediate function evaluations but disregard the previous steps. In contrast, *multi-step* methods implement information from earlier steps at each new step. Prominent examples are, e.g., the families of Adams-Moulton algorithms and backward-differentiation-formula (BDF) algorithms [41, 42]. The latter is particularly suited for *stiff* ODEs, a category of equations that require tiny step sizes, sometimes even below numerical precision, for numerically stable solving. Both families have implementations in the framework of SUNDIALS CVODES [43, 44], which will be employed for solving the ODEs in this thesis through the Python interface of AMICI [45].

2.2.2 Partial Differential Equations

For notational reasons, we will only mention PDEs up to the second order in this section. However, everything mentioned here also holds for higher-order equations. Let us first define the setup:

Definition 2. Let $S \subset \mathbb{R}^{n_s}$ be open, $x_0 : S \to \mathbb{R}^{n_x}$ and $t_0, T \in \mathbb{R}_0^+$, such that $t_0 \leq T$. Then for a mapping $f : [t_0, T] \times S \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_s \times n_x} \times \mathbb{R}^{n_s \times n_s \times n_x} \to \mathbb{R}^{n_x}$ and an operator B the equation system

$$\begin{aligned} \frac{d}{dt}x(t,s) &= f(t,s,x,\nabla x,\nabla^2 x) && \forall t \in [t_0,T], \forall s \in S \\ Bx(t,s) &= g(t,s) && \text{on } \partial S \\ x(t_0,s) &= x_0(s) && \text{in } S \end{aligned}$$

is called a *boundary value problem* (BVP).

The boundary operator, B, is usually the identity, corresponding to Dirichlet boundary conditions, the derivative, corresponding to Neumann boundary conditions, or a combination of the two, corresponding to Robin boundary conditions. Unlike for ODEs, there is no general existence and uniqueness proof for PDEs, and the applicability of different proof approaches like variational techniques or fixed point theorems depend on the shape of f [46]. A specific type of PDE, which will be put to use in this thesis, is the reaction-diffusion-drift equation, where f is of the following form:

$$f(t,s,x,\nabla x,\nabla^2 x) = \nabla_s(D(t,s)\nabla_s x) - \nabla_s(v(t,s)x) + h(t,s,x).$$

Here, x could represent expression values of a particular biological or chemical component at time t and position or state s. Its change over time is governed by diffusion D, advection or drift v, and the reaction term h. Since for PDEs, finding an analytical solution is even more improbable than for ODEs, we again resolve to numerical schemes. Various approaches exist, like finite elements, finite differences, or finite volumes. In this thesis, we only use the finite volume method (FVM), implemented in the Python framework FiPy [47]. This method approximates the solution by solving the BVP for integrated averages on control volumes V_i, such that

$$\dot{\bigcup_i} V_i = S.$$

After integrating the BVP, the derivatives can be transformed into boundary terms with the help of the divergence theorem. By setting $\tilde{x}_i := \int_{V_i} x \, ds$ and denoting the unit normal

vector pointing outward at a.e. point on the boundary ∂V_i as $\mathbf{n}$ we obtain the system

$$\frac{d}{dt}\tilde{x}_i = \int_{\partial V_i} D(t,s)\partial_s x \mathbf{n}\, ds - \int_{\partial V_i} v(t,s)x\mathbf{n}\, ds + \int_{V_i} h(t,s,x)\, ds.$$

The remaining derivative can be approximated by finite differences, and x in the boundary integral terms can be substituted by an average of neighboring $\tilde{x}_k$ (or g if it is on the boundary of S). Finally, unless h is linear and independent of s, the last integral must be approximated by some transformation of $\tilde{x}_i$. Then, we are left with an ODE system, which we can solve via the methods described in the previous section. More rigorous calculations for various examples can be found in literature [48, 49].

2.3 Parameter Estimation Techniques

As we have just seen, during the numerical solution of the PDE, the FVM reduces the PDE to a system of ODEs. Hence, to simplify notation and without loss of generality, we introduce the concepts of statistical inference in this subsubsection only for ODEs. Usually, those models depend on unknown quantities $\theta \in \mathbb{R}^{n_\theta}$, like rates of a reaction or reproduction of a population, which are also called *parameters*. This now leads to the parameter-dependent ODE

$$\begin{aligned} \frac{d}{dt}x(t,\theta) &= f(t,x,\theta), \\ x(t_0,\theta) &= x_0(\theta), \end{aligned} \tag{4}$$

for which the theory described above can be applied. In the following subsections, we will show how these states can be mapped to data, i.e., how one defines an appropriate objective function, and how we obtain a θ such that the model with these parameters optimally describes our data. In reality, f often also depends on experimental conditions $c \in \mathbb{R}^{n_c}$, which could be, e.g., drug dosages administered to a cell culture. For notational clarity, we will disregard this dependency in the following.

2.3.1 Observable mapping

We want to infer unknown parameters $\theta \in \mathbb{R}^{n_\theta}$ from data $\bar{y}$. Since states are usually not observed directly, we have to define an observable mapping h such that

$$\bar{y}(t) = h(t,x,\theta). \tag{5}$$

This could include, e.g., a scaling by $s \in \mathbb{R}_{>0}$ of a state x_1 to account for relative measurements of some quantities, as is the case, for example, for data from Western blots or carbon-13 isotopes

$$h_1(t,x,\theta) = x_1 * s,$$

or the sum of two states x_2, x_3 if two isoforms that are part of the model can not be distinguished during the measurement process

$$h_2(t,x,\theta) = x_2 + x_3.$$

Moreover, one usually assumes that the observation process is noisy and includes some error assumption in the observable. There are various possibilities for how the noise could

be distributed, like Gaussian, Laplacian, or log-normal, and several ways to incorporate it into the observable mapping, like additive or multiplicative noise. A widespread error assumption is that one has conditional independent additive Gaussian noise,

$$h(t, x, \theta) = \tilde{h}(t, x, \tilde{\theta}) + \xi(t), \qquad \text{where } \xi \sim \mathcal{N}(0, \sigma(t)^2), \tag{6}$$

where $\tilde{h}$ now denotes the theoretical observable function without any noise, h denotes the noisy observable function. Moreover, in this notation $\tilde{\theta}$ denotes model parameters and observable parameters without potential noise parameters related to σ, and θ refers to the full parameter vector. For appropriate choices of $\tilde{h}$, this scenario also covers other noise distributions like log-normal. The potentially time-dependent observable parameters like noise parameters σ, scalings, or offsets can sometimes be deduced from literature or the data itself. But generally, one wants to infer the full parameter vector, which is made up of model parameters and observable parameters

$$\theta = \begin{pmatrix} \theta^{\text{model}} \\ \theta^{\text{obs}} \end{pmatrix}.$$

2.3.2 Objective Function

The goal is to find θ, such that (5) is satisfied, or at least the left and right-hand sides are as close as possible. One could try to minimize the difference between observations $\bar{y}$ and model simulations h for measurement time points t_i. In particular, the L^2-difference is commonly used, which is known as *least squares* objective:

$$\hat{\theta} = \arg\min_{\theta} \sum_i \left\| \bar{y}(t_i) - \tilde{h}(t_i, x(t_i), \theta) \right\|_{L^2}^2. \tag{7}$$

Since, in general, this problem setup has no closed-form solution, one resolves to, often gradient-based, numerical optimization approaches. That is why the differentiability of the L^2-norm makes least squares advantageous compared to minimizing L^1-differences.
Before discussing the optimization approaches in detail in the next section, we will first have a look at a statistically motivated objective function called the *likelihood* $\mathcal{L}$, which is frequently employed in biological modeling [27]. Here, the idea is to find the parameters θ that maximize the likelihood of the observed data

$$\hat{\theta} = \arg\max_{\theta} \mathcal{L}(\theta) = \arg\max_{\theta} p(\bar{y}|\theta). \tag{8}$$

Under certain regularity conditions, one can show that maximum likelihood estimation is consistent and asymptotically efficient [50]. Consistency means, that assuming the data is generated by h for a large enough data set, the maximum likelihood estimate $\hat{\theta}$ will be arbitrarily close to the parameters used for generating the data. Logically, one of the conditions required for consistency is that the optimum $\hat{\theta}$ is unique, called *identifiability*. Asymptotic efficiency means that there exists no consistent estimator which has a lower asymptotic mean squared error.
In the common case of conditionally independent additive Gaussian noise, we can rewrite (6) as

$$h(t_i, x, \theta) \sim \mathcal{N}(\tilde{h}(t_i, x, \tilde{\theta}), \sigma_i^2)$$

and plug the multivariate normal density into (8) which leads to

$$\hat{\theta} = \arg\max_{\theta} \prod_i \prod_j \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} \exp\left(-\frac{\left(\bar{y}_j(t_i) - \tilde{h}_j\left(t_i, x(t_i), \tilde{\theta}\right)\right)^2}{2\sigma_{ij}^2} \right),$$

where i is the index for time points and j for observations. In the case of multiple experimental conditions, one would also multiply over the condition indices. Since sums are numerically more stable than products and the logarithm is invariant with respect to the minimum, due to its monotonicity, one usually optimizes the logarithmically transformed likelihood, the so-called *log-likelihood* $\mathcal{LL}$. Moreover, in many optimization packages, the default optimization is minimization. That is without loss of generality, since multiplication by -1 can convert every maximization problem to a minimization problem. Hence, we will minimize the negative log-likelihood

$$\begin{aligned}
\arg\min_{\theta} -\mathcal{LL}(\theta) &= \arg\min_{\theta} \sum_{ij} - \left(\log \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} - \frac{\left(\bar{y}_j(t_i) - \tilde{h}_j\left(t_i, x(t_i), \tilde{\theta}\right)\right)^2}{2\sigma_{ij}^2} \right) \\
&= \arg\min_{\theta} \sum_{ij} \left(\log \sigma_{ij} + \frac{\left(\bar{y}_j(t_i) - \tilde{h}_j\left(t_i, x(t_i), \tilde{\theta}\right)\right)^2}{2\sigma_{ij}^2} \right).
\end{aligned}$$

In the case where noise parameters σ are known and therefore not estimated, i.e., $\theta = \tilde{\theta}$, we can drop the constant $\log \sigma_{ij}$ from the sum and the negative log-likelihood reduces to a weighted least squares objective.

Another key likelihood objective used in this thesis is the *multinomial likelihood*, which is well-suited for modeling discrete distributions over continuous domains, such as densities over cell state bins. This approach was introduced in [22] and will be applied to the PDE models studied here. In these models, the system's state variable $x(t_i, \cdot)$ at each time point t_i describes a distribution of samples over the state space.

The motivation for using the multinomial likelihood arises from the fact that in solving the PDE with the FVM we are discretizing a continuous state space into finite control volumes, and for each time point, we model the distribution of samples across these volumes. The multinomial likelihood provides a natural way to represent the probabilities of samples falling into different control volumes, thus aligning with our need to capture the distribution of the state variable over time.

Consider the case of a 1-dimensional state space discretized into evenly distributed intervals. Without loss of generality, we can take the interval from 0 to 1:

$$S = [0, 1] = \bigcup_{k=0}^{n-1} \left[\frac{k}{n}, \frac{k+1}{n} \right] = \bigcup_k V_k, \tag{9}$$

where V_k represents the control volume for the k-th bin. As described in Section 2.2.2, the FVM computes the integrated averages $\tilde{x}_k$ of x over each V_k. By choosing, or rather rescaling an already chosen, observable mapping h such that the sum of these averages satisfies

$$\sum_k h(t, x_k, \theta) = 1, \tag{10}$$

we can treat these values as probabilities that a sample lies within each bin V_k.
Given m_i samples at each observed time point t_i, we can compute the histogram $\bar{y}_k(t_i)$ over bins V_k. These histogram values, together with the computed probabilities, can then be plugged into the *multinomial probability mass function* P_{mult} to obtain the likelihood of parameters θ:

$$\begin{aligned}\mathcal{L}(\theta) &= \prod_i P_{\text{mult}}\left(\bar{y}_0(t_i), \ldots, \bar{y}_{n-1}(t_i), m_i \mid h(t, x_0, \theta), \ldots, h(t, x_{n-1}, \theta)\right) \\ &= \prod_i \frac{m_i!}{\bar{y}_0(t_i)! \cdots \bar{y}_{n-1}(t_i)!} h(t, x_0, \theta)^{\bar{y}_0(t_i)} \cdots h(t, x_{n-1}, \theta)^{\bar{y}_{n-1}(t_i)}.\end{aligned}$$

In practice, we will again use the negative log-transformed likelihood for optimization:

$$\hat{\theta} = \arg\min_\theta - \sum_i \sum_k \bar{y}_k(t_i) \log h(t, x_k, \theta).$$

2.3.3 Gradient-Based Optimization

In this subsection, we will discuss how to numerically solve an optimization problem

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \mathcal{J}(\theta), \tag{11}$$

where $\mathcal{J}$ denotes a generic objective function and $\Theta \subset \mathbb{R}^{n_\theta}$ is the bounded set of feasible parameters, which could, e.g., enforce non-negativity for biological rates.
Only in straightforward cases is it feasible to calculate an analytical solution to an optimization problem (11); hence, we have to apply numerical optimization methods. A very straightforward method would be the Monte-Carlo-like approach of *random search* where one draws samples $(\theta^k)_k$ from a prior distribution, plugs them into $\mathcal{J}(\theta^k)$ and compares the values until convergence criteria are met. Here, the prior distribution can be informed, e.g., normally distributed around some literature value, or be uninformed, i.e., the uniform distribution on Θ. For many applications, it is difficult to find good prior distributions of the same scenario, and covering the entire parameter space Θ in the uninformed approach can be computationally very demanding or even infeasible. Hence, there are iterative *gradient-based* methods which make use of the shape of the objective function landscape $\mathcal{J}(\Theta) = \{\mathcal{J}(\theta) | \theta \in \Theta\}$ (Figure 1).
One idea of how to navigate through the objective function landscape towards a minimum is the method of *gradient descent* [51], where at each iteration, one takes the next step in the direction of the steepest descent (red arrows in Figure 1). To formalize this, one starts at a, usually random, initial point θ^0 in the parameter space Θ, calculates $\partial_\theta \mathcal{J}(\theta^0)$ and, since the gradient shows in the direction of the steepest ascend, sets $\theta^1 = \theta^0 - \tau \partial_\theta \mathcal{J}(\theta^0)$ for a step size $\tau > 0$, which is called *learning rate*, and repeats the process until some termination condition is satisfied (Algorithm 1). Usually, this condition ensures convergence, e.g., via $|\mathcal{J}(\theta_{k+1}) - \mathcal{J}(\theta_k)| < \epsilon$ or $\|\Delta\theta\| < \epsilon$, but the condition can also set an upper limit on the number of steps.
For Algorithm 1, we assume that $\mathcal{J}$ is differentiable with respect to θ, which holds for the objective functions presented in the previous section as long as the observable function h is differentiable with respect to θ. The ODE solution itself $x(t, \theta)$ is differentiable with respect to θ as long as $f(t, x, \theta)$ and $x_0(\theta)$ are sufficiently smooth, since from the IVP we

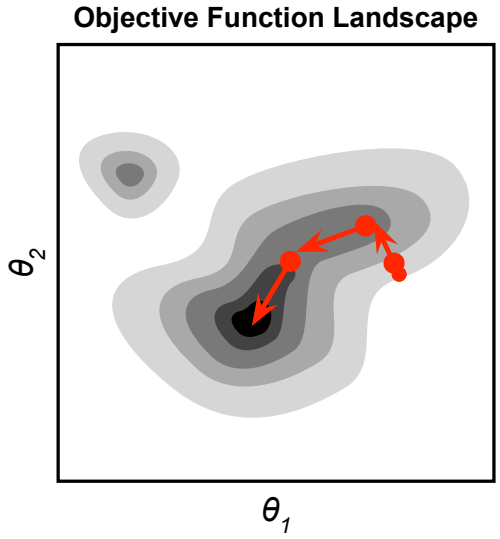


Figure 1: Objective function landscape $\mathcal{J}(\Theta)$ for a 2-dimensional parameter optimization problem. Function values highlighted by contour lines and steepest descent steps indicated by red arrows.

Algorithm 1 Gradient Descent

Require: $\mathcal{J}(\theta)$ differentiable, $\theta^0 \in \Theta$, $\tau > 0$
  while Termination condition not satisfied **do**
    Compute $\partial_\theta \mathcal{J}(\theta^k)$
    $\theta^{k+1} \leftarrow \theta^k - \tau \partial_\theta \mathcal{J}(\theta^k)$
  end while

obtain

$$\frac{d}{dt} x_\theta = f_\theta(t, x(t,\theta), \theta) + f_x(t, x(t,\theta), \theta) x_\theta(t,\theta)$$
$$x_\theta(0) = x_{0\theta},$$

where x_θ denotes the Jacobian of x with respect to θ. This equation has a unique solution following analog reasoning as in Section 2.2.1.
The proof of why subtracting the gradient points yields the steepest direction is commonly conducted via Taylor approximation, which states that

$$\mathcal{J}(\theta^k - \tau \Delta\theta) \approx \mathcal{J}(\theta^k) - \partial_\theta \mathcal{J}(\theta^k) \tau \Delta\theta \tag{12}$$

for sufficiently small τ and unit vectors $\Delta\theta$. Looking at (12), the unit vector minimizing the right-hand side is exactly $\partial_\theta \mathcal{J}(\theta^k)/\|\partial_\theta \mathcal{J}(\theta^k)\|$, which proofs the claim.
The choice of learning rate τ in Algorithm 1 is critical, and one can choose it differently at each iteration. More sophisticated algorithms compute an optimal learning rate via *line-search* where an additional minimization problem

$$\hat{\tau}_k = \arg\min_\tau \mathcal{J}(\theta^k - \tau \partial_\theta \mathcal{J}(\theta^k))$$

is solved at each step. Usually, the polynomial approximation (12) is employed, and optimal τ is found via backtracking, where one starts at rather large τ and iteratively reduces it until optimality criteria are met[52, 53].

Before we explore further enhancements of this method, we want to briefly point out that, unless we have the quite rare case of a convex J, one can end up at local minima or saddle points. To avoid this, one usually performs a *multistart* optimization, i.e., multiple optimization runs for different initial vectors θ^0. One can, e.g., sample these start points uniformly from Θ or employ more sophisticated strategies like Latin-hypercube-sampling, where the sample space is explored by enforcing that the samples are well-distributed across a pre-specified hypercube of the parameters. Convergence of these starts to the global minimum is assessed by comparing all optimization runs' final objective function values of all optimization runs and verifying that enough starts ended with the same minimal value.

Gradient descent already provides a reliable way to reach minima in the objective function landscape. However, it might not always take the shortest path downhill. To improve the direction in which to turn at each step, ideally, one does not only look at the steepness but also at the curvature of the landscape. This is precisely what *Newton's method* does. It can be motivated by the first-order Taylor approximation performed in (12). Since we now want to consider curvature, we will look at the second-order Taylor approximation, which is, in particular, the best local second-order polynomial approximation. Here, we will disregard the unit vector view we imposed above and look at general vectors $\Delta\theta$ sufficiently close to θ^k:

$$\mathcal{J}(\theta^k - \Delta\theta) \approx \mathcal{J}(\theta^k) - \partial_\theta \mathcal{J}(\theta^k)\Delta\theta + \frac{1}{2}(\Delta\theta)^T \partial_\theta^2 \mathcal{J}(\theta^k)\Delta\theta. \tag{13}$$

We can find the minimum of (13) by differentiating the right-hand side with respect to $\Delta\theta$ and setting it to 0. This implies $\Delta\theta = \partial_\theta \mathcal{J}(\theta^k)(\partial_\theta^2 \mathcal{J}(\theta^k))^{-1}$ and thus we obtain Newton's method, which is formally stated in Algorithm 2.

Algorithm 2 Newton's Method

Require: $\mathcal{J}(\theta)$ twice differentiable, $\theta^0 \in \Theta$, $\tau > 0$
  while Termination condition not satisfied **do**
    Compute $\partial_\theta \mathcal{J}(\theta^k)$
    $\theta^{k+1} \leftarrow \theta^k - \tau \partial_\theta \mathcal{J}(\theta^k)(\partial_\theta^2 \mathcal{J}(\theta^k))^{-1}$
  end while

Since computing the inverse of the Hessian $(\partial_\theta^2 \mathcal{J}(\theta^k))^{-1}$ can be computationally very demanding, in reality, *quasi-Newton methods* are often employed. In these methods the Hessian is approximated by solving the secant equation obtained from the Taylor approximation of the gradient of the objective function:

$$\partial_\theta \mathcal{J}(\theta^k - \Delta\theta) \approx \partial_\theta \mathcal{J}(\theta^k) - \partial_\theta^2 \mathcal{J}(\theta^k)\Delta\theta.$$

There exist algorithms that update this Hessian approximation efficiently along with the optimization steps, like *Broyden–Fletcher–Goldfarb–Shanno* (BFGS) and *Symmetric rank-one* (SR1), where the former is more robust due to its preservation of positive definiteness. At the same time, the latter can achieve higher accuracy and better convergence [54, 55]. There are versions for limited memory scenarios, L-BFGS and L-SR1, where only the most recent steps are saved and used for the next update.

There is also the, in some sense, dual to line-search approach of *trust regions* [56]. Here, the step size will be fixed after each step, and inside of this radius around the current parameter vector, the objective function will be approximated, e.g., quadratically, and

with this, the optimal direction for the next step is computed. If the approximation is not good enough in this radius, the step size will be decreased.
Moreover, for convex Θ, which is usually satisfied, e.g., if Θ is the product of closed intervals, there is also the family of *interior point* or *barrier* methods. We will briefly introduce them via one representative, the *primal-dual* method [57]. The original optimization problem (11) can be reformulated as

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^{n_\theta},\, c_j(\theta) \geq 0}{\arg\min} \mathcal{J}(\theta),$$

for suitable $c_j : \mathbb{R}^{n_\theta} \to \mathbb{R}$. Inspired by Lagrange multipliers, one now introduces an additional term, the barrier function

$$\mathcal{J}_\mu(\theta) = \mathcal{J}(\theta) - \mu \sum_j \log(c_j(\theta))$$

for a small $\mu > 0$. Now, one aims to find an optimal pair (θ, λ) such that

$$0 = \partial_\theta \mathcal{J}_\mu(\theta) = \partial_\theta \mathcal{J}(\theta) - \mu \sum_j \frac{\partial_\theta c_j(\theta)}{c_j(\theta)}$$
$$c_j(\theta)\lambda_j = \mu.$$

Again, this optimum is computed iteratively, and the concepts introduced above, like line search and the Newton method, can be applied.
A final aspect essential to this section is the calculation of the gradients of the objective function. Unless one has x as a closed-form solution, calculating the gradients of h, and hence $\mathcal{J}$, with respect to θ, is not feasible. Therefore, numeric gradient calculation methods are usually employed, like the rather straightforward approach of *finite differences*. It is motivated by the definition of the partial derivatives

$$\partial_k \mathcal{J}(\theta) = \lim_{\epsilon \to 0} \frac{\mathcal{J}(\theta + \epsilon e_k) - \mathcal{J}(\theta)}{\epsilon},$$

where e_k is the k-th unit vector. Finite difference approximates this by taking a fixed ϵ_0 and dropping the limit on the right-hand side of the equation. For $\epsilon_0 > 0$, this is called *forward* finite difference, and for $\epsilon_0 < 0$, it is called *backward* finite difference. However, due to better error control [25], one usually employs the *central* finite difference approximation for $\epsilon_0 > 0$:

$$\partial_k \mathcal{J}(\theta) \approx \frac{\mathcal{J}(\theta + \frac{\epsilon}{2} e_k) - \mathcal{J}(\theta - \frac{\epsilon}{2} e_k)}{\epsilon}.$$

A more accurate approach is making use of the original ODE problem (4), which is differentiated with respect to θ to obtain the *forward sensitivity equations*

$$\frac{d}{dt} x_\theta(t, \theta) = f_\theta(t, x, \theta) + f_x(t, x, \theta) x_\theta(t, \theta),$$

where again x_θ denotes the Jacobian of x with respect to θ, which is also called sensitivities. Since one can usually calculate the Jacobians of f with respect to θ, f_θ, and with respect to x, f_x, analytically and x will be obtained by solving the original ODE anyways, the computational demand of solving sensitivity equations comes down to solving an $n_\theta n_x$-dimensional ODE system. Compared to finite differences, this will be more

expensive. Still, since the gradients will induce no additional error on top of the ODE solving inaccuracy, one assumes that, in the end, the gradient-based optimization will require fewer steps. Hence, overall computational demand will be lower.
Since the dependency of J on h and h on x is usually not overly complex, one can use x_θ to compute the derivatives of the objective function easily. If we have, for example, a negative log-likelihood with additive Gaussian noise, whose variance we estimated a priori from the data, as the objective function, we obtain

$$\partial_k \mathcal{J} = -\sum_{ij} \frac{\bar{y}_j(t_i) - \tilde{h}_j\left(t_i, x(t_i), \theta\right)}{\sigma_{ij}^2} \Big(\partial_{\theta_k} \tilde{h}_j\left(t_i, x(t_i), \theta\right) + \partial_x \tilde{h}_j\left(t_i, x(t_i), \theta\right) x_{\theta_k}(t_i) \Big).$$

In this thesis, we use the interior point method implemented in IPOPT [58] for the parameter estimation problems involving PDEs, where gradients are approximated via finite differences. For the purely ODE problems, we employ the trust region optimizer FIDES [59], which uses a hybrid of BFGS and SR1, along with forward sensitivities provided by AMICI/CVODES. We call both optimizers through the Python interface of pyPESTO [7].

2.4 Uncertainty Analysis via Sampling

After finding an optimal parameter vector $\hat{\theta}$ via gradient-based multistart optimization, the question is how reliable these parameters are and how sensitive the model is towards small deviations from the optimum [60]. To achieve this, we consider an ensemble of parameter vectors instead of a single one. Since we only want to compare relevant parameters, in the sense that they are optimal or close to it, such an ensemble $\mathcal{E}$ should satisfy

$$\mathcal{E}(\hat{\theta}) \subset \{\theta \in \Theta | \mathcal{J}(\theta) \approx \mathcal{J}(\hat{\theta})\}. \tag{14}$$

A straightforward approach is to define the ensemble as the top K results of the multistart optimization with N starts, where $K << N$, and it should be assured that (14) holds.
However, since optimization runs are usually computationally very demanding and the initial points are chosen randomly, it might prove difficult to provide a thorough uncertainty analysis purely through this ensemble method. A variety of techniques have been developed to address this issue. For this thesis, we focus on sampling via Markov Chain Monte Carlo, which builds on the Bayesian inference theory, which we will briefly introduce.

2.4.1 Bayesian inference

In contrast to the optimization theory we have seen above, sometimes labeled *frequentist inference*, where we end up with one optimal vector θ, the so-called *point estimate*, *Bayesian inference* aims at obtaining a whole distribution of parameters, to account for the inherent stochasticity of looking only at one particular data set with finite sample size [61]. As the name implies, Bayesian inference grounds in Bayes' Theorem

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}, \tag{15}$$

where $p(\theta|\mathcal{D})$ is called the *posterior* distribution and $p(\theta)$ the *prior* distribution. From the frequentist theory, we recognize the likelihood of our data $p(\mathcal{D}|\theta)$, where it was one of the possible objective functions. The probability $p(\mathcal{D})$ of the data $\mathcal{D}$ is in general unknown, but since it is constant, we can circumvent it as we will show for the following approaches. The Bayesian counterpart to maximum likelihood estimation is the *maximum-a-posteriori estimation* (MAP). It extends the maximum likelihood by including the prior distribution via

$$\hat{\theta} = \arg\max_{\theta\in\Theta} p(\theta|\mathcal{D}) = \arg\max_{\theta\in\Theta} p(\mathcal{D}|\theta)p(\theta) = \arg\min_{\theta\in\Theta} -(\log p(\mathcal{D}|\theta) + \log p(\theta)).$$

We can see immediately that this is equivalent to minimizing the negative log-likelihood with an additional penalty term introduced by the prior distribution. In the case of uniformly distributed prior information, $-\log p(\theta)$ becomes 0 if θ is inside of the bounds and ∞ if θ is not in the bounds, effectively resulting in the frequentist optimization of the log-likelihood within parameter bounds. For independent Gaussian priors with mean vector μ and standard deviations σ_j we have

$$\begin{aligned} -\log p(\theta) &= -\log \prod_j \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{(\theta_j-\mu_j)^2}{2\sigma_j^2}\right) \\ &= \sum_j \left(\frac{1}{2}\log(2\pi\sigma_j^2) + \frac{(\theta_j-\mu_j)^2}{2\sigma_j^2}\right), \end{aligned}$$

which, after dropping the constant first term, effectively becomes an L^2-penalty term scaled by two times the standard deviations. In particular, if we normalize to zero-mean and standard deviation to 1, we obtain L^2-regularization, also known as ridge regularization. Similarly, one can show that Laplacian priors effectively result in an L^1 penalty term and, in the normalized case, L^1-regularization, also known as LASSO [62].

2.4.2 Markov Chain Monte Carlo

Although based on the Bayesian framework, MAP yields a point estimate and not the full posterior distribution $p(\cdot|\mathcal{D})$. To obtain the posterior distribution computationally, we employ *Markov Chain Monte Carlo* (MCMC) methods, which construct Markov chains so that their equilibrium distribution yields the true posterior distribution [63]. In particular, for a finite number of steps, the chain elements approximate the equilibrium distribution. Since the distribution of data, $p(\mathcal{D})$, in Bayes Theorem (15) is unknown, we make use of the *Metropolis-Hastings algorithm*. This algorithm requires that we know a distribution proportional to the desired distribution. Since the distribution of data is independent of parameters, we see in Bayes Theorem that we can use $p(\mathcal{D}|\theta)p(\theta)$ as a known proportional distribution of the desired posterior $p(\theta|\mathcal{D})$. Furthermore, Metropolis-Hastings requires a symmetric proposal distribution, $g(x|y)$, often considered a Gaussian distribution of around y. The chain is initialized at θ^0, and the proposed next step θ' is drawn from $g(\cdot|\theta^0)$. Then, the acceptance ratio

$$\alpha = \frac{p(\mathcal{D}|\theta')p(\theta')}{p(\mathcal{D}|\theta^0)p(\theta^0)}$$

is computed and a number u drawn from $\mathcal{U}([0,1])$. If $u \leq \alpha$ we accept θ' and set $\theta^1 = \theta$, and, if $u > \alpha$, we reject the proposal and set $\theta^1 = \theta^0$. These steps are repeated until we

reach the equilibrium distribution. The process is summarized in Algorithm 3. There are extensions to this algorithm, in particular choosing a proposal distribution, which adapts at each step based on previous samples, the *adaptive Metropolis-Hastings algorithm* [64].

Algorithm 3 Metropolis-Hastings

Require: Symmetric proposal distribution $g(x|y)$ and initial $\theta^0 \in \Theta$.
  while Chain $(\theta^k)_k$ not converged **do**
    Draw θ' from $g(\cdot|\theta^{k-1})$.
    Compute $\alpha = \frac{p(\mathcal{D}|\theta')p(\theta')}{p(\mathcal{D}|\theta^k)p(\theta^k)}$.
    Draw u from $\mathcal{U}$.
    if $\alpha \leq u$ **then** Set $\theta^k = \theta'$
    else Set $\theta^k = \theta^{k-1}$.
    end if
  end while

In our cases, where sampling is feasible, we combine frequentist (or MAP) and Bayesian approaches by first performing a (regularized) gradient-based optimization and using its result as the starting point for a MCMC chain. In this sense, MCMC is used as an uncertainty quantification measure since we obtain distributions around already optimized parameters. However, one can generally perform a global parameter optimization by computing multiple MCMC chains starting at randomly selected initialization points.

3 Integrating an Antibody Study into SEIR-Modeling to predict SARS-CoV-2 Spread in Ethiopia

Compartmental modeling is crucial in epidemiological modeling for predicting and analyzing the spread of infectious diseases [5]. In particular, for SARS-CoV-2, Susceptible-Exposed-Infectious-Recovered (SEIR) models have shown to be reliable. This section provides an overview of the study titled *Seroepidemiology and model-based prediction of SARS-CoV-2 in Ethiopia: longitudinal cohort study among frontline hospital workers and communities* [1]. The study conducted an antibody prevalence survey on SARS-CoV-2 in Ethiopia. We integrated the data into three compartmental population models, described by ODEs: a standard SEIR model, an extended SEIR model, accounting for two populations, and a SEIR-based model accounting for an additional virus variant. Serological data from multiple testing rounds of frontline healthcare workers (HCWs) and communities in Jimma and Addis Ababa were integrated into the different models. The novelty of this work lies in the application and extension of SEIR models to scenarios of underreporting and their specific adaptation to healthcare workers in Ethiopia, addressing critical gaps in data and disease dynamics in this context. The co-authors performed the antibody study and preliminary data analysis, while the thesis author conducted the modeling and parameter estimation and provided the corresponding descriptions and figures for the original manuscript.
In this summary, we will concentrate on the modeling and parameter estimation. The original publication is reprinted in Appendix A. It relies mainly on the mathematical frameworks introduced in Sections 2.1.5, 2.2.1, 2.3, and 2.4.

3.1 Data

The study sampled data from three rounds of SARS-CoV-2 anti-nucleocapsid antibody surveys conducted between August 2020 and April 2021 involving 1,104 HCWs and 1,229 residents from Addis Ababa and Jimma, alongside national positivity rates from the same period. The data was aggregated by collection site, participant group, sampling round, and, for multi-month rounds, by month. Monthly national test positivity rates were also aggregated. Errors were estimated by sampling from a binomial distribution and later used in model fitting.

3.2 Models

SEIR Model The basic SEIR model divides the population into four compartments:

- Susceptible (S): Individuals who can contract the disease.
- Exposed (E): Individuals who have been infected but are not yet infectious.
- Infectious (I): Individuals who can transmit the disease.
- Recovered (R): Individuals who have recovered and are immune.

The transitions between these compartments are governed by the following ordinary differential equation (ODE) system:

$$\begin{aligned}
\frac{dS}{dt} &= -\beta \frac{I}{N} S \\
\frac{dE}{dt} &= \beta \frac{I}{N} S - \kappa E \\
\frac{dI}{dt} &= \kappa E - \gamma I \\
\frac{dR}{dt} &= \gamma I,
\end{aligned}$$

where

- β is the transmission rate.
- κ is the rate at which exposed individuals become infectious.
- γ is the recovery rate.
- $N = S + E + I + R$ is the total population.

Furthermore, the initial time $t = 0$ was set to the date when SARS-CoV-2 was first observed in Ethiopia, as reported by the WHO. The initial number of susceptible individuals $S(0)$ was set to 510, approximately reflecting the number of participants in each round and site. The initial numbers of exposed and recovered individuals were both set to zero. The initial number of infected individuals was treated as a site-specific parameter, which we later estimated. This approach, combined with the memorylessness of the exponential transition times, i.e., the Markov property of the underlying stochastic process, allowed the model to account for different entry times for each site and from the national entry date. A later entry point is equivalent to an upscaled initial number of infected individuals.

Extended SEIR Model The extended SEIR model addresses interactions between HCWs (denoted by index H) and the community (denoted by index C), incorporating a potentially higher transmission probability from community members to HCWs, modeled by a factor α. The ODEs incorporate inter-population interactions:

$$\begin{aligned}
\frac{dS_H}{dt} &= -\beta \frac{I_H + \alpha I_C}{N} S_H \\
\frac{dE_H}{dt} &= \beta \frac{I_H}{N} S_H - \kappa E_H \\
\frac{dI_H}{dt} &= \kappa E_H - \gamma I_H \\
\frac{dR_H}{dt} &= \gamma I_H \\
\frac{dS_C}{dt} &= -\beta \frac{I_H + I_C}{N} S_C \\
\frac{dE_C}{dt} &= \beta \frac{I_C}{N} S_C - \kappa E_C
\end{aligned}$$

$$
\begin{aligned}
\frac{dI_C}{dt} &= \kappa E_C - \gamma I_C \\
\frac{dR_C}{dt} &= \gamma I_C \\
N &= S_H + E_H + I_H + R_H + S_C + E_C + I_C + R_C
\end{aligned}
$$

Here, we kept 510 as the initial susceptible HCW and set $S_C(0)$ to 100,000 to reflect a realistic ratio of HCW to community members. The entrance of the virus was assumed to happen in the community, $I_C(0) = I_0$, and all other initial values were set to 0.

Virus Variant Model This model considers a new virus strain with increased transmissibility. Key assumptions are:

- Increased reproduction rate by a literature-derived factor of 1.35.
- Previous infections with the variant confer immunity to the wildtype but not vice versa.

This yields the following ODE system:

$$
\begin{aligned}
\frac{dS}{dt} &= -\beta \frac{I_{wt}}{N} S - \beta \frac{I_{va} + I_{va,wt}}{N} S \\
\frac{dE_{wt}}{dt} &= \beta \frac{I_{wt}}{N} S - \kappa E_{wt} \\
\frac{dE_{va}}{dt} &= \beta \frac{I_{va} + I_{va,wt}}{N} S - \kappa E_{va} \\
\frac{dE_{va,wt}}{dt} &= \beta \frac{I_{va} + I_{va,wt}}{N} R_{wt} - \kappa E_{va,wt} \\
\frac{dI_{wt}}{dt} &= \kappa E_{wt} - \gamma I_{wt} \\
\frac{dI_{va}}{dt} &= \kappa E_{va} - \frac{\gamma}{1.35} I_{va} \\
\frac{dI_{va,wt}}{dt} &= \kappa E_{va,wt} - \frac{\gamma}{1.35} I_{va,wt} \\
\frac{dR_{wt}}{dt} &= \gamma I_{wt} - \beta \frac{I_{va} + I_{va,wt}}{N} R_{wt} \\
\frac{dR_{va}}{dt} &= \frac{\gamma}{1.35} I_{va} \\
\frac{dR_{va,wt}}{dt} &= \frac{\gamma}{1.35} I_{va,wt}
\end{aligned}
$$

with initial values $S(0) = 510$, $I_{wt}(0) = I_0$ and $I_{va}(t_0) = 1$ and all others set to 0. As for I_0, we allowed t_0 to be site-specific and estimated it.

3.3 Parameter Estimation

The SEIR and extended SEIR models were calibrated using data from the first two rounds of the antibody study. The SEIR model was fitted separately to the data from HCW and community members, while the extended SEIR model simultaneously integrated both datasets. To enhance statistical power while maintaining precision, samples were aggregated by month. The model fitting incorporated Addis Ababa and Jimma data with

site-specific initial values. Antibody prevalence was mapped to the recovered fraction R/N. The third round of data was subsequently used to validate the model. Although data on virus variants were unavailable, national test positivity rates were used to infer variant dynamics by mapping them to the fraction of the combined infectious population. All three sampling rounds of community members were utilized to calibrate the virus variant model, which was not used for predictions but highlighted the need for further research. For estimation, we assumed additive Gaussian noise as an approximation of the binomial model and included knowledge from literature about incubation and recovery as priors. The models, the data, and the parameter estimation setup were saved in the standardized parameter estimation format PEtab, to whose development the author of this thesis contributed. Parameter values were derived using AMICI with CVODE backend for simulations and pyPESTO for optimization and sampling. To the latter's development the author of this thesis also contributed. A point estimate obtained with FIDES was then used as the starting point for MCMC sampling via pyPESTO's adaptive Metropolis-Hastings implementation.

3.4 Key Insights

Model predictions revealed differences in seroprevalence between HCWs and the community. Due to community interactions, the extended SEIR model indicated a higher exposure risk for HCWs. The models without a variant predicted nearing herd immunity post-study, while the variant model suggested a continuing rise in antibody prevalence. The data analysis indicated significant underreporting in Ethiopia, with most people encountering SARS-CoV-2.

4 Multivariant and antibody level models of antibody data, variant sequences and vaccination of SARS-CoV-2 in Ethiopia

Building upon the study presented in the previous section, we extended the SEIR framework to a more complex multivariant model that captures the dynamics and cross-immunities of different SARS-CoV-2 variants over time: *Long-term monitoring of SARS-CoV-2 seroprevalence and variants in Ethiopia provides prediction for immunity and cross-immunity* [2]. Moreover, this study also introduces an antibody-level model, which provides a detailed understanding of how the levels of anti-nucleocapsid antibodies (Anti-N) and anti-spike antibodies (Anti-S) develop and decline across the population. Notably, the development of large-scale COVID-19 models and models including sequencing of viral variants had not been conducted for Ethiopia before, making this work a significant advancement. This research deepens our understanding of variant-specific immunity by incorporating a longitudinal antibody dataset spanning two years, including viral sequencing and national test positivity rates. It evaluates the effectiveness of public health interventions and vaccination strategies in a resource-limited setting like Ethiopia. The author of this thesis led the modeling and parameter estimation for both models, which are core to the study, as well as clustering analysis and error estimates of the data and drafted the related sections, introduction, and discussion of the original manuscript. The original publication is reprinted in Appendix B. It relies mainly on the mathematical frameworks introduced in Sections 2.1.5, 2.2.1, 2.3, and 2.4.

4.1 Data

The antibody surveys derived from serological surveys conducted in Addis Ababa and Jimma were extended by two additional rounds, enhancing the original data set by 3,384 new samples collected between August 2021 and May 2022. For the original and the latest samples, both Anti-N and Anti-S data were now available. Moreover, study participants provided information on their vaccination status and vaccination dates. We categorized the antibody data using k-means clustering. This clustering was applied to positive antibody data, resulting in the categories of medium and high levels, assumed to represent one infection or vaccination and multiple infections or vaccinations, respectively. Low antibody levels, corresponding to responses below the detection cutoff, were classified as negative.

Additionally, we sequenced 1,873 positive PCR tests, yielding 574 sequences of sufficient quality. We aggregated the resulting variant strains over their mutations of concern (MOC) into eight lineage groups and computed mutational distances between these variants as Hamming distance of MOC.

As in the previous study, we also utilized the nationally reported test positivity rates of our sampling period.

4.2 Models

Multivariant Model The SEIR model serves as the foundation for describing the transmission dynamics of SARS-CoV-2, with compartments representing the different stages of infection: Susceptible (S), Exposed (E), Infectious (I), and Recovered (R). The transitions between these compartments are governed by ODEs. The study extends this

basic model to a multivariant framework that tracks the wildtype virus and different variants of concern (VOC) over time. This model accounts for multiple sequences of infections and vaccinations, allowing for a detailed analysis of variant-specific immunity and cross-immunity. Moreover, it incorporates the impact of vaccination, which, based on the vaccines available in Ethiopia, is handled as being recovered from wildtype. The dynamics of a first infection with variant $i = 1, \ldots, 8$ are described by the following system:

$$\begin{aligned}
\dot{S} &= -\frac{\beta_i \hat{I}_i S}{N} - v_1 S \\
\dot{E}_i &= \frac{\beta_i \hat{I}_i S}{N} - \kappa E_i \\
\dot{I}_i &= \kappa E_i - \gamma I_i \\
\dot{R}_i &= \gamma I_i - \sum_{j \in P_i} \frac{\beta_{ij} \hat{I}_j R_i}{N} - v_1 R_i \\
\dot{R}_v &= v_1 S - \sum_{j=1,\ldots,8} \frac{\beta_j \hat{I}_j R_v}{N} - v_2 R_v
\end{aligned}$$

where:

- the transmission rate associated with variant i, is denoted by β_i if there was no previous infection and β_{ji} after recovery from variant j,
- κ is the rate at which exposed individuals become infectious,
- γ is the recovery rate,
- v_k denote the vaccination rates for the k-th vaccination.
- P_i is the set of potential reinfections, which we reduced to pathways reflecting the worldwide disease dynamics, e.g., by excluding wildtype infections after omicron infections,
- N is the total population,
- and $\hat{I}_j$ the sum of all currently infected with variant j.

We set the initial number of susceptible $S(0)$ to $120.3e6$, roughly reflecting the total population of Ethiopia at that time. Initial appearances of variants were implemented by $I_i(t_{0i} = 1)$, and all other initial values were set to zero.
The multivariant model refines the above equations by incorporating compartments for second infections and vaccinations. For $i = 1, \ldots, 8, v$ (numbers for infections, v for vaccination) and $j = 1, \ldots, 8$ we have

$$\begin{aligned}
\dot{E}_{ij} &= \frac{\beta_{ij} \hat{I}_j R_i}{N} - \kappa E_{ij} \\
\dot{I}_{ij} &= \kappa E_{ij} - \gamma I_{ij}
\end{aligned}$$

$$\dot{R}_{ij} = \gamma I_{ij} - \sum_{k=7,8} \frac{\beta_{ijk}\hat{I}_k R_{ij}}{N} - v_{n(i,j)+1} R_{ij}$$

$$\dot{R}_{iv} = v_{n(i,v)} R_i - \sum_{k=7,8} \frac{\beta_{ivk}\hat{I}_k R_{iv}}{N} - v_{n(i,v)+1} R_{iv},$$

where $n(\mathbf{Idx}) := \#\{v \in \mathbf{Idx}\}$. For $i, j = 1, \ldots, 8, v$ and $k = 7, 8$ we obtain the third infection or vaccination equations

$$\begin{aligned}
\dot{E}_{ijk} &= \frac{\beta_{ijk}\hat{I}_k R_{ij}}{N} - \kappa E_{ijk} \\
\dot{I}_{ijk} &= \kappa E_{ijk} - \gamma I_{ijk} \\
\dot{R}_{ijk} &= \gamma I_{ijk} - \frac{\beta_{ijk8}\hat{I}_8 R_{ijk}}{N} - v_{n(i,j,k)+1} R_{ijk}.
\end{aligned}$$

Furthermore, for highly immune evasive Omicron BA.4/5 variant we also implemented a fourth infection, i.e. for $i, j = 1, \ldots, 8, v$ and $k = 7, 8, v$:

$$\begin{aligned}
\dot{E}_{ijk8} &= \frac{\beta_{ijk8}\hat{I}_8 R_{ijk}}{N} - \kappa E_{ijk8} \\
\dot{I}_{ijk8} &= \kappa E_{ijk8} - \gamma I_{ijk8} \\
\dot{R}_{ijk8} &= \gamma I_{ijk8}.
\end{aligned}$$

The effective infection rates $\beta_{\mathbf{Idx}}$ are split into three parts

$$\beta_{\mathbf{Idx}} = s_{\text{seas}} s_{\text{reinf}}(\mathbf{Idx}) \hat{\beta}_{\mathbf{Idx}[-1]},$$

the seasonality factor s_{seas}, the reinfection factor s_{reinf} and the transmission rate $\hat{\beta}_{\mathbf{Idx}[-1]}$ of the currently encountered variant $\mathbf{Idx}[-1]$, i.e. variant corresponding to last index entry of $\mathbf{Idx}$.
The reinfection factor depends on the previously encountered variants encoded in all but the last index entries $\mathbf{Idx}[:-1]$ and the currently encountered variant encoded in the last index entry $\mathbf{Idx}[-1]$ and is formulated as follows

$$s_{\text{reinf}}(\mathbf{Idx}) = \begin{cases} 1, & \text{if } |\mathbf{Idx}| = 1 \\ (1 - s_0)(1 - s)^{d(\mathbf{Idx}[:-1], \mathbf{Idx}[-1])}, & \text{otherwise.} \end{cases}$$

Here $d(x, y)$ is the Hamming distance between MOC observed in variant y and MOC observed in variant or combination of variants x.

Antibody-Level Model The antibody-level model described the distribution of individuals with a certain combination of Anti-S and Anti-N levels. For each antibody type, we consider three discrete categories, with index $i = 0$ (low), 1 (medium), 2 (high) being used for Anti-S categories and index $j = 0$ (low), 1 (medium), 2 (high) being used for Anti-N categories. The time evolution of individuals in each category, A_{ij}, is governed by the following equations for $i, j = 0, 1, 2$:

$$
\begin{aligned}
\dot{A}_{ij} =& -\frac{\beta_{ij}\hat{I}A_{ij}}{N} - vA_{ij}\chi_{i\leq 1} \\
&+ \gamma\left(I_{i,j-1}\chi_{i=2} + I_{i-1,j}\chi_{j=2} + I_{i,j}\chi_{i=2}\chi_{j=2} + (1-\theta)^{\chi_{j=1}} I_{i-1,j-1}\right. \\
&\qquad \left. +\theta I_{i-1,j-2}\chi_{j=2}\right)\chi_{i\geq 1}\chi_{j\geq 1} \\
&+ \delta_N A_{i+1,j}\chi_{i\leq 1} + \delta_S A_{i,j+1}\chi_{j\leq 1} + \delta_{\mathrm{SN}} A_{i+1,j+1}\chi_{i\leq 1}\chi_{j\leq 1} \\
&+ vA_{i-1,j}\chi_{i\geq 1} \\
&- \left(\delta_N\chi_{i\geq 1} + \delta_S\chi_{j\geq 1} + \delta_{\mathrm{SN}}\chi_{i\geq 1}\chi_{j\geq 1}\right) A_{ij} \\
\dot{E}_{ij} =& \frac{\beta_{ij}\hat{I}A_{ij}}{N} - \kappa E_{ij} \\
\dot{I}_{ij} =& \kappa E_{ij} - \gamma I_{ij},
\end{aligned}
$$

where

- β_{ij} represent antibody-level dependent exposure rates
- χ is the indicator function,
- v represents vaccination rate,
- γ and κ account for recovery and incubation,
- θ is the fraction of population experiencing a boosting effect of Anti-N levels after recovery.

We set the initial values to

$$
\begin{aligned}
A_{ij}(0) &= \begin{cases} 120.3\text{e}6 & \text{if } i=j=0 \\ 0 & \text{otherwise} \end{cases} \\
E_{ij}(0) &= 0 \\
I_{ij}(t_0) &= \begin{cases} 1 & \text{if } i=j=0 \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}
$$

Moreover, the effective transmission rates β_{ij} are defined as

$$
\beta_{ij} = s_{\mathrm{seas}}(1-s_1)^{\chi_{i\geq 1 \text{ or } j\geq 1}}(1-s_2)^{\chi_{i=2 \text{ or } j=2}} \sum_{k=1}^{8} \alpha_k \hat{\beta}_k,
$$

with immunity factors s_1 and s_2, variant distributions $\alpha_k(t)$ and variants' transmission rates $\hat{\beta}_k$.

4.3 Parameter Estimation

The model parameters were estimated using the comprehensive dataset of antibody levels, virus variant distributions, and national test positivity rates. Since sequencing data was only available for community members, we used the healthcare worker antibody data to verify our clustering approach conceptually. We performed the estimations with the antibody data for community members.

For the multivariant model, only Anti-S levels were utilized to manage computational feasibility. The antibody model incorporates the variant distributions via weighting factors for each exposure rate. The weights are computed as normalized Gaussian fits to the distributions a priori to the estimation, while the transmission rates are later estimated along with the other model parameters. Vaccination rates were calculated in advance based on participants' responses to the antibody survey.
As in the previous study, antibody levels were mapped to the fraction of the corresponding recovered population relative to the total population. Errors for all data types used in estimation were inferred using multinomial and binomial models implemented in PyMC3 [65]. The models, the data, and the parameter estimation setup were saved in the standardized parameter estimation format PEtab, to whose development the author of this thesis contributed. Estimation was performed with additive Gaussian noise as an approximation to these error models and implemented in the Python frameworks AMICI and pyPESTO, with CVODE and FIDES providing the simulation and estimation backends. The author of this thesis also contributed to the development of pyPESTO.

4.4 Key Insights

The multivariant model provides insights into the infection history in Ethiopia, revealing that most individuals experienced multiple exposures to different variants. The model identified three major infection waves corresponding to the Wildtype, Delta, and Omicron BA.4/5 variants. It also highlights the role of cross-immunity in reducing the risk of reinfection, with reinfection risks varying based on the genetic distance, represented as a difference in mutations of concern (MOC) between variants.
The antibody-level model predicted that early-on widespread vaccination might have significantly mitigated delta and omicron waves. However, following the Omicron wave, it predicts up to 100% of the population to be in the high antibody category for both Anti-N and Anti-S, suggesting further vaccinations might have a limited impact on overall immunity, given the already high levels of antibody saturation.

5 PDE Modeling of Ligand Feedback in Immune Cell Activation of Mural Dendritic Cells

This section provides an overview of the mathematical modeling of cell-to-cell communication within the framework of Waddington's developmental potential landscape, as presented in the paper *A Dynamic Model for Waddington's Landscape Accounting for Cell-to-Cell Communication* [3]. The study extends the classical mathematical descriptions of Waddington's landscape, traditionally used to describe cell differentiation, by introducing a coupled system of partial and ordinary differential equations (PDE-ODE) that accounts for the dynamics of cell populations and ligand concentrations. This enhanced model provides a more accurate depiction of cellular processes, including the effects of cell-to-cell communication on developmental pathways. The author of this thesis contributed to the formulation of the mathematical model and formal analysis of existence and uniqueness. Moreover, he implemented the numerical simulation and the parameter estimation for the immune cell activation application and drafted the related sections, introduction, and discussion of the original manuscript. The original preprint is provided in Appendix C. It relies mainly on the mathematical frameworks introduced in Sections 2.1.1, 2.1.2, 2.1.3, 2.1.4, 2.1.6, 2.2.2 and 2.3.

5.1 Mathematical Model

The model describes a population of cells communicating through ligands, a key aspect of biological systems. The state of a cell is denoted by $s(t) \in \mathbb{R}^{n_s}$, and the ligand concentration by $l(t) \in \mathbb{R}^{n_l}_{\geq 0}$. The dynamics of the cell state are governed by drift $v(s,l,t)$, which corresponds to a directed change of state, and diffusion $D^{1/2}(s,l,t)$, which corresponds to random changes of the state. Together, this is denoted by the stochastic differential equation (SDE)

$$ds = v(s,l,t)dt + D^{1/2}(s,l,t)dB_t.$$

From this, we obtain in the limit of large cell numbers a PDE, the population balance model, which captures the time- and state-dependent number density function $u(s,t)$:

$$\frac{\partial u(s,t)}{\partial t} = \frac{\partial}{\partial s}\left(D(s,l,t)\frac{\partial u(s,t)}{\partial s}\right) - \frac{\partial}{\partial s}\left(v(s,l,t)u(s,t)\right) + g(s,l,t)u(s,t),$$

where $g(s,l,t)$ represents the effective proliferation rate. The initial and boundary conditions are specified to ensure the model captures the biological reality of cell dynamics and assumes that all possible cell states are observable, i.e., a nonnegative initial condition and no-flux boundary conditions.
The ligand dynamics are governed by the following ODE:

$$\frac{dl(t)}{dt} = \int_\Omega \alpha(s,t)u(s,t)\,ds - \left(\int_\Omega \beta(s,t)u(s,t)\,ds\right) l(t) - \gamma(t)l(t)$$

Here, $\alpha(s,t)$ denotes the ligand secretion rate, $\beta(s,t)$ the binding rate, and $\gamma(t)$ the degradation rate of the ligand. The coupling of these equations enables the model to dynamically represent the interaction between cells in different states via ligand-mediated communication.

5.2 Mathematical Analysis of Model

The proof of existence and uniqueness provides the theoretical foundation for the coupled PDE-ODE model describing cell population dynamics and ligand-mediated cell-to-cell communication. The key idea is to demonstrate that, under appropriate conditions, the system admits a unique global solution. The proof proceeds in several steps. First, it shows that the ligand concentration, governed by the first-order ODE, has a unique solution for a fixed cell density distribution. Next, it establishes that for a given ligand concentration, the cell population dynamics, represented by the parabolic PDE, also have a unique weak solution using energy estimates and Galerkin approximations. Finally, it is proven that the mapping from cell population to ligand concentration and back to cell population forms a contraction, which, with the Banach Fixed Point Theorem, ensures the existence of a unique local solution for the entire system. This result holds globally by extending the local solution over any finite time interval. The proof relies on regularity, boundedness, and Lipschitz continuity in model parameters, ensuring the applicability of standard PDE and ODE theory.

5.3 Data of Application Study

We applied the model to single cell RNA-seq data of dendritic cell activation after stimulation with lipopolysaccharide from Shalek et al. [66]. They investigate dendritic cell behavior under two experimental conditions: with communication (in-tube) and without communication (on-chip). The in-tube setup, where cells communicated through a ligand (IFN-β), showed progressive cell activation over time, while the on-chip condition, lacking communication, resulted in limited activation. We confirmed this by performing a UMAP on the single cell data, which produced two visually distinct clusters: one consisting of on-chip and early in-tube cells, the inactivated cluster, and one consisting of the later in-tube cells, the activated cluster. Then, we employed trajectory inference via diffusion pseudotime for each measurement time point and condition to obtain a 1-dimensional representation of their data.

5.4 Parameter Estimation

The parameter estimation process ensures that the mathematical model accurately reflects biological processes. The following steps were undertaken to estimate the model parameters in this study.

Parametric Functions To formulate a well-posed inverse problem, the coefficient functions of the equations must be parameterized appropriately. Given the short duration of the experiment (6 hours), cell growth was assumed to be negligible. It was assumed that the coefficients do not inherently change over time for the remaining cell dynamics—drift and diffusion—and the ligand dynamics. Any time-dependent changes in these coefficients are solely induced by variations in ligand concentrations. The drift and diffusion terms were parameterized using splines to capture the baseline dynamics, with additional ligand- and space-dependent components described via Hill functions. Ligand secretion and binding rates were modeled as Hill functions, while ligand degradation, independent of time, required no further parameterization. The parameters are denoted by θ and are bounded within the set Θ.

Simulation The finite volume method (FVM) was applied specifically to the PDE component of the coupled PDE-ODE system modeling cell population dynamics. The FVM discretizes the spatial domain into control volumes, and the integral form of the PDE is solved over each volume. Standard numerical solvers were used for the ODE governing ligand concentration. The initial cell density for the PDE was derived by performing a kernel density estimation on the experimental data of measurement time point 0, providing a smooth approximation of the cell state distribution as the starting condition for the simulation. The initial ligand concentration was set to 0.

Maximum Likelihood Estimation The model parameters θ were estimated by maximizing the multinomial likelihood of observing the experimental data given the model. The optimization problem was formulated:

$$\theta_{\mathrm{ml}} = \arg\max_{\theta \in \Theta} \prod_{k=1}^{n_t} \prod_{j=1}^{n_c} p(y_{kj} | \phi_u(t_k, \cdot; c_j, \theta)),$$

where $\phi_u(t, \cdot; c, \theta)$ denotes the solution operator for the population density u under condition c and parameters θ. The likelihood was based on the multinomial probability mass function suitable for the population-level histogram data, which we computed from the diffusion pseudotime representation of the single-cell measurements.

Discretize-Optimize Strategy The model followed a discretize-optimize strategy, assuming that the numerical simulation algorithm provides an accurate solution to the coupled PDE-ODE system. The fractions of cell states required as input to the multinomial probability mass function were computed directly from the finite volume approximation of $u(t, s; c, \theta)$.

pyPESTO Framework Parameter estimation was implemented using the pyPESTO framework, to which the author of this thesis also contributed and which offers a wide range of local and global optimization methods. A multi-start local optimization was performed using the gradient-based interior point algorithm IPOPT. The gradients were computed using finite differences.

Quality-of-Fit Assessment The quality-of-fit for the maximum likelihood estimate was assessed by comparing the experimental data with the expected distribution of measurements from the model. Gaussian kernel density estimates were calculated based on the states of experimentally observed cells, and these were compared with samples from the model's predicted population density $u(t_k, \cdot; c_j, \theta_{ml})$.

Uncertainty Analysis The uncertainty of the parameter estimates was evaluated using an ensemble method. The top K results from a multi-start optimization were selected as representatives of the parameter set, providing insights into the confidence intervals for each parameter.

5.5 Key Insights

The study provides a rigorous mathematical analysis of the model, proving the existence and uniqueness of solutions to the coupled PDE-ODE system. This theoretical foundation ensures the reliability of the model under various biological conditions.

We applied the model to a dataset involving dendritic cell activation upon LPS stimulation to validate it. The model successfully captured cell activation dynamics, highlighting the critical role of cell-to-cell communication in immune response processes. Moreover, it was able to describe the effect of cell-to-cell communication on Waddington's potential landscape in this scenario. Interestingly, the estimated model parameters show that the impact of increasing ligand concentrations on the drift is relatively small, while the effect on diffusion is quite large. In the context of Waddington's landscape, this suggests that instead of altering the landscape itself, increasing ligand concentration enhances the random change of cell state in the initial stable state, making it more likely for them to leave this state and move into energetically more favorable potential states.

6 Discussion of Results

Integrating complex biomedical datasets into differential equation models poses significant challenges due to the complexity of biological phenomena and the need to balance computational efficiency with model accuracy. In addressing the first research question—How can mechanistic models enhance classical cohort studies to understand disease dynamics better?—, we proposed advanced SEIR models enriched with longitudinal antibody studies. Specifically in Section 3, we demonstrated how these models could be used to gain valuable insights into the dynamics of SARS-CoV-2 spread in Ethiopia. The models provided a robust framework for understanding and predicting the progression of the pandemic by distinguishing between healthcare workers and the general community. Using Bayesian parameter estimation ensured the predictions were accurate and reliable, potentially aiding in practical public health planning and response. However, including a hypothetical viral variant in the model underscored the need for more comprehensive data on virus variants, as increased exposure rates and immunity-evasion properties significantly impacted herd immunity levels.

Building on this foundation, Section 4 extended the SEIR framework to include multiple SARS-CoV-2 variants and vaccination effects, providing a more detailed understanding of variant-specific immunity and cross-immunity. This section showcased the integration of newly obtained viral sequencing data into the modeling framework, allowing the capture of infection pathways across multiple variants and vaccination events. By incorporating a cross-immunity factor based on the genetic differences between variants, we managed to maintain computational feasibility while still gaining insights into variant interactions. Additionally, an antibody-level model was developed to assess the dynamics of the population's anti-nucleocapsid and anti-spike antibody levels. The model predicted that early and widespread vaccination could have significantly mitigated the impact of the Delta and Omicron waves, and it provided valuable forecasts about the long-term prevalence of high antibody levels in the population.

Addressing the second research question—How can mechanistic models be utilized to capture communication processes derived from single-cell data?— in Section 5, the focus shifted from whole-population level modeling to single-cell level processes and from infectious diseases to a more general immune response. We introduced a novel PDE-ODE system that extends models of the classical Waddington's landscape by incorporating cell-to-cell communication via ligand dynamics. This model successfully described the distribution of cells and ligand concentrations simultaneously, offering a comprehensive view of cellular communication processes. By fitting the model to experimental data from an immune cell activation study, we demonstrated the critical role of cell-to-cell communication in immune response. The model's ability to capture these complex interactions highlighted its potential for broader applications in developmental biology and immunology.

The scientific results presented in this thesis have significantly advanced our understanding of integrating complex biomedical datasets into mechanistic models, effectively addressing both macro-level population dynamics and micro-level cellular processes. The SEIR models developed in this work have improved our ability to predict disease dynamics and assess the impact of various factors, such as healthcare worker exposure and viral variant evolution, on epidemic progression. Although compartmental modeling is a well-established approach, the models introduced in the first study demonstrated the advantages of integrating healthcare workers with community members, particularly in

regions where insufficient testing infrastructure and demographic factors contribute to severe under-reporting. These models have proven especially valuable in such contexts, offering a more accurate reflection of the pandemic's true scale.
The more complex multivariant and antibody-level models developed in this thesis have provided new insights into the interplay between different SARS-CoV-2 variants and the role of vaccination in shaping immunity landscapes. This research underscores the potential importance of early vaccination in future pandemics. When early vaccination is not feasible, its insights into the spread of multiple infections and antibody levels could imply the adoption of sophisticated vaccination strategies, such as adjusting the number of administered doses based on previous infections. This approach is further investigated by a cost-effectiveness analysis from a study conducted in Jimma, Ethiopia, by Gudina et al. [8], to which the author of this thesis also contributed. It builds on top of the findings presented here and enhances them by additional antibody survey of persons before and after vaccination. The methodologies established are not limited to SARS-CoV-2 but can be generalized to other infectious diseases exhibiting similar transmission dynamics or immune response mechanisms.
One of the challenges encountered was the limited availability of comprehensive genetic sequencing data for SARS-CoV-2 variants in Ethiopia. This scarcity necessitated assumptions and estimations that could introduce uncertainties into the models. Addressing this limitation requires strengthening local sequencing capabilities and establishing data-sharing collaborations to enhance model accuracy and reliability. Moreover, the interdisciplinary nature of this research highlights the value of collaborative efforts in addressing complex biomedical challenges. By integrating diverse expertise, the developed models offer a more holistic understanding of disease dynamics and immune responses.
On the modeling side, future research could explore the integration of within-host viral dynamics and immune system dynamics, e.g., via a target cell-limited model [67], with population-level models, potentially bridging the gap between individual-level and population-level understanding of disease dynamics. Integrating different scales could also address the challenge of linking antibody levels to actual immunity, providing a more comprehensive understanding of protective immunity within populations. Before the outbreak of COVID-19, the need for such models was already discovered for other infectious diseases and the subject of special issues, reviews, and opinion letters [68–70]. An excellent example of how one can bridge the scales can be found in Almocera et al. [71], who investigate mathematically a model connecting a micro-scale cell model describing the interaction between virus concentration and T-cells to a macro-scale model describing populations of susceptible and infected via a viral load dependent transmission rate. A similar approach tailored to HIV with more compartments can be found in Manda and Chirove [72]. However, for SARS-CoV-2, there appear to be few attempts to apply such multi-scale modeling approaches.
However, most of the studies mentioned above only investigate model properties using parameters from the literature. Since the macro-scale multivariant model described above is already quite complex, one would have to simplify it to retain computational feasibility and include more data to inform all parameters during calibration.
Another refinement of the epidemiological models presented in this thesis could involve incorporating more detailed genetic and immunological data to enhance the accuracy and applicability of the predictions. However, one would require even more data on the virus variants to inform such sophisticated models, which is not straightforward for countries with scarce sequencing infrastructure. Since virus variants constantly mutate and their

competition for susceptible hosts influences the evolution of new variants, these genetic pressures are an exciting subject of study [73]. If one could incorporate a reliable representation of mutation drivers and virus evolution into mechanistic models, the insights gained would have substantially more predictive power than what can be inferred from models describing already existing variants.
On the single-cell level, the PDE-ODE model for cell-to-cell communication has opened new avenues for exploring cellular interactions in developmental and immune processes. The rigorous mathematical foundation and successful application of this model to immune cell activation underscore its potential for future research in other areas of biology.
Applying this framework to more complex developmental processes that fully exploit its versatility, such as scenarios involving multiple ligands or multiple branching pathways, presents an exciting direction for future research. A particularly compelling use case would be, e.g., to describe the entire differentiation landscape of hematopoiesis, where blood stem cells undergo a highly regulated process of specialization into various blood cell types, including red blood cells, white blood cells, and platelets. This system involves multiple stages of differentiation and is influenced by a wide array of signaling molecules [74]. Modeling such a system using the proposed framework would allow a more comprehensive understanding of how different signaling pathways interact and drive cell fate decisions in a multi-lineage context.
However, a key challenge lies in advocating for experiments that cover multiple time points and compare communication and non-communication scenarios. Such experiments are crucial for gaining deeper insights into these critical communication processes and validating the model across biological contexts.
Another promising direction for enhancing the framework's applicability lies in incorporating neural networks instead of traditional splines and Hill functions as model coefficients and dependency terms. By formulating such a model combining differential equations and neural networks, which is also known as a universal differential equation (UDE), it could achieve even greater generality and flexibility, potentially covering a more comprehensive range of biological scenarios with reduced need for manual specification of functional forms [75]. The UDE approach would allow the model to learn complex relationships directly from data, thereby decreasing the reliance on user-defined inputs and improving its adaptability to various developmental and immune processes.
Since the parameter calibration of the PDE-ODE system is computationally very expensive, research on amortized inference could also prove valuable. Amortized inference directly learns the posterior distribution from model simulations with parameters sampled from the prior via an invertible neural network [76]. While the training of the neural network is computationally demanding, the invertibility of the network can be used to infer parameters from measurements with negligible computational effort. Since the initial training cost only amortizes if sufficient inference tasks are performed using the trained network, it works best with standardized experimental setups, where time points of data collection, measured properties, and observations quantities are comparable.
Overall, this thesis has made significant contributions to computational life sciences by developing and applying sophisticated mathematical models to real-world biomedical data, providing a deeper understanding of the complex processes that govern population-level disease dynamics and single-cell communication. The research presented in this thesis highlights the importance of integrating complex biomedical datasets into differential equation models to gain a deeper understanding of epidemiological patterns and cellular communication processes. The advancements in modeling techniques, parameter estima-

tion, and data integration in epidemics and cell-to-cell communication demonstrated in this work contribute to a more nuanced and comprehensive approach to studying biological systems, with potential applications in public health, immunology, and beyond.

Acronyms

Anti-N anti-nucleocapsid antibodies. 26, 28–30

Anti-S anti-spike antibodies. 26, 28, 30

BDF backward-differentiation-formula. 11

BFGS Broyden–Fletcher–Goldfarb–Shanno. 17, 19

BVP boundary value problem. 11

COVID-19 Coronavirus Disease 2019. 1, 36

FVM finite volume method. 11, 12, 14, 33

HCW healthcare worker. 22–25

IVP initial value problem. 10, 15

MAP maximum-a-posteriori estimation. 20, 21

MCMC Markov Chain Monte Carlo. 20, 21, 25

MOC mutations of concern. 26, 28, 30

mRNA messenger RNA. 2

ODE ordinary differential equation. 2, 3, 10–12, 15, 18, 19, 22–24, 26, 31–33, 35, 37

PCA Principal Component Analysis. 5, 6

PDE partial differential equation. 2, 3, 10–12, 14, 19, 31–33, 35, 37

RK4 Runge-Kutta method of fourth-order. 10

RNA ribonucleic acid. 39

RNA-seq RNA-sequencing. 32

SARS-CoV-2 Severe acute respiratory syndrome coronavirus type 2. 1–3, 22, 23, 25, 26, 35, 36

SDE stochastic differential equation. 31

SEIR Susceptible-Exposed-Infectious-Recovered. 22–26, 35

SR1 Symmetric rank-one. 17, 19

SVD singular value decomposition. 5

UDE universal differential equation. 37

UMAP Uniform Manifold Approximation and Projection. 7, 8, 32

VOC variants of concern. 27

WHO World Health Organization. 1, 23

Glossary

adaptive Metropolis-Hastings algorithm Metropolis-Hastings algorithm with adaptive proposal distribution. 21

backward finite difference approximation Finite differences with negative small step away from differentiation point. 18

bandwidth Hyperparameter of kernel density estimation. 9

Bayesian inference Statistical inference quantifying parameter uncertainty. 19

central finite difference approximation Finite differences with mixture of positive and negative small steps away from differentiation point. 18

diffusion distance Distance measure based on a random walk. 6

diffusion maps Nonlinear dimension reduction method. 6

diffusion pseudotime A trajectory inference method based on diffusion maps. 7

empirical model Model which is mainly data-driven. 2

finite difference approximation Gradient approximation method substituting the limit to zero in the definition of gradient by small value. 18

forward finite difference approximation Finite differences with positive small step away from differentiation point. 18

forward sensitivity equations Equations for computing the derivatives of state variables with respect to parameters. 18

frequentist inference Statistical inference assuming parameters to be fixed. 19

gradient descent Optimization method of iteratively going in the direction of steepest descent. 15

gradient-based optimization Optimization methods utilizing gradients of the objective function. 15

identifiability Property of a parameter estimation problem stating that the optimum is unique. 13

interior point or barrier methods Class of optimization methods for convex parameter spaces. 18

kernel density estimation Method which tries to capture the underlying density function of independent identically sampled data. 9

k-means Clustering method based on iteratively assigning data points to closest cluster means. 8

learning rate Step size of gradient descent. 15

least squares objective Objective function based on minimizing the sum of squared differences. 13

likelihood Probability of parameters conditional on data. 13

line-search Additional optimization problem in each gradient-descent step for choosing learning rate. 16

log-likelihood $\mathcal{LL}$ Log-transformed likelihood. 14

mechanistic model Model implementing apriori known mechanisms. 2

Metropolis-Hastings algorithm Algorithm for approximating an unknown distribution using a proposal distribution and a known distribution proportional to the target distribution. 20

multinomial likelihood A likelihood based on the multinomial distribution. 14

multinomial probability mass function $P_{\mathbf{mult}}$ Probability mass function of the multinomial distribution. 15

multistart optimization Method of obtaining global optimum by multiple local optimizations starting at sufficiently many different initial values. 17

multi-step solvers ODE solvers using multiple function evaluations per step. 11

Newton's method Optimization method using gradient and Hessian of objective function. 17

parameters Unknown quantities of a model like reaction rates. 12

point estimate Estimation result of frequentist inference. 19

posterior distribution Distribution of information on parameters after updating the prior distribution by the information contained in the data. 20

primal-dual optimization A interior point optimization method. 18

principal components Orthogonal axes of greatest variance in the data. 5

prior distribution Distribution of parameters prior to informing them by data. 20

quasi-Newton methods Methods using approximations of Hessian in Newton's method. 17

random search Method of choosing parameters by drawing randomly from distribution and comparing objective values. 15

silhouette scores Score to evaluate clustering quality by comparing mean inner to in-between cluster distances. 9

single-step solvers ODE solvers using one function evaluation per step. 11

stiff equation ODE requiring tiny step sizes for numerical solution. 11

trajectory inference Reduction methods to a one-dimensional space including an ordering representing cell developmental stage. 7

trust region optimization Method where at each step a local approximation of objective function is optimized analytically. 17

References

1. Gudina, E. K. *et al.* Seroepidemiology and model-based prediction of SARS-CoV-2 in Ethiopia: longitudinal cohort study among front-line hospital workers and communities. *The Lancet Global Health* **9,** e1517–e1527. doi:10.1016/S2214-109X(21)00386-7 (2021).
2. Merkt, S. *et al.* Long-term monitoring of SARS-CoV-2 seroprevalence and variants in Ethiopia provides prediction for immunity and cross-immunity. *Nature Communications* **15,** 3463. doi:10.1038/s41467-024-47556-2 (2024).
3. Merkt, S. *et al. A Dynamic Model for Waddington's Landscape Accounting for Cell-to-Cell Communication* Preprint available on *SSRN*. 2024. doi:10.2139/ssrn.5051345.
4. Schmiester, L. *et al.* PEtab—Interoperable specification of parameter estimation problems in systems biology. *PLOS Computational Biology* **17,** 1–10. doi:10.1371/journal.pcbi.1008646 (2021).
5. Raimúndez, E. *et al.* COVID-19 outbreak in Wuhan demonstrates the limitations of publicly available case numbers for epidemiological modeling. *Epidemics* **34,** 100439. doi:10.1016/j.epidem.2021.100439 (2021).
6. Stapor, P. *et al.* Mini-batch optimization enables training of ODE models on large-scale datasets. *Nature Communications* **13,** 34. doi:10.1038/s41467-021-27374-6 (2022).
7. Schälte, Y. *et al.* pyPESTO: a modular and scalable tool for parameter estimation for dynamic models. *Bioinformatics* **39,** btad711. doi:10.1093/bioinformatics/btad711 (2023).
8. Gudina, E. K. *et al.* Tailoring COVID-19 Vaccination Strategies in High-Seroprevalence Settings: Insights from Ethiopia. *Vaccines* **12.** doi:10.3390/vaccines12070745 (2024).
9. Magner, L. N. *A History of the Life Sciences, Revised and Expanded* 3rd. ISBN: 9780429213342 (CRC Press, New York, 2002).
10. Lison, A. *et al.* Effectiveness assessment of non-pharmaceutical interventions: lessons learned from the COVID-19 pandemic. *The Lancet Public Health* **8,** e311–e317. doi:10.1016/S2468-2667(23)00046-4 (2023).
11. Beladiya, J. *et al.* Safety and efficacy of COVID-19 vaccines: A systematic review and meta-analysis of controlled and randomized clinical trials. *Reviews in Medical Virology* **34,** e2507. doi:10.1002/rmv.2507 (2024).
12. Lucero-Prisno III, D. E. *et al.* Top 10 public health challenges to track in 2023: Shifting focus beyond a global pandemic. *Public Health Challenges* **2,** e86. doi:10.1002/puh2.86 (2023).
13. Napolitano, F., Xu, X. & Gao, X. Impact of computational approaches in the fight against COVID-19: an AI guided review of 17 000 studies. *Briefings in Bioinformatics* **23,** bbab456. doi:10.1093/bib/bbab456 (2021).
14. Thakur, A. K. in *New Trends in Pharmacokinetics* (eds Rescigno, A. & Thakur, A. K.) 41–51 (Springer US, Boston, MA, 1991). ISBN: 978-1-4684-8053-5. doi:10.1007/978-1-4684-8053-5_3.

15. Baker, R. E., Peña, J.-M., Jayamohan, J. & Jérusalem, A. Mechanistic models versus machine learning, a fight worth fighting for the biological community? *Biology Letters* **14,** 20170660. doi:10.1098/rsbl.2017.0660 (2018).

16. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596,** 583–589. doi:10.1038/s41586-021-03819-2 (2021).

17. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* **17,** 333–351 (2016).

18. Klipp, E., Liebermeister, W., Wierling, C. & Kowald, A. *Systems biology a textbook* Second, completely revised and enlarged edition. ISBN: 9783527336364 (Wiley-VCH, Weinheim, 2016).

19. Kermack, W. O. & McKendrick, A. G. 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,** 700–721. doi:10.1098/rspa.1927.0118 (1927).

20. Hethcote, H. W. The mathematics of infectious diseases. *SIAM review* **42,** 599–653. doi:10.1137/S0036144500371907 (2000).

21. Brauer, F. & Castillo-Chavez, C. in *Mathematical Models in Population Biology and Epidemiology* 345–409 (Springer New York, New York, NY, 2012). ISBN: 978-1-4614-1686-9. doi:10.1007/978-1-4614-1686-9_9.

22. Hross, S. & Hasenauer, J. Analysis of CFSE time-series data using division-, age- and label-structured population models. *Bioinformatics* **32,** 2321–2329. doi:10.1093/bioinformatics/btw131 (2016).

23. Fischer, D. S. *et al.* Inferring population dynamics from single-cell RNA-sequencing time series data. *Nature Biotechnology* **37,** 461–468. doi:10.1038/s41587-019-0088-0 (2019).

24. Cho, H., Kuo, Y.-H. & Rockne, R. C. Comparison of cell state models derived from single-cell RNA sequencing data: graph versus multi-dimensional space. *Mathematical Biosciences and Engineering* **19,** 8505–8536. doi:10.3934/mbe.2022395 (2022).

25. Butcher, J. C. *Numerical Methods for Ordinary Differential Equations* Third edition. ISBN: 9781119121534. doi:10.1002/9781119121534 (Wiley, Chichester, UK, 2016).

26. Tadmor, E. A review of numerical methods for nonlinear partial differential equations. *Bulletin (New Series) of the American Mathematical Society* **49.** doi:10.1090/S0273-0979-2012-01379-4 (2012).

27. Raue, A. *et al.* Lessons Learned from Quantitative Dynamical Modeling in Systems Biology. *PLOS ONE* **8,** 1–17. doi:10.1371/journal.pone.0074335 (2013).

28. Jolliffe, I. T. & Cadima, J. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **374,** 20150202. doi:10.1098/rsta.2015.0202 (2016).

29. Schölkopf, B., Smola, A. & Müller, K.-R. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation* **10,** 1299–1319. doi:10.1162/089976698300017467 (1998).

30. Coifman, R. R. *et al.* Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences* **102,** 7426–7431. doi:10.1073/pnas.0500334102 (2005).

31. Haghverdi, L., Buettner, F. & Theis, F. J. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* **31,** 2989–2998. doi:10.1093/bioinformatics/btv325 (2015).

32. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology* **19,** 15. doi:10.1186/s13059-017-1382-0 (2018).

33. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods* **13,** 845–848. doi:10.1038/nmeth.3971 (2016).

34. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* **3,** 861. doi:10.21105/joss.00861 (2018).

35. MacQueen, J. *Some methods for classification and analysis of multivariate observations* in *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability* (eds Cam, L. M. L. & Neyman, J.) (University of California Press, 1967).

36. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20,** 53–65. doi:10.1016/0377-0427(87)90125-7 (1987).

37. Rosenblatt, M. Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics* **27,** 832 –837. doi:10.1214/aoms/1177728190 (1956).

38. Parzen, E. On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics* **33,** 1065 –1076. doi:10.1214/aoms/1177704472 (1962).

39. Silverman, B. *Density Estimation for Statistics and Data Analysis* ISBN: 9780412246203 (Taylor & Francis, London, New York, 1986).

40. Amann, H. & Escher, J. *Analysis II* Zweite, korrigierte Auflage. ISBN: 9783764374020 (Birkhäuser Verlag, Basel, 2006).

41. Hairer, E., Nørsett, S. P. & Wanner, G. in *Solving Ordinary Differential Equations I: Nonstiff Problems* 303–432 (Springer Berlin Heidelberg, Berlin, Heidelberg, 1987). ISBN: 978-3-662-12607-3. doi:10.1007/978-3-662-12607-3_3.

42. Hairer, E. & Wanner, G. in *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems* 255–398 (Springer Berlin Heidelberg, Berlin, Heidelberg, 1991). ISBN: 978-3-662-09947-6. doi:10.1007/978-3-662-09947-6_2.

43. Hindmarsh, A. C. *et al.* SUNDIALS: Suite of nonlinear and differential/algebraic equation solvers. *ACM Transactions on Mathematical Software (TOMS)* **31,** 363–396. doi:10.1145/1089014.1089020 (2005).

44. Gardner, D. J., Reynolds, D. R., Woodward, C. S. & Balos, C. J. Enabling new flexibility in the SUNDIALS suite of nonlinear and differential/algebraic equation solvers. *ACM Transactions on Mathematical Software (TOMS).* doi:10.1145/3539801 (2022).

45. Fröhlich, F. *et al.* AMICI: High-Performance Sensitivity Analysis for Large Ordinary Differential Equation Models. *Bioinformatics.* btab227. doi:10.1093/bioinformatics/btab227 (2021).

46. Evans, L. C. *Partial differential equations / Lawrence C. Evans.* ISBN: 0821807722 (American Mathematical Society, Providence, R.I, 1998).

47. Guyer, J. E., Wheeler, D. & Warren, J. A. FiPy: Partial Differential Equations with Python. *Computing in Science & Engineering* **11,** 6–15. doi:10.1109/MCSE.2009.52 (2009).

48. Lazarov, R. D., Mishev, I. D. & Vassilevski, P. S. *Finite Volume Methods for Convection-Diffusion Problems* 31–55 (Society for Industrial and Applied Mathematics, 1996).

49. Eymard, R., Gallouët, T. & Herbin, R. in *Solution of Equation in* $\mathbb{R}^n$ *(Part 3), Techniques of Scientific Computing (Part 3)* 713–1018 (Elsevier, 2000). doi:10.1016/S1570-8659(00)07005-8.

50. Newey, W. K. & McFadden, D. in *Handbook of Econometrics* 2111–2245 (Elsevier, 1994). doi:10.1016/S1573-4412(05)80005-4.

51. Boyd, S. P. & Vandenberghe, L. *Convex Optimization* ISBN: 9780521833783 (Cambridge University Press, Cambridge, New York, 2004).

52. Armijo, L. Minimization of functions having Lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics* **16,** 1 –3. doi:10.2140/pjm.1966.16.1 (1966).

53. Wolfe, P. Convergence Conditions for Ascent Methods. *SIAM Review* **11,** 226–235. doi:10.1137/1011036 (1969).

54. Fletcher, R. *Practical methods of optimization* 2. ed., reprint. in paperback. ISBN: 0471494631 (Wiley, Chichester [u.a, 2000).

55. Conn, A. R., Gould, N. I. M. & Toint, P. L. Convergence of quasi-Newton matrices generated by the symmetric rank one update. *Mathematical Programming* **50,** 177–195. doi:10.1007/BF01594934 (1991).

56. Sorensen, D. C. Newton's Method with a Model Trust Region Modification. *SIAM journal on numerical analysis* **19,** 409–426. doi:10.1137/0719026 (1982).

57. Wright, S. *Primal-Dual Interior-Point Methods* ISBN: 9780898713824 (Society for Industrial and Applied Mathematics, Philadelphia, 1997).

58. Wächter, A. & Biegler, L. T. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming* **106,** 25–57. doi:10.1007/s10107-004-0559-y (2006).

59. Fröhlich, F. & Sorger, P. K. Fides: Reliable trust-region optimization for parameter estimation of ordinary differential equation models. *PLOS Computational Biology* **18,** 1–28. doi:10.1371/journal.pcbi.1010322 (2022).

60. Villaverde, A. F., Raimúndez, E., Hasenauer, J. & Banga, J. R. Assessment of Prediction Uncertainty Quantification Methods in Systems Biology. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **20,** 1725–1736. doi:10.1109/TCBB.2022.3213914 (2023).

61. Box, G. & Tiao, G. *Bayesian Inference in Statistical Analysis* ISBN: 9780471574286 (Wiley, New York, 1992).

62. Melkumova, L. & Shatskikh, S. Comparing Ridge and LASSO estimators for data analysis. *Procedia Engineering* **201,** 746–755. doi:10.1016/j.proeng.2017.09.615 (2017).

63. Tierney, L. Markov Chains for Exploring Posterior Distributions. *The Annals of Statistics* **22,** 1701 –1728. doi:10.1214/aos/1176325750 (1994).

64. Ballnus, B. *et al.* Comprehensive benchmarking of Markov chain Monte Carlo methods for dynamical systems. *BMC Systems Biology* **11,** 63. doi:10.1186/s12918-017-0433-1 (2017).

65. Salvatier, J., Wiecki, T. V. & Fonnesbeck, C. Probabilistic programming in Python using PyMC3. *PeerJ Comput. Sci.* **2,** e55. doi:10.7717/peerj-cs.55 (2016).

66. Shalek, A. K. *et al.* Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510,** 363–369. doi:10.1038/nature13437 (2014).

67. Hernandez-Vargas, E. A. & Velasco-Hernandez, J. X. In-host Mathematical Modelling of COVID-19 in Humans. *Annual Reviews in Control* **50,** 448–456. doi:10.1016/j.arcontrol.2020.09.006 (2020).

68. Heesterbeek, H. *et al.* Modeling infectious disease dynamics in the complex landscape of global health. *Science* **347,** aaa4339. doi:10.1126/science.aaa4339 (2015).

69. Handel, A. & Rohani, P. Crossing the scale from within-host infection dynamics to between-host transmission fitness: a discussion of current assumptions and knowledge. *Philosophical Transactions of the Royal Society B: Biological Sciences* **370,** 20140302. doi:10.1098/rstb.2014.0302 (2015).

70. Gutierrez, J. B., Galinski, M. R., Cantrell, S. & Voit, E. O. From within host dynamics to the epidemiology of infectious disease: Scientific overview and challenges. *Mathematical Biosciences* **270,** 143–155. doi:10.1016/j.mbs.2015.10.002 (2015).

71. Almocera, A. E. S., Nguyen, V. K. & Hernandez-Vargas, E. A. Multiscale model within-host and between-host for viral infectious diseases. *Journal of Mathematical Biology* **77,** 1035–1057. doi:10.1007/s00285-018-1241-y (2018).

72. Manda, E. C. & Chirove, F. Modelling coupled within host and population dynamics of R_5 and X_4 HIV infection. *Journal of Mathematical Biology* **76,** 1123–1158. doi:10.1007/s00285-017-1170-1 (2018).

73. Markov, P. V. *et al.* The evolution of SARS-CoV-2. *Nature Reviews Microbiology* **21,** 361–379. doi:10.1038/s41579-023-00878-2 (2023).

74. Laurenti, E. & Göttgens, B. From haematopoietic stem cells to complex differentiation landscapes. *Nature* **553,** 418–426. doi:10.1038/nature25022 (2018).

75. Rackauckas, C. *et al. Universal Differential Equations for Scientific Machine Learning.* Preprint available on *arXiv*. 2021. doi:10.48550/arXiv.2001.04385.

76. Radev, S. T. *et al. BayesFlow: Amortized Bayesian Workflows With Neural Networks.* Preprint available on *arXiv*. 2023. doi:10.48550/arXiv.2306.16015.

Appendices

A Seroepidemiology and model-based prediction of SARS-CoV-2 in Ethiopia: longitudinal cohort study among front-line hospital workers and communities

This publication is reprinted as part of this thesis according to Elsevier guidelines on sharing published journal articles. Material from:
Gudina, E. K. *et al.* Seroepidemiology and model-based prediction of SARS-CoV-2 in Ethiopia: longitudinal cohort study among front-line hospital workers and communities. en. *The Lancet Global Health* **9,** e1517–e1527. doi:10.1016/S2214-109X(21)00386-7 (2021)

Seroepidemiology and model-based prediction of SARS-CoV-2 in Ethiopia: longitudinal cohort study among front-line hospital workers and communities

Esayas Kebede Gudina, Solomon Ali*, Eyob Girma, Addisu Gize, Birhanemeskel Tegene, Gadissa Bedada Hundie, Wondewosen Tsegaye Sime, Rozina Ambachew, Alganesh Gebreyohanns, Mahteme Bekele, Abhishek Bakuli, Kira Elsbernd, Simon Merkt, Lorenzo Contento, Michael Hoelscher, Jan Hasenauer, Andreas Wieser*, Arne Kroidl**

Summary

Background Over 1 year since the first reported case, the true COVID-19 burden in Ethiopia remains unknown due to insufficient surveillance. We aimed to investigate the seroepidemiology of SARS-CoV-2 among front-line hospital workers and communities in Ethiopia.

Methods We did a population-based, longitudinal cohort study at two tertiary teaching hospitals involving hospital workers, rural residents, and urban communities in Jimma and Addis Ababa. Hospital workers were recruited at both hospitals, and community participants were recruited by convenience sampling including urban metropolitan settings, urban and semi-urban settings, and rural communities. Participants were eligible if they were aged 18 years or older, had provided written informed consent, and were willing to provide blood samples by venepuncture. Only one participant per household was recruited. Serology was done with Elecsys anti-SARS-CoV-2 anti-nucleocapsid assay in three consecutive rounds, with a mean interval of 6 weeks between tests, to obtain seroprevalence and incidence estimates within the cohorts.

Findings Between Aug 5, 2020, and April 10, 2021, we did three survey rounds with a total of 1104 hospital workers and 1229 community residents participating. SARS-CoV-2 seroprevalence among hospital workers increased strongly during the study period: in Addis Ababa, it increased from 10·9% (95% credible interval [CrI] 8·3–13·8) in August, 2020, to 53·7% (44·8–62·5) in February, 2021, with an incidence rate of 2223 per 100 000 person-weeks (95% CI 1785–2696); in Jimma Town, it increased from 30·8% (95% CrI 26·9–34·8) in November, 2020, to 56·1% (51·1–61·1) in February, 2021, with an incidence rate of 3810 per 100 000 person-weeks (95% CI 3149–4540). Among urban communities, an almost 40% increase in seroprevalence was observed in early 2021, with incidence rates of 1622 per 100 000 person-weeks (1004–2429) in Jimma Town and 4646 per 100 000 person-weeks (2797–7255) in Addis Ababa. Seroprevalence in rural communities increased from 18·0% (95% CrI 13·5–23·2) in November, 2020, to 31·0% (22·3–40·3) in March, 2021.

Interpretation SARS-CoV-2 spread in Ethiopia has been highly dynamic among hospital worker and urban communities. We can speculate that the greatest wave of SARS-CoV-2 infections is currently evolving in rural Ethiopia, and thus requires focused attention regarding health-care burden and disease prevention.

Funding Bavarian State Ministry of Sciences, Research, and the Arts; Germany Ministry of Education and Research; EU Horizon 2020 programme; Deutsche Forschungsgemeinschaft; and Volkswagenstiftung.

Introduction

Despite the initial prediction that the COVID-19 pandemic would hit Africa hard, the feared humanitarian crisis from COVID-19 has so far largely been avoided.[1,2] The total reported numbers from Africa represent only 2·9% of COVID-19 cases and 3·7% of COVID-19 deaths reported globally.[3,4] However, the true number of cases and the impact of COVID-19 remains largely unknown due to insufficient testing and weak surveillance and reporting systems.[5–7]

Seroepidemiological evidence from various African countries showed high prevalence of SARS-CoV-2 antibodies among health-care workers: 41·2% in Democratic Republic of the Congo[8] and 45·1% in Nigeria[9] and 60% among blood donors in South Africa.[10] These findings pose a serious question of the true extent to which Africa has been affected by COVID-19, and whether this is largely unknown due to underdiagnosis and under-reporting.

Ethiopia, which reported its first case on March 13, 2020,[11] has implemented a targeted testing strategy that focuses on individuals who are symptomatic, contacts of confirmed cases, and high-risk groups.[12] This approach neglects most cases with mild or no symptoms.[13] With

Lancet Glob Health 2021; 9: e1517–27

See **Comment** page e1477

For the Amharic translation of the abstract see **Online** for appendix 1

For the Afan Oromo translation of the abstract see **Online** for appendix 2

*Contributed equally

Jimma University Institute of Health, Jimma, Ethiopia (Prof E K Gudina PhD, E Girma MPH); **Saint Paul's Hospital Millennium Medical College, Addis Ababa, Ethiopia** (S Ali PhD, A Gize MSc, B Tegene MSc, G B Hundie PhD, W T Sime PhD, R Ambachew MSc, A Gebreyohanns MSC, M Bekele MD); **Division of Infectious Diseases and Tropical Medicine, Medical Center of the University of Munich, Munich, Germany** (A Bakuli PhD, K Elsbernd MPH, Prof M Hoelscher MD, A Wieser MD, A Kroidl MD); **German Center for Infection Research, partner site Munich, Munich, Germany** (A Bakuli, Prof M Hoelscher, A Wieser, A Kroidl); **Institute for Medical Information Processing, Biometry, and Epidemiology—IBE, Ludwig Maximilian University of Munich, Munich, Germany** (K Elsbernd); **Institute of Computational Biology, Helmholtz Zentrum München—German Research Center for Environmental Health, Neuherberg, Germany** (Prof J Hasenauer PhD); **Center for Mathematics, Technische Universität München, Garching, Germany** (Prof J Hasenauer); **Faculty of Mathematics and Natural Sciences, University of Bonn, Bonn, Germany** (S Merkt MSc, L Contento PhD, Prof J Hasenauer)

Correspondence to:
Prof Esayas Kebede Gudina, Jimma University, Jimma, Ethiopia
esayas.gudina@ju.edu.et

Research in context

Evidence before this study

The burden of COVID-19 in Africa was not as overwhelming as in other regions of the world during the so-called first wave of the pandemic. However, an apparent second wave characterised by a greater impact on African health systems has been observed since the end of 2020. This observation was supported by a few cross-sectional serological studies indicating high infection rates, mainly among health-care workers. Ethiopia reported its first case in March, 2020, but the true burden of the pandemic remains unknown due to insufficient testing and weak surveillance. We searched PubMed from database inception to May 31, 2021, for peer-reviewed articles using the terms "COVID-19" OR "SARS-CoV-2" AND "Ethiopia", with no language restrictions. Additionally, we searched bibliographies of identified studies and Google for manuscripts and unpublished reports. We identified three studies, one in preprint version, reporting seroprevalence of SARS-CoV-2 from Ethiopia. All three were cross-sectional studies with sample sizes ranging from 99 to 1856 individuals and focused mainly on the general population of Addis Ababa. Only one study additionally involved rural communities, and none involved health-care workers.

Added value of this study

To our knowledge, we provide the first report of prospective longitudinal SARS-CoV-2 transmission dynamics and incidence rates from an African country, derived from front-line health-care workers at major tertiary referral hospitals, urban residents, and rural communities. The sampling period of this repeated seroprevalence survey fell within the transmission period between the first and second COVID-19 wave in Africa and reveals a strong increase in SARS-CoV-2 transmission within different populations. Our data coincided with national reports of increased burden of critical patient care and PCR test positivity rates. On the basis of our seroprevalence data, we additionally provide a modelling analysis predicting possible SARS-CoV-2 herd immunity first for the wild-type virus, and then assuming introduction of variant strains.

Implications of all the available evidence

This study illustrates current COVID-19 disease dynamics in an African population, indicating a predominance of SARS-CoV-2 transmission in urban settings. It can be speculated that the greatest wave of SARS-CoV-2 infection in rural Ethiopia is currently evolving, and thus requires focused attention regarding health-care burden and outbreak control. Approaching peak herd immunity level, either through natural disease exposure or SARS-CoV-2 vaccination, is widely investigated regarding not only health-care burden, but also emerging viral escape variants. COVID-19 vaccinations are currently scaled out in Africa and will, for most individuals, represent a booster immunisation after previous SARS-CoV-2 exposure. Vaccination strategies should be adapted; for instance, assessing the serostatus before vaccination and providing boosting only with one dose. On the basis of our data on disease dynamics and modelling analysis, we expect valuable follow-up information on pandemic disease control strategies applicable for Africa.

this strategy, fewer than 3% of the population has been tested, and only 273 175 cases have been detected.[4,14] As a result, the true number of SARS-CoV-2 infections in the larger community is completely unknown.

Ethiopia lifted most of its COVID-19-related restrictions on Sept 8, 2020, and the daily testing capacity declined sharply due to insufficient laboratory infrastructure, supplies, and trained workforce.[6] The country saw an increase in the number of cases, test positivity rate, severe disease, and COVID-19-related deaths from the second half of 2020 to March, 2021.[14] As for most African countries, Ethiopia does not have routine death registration and baseline vital statistics. Therefore, it is extremely difficult to estimate excess deaths due to COVID-19 by use of mortality data. Nevertheless, more deaths compared with similar periods during previous years have been reported from cemeteries in Addis Ababa.[15]

Serological studies remain the only option to identify the burden of infection in the community and assess outbreak dynamics.[16,17] Therefore, in this study, we aimed to determine the seroprevalence and seroincidence of SARS-CoV-2 across time and model the COVID-19 epidemic among communities and health-care workers in Ethiopia.

Methods

Study design and settings

This population-based, longitudinal, exploratory cohort study was done at two tertiary teaching hospitals (Jimma Medical Center [JMC] and St Paul's Hospital), in Jimma Town and surrounding rural communities and Addis Ababa (figure 1).

JMC is the only tertiary referral centre in southwest Ethiopia, with a catchment population of more than 20 million, 800 inpatient beds, and about 3000 hospital workers. It is located in Jimma Town, the biggest city in southwest Ethiopia, with a population of 300 000.

St Paul's Hospital is one of several public tertiary teaching hospitals in Addis Ababa, with 700 beds and more than 2800 hospital workers. Addis Ababa is the capital and largest city of Ethiopia, with an estimated population of 5 million.

This research was approved by the Institutional Review Boards of Jimma University Institute of Health (RPGD/978/2020), St Paul's Hospital Millennium Medical College (PM23/239), and Ludwig Maximilian University of Munich (21–0293). Additional approvals were obtained from Addis Ababa and Oromia Regional Health Bureaus (BEFO/KBTFU/1-16/488). Written informed consent in local languages was obtained from

all participants. For participants who could not read and write, an impartial witness was involved during the consenting process to ensure the provision of all necessary information before obtaining the participant's fingerprint for consent. Preliminary results were communicated to the Ethiopian Public Health Institute, Federal Ministry of Health of Ethiopia, and Ethiopian Medical Association.

Selection of study participants

Front-line hospital workers from outpatient and inpatient units—including clinical staff, medical interns, cleaners, guards, food handlers, and administrative personnel—were recruited at both hospitals. A sample size of 499 hospital workers per hospital was targeted on the basis of an estimated seroprevalence of 50% (95% CI, 5% margin of error) and a design effect of two clusters (JMC and St Paul's Hospital). A non-response rate of 10% for each round (a total of 30% for all three rounds) was assumed.

The recruitment of community participants was guided by convenience sampling and included urban metropolitan settings (Addis Ababa), urban and semi-urban settings (Jimma Town), and rural communities (four rural districts in Jimma Zone; figure 1). In Addis Ababa, we intentionally selected two subcities on the basis of their population density: Addis Ketema (most densely populated) and Yeka (sparsely populated). In Jimma Town, we recruited participants from all areas of the city. Rural residents were recruited from four rural districts located along four main roads connecting to Jimma Town. Households were selected randomly in a way that avoided frequent interaction from the next candidate household to prevent cross-contamination. The sample size calculation was done in July, 2020, when not much baseline data was available. At the time, we planned to include 664 households (332 in Jimma and 332 in Addis Ababa). However, we later became flexible as more data became available. Moreover, as the rate of dropout was more than 30% (our initial expectation), we recruited more participants to compensate for the dropouts. As a result, we included more participants than initially calculated. During data collection, data collectors included the next nearest household if the candidate household was closed or no eligible participant was available in the index household. Only one person from each household was recruited. Inclusion criteria were age 18 years or older, written informed consent, and willingness to provide blood samples by venepuncture.

This study was done between Aug 5, 2020, and April 10, 2021, and data collection was spaced with a minimum of 4 weeks between each round. All participants were enrolled before the introduction of COVID-19 vaccines in Ethiopia.

Data collection and laboratory procedures

We collected demographic data, COVID-19-related symptoms in the preceding 6 months, and prevention practices at the first round. During subsequent rounds, participants were asked about new onset of symptoms and contact with individuals with confirmed or suspected COVID-19. We collected about 3 mL of venous blood for serology at each round using standard serum tubes. Serum specimens were processed daily and stored at –20 °C in aliquots. To ensure best reproducibility and a cost-effective operation, one aliquot was subsequently thawed and serology testing was done in batches. We did measurements with Elecsys anti-SARS-CoV-2 anti-nucleocapsid assay using the Cobas 6000 module e601 system (Roche Diagnostics, Basel, Switzerland).[18] This assay has an in-solution double-antigen sandwich format, with a reported specificity higher than 99·8% and sensitivity of 100%. It received emergency use authorisation from the US Food and Drug Administration in May, 2020.[19] Results of the serology test were communicated to all participants during all rounds via text message containing a reminder to practice the recommended COVID-19 prevention methods regardless of the result.

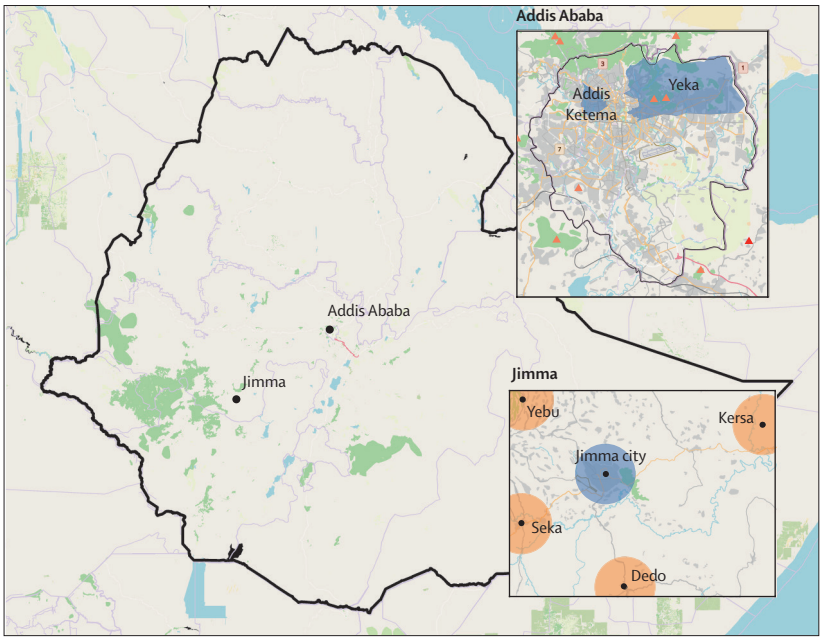


Figure 1: **Map of Ethiopia showing the study sites**
Base map reproduced from OpenStreetMap and OpenStreetMap Foundation, under the Creative Commons Attribution-ShareAlike 4.0 International License. Blue represents urban areas; orange represents rural areas.

Statistical analysis

Data were double entered into a study-specific database (EpiData Manager, version 4.6.0.0) and linked with serology data from analyser extracts. Data analysis was done in R and Python (details in appendix 3 p 1).

See Online for appendix 3

We calculated the seroprevalence of anti-SARS-CoV-2 antibodies as the number of positive cases divided by the total number of individuals tested per round. The incidence rate (IR) was calculated as the number of newly

positive cases divided by those still at risk of infection. The IR is presented as rate per 100 000 person-weeks. Only participants with at least two timepoints were included in incidence calculations (appendix 3 pp 1–2). Prevalence and IR are given along with 95% credible interval (CrI). We used the national COVID-19 daily official report of the Federal Ministry of Health of Ethiopia for COVID-19 to compare the seroprevalence changes over time.

SEIR model

We developed compartment models using a SEIR (susceptible, exposed, infectious, and recovered) approach to analyse and predict the dynamics of the pandemic, encoded using the systems biology markup language format and simulated using the software toolbox AMICI. The model parameters were inferred with a Bayesian approach, integrating our seroprevalence data with previous knowledge from the literature on the rates of disease progression. We estimated model parameters using an adaptive Metropolis Hastings algorithm implemented in the Python Parameter Estimation Toolbox. For rounds with long recruitment periods, the seroprevalence datasets were split into an early and late phase. The resulting samples from the posterior distribution were used to derive prediction and prediction CrIs (more detail about modelling is provided in appendix 3 pp 3–5).

	Hospital workers		General population			
	Jimma Medical Center (n=510)	St Paul's Hospital (n=487)	Jimma urban (n=297)	Jimma rural (n=238)	Yeka subcity (n=224)	Addis Ketema subcity (n=218)
Age	26 (24–29)	28 (25–31)	31 (25–45)	30 (25–39)	33 (28–40)	38 (30–50)
Sex						
Men	239 (46·9%)	233 (47·8%)	117 (39·4%)	158 (66·4%)	58 (25·9%)	42 (19·3%)
Women	271 (53·1%)	254 (52·2%)	180 (60·6%)	80 (33·6%)	162 (72·3%)	173 (79·4%)
Missing	0	3 (0·6%)	0	0	4 (1·8%)	3 (1·4%)
Education						
No formal education	0	0	18 (6·1%)	49 (20·6%)	32 (14·3%)	31 (14·2%)
Primary school	15 (2·9%)	44 (9·0%)	70 (23·6%)	50 (21·0%)	66 (29·5%)	113 (51·8%)
High school	134 (26·3%)	93 (19·1%)	85 (28·6%)	126 (52·9%)	58 (25·9%)	39 (17·9%)
College graduate	361 (70·8%)	350 (71·9%)	124 (41·8%)	13 (5·5%)	62 (27·7%)	26 (11·9%)
Missing	..	..	0	0	6 (2·7%)	9 (4·1%)
Profession						
Medical doctor	230 (45·1%)	199 (40·9%)	NA	NA	NA	NA
Nurse or midwife	164 (32·2%)	112 (23·0%)	NA	NA	NA	NA
Other health professionals	38 (7·5%)	62 (12·7%)	NA	NA	NA	NA
Non-clinical staff	78 (15·3%)	113 (23·2%)	NA	NA	NA	NA
Missing	0	1 (0·2%)	NA	NA	NA	NA
Routine PPE or mask use						
PPE at work	507 (99·4%)	..*	NA	NA	NA	NA
Mask use in public	479 (93·9%)	..*	244 (82·2%)	137 (57·6%)	178 (79·5%)	183 (83·9%)
Missing	0	NA	NA	NA	5 (2·2%)	10 (4·6%)

Data are median (IQR) or n (%). NA=not applicable. PPE=personal protective equipment. *Such data were not collected at baseline for hospital workers at St Paul's Hospital.

Table: **Demographic characteristics of study participants**

Role of the funding source

The funders had no role in study design, data collection, data analysis, data interpretation, writing of the manuscript, or the decision to publish.

Results

Between Aug 5, 2020, and April 10, 2021, we did three rounds of seroprevalence surveys. 1104 hospital workers and 1229 community residents participated in the study; demographic characteristics are provided in the table. Flow diagrams for recruitment and follow-up of participants are provided in figure 2.

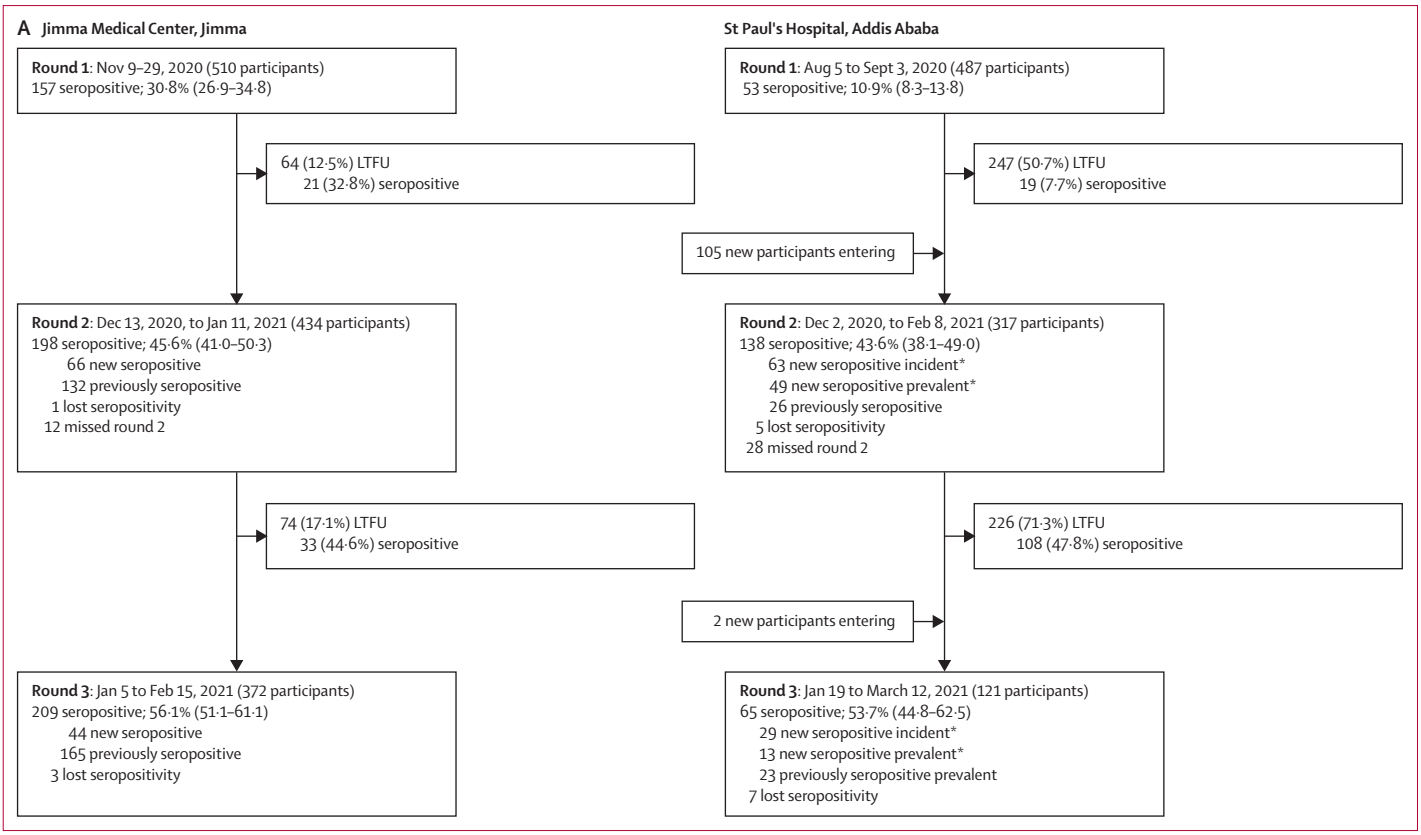
At St Paul's Hospital, serosurvey rounds were done in August and September, 2020; December, 2020, and January, 2021 (mean interval 17·7 weeks [SD 2·1] from round 1); and February and March, 2021 (mean interval 9·3 weeks [SD 1·9] from round 2). There was a high proportion of dropouts, potentially due to long survey intervals; only 51 (10·5%) of 487 individuals included in the first round completed all three rounds. Nevertheless, we did not observe significant differences in the proportions of dropouts regarding seropositivity across all cohorts and survey rounds, indicating that dropouts did not result in a sampling bias (more detail on missing data is provided in appendix 3, pp 5–7). At JMC, survey rounds were done in November, 2020; December, 2020, and January, 2021 (mean interval 5·2 weeks [SD 0·8] from round 1); and January and February, 2021 (mean interval 6·3 weeks [SD 0·9] from round 2). Dropout rates were lower than that in St Paul's Hospital—360 (70·6%) of 510 participants completed all three rounds.

Recruitment for the general population started with 297 participants from urban communities and 238 from rural communities in Jimma. The survey rounds were done in December, 2020; January and February, 2021 (mean interval 5·4 weeks [SD 0·6] in urban communities and 7·2 weeks [SD 0·9] in rural communities); and February and March, 2021 (mean interval 6·6 weeks [SD 0·9] in urban communities and 5·1 weeks [SD 1·0] in rural communities). General population survey rounds in Addis Ababa were done in December, 2020, and January, 2021; February, 2021 (mean interval 4·7 weeks [SD 1·9] in Addis Ketema and 5·1 weeks [SD 0·7] in Yeka); and April, 2021 (mean interval 8·0 weeks [SD 1·8] in Addis Ketema and 6·6 weeks [SD 0·5] in Yeka). At baseline, 224 participants from Yeka and 218 from Addis Ketema were included; however, new participants were added at later survey rounds (figure 2B).

The evolution of seroprevalence over time in different cohorts is depicted in figure 2 and figure 3A, and the corresponding incidence data are reported in appendix 3 (p 1). Differences in seroprevalence for each survey period and for seroincidence are summarised in appendix 3 (p 2). In August, 2020, SARS-CoV-2 seroprevalence among hospital workers at St Paul's Hospital was 10·9% (95% CrI 8·3–13·8), increasing to 53·7% (44·8–62·5) by February, 2021 (figure 2A). The IR over this period was 2223 per 100 000 person-weeks (95% CI 1785–2696). At JMC, the seroprevalence increased from 30·8% (95% CrI 26·9–34·8) in November, 2020, to 56·1% (51·1–61·1) in February, 2021, with an IR of 3810 per 100 000 person-weeks (95% CI 3149–4540). The seroincidence in hospital workers from Addis Ababa was significantly lower than that in Jimma (risk ratio 0·6, 95% CrI 0·4–0·7).

In the most populous area of the general population surveyed, Addis Ketema, an initial seroprevalence of 54·2% (47·5–60·7) in January, 2021, increased to 72·7% (65·9–79·1) in April, 2021 (figure 2B); in Yeka subcity, we observed an increase from 39·7% (33·4–46·3) to 54·8% (47·7–61·9) during the same timepoints. The seroprevalence in Addis Ketema was not only significantly higher than in Yeka during all rounds, but also higher than that of hospital workers at St Paul's Hospital, for example, during the December, 2020, to January, 2021 survey (odds ratio [OR] 1·5, 95% CI 1·1–2·1; appendix 3 p 2). The combined IR from both subcities was 4535 (95% CI 3372–5906) per 100 000 person-weeks, and the overall incidence was significantly higher compared with that of hospital workers at St Paul's Hospital (OR 2·0, 1·4–2·8). In Jimma Town, a seroprevalence of 32·3% (95% CrI 27·0–37·9) in December, 2020, increased to 45·2% (37·7–52·7) in February, 2021. The seroprevalence in rural communities was 18·0% (13·5–23·2) from November to December, 2020, and 31·0% (22·3–40·3) by March, 2021, which was significantly lower than in the city for the first two rounds and lower than that among hospital workers during all rounds. IRs were similar between urban and rural populations in Jimma, with a combined IR of 1720 (95% CI 1258–2258) per 100 000 person-weeks, and the overall Jimma community incidence was lower than that at JMC (OR 0·4, 95% CI 0·3–0·6). The seroincidence in communities from Addis Ababa was significantly higher than that in Jimma (2·6, 1·6–3·8; appendix 3 p 2).

When comparing the differences between rounds, we observed significant differences overall between round 2 and round 3 (OR 1·92, 95% CI 1·21–3·05). Differences between round 1 and round 2 were not significant except in St Paul's Hospital, where round 1 was done much



(Figure 2 continues on next page)

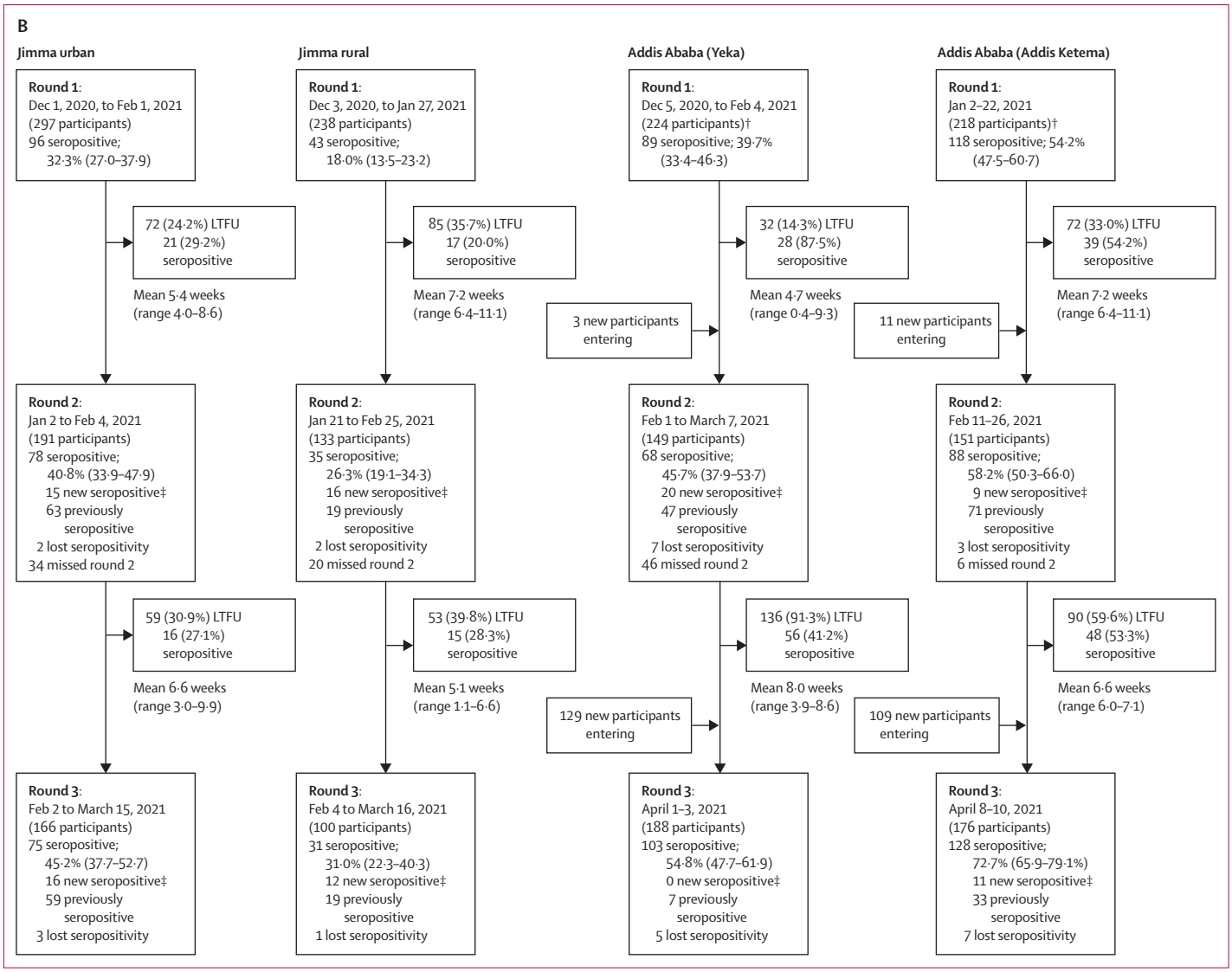


Figure 2: **Study flow and point prevalence for SARS CoV-2 seropositivity in hospital workers recruited in Jimma and Addis Ababa (A) and in participants recruited from the general population in urban and rural Jimma and Addis Ababa (B)**

Data are n, n (%), seroprevalence (% and 95% credible interval), or mean (range). LTFU=lost to follow-up. *New seropositive incident refers to seropositive cases with previous negative serology result during round 1; new seropositive prevalent refers to seropositive cases that entered the study without a preceding diagnosis. †Additional 13 participants from Addis Ababa, nine of whom participated in two rounds and four of whom only participated in one round, were included but did not have data available for subcity. ‡New seropositive refers only to participants who were negative in one or more previous rounds, but became seropositive in the subsequent round, excluding new participants entering; therefore, the sum of new and previously seropositive participants does not always equal the total number of seropositive participants in that round.

earlier than in all other cohorts, and thus the finding is likely to result from a study design effect (appendix 3 p 2).

On the basis of the results for the first two rounds, we constructed an SEIR model for the progression of the SARS-CoV-2 epidemic in Ethiopia (figure 4A). We started estimating the model parameters with the data for the hospital workers because it provided better coverage for the early dynamics and more well determined parameter estimates. The resulting model for JMC and St Paul's Hospital provided a good description of the available data for round 1 and round 2 (figure 4B) and reliable parameter estimates (appendix 3 p 3). Particularly, we obtained a median exposure rate of 0·08 per day (IQR 0·06–0·13), a median incubation period of 5·6 days (2·2–13·6) and a median recovery time of 19·3 days (11·4–28·9).

The model showed a seroprevalence approaching a predicted saturation level of 50–70%. These predictions based on the first two rounds agreed with the findings in round 3, which was found to be a seroprevalence of 53·7% for hospital workers at St Paul's Hospital from mid-January to mid-March, 2021, and 56·1% at JMC from January to February, 2021.

In addition to the standard SEIR model, we constructed a combined model using data from hospital workers and

the community (figure 4C). This model simultaneously described both groups and allowed for cross-infections. Infection of hospital workers by community members is considered more likely due to contact patterns. The extended model based on the first two rounds provided a good description of the joint datasets and also predicted saturation (figure 4D, appendix 3 p 3). As expected, the seroprevalences for the hospital workers were predicted to be higher than those in the community. Again, the model predictions of the extended model were supported by the observations in round 3. Moreover, we constructed a model considering the possible entry of a SARS-CoV-2 variant, which will be further debated in the Discussion section.

Because of the highly dynamic nature of the seroepidemiological change observed in this study, we sought to compare it with corresponding clinical effects of COVID-19 on the health-care system in Ethiopia. A strong increase in the test positivity rate for SARS-CoV-2 RT-PCR since February, 2021—reaching a high of 28·6% on April 1, 2021—was reported by the Ministry of Health. Similarly, numbers of admissions to intensive care units (ICUs) across Ethiopian hospitals passed 500 for the first time in March, 2021, and reached a peak of 1059 on April 21, 2021 (figure 3B). Clinical data for signs and symptoms of COVID-19 was collected from 1909 participants; however, only 721 (37·8%) of these participants reported having had COVID-19-related symptoms—371 (45·8%) of 810 seropositive cases and 350 (31·8%) of 1099 seronegative individuals (p<0·0001) and none were admitted to hospital due to COVID-19.

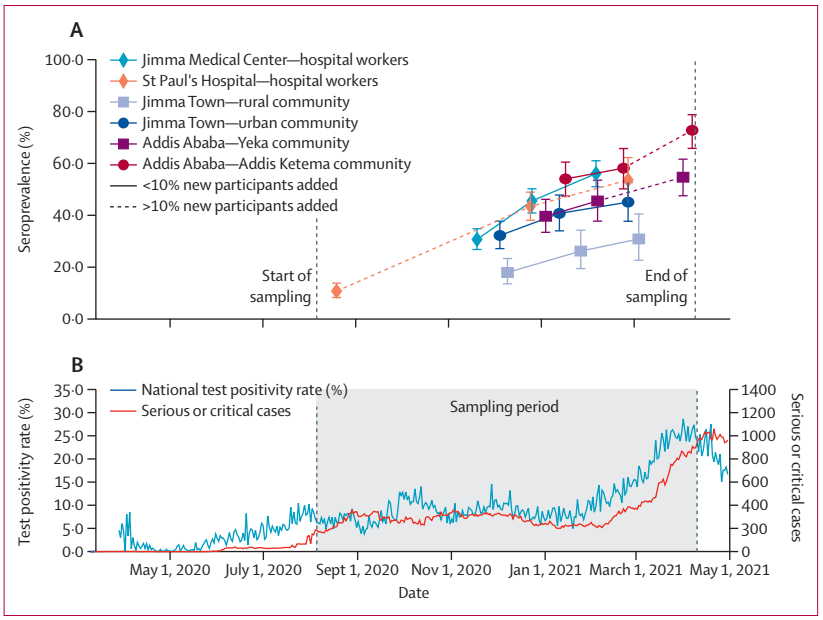


Figure 3: **Seroprevalence over time for all six cohorts investigated in the study (A), and PCR test positivity rates and number of admissions to intensive care units due to COVID-19 in Ethiopia (B)**

Discussion

Here, we provide the first data from a seroepidemiological investigation for SARS-CoV-2 infection in a population-based, longitudinal, exploratory cohort study from Ethiopia. This study revealed a striking increase in seroprevalence of SARS-CoV-2 among front-line hospital workers and communities in Ethiopia over the last months of 2020 and the first quarter of 2021. A SEIR model predicted a seroprevalence approaching saturation for hospital workers and urban communities. Although no COVID-19-related severe disease (as defined by hospitalisation) was reported among our cohorts, the substantial change in seroepidemiology in our study aligns with increased caseloads and ICU admissions in Ethiopia during the same period (figure 3).

After detection of the first few cases, Ethiopia declared a state of emergency on April 8, 2020, to contain the COVID-19 outbreak and mitigate its impact.[20] Various restrictions and prohibitions were imposed for 5 months to reinforce this, and the spread of infection appeared to be halted during that period. Two serosurveys done in April and May, 2020, among communities and outpatients in Addis Ababa reported seroprevalence of 8%[21] and 3%.[22] A seroprevalence survey done between July and September, 2020, among the general population indicated a seroprevalence lower than 1% in both Jimma Town and rural areas and 2–5% in Addis Ababa.[23]

Our first serosurvey, done from August to September, 2020, among hospital workers in Addis Ababa showed a seroprevalence of 10·9%, indicating a slow but steady spread of SARS-CoV-2 in Ethiopia even when restrictions were in place.

Ethiopia lifted the state of emergency and relaxed most restrictions on Sept 8, 2020.[24] We subsequently observed a strong increase in seroprevalence among hospital workers to 53·7% in Addis Ababa and 56·1% in Jimma Town by March, 2021. Likewise, our community seroepidemiological data from two subcities in Addis Ababa indicated an increment of combined seroprevalence to 63·7% by April, 2021, and to 45·2% by March, 2021, in Jimma Town's urban community. Notably, a lower seroprevalence was observed among rural residents during all three rounds compared with that among urban communities. Seroprevalence among hospital workers and the surrounding urban communities were similar, except for the densely populated Addis Ketema, where seroprevalence was significantly higher than that for hospital workers.

It can be speculated that the Ethiopian Government's disease control restrictions during the first few months helped in slowing down the spread of the disease. It is widely believed that the COVID-19 burden was not as heavy in African countries as in other world regions because of a younger population being less susceptible to severe disease.[25] However, the sheer increase of SARS-CoV-2 infections as observed in the second wave of the African COVID-19 pandemic probably inflicts greater health-care challenges.[26] In this respect, our data

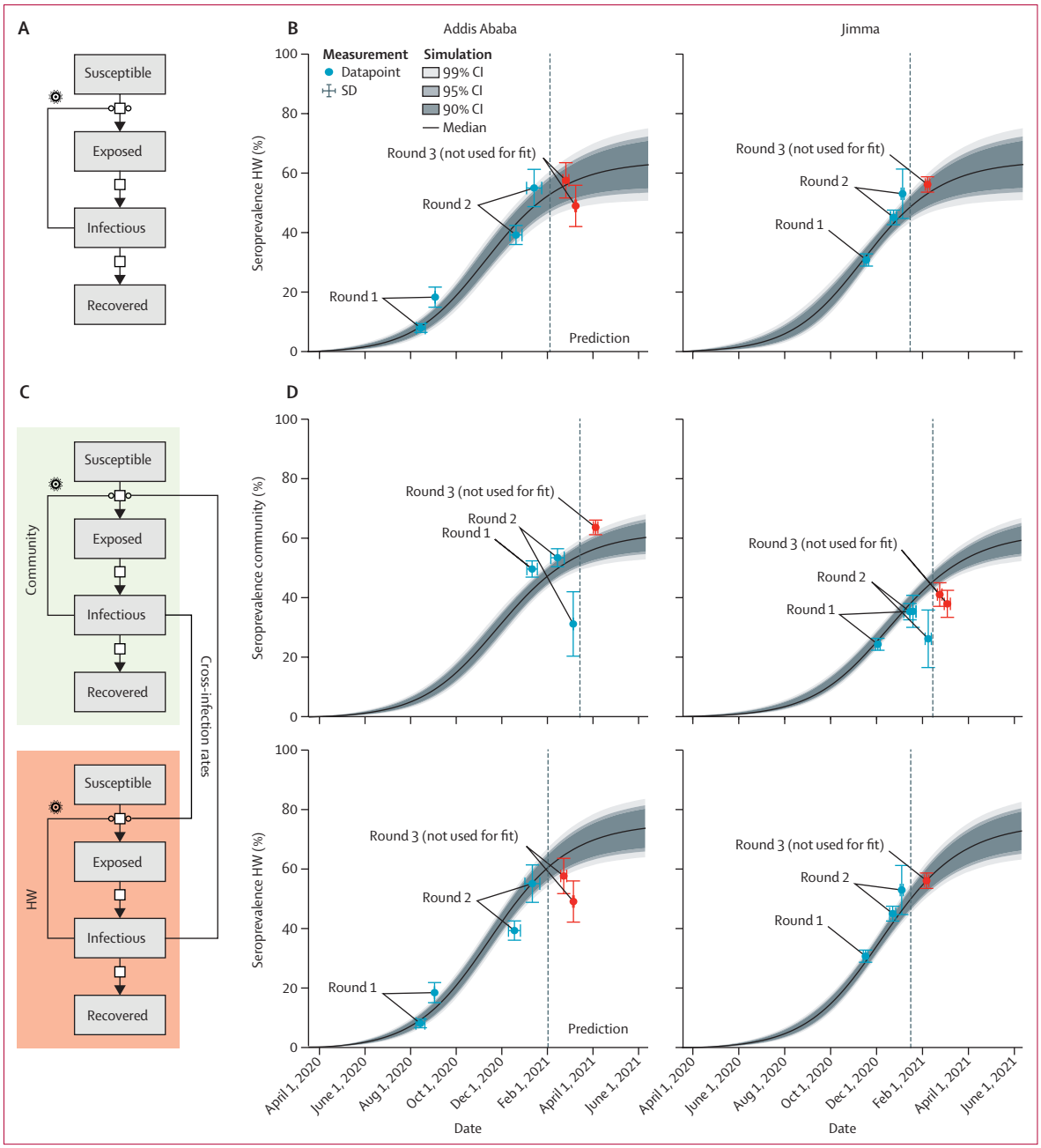


Figure 4: **SEIR model of SARS-CoV-2 epidemic in Ethiopia**
(A) Compartments of the SEIR models and possible transition. (B) Model simulation of SEIR model for HW in Jimma Medical Center and St Paul's Hospital; data from round 1 and 2 were used for model training; later points, including round 3, were predictions. (C) Compartments of the extended SEIR models and possible transition; data from round 1 and 2 were used for model training; later points, including round 3, were predictions. (D) Model simulation of extended SEIR model for HW in Jimma Medical Center and St Paul's Hospital and community members in Jimma (combined) and Addis Ababa (combined); data from round 1 and 2 were used for model training; later points, including round 3, were predictions. HW=hospital workers. SEIR=susceptible, exposed, infectious, and recovered.

are supported by the notification data obtained from the Ethiopian Government, showing great increases in SARS-CoV-2 RT-PCR test positivity rates and unprecedented numbers of ICU admissions (figure 3).

Despite a strong increment in seroprevalence, most individuals in our cohorts did not report COVID-19-related symptoms or hospital admissions. Therefore, silent transmission of SARS-CoV-2 infections in Ethiopia might be assumed for most of the population, considering also the younger age demographics compared with other world regions. However, the observed high seroincidence with no serious clinical impact can be a blessing in disguise because individuals who are asymptomatic but possibly infectious probably continue with their

working, private, and social interactions, thus creating a risk for people with predisposing risk factors for COVID-19. Conversely, the observed high transmission of SARS-CoV-2 in the community with a low number of deaths, critical cases, and hospital admissions could lead to achieving herd immunity, given that the probability of repeated infection is low.[27,28]

In this study, we were unable to determine the expected herd immunity threshold for COVID-19 in Ethiopia due to no data on the basic reproductive ratio. Instead, we did SEIR modelling to show the epidemic trajectories and indicate possible saturation points in time. The model assumed constant parameters because the intervention measures in Ethiopia were limited. Possible changes in behaviour of hospital workers and community members over the course of the pandemic and the seasonality component were not considered because they would have required additional data. However, despite these limitations, our model, based on the first two survey rounds, predicted disease saturation and assumed a related herd immunity by April, 2021. This finding was supported by the third-round serosurvey and will be further followed up in 2021 and 2022. However, the assumed herd immunity in Manaus, Brazil (suggested in September, 2020) did not prevent high COVID-19 disease burden during a subsequent wave from December, 2020, to January, 2021, possibly due to immune escape of the newly emerging gamma (P1) strain.[29]

Although our SEIR model was able to describe and even predict seroprevalence observations, it did not explain the recent surge in the positivity rate of PCR tests. Because the test strategy did not change, we speculate that this fraction should, to some degree, reflect the current number of infectious individuals. We also considered the possible entry of a SARS-CoV-2 variant capable of re-infecting individuals who had recovered from COVID-19, and we developed a compartment model describing this scenario (figure 5, appendix 3, p 4). The model provided a good description of the seroprevalence data from rounds 1 to 3 for community members and hospital workers, as well as the test rates. By contrast with the basic SEIR model, it predicted a substantially higher number of infectious individuals over the months after round 3, as well as a final seroprevalence in the range of 80–90%. This prediction suggests that herd immunity is not easily reached if re-infections with possible secondary transmissions occur.

Although infection-blocking immunity might wane rapidly or be challenged by immune escape variants, disease-reducing cross-immunity should be long lived, according to 2021 models.[30] This effect would be even stronger when providing booster vaccinations to individuals previously exposed to SARS-CoV-2. Therefore, for individuals who have recovered from COVID-19, one booster vaccination dose might be sufficient to provide longer protection. Depending on availability of test systems or shortage of vaccines in some parts of the world, it might be cost-effective and reasonable to test the population serologically before administering vaccines.

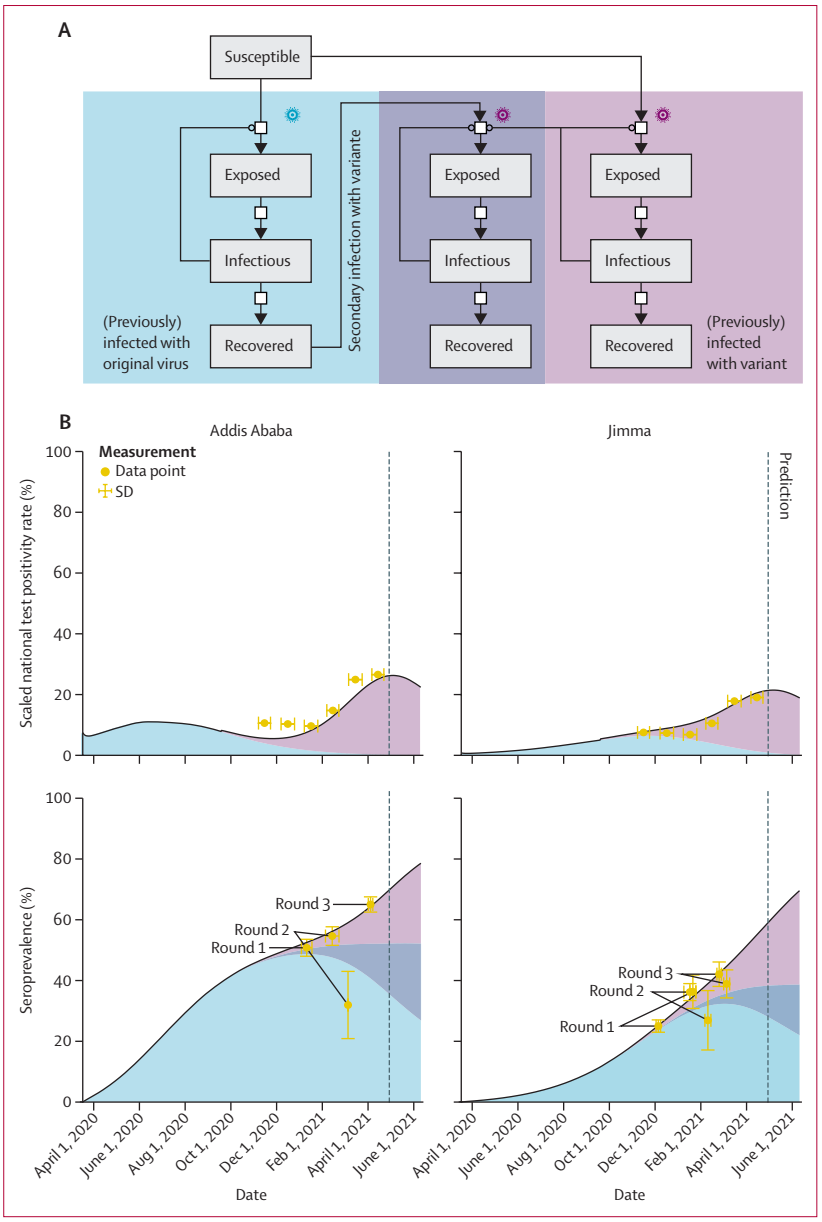


Figure 5: **SEIR model of SARS CoV-2 epidemic in Ethiopia integrating the potential effect of exposure to a SARS-CoV-2 variant with immune escape potential**

(A) Topology of compartment model that allows for the infection with the variant of individuals who were exposed to the original virus. (B) Scaled test positivity rate (mapped from the complete country to the individual cities) and seroprevalence. The contribution of different variants is indicated, as well as the proportion of individuals exposed to both. SEIR=susceptible, exposed, infectious, and recovered.

Our study was based on hospital workers at major tertiary hospitals and on residents in typical metropolitan, semi-urban, and rural settings in Ethiopia. Because of the nature of the design, our study had significant dropout during round 3 among community participants. We recruited additional participants with similar characteristics to replace those who dropped out so that

prevalence could be compared with the first two rounds. Furthermore, findings might not be generalisable to primary-level health facilities, which are the most common points of interaction, and to rural communities representing the majority of the Ethiopian population. Interpretation of our serosurvey data would have been more informative in the context of circulating viral variant characteristics. However, this information is not yet available due to the absence of tests for new variants in Ethiopia.

In conclusion, this study has shown that SARS-CoV-2 infection among hospital workers at tertiary hospitals and community residents in Ethiopia has been widespread and highly dynamic. Our SEIR model, fitted on the basis of the current trend of seroincidence and poor adherence to mitigation strategies, has shown that front-line hospital workers at tertiary hospitals were approaching a threshold for herd immunity, even before the start of the vaccination initiative. However, this pattern of disease spread poses a substantial risk to the community because silent spread among a mostly young population might ultimately put infectious pressure on highly vulnerable groups of society, leading to increased ICU admissions and deaths in the following few months. Hence, mitigation measures should target safeguarding the most vulnerable, including older people and those with underlying medical conditions.

Contributors

EKG, SA, AK, AW, and MH conceived of and designed the study. EKG, SA, EG, AGi, BT, GBH, WTS, RA, AGe, and MB participated in data collection. KE, AB, and SM summarised, cleaned, and analysed the data. SM, LC, and JH did the modelling and parameter estimation. AGi, GBH, and BT did laboratory studies. EKG, SA, AK, AW, MH, and JH interpreted the results. EKG, SA, and AK drafted the Article. All authors contributed to the writing of the final version of the manuscript. EKG, SA, KE, AB, and AK have accessed and verified the data; all authors accepted responsibility for the decision to submit for publication.

Declaration of interests

We declare no competing interests.

Data sharing

All data used are publicly available, and sources are cited throughout. The data can be accessed on https://zenodo.org/record/4885064#.YU2tGbdCSyU.

Acknowledgments

We are grateful for research funding provided by the Bavarian State Ministry of Sciences, Research and the Arts (Bayerisches Staatsministerium, F.4-V0122.4/3/20); the Germany Ministry of Education and Research (MoKoCo19; 01KI20271); the EU Horizon 2020 programme (ORCHESTRA; 101016167); Deutsche Forschungsgemeinschaft (SEPAN; HA 7376/3-1); and Volkswagenstiftung (E2; 99 450). We thank participants, study teams, Jimma Medical Center, Oromia Regional Health Bureau, St Paul's Hospital, and Addis Ababa Health Bureau for the support provided during data collection. We would also like to thank Michel Pletschette from Ludwig Maximilian University for his support in the literature review and revision.

Editorial note: the *Lancet* Group takes a neutral position with respect to territorial claims in published maps and institutional affiliations.

References

1 Massinga Loembé M, Tshangela A, Salyer SJ, Varma JK, Ouma AEO, Nkengasong JN. COVID-19 in Africa: the spread and response. *Nat Med* 2020; **26:** 999–1003.
2 Gudina EK, Gobena D, Debela T, et al. COVID-19 in Oromia Region of Ethiopia: a review of the first 6 months' surveillance data. *BMJ Open* 2021; **11:** e046764.
3 African Centers for Disease Control and Prevention. Coronavirus disease 2019 (COVID-19). https://africacdc.org/covid-19/ (accessed May 30, 2020).
4 Johns Hopkins University. COVID-19 dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. https://coronavirus.jhu.edu/map.html (accessed June 9, 2021).
5 Gudina EK, Tesfaye M, Siraj D, Haileamilak A, Yilma D. COVID-19 in Ethiopia in the first 180 days: lessons learned and the way forward. *Ethiop J Health Dev* 2020; **34:** 6.
6 Mulu A, Bekele A, Abdissa A, et al. The challenges of COVID-19 testing in Africa: the Ethiopian experience. *Pan Afr Med J* 2021; **38:** 6.
7 Colombo S, Scuccato R, Fadda A, Cumbi AJ. COVID-19 in Africa: the little we know and the lot we ignore. *Epidemiol Prev* 2020; **44** (suppl 2)**:** 408–22.
8 Mukwege D, Byabene AK, Akonkwa EM, et al. High SARS-CoV-2 seroprevalence in healthcare workers in Bukavu, eastern Democratic Republic of Congo. *Am J Trop Med Hyg* 2021; **104:** 1526–30.
9 Olayanju O, Bamidele O, Edem F, et al. SARS-CoV-2 seropositivity in asymptomatic frontline health workers in Ibadan, Nigeria. *Am J Trop Med Hyg* 2021; **104:** 91–94.
10 Sykes W, Mhlanga L, Swanevelder R, et al. Prevalence of anti-SARS-CoV-2 antibodies among blood donors in Northern Cape, KwaZulu-Natal, Eastern Cape, and Free State provinces of South Africa in January 2021. *Res Sq* 2021; published online Feb 12. https://doi.org/10.21203/rs.3.rs-233375/v1 (preprint).
11 WHO. First case of COVID-19 confirmed in Ethiopia. 2020. https://www.afro.who.int/news/first-case-covid-19-confirmed-ethiopia (accessed May 30, 2021).
12 Ethiopian Public Health Institute. Interim national strategy and guidance for the laboratory diagnosis of COVID-19 in Ethiopia. Addis Ababa: Ethiopian Public Health Institute, 2020.
13 Alene M, Yismaw L, Assemie MA, et al. Magnitude of asymptomatic COVID-19 cases throughout the course of infection: a systematic review and meta-analysis. *PLoS One* 2021; **16:** e0249090.
14 National Public Health Emergency Operation Center of Ethiopia. COVID-19 pandemic preparedness and response in Ethiopia. Ethiopian Public Health Institute. https://ephi.gov.et/download/pheoc/ (accessed May 30, 2021).
15 Endris BS, Saje SM, Metaferia ZT, et al. Excess mortality in the face of COVID-19: evidence from Addis Ababa Mortality Surveillance Program. *SSRN* 2021; published online Feb 17. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3787447 (preprint).
16 Peeling RW, Wedderburn CJ, Garcia PJ, et al. Serology testing in the COVID-19 pandemic response. *Lancet Infect Dis* 2020; **20:** e245–49.
17 Winter AK, Hegde ST. The important role of serology for COVID-19 control. *Lancet Infect Dis* 2020; **20:** 758–59.
18 Roche Diagnostics. Elecsys® Anti-SARS-CoV-2: immunoassay for the qualitative detection of antibodies (incl IgG) against SARS-CoV-2. 2020. https://diagnostics.roche.com/global/en/products/params/elecsys-anti-sars-cov-2.html (accessed May 28, 2021).
19 Roche Diagnostics. Roche's COVID-19 antibody test receives FDA Emergency Use Authorization and is available in markets accepting the CE mark. 2020. https://www.roche.com/media/releases/med-cor-2020-05-03.htm (accessed May 28, 2021).
20 Council of Minister of Ethiopia. State of Emergency Proclamation No. 3/2020: Implementation Regulation No. 466/2020. A regulation issued to implement the state of emergency proclamation enacted to counter and control the spread of COVID-19 and mitigate its impacts. 2020. https://www.moh.gov.et/ejcc/sites/default/files/2020-04/negarit.pdf (accessed June 2, 2021).
21 Alemu BN, Addissie A, Mamo G, et al. Sero-prevalence of anti-SARS-CoV-2 antibodies in Addis Ababa, Ethiopia. *bioRxiv* 2020; published online Oct 13. https://doi.org/10.1101/2020.10.13.337287 (preprint).
22 Kempen JH, Abashawl A, Suga HK, et al. SARS-CoV-2 serosurvey in Addis Ababa, Ethiopia. *Am J Trop Med Hyg* 2020; **103:** 2022–23.

23 Abdella S, Riou S, Tessema M, et al. Prevalence of SARS-CoV-2 in urban and rural Ethiopia: randomized household serosurveys reveal level of spread during the first wave of the pandemic. *EClinicalMedicine* 2021; **35:** 100880.

24 Ethiopian Public Health Institute. A directive issued for the prevention and control of the COVID-19 pandemic No. 30/2020. 2020. https://www.ephi.gov.et/images/Registerd-COVID-19-Directive-2013_Final_051020.pdf (accessed May 30, 2021).

25 Diop BZ, Ngom M, Pougué Biyong C, Pougué Biyong JN. The relatively young and rural population may limit the spread and severity of COVID-19 in Africa: a modelling study. *BMJ Glob Health* 2020; **5:** e002699.

26 Salyer SJ, Maeda J, Sembuche S, et al. The first and second waves of the COVID-19 pandemic in Africa: a cross-sectional study. *Lancet* 2021; **397:** 1265–75.

27 Lumley SF, O'Donnell D, Stoesser NE, et al. Antibody status and incidence of SARS-CoV-2 infection in health care workers. *N Engl J Med* 2021; **384:** 533–40.

28 Hansen CH, Michlmayr D, Gubbels SM, Mølbak K, Ethelberg S. Assessment of protection against reinfection with SARS-CoV-2 among 4 million PCR-tested individuals in Denmark in 2020: a population-level observational study. *Lancet* 2021; **397:** 1204–12.

29 Sabino EC, Buss LF, Carvalho MPS, et al. Resurgence of COVID-19 in Manaus, Brazil, despite high seroprevalence. *Lancet* 2021; **397:** 452–55.

30 Lavine JS, Bjornstad ON, Antia R. Immunological characteristics govern the transition of COVID-19 to endemicity. *Science* 2021; **371:** 741–45.

THE LANCET
Global Health

Supplementary appendix 3

This appendix formed part of the original submission and has been peer reviewed. We post it as supplied by the authors.

Supplement to: Gudina EK, Ali S, Girma E, et al. Seroepidemiology and model-based prediction of SARS-CoV-2 in Ethiopia: longitudinal cohort study among front-line hospital workers and communities. *Lancet Glob Health* 2021; **9:** e1517–27.

Supplementary Material

1. Prevalence and Incidence Estimation

Supplementary methods

A two-way, repeated measures ANOVA model was used to examine the effect of population group and time point on the number of seropositive individuals. The variability explained by the model is divided into two factors: Group (Jimma vs. Addis Ababa and HCW vs Community; between-subjects factor indicating population group) and Round (1, 2, 3; within-subjects factor denoting time point of the serology test); and an interaction term Group:Round testing whether the effect of Round and Group jointly influences the seropositive count, i.e. if some groups have a differential effect in specific rounds.

$$y = \alpha + \beta_1 Group + \beta_2 Round + \beta_3 Group{:}Round + \varepsilon$$

In this case, y refers to the count of seropositives (pos) within each group and survey round and we have the following equations to estimate p_i which is the probability of positive in a group and round.

$$pos_i \sim Binomial(n_i, p_i)$$

$$logit(p_i) = \alpha + \beta_1 Group + \beta_2 Round + \beta_3 Group{:}Round$$

$$\alpha, \beta_i \sim Normal(0{,}10)$$

In addition to considering binomial outcomes, we also examined the count of seropositives assuming a Poisson outcome distribution. The equations are similar to the binomial distribution except for the need to have an offset variable adjusting for the denominator to estimate the rate λ_i for being positive.

$$pos_i \sim Normal(\lambda_i)$$

$$log(\lambda_i) = \alpha + \beta_1 Group + \beta_2 Visit + \beta_3 Group{:}Visit + offset(n_i)$$

$$\alpha, \beta_i \sim Normal(0{,}10)$$

Estimates of the counts along with the 95% Credible Intervals were obtained using non-informative priors (normal distribution with mean zero and standard deviation 10) with 5000 warm-up samples followed by 5000 MCMC samples for the posterior outcome of a generalized linear model using the brms (Bayesian Regression Models using 'Stan') package in R.[1,2,3] The prevalence estimates are obtained by dividing the estimated count for positives by the observed samples. In case of the sero-incidence measures we have the count of new positives instead of positives and there is no component of round. Instead, the denominator is person-weeks of being observed within the study. The above models were also used to estimate the contrasts to check group wise and/or round wise differences. We published the code and tables used in this paragraph at Zenodo.[4]

Supplementary results for incidence and prevalence estimation

Table S1: SARS-CoV-2 seroincidence rates per person-weeks for HCW at Jimma Medical Center and St. Paul's Hospital, and communities from Jimma and Addis Ababa

	New seropositives (N)	Person-weeks	Seroincidence rates per 100,000 person-weeks (95% CI)
HCW Jimma Medical Center	111	2913	3810 (3149, 4540)
HCW St· Paul's Hospital	90	4051	2223 (1785, 2696)
Jimma Community Combined	44	2556	1720 (1258, 2258)
Jimma Rural	23	1261	1824 (1157, 2727)
Jimma Urban	21	1295	1622 (1004, 2479)
Addis Ababa Community Combined*	46	1017	4535 (3372, 5906)
Yeka sub-city	24	557	4309 (2761, 6412)
Addis Ketema sub-city	19	409	4646(2797, 7255)

**New seropositives and person-weeks from Yeka and Addis Ketema sub-cities do not add up due to missing data for sub-city.*
CI – Credible Interval; HCW – Healthcare worker

Between cohorts, we observed statistically significant differences for seroincidence and seroprevalence during different survey periods (Table S2). For seroprevalence over time, we do not see much difference between Round 1 and Round 2 except for Addis HCW, which is by design and expected. However, the difference to Round 3 is statistically significant (Table S3).

Table S2: Difference in the seroincidence and seroprevalence during survey periods between communities and health care workers (HCW) observed in Addis Ababa and Jimma.

Seroincidence (HCW)	**RR (95% CI)†**
Addis community versus Jimma community	2·6 (1·6; 3·8)*
Addis HCW versus Jimma HCW	0·6 (0·4; 0·7)*
Addis community versus Addis HCW	2·0 (1·4; 2·8)*
Jimma community versus Jimma HCW	0·4 (0·3; 0·6)*
Seroprevalence (Addis Ababa)	**OR (95% CI)†**
December 2020/January 2021	
Addis Ketema (R1) versus Addis Yeka (R1)	1·8 (1·2; 2·6)*
Addis Ketema (R1) versus Addis HCW (R2)‡	1·5 (1·1; 2·1)*
Addis Yeka (R1) versus Addis HCW (R2)‡	0·8 (0·6; 1·2)
February 2021/March 2021	
Addis Ketema (R2) versus Addis Yeka (R2)	1·6 (1·0; 2·5)*
Addis Ketema (R2) versus Addis HCW (R3)‡	1·2 (0·8; 2·1)
Addis Yeka (R2) versus Addis HCW (R3)‡	0·7 (0·4; 1·1)
April 2021	
Addis Ketema (R3) versus Addis Yeka (R3)‡	2·2 (1·3; 3·3)*
Seroprevalence Jimma	
November 2020/December 2021	
Jimma City (R1) versus Jimma Rural (R1)	2·2 (1·4; 3·2)*
Jimma City (R1) versus HCW (R1)	1·1 (0·7; 1·4)
Jimma Rural (R1) versus HCW (R1)	0·5 (0·3; 0·7)*
January 2021/February 2021	
Jimma City (R2) versus Jimma Rural (R2)	1·9 (1·1; 3·0)*
Jimma City (R2) versus Jimma HCW (R2)	0·8 (0·6; 1·1)
Jimma Rural (R2) versus Jimma HCW (R2)	0·4 (0·3; 0·6)*
February 2021/March	
Jimma City (R3) versus Jimma Rural (R3)	1·8 (0·9; 2·9)
Jimma City (R3) versus Jimma HCW (R3)	0·6 (0·4; 0·9)*
Jimma Rural (R3) versus Jimma HCW (R3)	0·3 (0·2; 0·5)*

†Estimate –ratio for the comparison of the contrasts, RR=risk ratio for seroincidence, OR=odds ratio for seroprevalence
* Statistically significant; R= survey round
Note: in order to compare seroprevalences between cohorts, we applied periods instead of round. This distinction was made as in Addis Ababa survey rounds in HCW did not match those of the corresponding communities in terms of time periods (initiated with ‡). In April, no matching HCW information from Addis was available.

Table S3: Difference in the seroprevalence over the different rounds for the overall population and by cohort

Effects	**Odds Ratio***	**Lower 95%CI**	**Upper 95% CI**	**Statistically significant difference**
Intercept	1·403	1·020	1·937	-
Yeka Sub-city	0·597	0·379	0·935	Yes
Jimma City	0·490	0·316	0·755	Yes
Jimma Rural	0·252	0·152	0·414	Yes
Jimma Medical Center	0·598	0·410	0·865	Yes
St· Paul's Hospital Addis	0·549	0·371	0·814	Yes
Overall Round 1	0·840	0·551	1·259	No
Yeka Sub-cityRound1 (interaction)	0·935	0·523	1·704	No
Jimma City Round1 (interaction)	0·825	0·470	1·443	No
Jimma Rural Round1 (interaction)	0·737	0·386	1·433	No
Jimma Medical Center Round1 (interaction)	0·631	0·387	1·031	No
St· Paul's Hospital Addis Round1 (interaction)	0·187	0·108	0·323	Yes
Overall Round 3	1·918	1·213	3·047	Yes
AddisYeka Round3 (interaction)	0·755	0·403	1·414	No
Jimma City Round3 (interaction)	0·624	0·337	1·161	No
Jimma Rural Round3 (interaction)	0·657	0·318	1·359	No
Jimma Medical Center Round3 (interaction)	0·798	0·464	1·369	No
St· Paul's Hospital Addis Round3 (interaction)	0·788	0·420	1·460	No

*Round 2 is reference category; Addis Ketema is reference site

In the above table, we see that the interaction effects are not significantly different except for the Round 1 at St. Paul's Hospital (Addis Ababa), which is a design effect. Overall, ignoring the interaction effect, we observed no significant difference between Round 1 and Round 2; however, Round 3 compared to Round 2 had an overall increase (OR 1·918 with 95% Credible Interval (1·213-3·047)). We also observe that within Round 2, Addis Ketema sub-city had the highest seroprevalence as compared all the other cohort groups.

2. The Models

We considered three different models for the analysis of the virus spread in Ethiopia: A simple SEIR model (which was applied separately to data for *healthcare workers (H)* or *community members (C)*), an extended SEIR model which simultaneously described the populations for healthcare worker and community members, and an SEIR model which

allows for the original virus (wt) and a virus variant (va). We chose SEIR models due to their widespread use for the study of the Covid-19 progression,[5–9] which facilitates a comparison to related work. Furthermore, we established earlier a comprehensive analysis pipeline for these types of models.[10] In all these models, the populations are split into *Susceptible (S), Exposed (E), Infectious (I)* and *Recovered (R).* To compare the model simulations to the observed seroprevalence, we compute the ratio of recovered to total population.

a) SEIR model

The model structure is depicted in Figure 4A and the corresponding ordinary differential equations (ODEs) for the time-dependent size of the compartments are:

$$\begin{aligned}
\dot{S} &= -\frac{\beta\, I}{N} S & \mathrm{S}(0) &= 510 \\
\dot{\mathrm{E}} &= \frac{\beta \mathrm{I}}{\mathrm{N}} \mathrm{S} - \kappa \mathrm{E} & \mathrm{E}(0) &= 0 \\
\dot{I} &= \kappa E - \gamma I & \mathrm{I}(0) &= I_0 \\
\dot{R} &= \gamma I & R(0) &= 0 \\
N &= S + E + I + R. &&
\end{aligned}$$

The parameters are listed in Table S4. This table includes the respective names in the PEtab model which we published at Zenodo.[4]

Table S4: Parameters of the SEIR model. Some depend on study site, i.e. Jimma and Addis Ababa.

Parameter	Description	Sampling result - Median (CI 95%)	Scale used for sampling	Prior (in scale)	Est. Start Sampling	Unit
β	Exp. rate	0·08 (0·06, 0·13)	log_{10}	$\mathcal{U}(-5,1)$	0·09	$\frac{1}{\text{day}}$
κ^{-1}	Inc. period	5·6 (2·2, 13·6)	log	$\mathcal{N}(1\cdot 63, 0\cdot 50)$	5·0	days
γ^{-1}	Rec. time	19·3 (11·4, 28·9)	linear	$\mathcal{N}(15\cdot 7, 6\cdot 7)$	15·0	days
I_0	Initial inf.	J: 1·1 (0·3, 3·1) A: 1·2 (0·4, 2·9)	log_{10}	$\mathcal{U}(-1,3)$	J: 0·74 A: 6·5	-

b) Extended SEIR model for two populations

In addition to the dynamics of the individual populations, we account for their interaction: Infectious healthcare workers can expose community members and vice versa. Virus transmission from community members to healthcare workers is supposed to be more probable, which is modeled by a factor $\alpha > 1$. The model structure can be seen in Figure 4C and the ODEs are:

$$\begin{aligned}
\dot{S_H} &= -\frac{\beta(I_H + \alpha I_C)}{N} S_H & S_H(0) &= 510 \\
\dot{E_H} &= \frac{\beta I_H}{N} S_H - \kappa E_H & E_H(0) &= 0 \\
\dot{I_H} &= \kappa E_H - \gamma I_H & I_H(0) &= 0 \\
\dot{R_H} &= \gamma I_H & R_H(0) &= 0 \\
\dot{S_C} &= -\frac{\beta(I_H + I_C)}{N} S_C & S_C(0) &= 100000 \\
\dot{E_C} &= \frac{\beta I_C}{N} S_C - \kappa E_C & E_C(0) &= 0 \\
\dot{I_C} &= \kappa E_C - \gamma I_C & I_C(0) &= I_0 \\
\dot{R_C} &= \gamma I_C & R_C(0) &= 0 \\
N &= S_H + E_H + I_H + R_H & & \\
&\quad + S_C + E_C + I_C + R_C. & &
\end{aligned}$$

The parameters are listed in Table S5. This table includes the respective names in the PEtab model which we published at Zenodo.[4.] All initial states which are not mentioned in the table are set to 0.

Table S5: Parameters of the extended SEIR model. Some depend on study site, i.e. Jimma and Addis Ababa.

Parameter	Description	Sampling result - Median (CI 95%)	Scale used for sampling	Prior (in scale)	Est. Start Sampling	Unit
β	Exp. rate	0·08 (0·06, 0·10)	log_{10}	$\mathcal{U}(-5,1)$	0·08	$\frac{1}{\text{day}}$
κ^{-1}	Inc. period	5·4 (2·6, 11·0)	log	$\mathcal{N}(1\cdot 63, 0\cdot 50)$	5·7	days
γ^{-1}	Rec. time	19·8 (14·9, 26·3)	linear	$\mathcal{N}(15\cdot 7, 6\cdot 7)$	18·5	days
α	Increased HCW risk	1·5 (1·3, 1·7)	log_{10}	$\mathcal{U}(-1,3)$	1·5	-
I_0	Initial inf.	J: 131·4 (56·8, 293·3) A: 204·3 (96·7, 428·2)	log_{10}	$\mathcal{U}(-1,3)$	J: 121·9 A: 189·9	-

c) Virus variant model

This model accounts for the possibility that a virus variant altered characteristics is present in Ethiopia. As sequencing data are missing, we assume the variant to appear at an unknown time t_0 and has a reproduction rate increased by a factor of 1·35, which is in the range of increase observed for variants such as B·1·1·7 and B·1·351. We account for the increase by reducing the recovery rate.[11] Moreover we assume previous variant infections make individuals immune to wild type infections but not the other way around.

The model structure is depicted Figure 5A and the ODEs are:

$$
\begin{aligned}
\dot{S} &= -\frac{\beta I_{wt}}{N} S - \frac{\beta (I_{va} + I_{va}^{wt})}{N} S && S(0) = 510 \\
\dot{E_{wt}} &= \frac{\beta I_{wt}}{N} S - \kappa E_{wt} && E_{wt}(0) = 0 \\
\dot{E_{va}} &= \frac{\beta (I_{va} + I_{va}^{wt})}{N} S - \kappa E_{va} && E_{va}(0) = 0 \\
\dot{E_{va}^{wt}} &= \frac{\beta (I_{va} + I_{va}^{wt})}{N} R_{wt} - \kappa E_{va}^{wt} && E_{va}^{wt}(0) = 0 \\
\dot{I_{wt}} &= \kappa E_{wt} - \gamma I_{wt} && I_{wt}(0) = I_0 \\
\dot{I_{va}} &= \kappa E_{va} - \frac{\gamma}{1.35} I_{va} && I_{va}(t_0) = 1 \\
\dot{I_{va}^{wt}} &= \kappa E_{va}^{wt} - \frac{\gamma}{1.35} I_{va}^{wt} && I_{va}^{wt}(0) = 0 \\
\dot{R_{wt}} &= \gamma I_{wt} - \frac{\beta (I_{va} + I_{va}^{wt})}{N} R_{wt} && R_{wt}(0) = 0 \\
\dot{R_{va}} &= \frac{\gamma}{1.35} I_{va} && R_{va}(0) = 0 \\
\dot{R_{va}^{wt}} &= \frac{\gamma}{1.35} I_{va}^{wt} && R_{va}^{wt}(0) = 0 \\
N &= S + E_{wt} + E_{va} + E_{va}^{wt} + I_{wt} \\
&\quad + I_{va} + I_{va}^{wt} + R_{wt} + R_{va} + R_{va}^{wt}.
\end{aligned}
$$

The parameters are listed in Table S6. This table includes the respective names in the PEtab model which we published at Zenodo.[4]

Table S6: Parameters of the virus variant model. Some depend on study site, i.e. Jimma and Addis Ababa.

Parameter	Description	Sampling result - Median (CI 95%)	Scale used for sampling	Prior (in scale)	Est. Start Sampling	Unit
β	Exp. rate	0·08 (0·06, 0·10)	log_{10}	$\mathcal{U}(-5, 1)$	0·08	$\frac{1}{\text{day}}$
κ^{-1}	Inc. period	5·0 (2·4, 10·0)	log	$\mathcal{N}(1 \cdot 63, 0 \cdot 50)$	5·3	days
γ^{-1}	Rec. time	16·7 (12·9, 22·1)	linear	$\mathcal{N}(15 \cdot 7, 6 \cdot 7)$	17·2	days
t_0	Entry va	184·5 (152·6, 231·3)	linear	$\mathcal{U}(150, 360)$	170·3	days
s_{TPR}	Scaling nat. TPR	J: 2·3 (1·5, 3·6) A: 2·8 (1·7, 4·3)	log_{10}	$\mathcal{U}(-1, 3)$	J: 2·3 A: 2·7	-
I_0	Initial inf.	J: 1·8 (0·6, 4·9) A: 13·8 (3·6, 42·5)	log_{10}	$\mathcal{U}(-1, 3)$	J: 2·2 A: 16·2	-

d) Calibration workflow

The models were encoded using the Systems Biology Markup Language (SBML)[12] and the Parameter estimation problems were formulated using the Parameter Estimation table (PEtab)[13] standard. The two community standards allow for the direct reproduction of the result in various software tools.

For parameter estimation, the seroprevalence data for each site, round and study group was each split by month of their collection and then accumulated on the mean date respectively. Standard deviations were calculated assuming binomial distribution in a similar way as described in the paragraph *Prevalence and Incidence Estimation* of this section. The seroprevalence measurement is assumed to not distinguish between infection with original virus or variant. In addition to seroprevalence information, we used for the virus variant model also information about national test positivity rates (TPR). As over a long time the number of test and test strategies remained unchanged, we assumed that the TPR is roughly proportional to the sum of exposed and infectious individuals in the different groups and location. For incubation and recovery times we used priors from literature.[14,15]

Bayesian parameter estimation was performed using the adaptive Metropolis-Hastings algorithm methods implemented in the parameter estimation toolbox pyPESTO[16]. Selected results were confirmed using pyMC3. Simulation was performed using the simulation toolbox AMICI[17]. The sampling results were post-processed, e.g. by removing the burn-in, and convergence was assessed visually and using the Geweke test.

Supplementary Results for model prediction

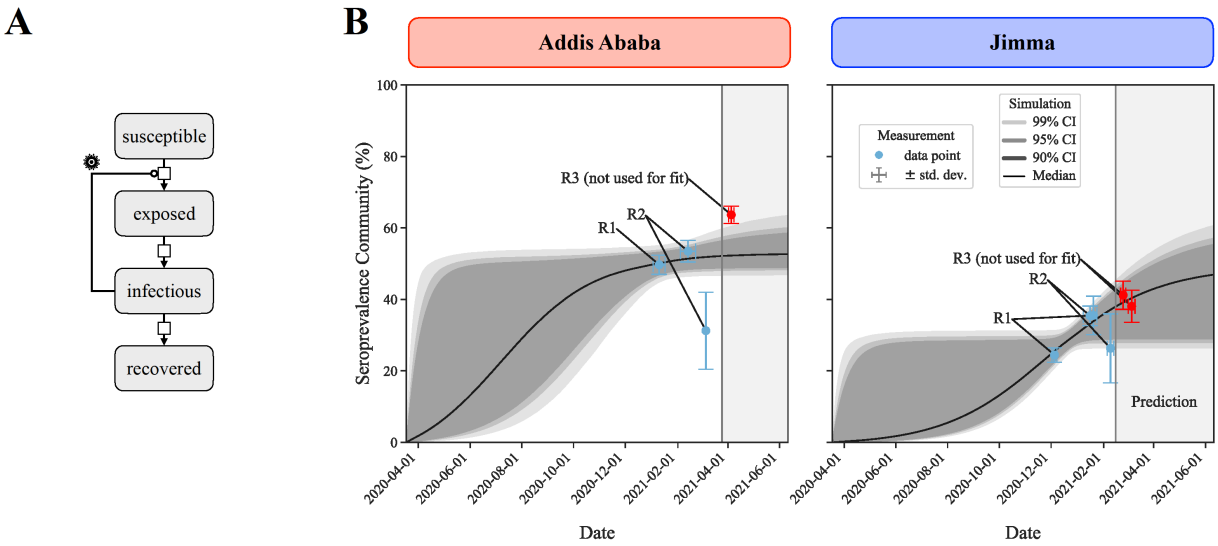
The parameter sampling for the SEIR model with healthcare workers data was performed with a sample size of 1e6. Convergence of parameters was achieved after a burn in of 5e4 samples.

The parameter sampling for the extended SEIR model for two populations with combined healthcare workers and community data was performed with a sample size of 1e5. Convergence of parameters was achieved without any burn.
The parameter sampling for the virus variant model with combined community and national TPR data was performed with a sample size of 1e5. Convergence of parameters was achieved after a burn in of 1e4 samples.

The parameter sampling for the SEIR model with combined community members data was performed with a sample size of 1e6. Since the parameters showed alternating behaviour between two models, we refrained from conducting prediction simulations based on this model-data combination.

The parameter sampling for the SEIR model with combined community members data was performed with a sample size of 1e6. Since the parameters showed alternating behaviour between two models, we refrained from conducting prediction simulations based on this model-data combination. For completeness we included these prediction results as Figure S1.

Figure S1: SEIR model of SARS-CoV-2 epidemic in Ethiopia.



(A) Compartments of the SEIR models and possible transition. (**B**) Model simulation for Community members in Jimma Medical Center and St. Paul's Hospital. Data from the 1st and 2nd round was used for model training. Later points, including the 3rd round, were predictions.

3. Information on missing data

The following tables describe the numbers and percentages of missing data between rounds (A. between Round 1 and Round 2; B (between Round 2 and Round 3); C. between Round 1 and Round 3) and for different cohorts (1. HCW Jimma, 2. urban and rural community combined for Jimma, C. HCW Addis Ababa, D. Addis community combined (Ketema and Yeka). Overall, dropout rates are higher, especially in Addis Ababa as compared to Jimma. However, dropout rates do not significantly differ between seropositive and seronegative population, which indicates that there was no sampling bias over the entire period of the study.

1. Jimma Health Care Workers (HCW) Missing Data by Result

A

	Round 2				
Round 1	**Negative**	**Positive**	**Missing**	**(all)**	**Round 2 Missing %**
Negative	235	66	52	353	14·73%
Positive	1	132	24	157	15·29%
(all)	236	198	76	510	

B

	Round 3				
Round 2	**Negative**	**Positive**	**Missing**	**(all)**	**Round 3 Missing %**
Negative	152	43	41	236	17·37%
Positive	3	162	33	198	16·67%
Missing	7	5	64	76	
(all)	162	210	138	510	

C

		Round 3				
Round 1	**Round 2**	**Negative**	**Positive**	**Missing**	**(all)**	**Round 3 Missing %**
Negative	Negative	151	43	41	235	17·45%
Negative	Positive	1	55	10	66	15·15%
Negative	Missing	7	2	43	52	
Negative	(all)	159	100	94	353	26·63%
Positive	Negative	1	0	0	1	
Positive	Positive	2	107	23	132	17·42%
Positive	Missing	0	3	21	24	
Positive	(all)	3	110	44	157	28·03%
(all)	(all)	162	210	138	510	

2. Jimma Community (combined Jimma City and Jimma urban)

A

	Round 2				Round 2 Missing %
Round 1	**Negative**	**Positive**	**Missing**	**(all)**	
Negative	207	31	158	396	39·90%
Positive	4	82	53	139	38·13%
(all)	211	113	211	535	

B

	Round 3				Round 3 Missing %
Round 2	Negative	Positive	Missing	(all)	
Negative	124	6	81	211	38·39%
Positive	4	78	31	113	27·43%
Missing	32	22	157	211	
(all)	160	106	269	535	

C

		Round 3				Round 3 Missing %
Round 1	**Round 2**	**Negative**	**Positive**	**Missing**	**(all)**	
Negative	Negative	121	6	80	207	38·65%
Negative	Positive	1	19	11	31	35·48%
Negative	Missing	32	7	119	158	
Negative	(all)	154	32	210	396	53·03%
Positive	Negative	3	0	1	4	
Positive	Positive	3	59	20	82	24·39%
Positive	Missing	0	15	38	53	
Positive	(all)	6	74	59	139	42·45%
(all)	(all)	160	106	269	535	

3. Addis Health Care Workers (HCW) Missing Data by Result

A

	Round 2				
Round 1	**Negative**	**Positive**	**Missing**	**(all)**	**% Missing Round 2**
Negative	103	53	275	431	63·81%
Positive	5	22	25	52	48·08%
Missing	56	48	0	104	
(all)	164	123	300	587	

B

	Round 3				
Round 2	**Negative**	**Positive**	**Missing**	**(all)**	**% Missing Round 3**
Negative	28	27	109	164	66·46%
Positive	6	22	95	123	77·24%
Missing	18	13	269	300	
(all)	52	62	473	587	

C

		Round 3				
Round 1	**Round 2**	**Negative**	**Positive**	**Missing**	**(all)**	**% Missing Round 3**
Negative	Negative	19	12	72	103	69·90%
Negative	Positive	4	4	45	53	84·91%
Negative	Missing	17	11	247	275	89·82%
Negative	(all)	40	27	364	431	84·45%
Positive	Negative	0	1	4	5	
Positive	Positive	0	4	18	22	81·82%
Positive	Missing	1	2	22	25	88·00%
Positive	(all)	1	7	44	52	84·62%
Missing	Negative	9	14	33	56	
Missing	Positive	2	14	32	48	
Missing	(all)	11	28	65	104	
(all)	(all)	52	62	473	587	

4. Addis Community (combined for Ketema and Yeka)

A

	Round2				
Round1	**Negative**	**Positive**	**Missing**	**(all)**	**% Missing Round 2**
Negative	84	22	62	168	36·90%
Positive	5	92	68	165	41·21%
Missing	48	36	259	343	
(all)	137	150	389	676	

B

	Round3				
Round2	**Negative**	**Positive**	**Missing**	**(all)**	**% Missing Round 3**
Negative	11	10	116	137	84·67%
Positive	12	40	98	150	65·33%
Missing	112	185	92	389	
(all)	135	235	306	676	

C

		Round3				
Round1	**Round2**	**Negative**	**Positive**	**Missing**	**(all)**	**% Missing Round 3**
Negative	Negative	9	6	69	84	82·14%
Negative	Positive	0	5	17	22	77·27%
Negative	Missing	14	15	33	62	53·23%
Negative	(all)	23	26	119	168	70·83%
Positive	Negative	1	1	3	5	
Positive	Positive	8	29	55	92	59·78%
Positive	Missing	3	6	59	68	86·76%
Positive	(all)	12	36	117	165	70·91%
Missing	Negative	1	3	44	48	
Missing	Positive	4	6	26	36	
Missing	Missing	95	164	0	259	
Missing	(all)	100	173	70	343	
(all)	(all)	135	235	306	676	

5. Seroprevalence among complete cases for Jimma

Complete cases	Round	Observed Individuals	Seropositivity	Estimated Seroprevalence reported in manuscript
Jimma HCW	1	360	30·60%	30·8% (26·9%, 34·8%)
	2	360	45·80%	45·6% (41·0%, 50·3%)
	3	360	56·90%	56·1% (51·1%, 61·1%)
Jimma Urban	1	132	38·60%	32·3% (27·0%, 37·9%)
	2	132	47·00%	40·8% (33·9%, 47·9%)
	3	132	47·00%	45·2% (37·7%, 52·7%)
Jimma Rural	1	80	17·50%	18·0% (13·5%, 23·2%)
	2	80	25·00%	26·3% (19·1%, 34·3%)
	3	80	27·50%	31·0% (22·3%, 40·3%)

Supplementary References

1 McElreath R. Statistical Rethinking: A Bayesian Course With Examples in R and Stan. Chapman & Hall/CRC, 2016 https://books.google.com/books/about/Statistical_Rethinking.html?hl=&id=d7fCsgEACAAJ.

2 Bürkner P-C. brms: An R Package for Bayesian Multilevel Models Using Stan. Journal of Statistical Software. 2017; **80**. DOI:10.18637/jss.v080.i01.

3 Bürkner P-C. Advanced Bayesian Multilevel Modeling with the R Package brms. The R Journal. 2018; **10**: 395–411.

4 Gudina EK, Ali S, Girma E, *et al.* Supplementary code and models for 'Silent spread of SARS-CoV-2 in Ethiopia: Longitudinal cohort study among frontline healthcare workers and community'. 2021; published online June 7. DOI:10.5281/zenodo.4885064.

5 Ahmad Z, Arif M, Ali F, Khan I, Nisar K S. A report on COVID-19 epidemic in Pakistan using SEIR fractional model. *Scientific Reports* 2020; **10**: 22268.

6 Annas S, Pratama M I, Rifandi M, Sanusi W, Side S. Stability analysis and numerical simulation of SEIR model for pandemic COVID-19 spread in Indonesia. *Chaos, Solitons & Fractals* 2020; **139**: 110072.

7 Grimm V, Mengel F, Schmidt M. Extensions of the SEIR model for the analysis of tailored social distancing and tracing approaches to cope with COVID-19. *Scientific Reports* 2021; **11**: 4214.

8 López L, Rodó X. A modified SEIR model to predict the COVID-19 outbreak in Spain and Italy: Simulating control scenarios and multi-scale epidemics. *Results in Physics* 2021; **21**: 103746.

9 Prem K, Liu Y, Russell T W, Kucharski A J, *et al*. The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study. *The Lancet Public Health* 2020; **5**: e261–e270.

10 Raimúndez E, Dudkin E, Vanhoefer J, *et al*. COVID-19 outbreak in Wuhan demonstrates the limitations of publicly available case numbers for epidemiological modeling. *Epidemics* 2021; **34**: 100439.
11 Kissler SM, Fauver JR, Mack C, *et al.* Densely sampled viral trajectories suggest longer duration of acute infection with B.1.1.7 variant relative to non-B.1.1.7 SARS-CoV-2. *medRxiv* 2021. DOI:10.1101/2021.02.16.21251535.
12 Keating SM, Waltemath D, König M, *et al.* SBML Level 3: an extensible format for the exchange and reuse of biological models. *Mol Syst Biol* 2020; **16**: e9110.
13 Schmiester L, Schälte Y, Bergmann FT, *et al.* PEtab—Interoperable specification of parameter estimation problems in systems biology. PLOS Computational Biology. 2021; **17**: e1008646.
14 Fang Z, Zhang Y, Hang C, Ai J, Li S, Zhang W. Comparisons of viral shedding time of SARS-CoV-2 of different samples in ICU and non-ICU patients. J. Infect. 2020; **81**: 147–78.
15 McAloon C, Collins Á, Hunt K, *et al.* Incubation period of COVID-19: a rapid systematic review and meta-analysis of observational research. *BMJ Open* 2020; **10**: e039652.
16 Schälte Y, Fröhlich F, Stapor P, *et al.* ICB-DCM/pyPESTO: pyPESTO 0.2.6. Zenodo, 2021; published online May DOI:10.5281/ZENODO.4768592.
17 Fröhlich F, Weindl D, Schälte Y, *et al.* AMICI: High-Performance Sensitivity Analysis for Large Ordinary Differential Equation Models. *Bioinformatics* 2021; published online April. DOI:10.1093/bioinformatics/btab227.

B Long-term monitoring of SARS-CoV-2 seroprevalence and variants in Ethiopia provides prediction for immunity and cross-immunity

This publication is reprinted as part of this thesis according to Springer Nature's permissions on author reuse. Material from:
Merkt, S. *et al.* Long-term monitoring of SARS-CoV-2 seroprevalence and variants in Ethiopia provides prediction for immunity and cross-immunity. *Nature Communications* **15,** 3463. doi:10.1038/s41467-024-47556-2 (2024)

nature communications

Article

https://doi.org/10.1038/s41467-024-47556-2

Long-term monitoring of SARS-CoV-2 seroprevalence and variants in Ethiopia provides prediction for immunity and cross-immunity

Received: 29 August 2023

Accepted: 3 April 2024

Published online: 24 April 2024

Check for updates

Simon Merkt[1,15], Solomon Ali[2,15], Esayas Kebede Gudina[3,15], Wondimagegn Adissu[3], Addisu Gize[2,4], Maximilian Muenchhoff[5,6], Alexander Graf[7], Stefan Krebs[7], Kira Elsbernd[8,9], Rebecca Kisch[8], Sisay Sirgu Betizazu[2], Bereket Fantahun[2], Delayehu Bekele[2], Raquel Rubio-Acero[8], Mulatu Gashaw[3], Eyob Girma[3], Daniel Yilma[3], Ahmed Zeynudin[3], Ivana Paunovic[8,10], Michael Hoelscher[6,8,10,11], Helmut Blum[7], Jan Hasenauer[1,12,13,16] ✉, Arne Kroidl[6,8,16] ✉ & Andreas Wieser[6,8,10,14,16] ✉

Under-reporting of COVID-19 and the limited information about circulating SARS-CoV-2 variants remain major challenges for many African countries. We analyzed SARS-CoV-2 infection dynamics in Addis Ababa and Jimma, Ethiopia, focusing on reinfection, immunity, and vaccination effects. We conducted an antibody serology study spanning August 2020 to July 2022 with five rounds of data collection across a population of 4723, sequenced PCR-test positive samples, used available test positivity rates, and constructed two mathematical models integrating this data. A multivariant model explores variant dynamics identifying wildtype, alpha, delta, and omicron BA.4/5 as key variants in the study population, and cross-immunity between variants, revealing risk reductions between 24% and 69%. An antibody-level model predicts slow decay leading to sustained high antibody levels. Retrospectively, increased early vaccination might have substantially reduced infections during the delta and omicron waves in the considered group of individuals, though further vaccination now seems less impactful.

The COVID-19 pandemic continues to have a significant global impact, with a substantial number of deaths continually being recorded worldwide (covid19.who.int). However, observations indicate a shift from the initial phase of the pandemic to an endemic stage, with reduced confirmed case numbers as well as deaths. Despite this, the emergence and evolution of more transmissible variants still pose a threat globally, necessitating ongoing monitoring by organizations such as the World Health Organization (WHO). In order to better prepare for future Sars-CoV-2 waves and potential pandemics, understanding the dynamics of the disease and the immune response protecting against infection as well as severe disease courses is crucial.

Policymakers rely on accurate data to inform vaccination strategies and intervention measures. However, these strategies may differ greatly depending on circumstances like information about the actual virus spread and public acceptance of policies. Especially within the African continent, comprehensive data is scarce. Even in July 2023, the

A full list of affiliations appears at the end of the paper. ✉e-mail: jan.hasenauer@uni-bonn.de; akroidl@lrz.uni-muenchen.de; andreas.wieser@lmu.de

WHO still only lists 9.5 million confirmed cases of SARS-CoV-2 in the whole of Africa. Given our data, serological evidence of past infection in Ethiopia alone suggests that by autumn 2022, there were ten times as many infections in Ethiopia as officially reported[1].

Besides the scarcity of data, African countries, including Ethiopia, face unique challenges in dealing with the pandemic, such as limited testing infrastructure[2], insufficient vaccine supplies[3], low vaccine acceptance[4], and being overlooked in global research efforts[5]. For Ethiopia in particular, research shows that though adequate pandemic prevention strategies have been enacted over time, shortages of medical supplies and equipment is an ongoing struggle[6].

In 2021, we demonstrated a severe under-reporting of COVID-19 cases in Ethiopia through an antibody prevalence study[1]. By employing epidemiological modeling, we predicted prevalence levels above 50% for the population. While this earlier phase of the pandemic has received some research attention, later phases of the SARS-CoV-2 pandemic, including the Delta and Omicron waves, remain inadequately investigated in Ethiopia[7,8]. Additionally, due to very limited access to sequencing facilities, the knowledge about circulating variants has been scarce. Previous publications touch upon this topic hypothetically, e.g. Gudina et al. by simulating a scenario with two variants[1], but longitudinal data on variant distribution has only recently become available for Ethiopia[9]. We have simultaneously acquired broad data to address the gaps for modeling and prediction of the epidemic in Ethiopia.

In this study, we obtained sequencing results for SARS-CoV-2 samples collected at various time points between October 2020 and July 2022 at two different sites in Ethiopia. This dataset enabled us to investigate the composition of variants of concerns (VOCs) between the initial appearance of COVID-19 in Ethiopia in March 2020 to the spread of Omicron variant BA.4/5 as the dominant genotype in fall 2022. Additionally, we extended our serology-based antibody survey by conducting two further sampling rounds to cover the time span between late fall 2020 to April 2022 in a total of five sampling rounds. In addition to the serological testing against Anti-nucleocapsid antibodies (Anti-N), all samples were re-tested against anti-spike antibodies (Anti-S), and questionnaires were used to explore vaccination- and potential infection status for all participants. Using this large and multidimensional dataset for analysis, we developed a large-scale multivariant model to characterize the infection pathways and to explore the cross-immunity properties among different variants circulating in Ethiopia. This analysis allowed us to gain insights into the interplay between the variants and their impact on the overall population's immune response.

Furthermore, we leveraged the information from multiple rounds of sampling, which provided Anti-N and Anti-S antibody levels of individuals. The resulting dataset was used for a detailed temporal analysis, comparing the antibody levels observed during the initial three rounds with those from the subsequent two rounds. We utilized a second epidemiological model to predict future antibody dynamics, providing insights into the expected long-term immunity landscape in the Ethiopian population. This might provide decision makers with information which is helpful for the assessment of the situation and the choice of appropriate measures.

In summary, this study expands upon previous findings and presents novel insights into the antibody dynamics and concurrent variant prevalence in Ethiopia. By integrating modeling techniques and broad datasets, we aim to contribute to a deeper understanding of SARS-CoV-2 infections and the implications for public health interventions and vaccination strategies in Ethiopia, other resource-limited settings, and beyond.

Results

Antibody data reveals majority had multiple infections

In our previous study, we assessed the dynamics of COVID-19 infection between August 2020 and April 2021 in Addis Ababa and Jimma, Ethiopia[1]. To understand how the COVID-19 pandemic evolved afterwards, we conducted two additional rounds of sampling. As our previous study predicted a complete transmission within the population for SARS-CoV-2 in Ethiopia by late 2021, we complemented the previous semi-quantitative analysis of Anti-N antibody levels by a quantitative analysis of the Anti-S antibody levels in the newly collected and historic samples to gain more detailed insight into possible reinfection occurrences. An overview of the demographics of the participants of the original three rounds and the two follow up rounds is shown in Table 1 (for healthcare workers Supplementary Table 1). Study flows are depicted in Supplementary Fig. 1.

Our SARS-CoV-2 specific antibody tests revealed that in April 2022, the majority of individuals (in Round 5: 95.9% of the healthcare workers and 94.8% of the community members), reacted positive for both Anti-S and Anti-N antibodies (Fig. 1a–e), suggesting an infection event. Based on a previous study, this result is unlikely to be explained by cross-reactivity[10]. In Round 3 (April 2021, Fig. 1c) and four (August 2021, Fig. 1d), significant numbers of samples were observed which showed isolated positivities for Anti-N or Anti-S. This can be explained by a delayed onset of either Anti-N or Anti-S response shortly after or during infection or, for Anti-S positivity, by vaccination. As large-scale vaccination campaigns started in Ethiopia rather late in November 2021, the data suggests that sampling in Round 3 coincided with waves of SARS-CoV-2 infections. First confirmed vaccinated individuals show up only in rounds four (August 2021, Fig. 1d) and five (April 2022, Fig. 1e). Interestingly, although the vaccines used in Ethiopia only induce Anti-S, most individuals vaccinated also showed reactivity for Anti-N (in Round 5: 94.8% of the healthcare workers and 96.4% of the community members), suggesting they had been exposed to the infection prior to or shortly after vaccination. By Round 4 all vaccines

Table 1 | Demographic characteristics of community members participating in study

	Jimma					Addis Ababa				
	R1 (Dec 20)	R2 (Jan 21)	R3 (Feb 21)	R4 (Aug 21)	R5 (Apr 22)	R1 (Jan 21)	R2 (Feb 21)	R3 (Apr 21)	R4 (Sep 21)	R5 (Mar 22)
Participants	536	325	267	539	575	361	314	721	424	461
Age	30 (19, 63)	30 (19, 62)	32 (19, 63)	33 (20, 65)	32 (19, 63)	36 (21, 68)	36 (22, 67)	35 (21, 67)	33 (19, 65)	38 (20, 68)
Sex										
Female	260 (48.5%)	166 (51.1%)	136 (50.9%)	331 (61.4%)	317 (55.1%)	279 (77.3%)	236 (75.2%)	360 (49.9%)	209 (49.3%)	162 (35.1%)
Male	276 (51.5%)	159 (48.9%)	131 (49.1%)	207 (38.4%)	258 (44.9%)	79 (21.9%)	70 (22.3%)	109 (15.1%)	71 (16.7%)	299 (64.9%)
Missing	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.2%)	0 (0.0%)	3 (0.8%)	8 (2.5%)	252 (35.0%)	144 (34.0%)	0 (0.0%)
Anti-N positive	139 (25.9%)	114 (35.1%)	107 (40.1%)	313 (58.1%)	543 (94.4%)	165 (45.7%)	150 (47.8%)	234 (32.5%)	286 (67.5%)	458 (99.3%)
Vaccinated	0 (0.0%)	0 (0.0%)	1 (0.4%)	47 (8.7%)	195 (33.9%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	28 (6.6%)	167 (36.2%)

Age denoted as median and 90% quantiles, and sex in absolute and relative numbers. Round 1-3 (R1-R3) are the previous study[1].

https://doi.org/10.1038/s41467-024-47556-2

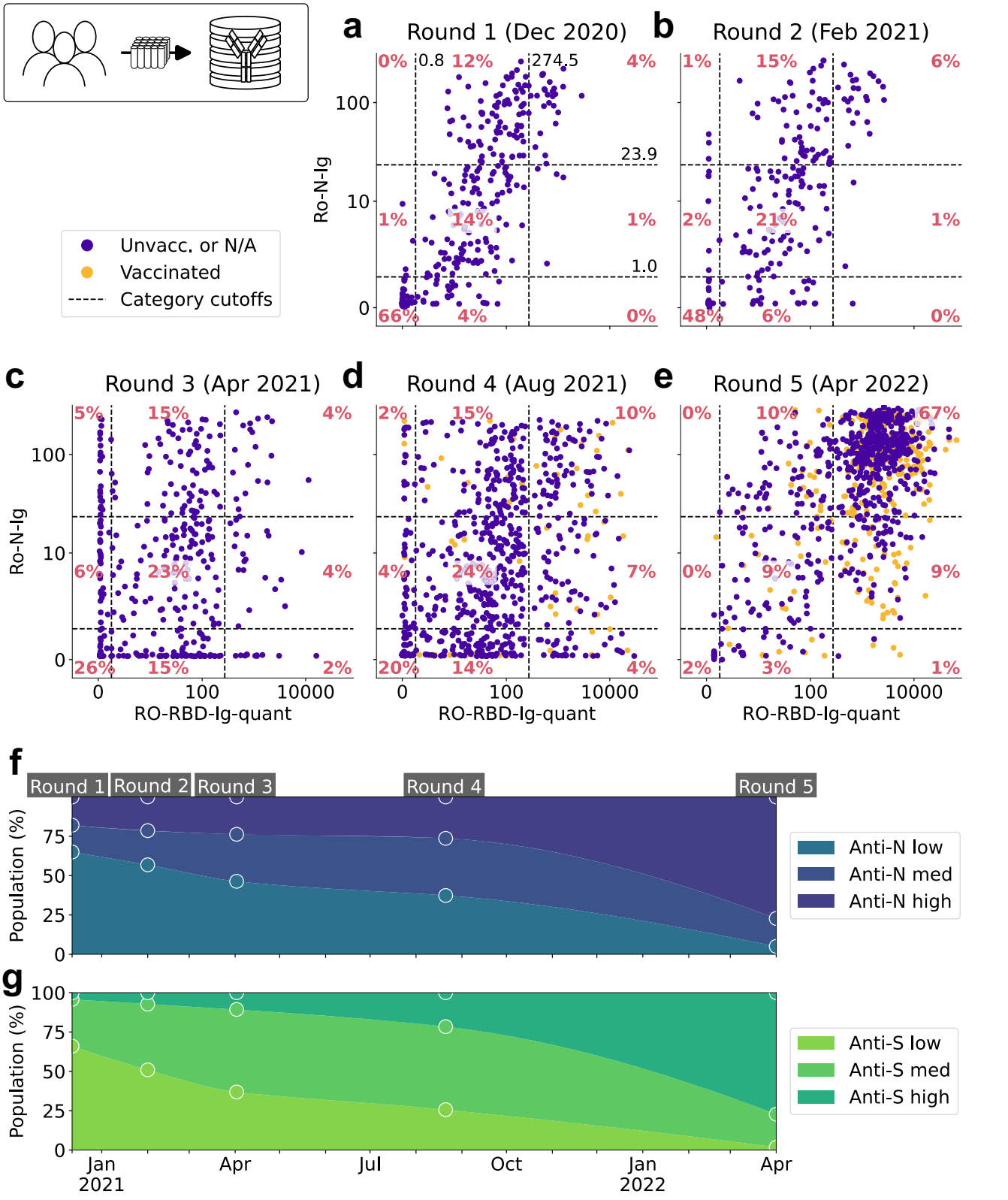


Fig. 1 | Ro-N-Ig and Ro-RBD-Ig-quant measurements of five rounds of convenience sampled community members. a–e Scatterplots displaying the relationship between levels of N- and S-specific antibodies across five rounds of measurement. Known vaccination status of each participant indicated by colors, cutoff levels indicated by dashed lines and percentages of people per category annotated in red. **f–g** Antibody levels over time between end of 2020 and April 2022. The observations are indicated by circles and the trend is indicated via smoothing splines constructed on the basis of these data. Source data are provided as a Source Data file.

in Ethiopia were Covishield (AstraZeneca type vaccine manufactured by Serum Institute of India) and by Round 5 Johnson & Johnson has become another major type of the vaccine. Although few doses of Sinavac/Sinopharm, Sputnik-V, Moderna, and Pfizer-BioNTech were reported to be donated to the country, they were very little and hence negligible. Therefore we can safely disregard the influence of mRNA vaccines in our study.

Analyzing the magnitude of the Anti-S responses considered positive (above the test threshold of 0.8), we observed two populations, separating positive samples into those with higher and lower levels (Fig. 1c–e, Supplementary Fig. 3c–e, Fig. SN1). Comparing the data in this study and with experience gathered in our population-based studies in Munich, Germany[11], it can be appreciated that one exposure to SARS-CoV-2 with a natural infection generally induces Anti-S values below a cutoff value centered in the middle of the antibody level range (shifted log-scale) as indicated with the vertical dashed line in Fig. 1a and b. Higher Anti-S levels are only reached after multiple exposures leading to a boosting effect. Employing 1-dimensional k-means clustering with two means on the S-positive samples from all five rounds, we determined the cutoff value for the groups with one or multiple exposures to be 274.5 (for more details see Supplementary Information's Supplementary Note 1 and Fig. SN2). Anti-S results are diluted and measured within the linear range to provide quantitative results for all samples as described in more detail in the Methods section.

For Anti-N values, a clear division into two populations is not as evident as for Anti-S, likely due to the semi-quantitative nature of the Anti-N measurements. A noticeable shift towards higher Anti-N values is observed between Round 4 (Fig. 1d) and Round 5 (Fig. 1e). However, we also performed k-means clustering on Anti-N values to determine distinct categories, similar to the process carried out for the Anti-S signals. Using the calculated cutoffs and positivity thresholds, we assigned the individual patients for each round into the categories *low* (negative, i.e. below threshold), *medium* (positive, i.e. above positivity threshold but below calculated category cutoff), and *high* (above category cutoff) for both Anti-N and Anti-S, respectively.

Moreover, we summarized the progression of Anti-N and Anti-S level categories separately over time (Fig. 1f–g). Remarkably, in the latest round of sample collection in April 2022, a substantial proportion (75-80%) of the sampled individuals exhibited *high* antibody levels for Anti-N as well as Anti-S. Since Anti-N is only induced after an infection due to the spike-protein nature of the vaccines used in Ethiopia, this suggests that a significant fraction of the population had already experienced at least two exposures for each antigen by that time.

Variant sequencing identifies all major substrains

The antibody data provide information about previous infections, but not about the SARS-CoV-2 variants which caused them. Moreover, up until very recently, there was no available data on virus variants in Ethiopia[9]. Hence, to better understand the pandemic, we sequenced a total of 1873 SARS-CoV-2 reverse transcription polymerase chain reaction (RT-PCR) positive swabs, collected in Jimma and Addis Ababa, between October 2020 and July 2022. Overall 574 sequences were of sufficient quality to allow full pangolin strain matching and were thus used for analysis.

The sequencing data revealed the presence of several variant strains, including wildtype (A and all without any "interesting" mutations, details below), wildtype* (B.1.480), alpha (B.1.1.7), beta (B.1.351), eta (B.1.525), delta (B.1.617.2 and AY.*), and the two omicron lineages BA.1 and BA.4/5 (Fig. 2a). At the beginning of the sampling period in autumn 2020, the wildtype strain was predominant (as expected) and accompanied by a notable presence of the wildtype* (B.1.480) strain. However, in late 2020 to January 2021, the alpha variant emerged and rapidly became the dominant strain, accounting for approximately 80% of the PCR-positive swabs by April 2021. During this time, the eta lineage also briefly appeared, which was previously reported as the predominant strain in Nigeria in early Spring 2021 (B.1.525 on cov-lineages.org). In Ethiopia, the eta lineage was unable to outcompete the alpha variant, and with the appearance of the delta variant in July 2021, both alpha and eta disappeared. In early 2022, the omicron BA.1 variant emerged and completely took over. Despite that we had only limited samples during the transition phase, it is evident that by June 2022, the BA.1 variant was subsequently substituted by omicron BA.4/5. The full and detailed results of the sequencing analysis can be found in the supplementary materials (Table SN1).

The mutational variety observed in our dataset is extensive, with mutations spanning from less than 10 to more than 90 mutations relative to the original wildtype variant that originated in Wuhan (Fig. 2b, c). As variations in the spike protein play a critical role for immune escape, we assessed this in more detail following the definition and mapping of outbreak.info's mutations of interest or concern (MOIC)[12,13]. For the observed strains, the presence and absence of MOICs are indicated in Fig. 2d. In previous studies[14–16], the overall number of mutations (Fig. 2b, c) was used as a measure for reinfection potential. Grouping our variants by MOIC allows us to maintain the statistical power of the lineage groups for subsequent analysis of potential cross-immunity while still retaining their relevant spike protein differences. The grid of distances of MOIC between observed lineage groups in Fig. 2e demonstrates that our dataset encompasses a range of distances up to 6, indicating diverse genetic distances between the variants. Moreover we see that our data set consists of variants which emerged earlier in other parts of the world, hence implies a continuous introduction of new variants to Ethiopia rather than a mutation of the wild-type inside of Ethiopia. We provide more information about these distances in the methods section of this paper.

Multivariant model describes antibody prevalence and strains

The long-term antibody and variant data from Addis Ababa and Jimma provide valuable information about the course of the pandemic. Yet, the observations themselves did not allow for a direct assessment of infection or reinfection risk, or (cross-)immunity. Challenges are: (i) most study participants contributed to less than three of five rounds of antibody testing and (ii) the participant groups for antibody testing and swab collection were disjoint. Therefore, it is not possible to map the data types to each other and to analyze individual disease history. To achieve a good understanding of the COVID-19 dynamics and the interactions of different variants in Ethiopia, we instead employ epidemiological modeling of population averages.

We constructed a multivariant model to investigate the temporal evolution of the SARS-CoV-2 pandemic in Ethiopia. The model accounts for different sequences of infections and vaccination events (Fig. 3a). The sequence of infections and vaccinations - to which we refer in the following as *pathways* - is tracked to determine the immunity status of individuals. Each infection follows the SEIR schematic, with individuals transitioning from being susceptible to exposed, then infected, and finally recovered. Due to official vaccine availability in Ethiopia only after Round 3[17] in combination with our previous observation that vaccinated individuals are more likely to answer questions on the vaccination status on the questionnaire than unvaccinated individuals, we considered individuals without an answer ("N/A") as "unvaccinated" for modeling. The structure of the multivariant model is outlined in Fig. 3a using a small number of possible pathways. The model has a total of 364 possible pathways, and possesses 950 compartments and more than 950 transitions.

We allowed for immunity and cross-immunity conferred by previous infections and vaccination in the multivariant model. As the precise dependencies are not known, we assumed a variant-specific risk reduction for reinfections with previously encountered variants.

https://doi.org/10.1038/s41467-024-47556-2

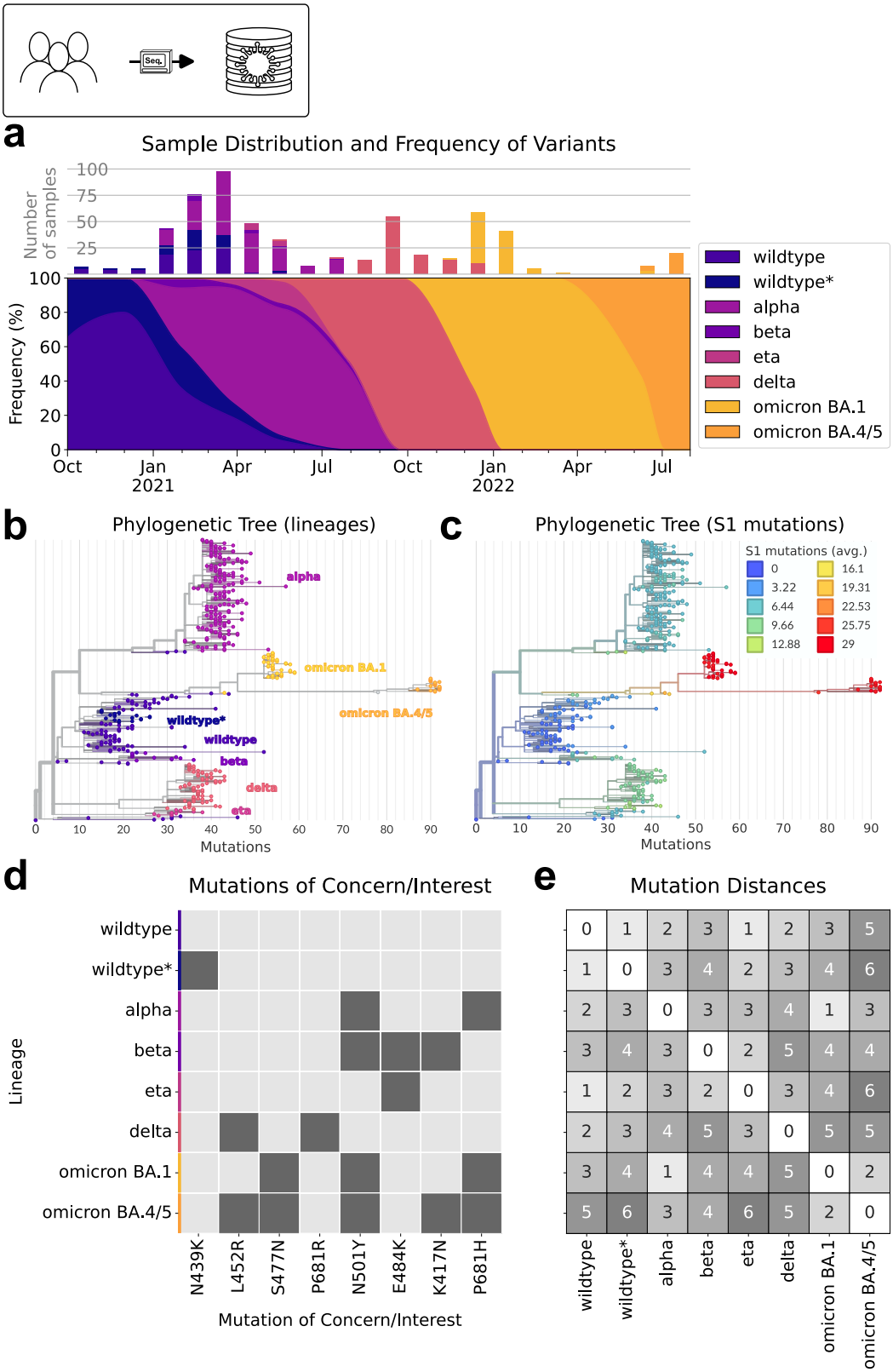


Fig. 2 | Sequencing results of samples obtained between October 2020 and July 2022. a Number of successfully sequenced samples, variant frequency and smoothed variant time-course. Variants are indicated using colors. **b,c** Phylogenetic tree of the sequenced samples, illustrating the relationships between variants and their sub-variants (full list of variants in Supplementary Information Table SN1). Distance between variants represented by overall difference in their mutations. Lineage groups **b** and number of mutations in the spike protein **c** highlighted by color. **d** Heatmap indicating which variants possess specific mutations of interest on their S1 protein. **e** Heatmap depicting MOIC mutation distances with respect to mutations of interest between different variants. Distance indicated by gray scale. Source data are provided as a Source Data file.

https://doi.org/10.1038/s41467-024-47556-2

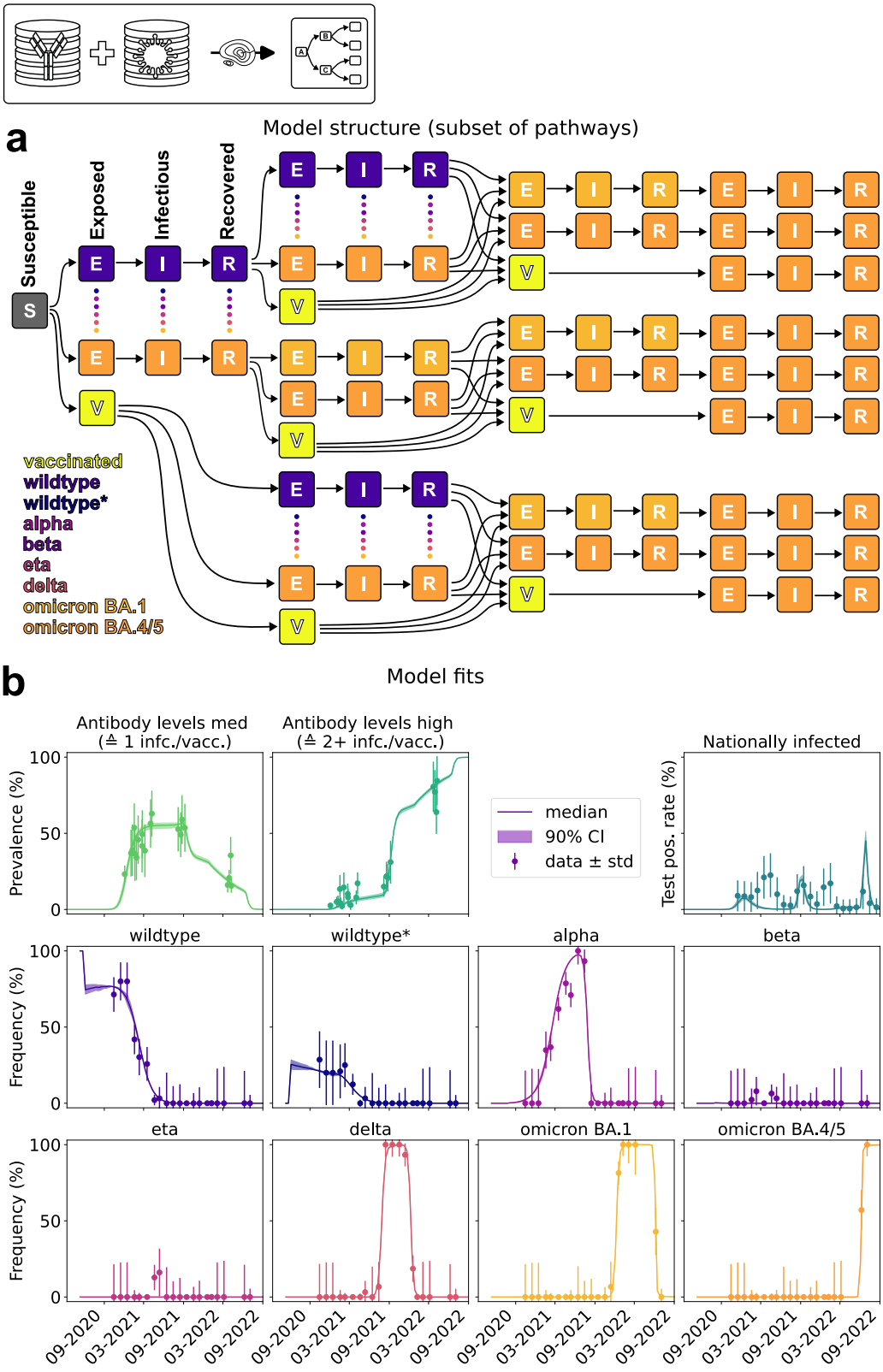


Fig. 3 | Structure and fitting results of the multivariant model. a Model structure depicting up to four consecutive infections/vaccinations. Potential infection pathways labeled by the stages S(usceptible), E(xposed), I(nfectious) and R(ecovered) and their respective variants highlighted by different colors. Only a small subset of the in total more than 350 possible paths is shown. **b** Model fitting results shown by progression of all observables against their respective (mean) measurements. Bayesian 90% credibility intervals for model simulation obtained by sampling included as well as the standard deviation of the measurements. Prediction simulations performed on $n = 6001$ parameter samples after burn-in from Markov chain Monte Carlo. Sample sizes of data points provided in Supplementary Note 2 (Table SN4). Source data are provided as a Source Data file.

For infection with a different variant, we assume that the infection risk depends on the difference of MOIC between the previously encountered variant and the variant to which individuals are exposed. In the case of multiple previous infections, the union of mutations from the previous variants is considered, and the distance to the new variant is calculated. This is based on the assumption that antibodies against regions with different MOIC can be developed. Vaccination is treated as a recovery from the wild-type infection. Exposure risk is also influenced by seasonality, which is incorporated using a 1-year-periodic factor. The unknown parameters of this seasonality factor and cross-immunity are estimated, along with the appearance times of the variants, incubation and recovery times, a basic exposure rate, and the exposure multipliers for the variants. A detailed mathematical description of the multivariant model and a complete list of its parameters is provided in Supplementary Information (Supplementary Note 2 and Table SN5).

To assess the evolution of the SARS-CoV-2 pandemic in Ethiopia, we parameterized the multivariant model using data on antibody levels, viral variant distribution, and national test positivity rate. The Anti-S antibody measurements were used to provide information on the fraction of individuals with a single infection or vaccination (medium level) and the fraction of individuals with at least two infections, vaccinations or a combination of both (high level). Since it is impossible to distinguish between vaccinations and infections from Anti-S levels we implemented observables corresponding to the medium and high levels without discriminating between vaccination or infection (c.f. Supplementary Note 2 for detailed equations). The viral variant data provided information on the relative levels of each of the eight variants, mapping the relative measurements to the percentage of individuals in an infectious state associated with each variant. The national PCR test positivity rate was used to determine the percentage of currently infected individuals, irrespective of the variant.

The parameterization of the model was performed using Markov chain Monte Carlo sampling. The sampling results revealed good agreement of the parameterized multivariant model with the observed data (Fig. 3b). The antibody levels and variant distributions (the primary focus of our investigation) are captured accurately. The national test positivity rate is described well up to two peaks (which might be caused by different regions in Ethiopia). In fact looking at the timing of the first peak, which is missed by our model, we see that our antibody data is already saturated and hence tells a different story than the nationally reported data. Most of the model parameters are well determined (Table SN5 in Supplementary Information) and in agreement with estimates provided in the literature. For a comprehensive description of estimation and uncertainty analysis results for specific parameters, as well as convergence information, we refer readers to the supplementary materials.

Overall, comparison of model simulation and data revealed that the proposed multivariant model provides a good description for the progression of the SARS-CoV-2 pandemic in Ethiopia. Furthermore, the assumed model for (cross-)immunity appears appropriate to accurately describe the data for Addis Ababa and Jimma.

Reconstruction of infection history and cross-immunities

As the multivariant model provides an accurate description of the observed data, we used it to study the population-level infection history in Addis Ababa and Jimma. This infection history is encoded in the time-dependent state of the parameterized model, which is informed by our broad datasets.

The analysis of the model predicted that the most common pathway of infections and vaccinations was: 1st infection with wildtype, 2nd infection with delta, vaccination, and 3rd infection with omicron BA.4/5 (Fig. 4a, b). In particular wildtype*, alpha, beta, eta, and omicron BA.1 are not part of it, of which omicron BA.1 appears in the second most common pathway (delta, omicron BA.1, omicron BA.4/5) and alpha appears in the third most common pathway (alpha, delta, vaccination, omicron BA.4/5). The estimates indicate that a median of 12.7% with 90% credible interval (CI) of (10.9%,14.4%) of the inhabitants of Addis Ababa and Jimma followed this pathway. As suggested by the low percentage of individuals following the most common pathway, there has been a large degree of pathway variability. Indeed, the 10 most common pathways account for only 59.0% (42.8, 69.8) of the overall pathways (Fig. 4b). The high variability is caused by a large number of different combinations of virus variants. Overall wildtype, delta, and omicron BA.4/5 variants are the primary contributors to the infection progression (Fig. 4c). They are followed by wildtype*, alpha, and omicron BA.1, which also exhibit notable contributions. The model predicts a negligible impact of beta and eta variants, which is consistent with the data used to parameterize it.

The analysis of the time of infections (Fig. 4c) indicates three distinct waves, which coincide with reports for wildtype, delta and omicron BA.4/5. Notably, the emergence of the delta variant marks a shift where second infections start playing a significant role, which aligns with findings from other published studies[18]. Furthermore, with the introduction of the omicron variants, third infections become more prevalent, resulting in nearly the entire population experiencing at least two infections. Until September 2022, the occurrence of fourth infections appears to be minimal, likely due to the influence of vaccination and pre-existing immunity.

To assess the impact of cross-immunity on the pandemic, we assessed the corresponding model parameters used to describe it (Fig. 4d, e). The statistical inference suggests that the reinfection risk with the same variant - corresponding to a MOIC mutation distance of 0 - is reduced to 10.0% (5.1, 14.7) of the risk of an initial infection. In contrast, reinfection with different variants demonstrates a range of probabilities, ranging from 24.5% (21.3, 27.8) for a MOIC mutation distance of 1 (e.g., wildtype to wildtype*) to 68.6% (63.2.3, 72.4) for a distance of 6 (e.g., wildtype* to omicron BA.4/5). The 90% CIs for all variant-variant combinations are displayed in Supplementary Fig. 2.

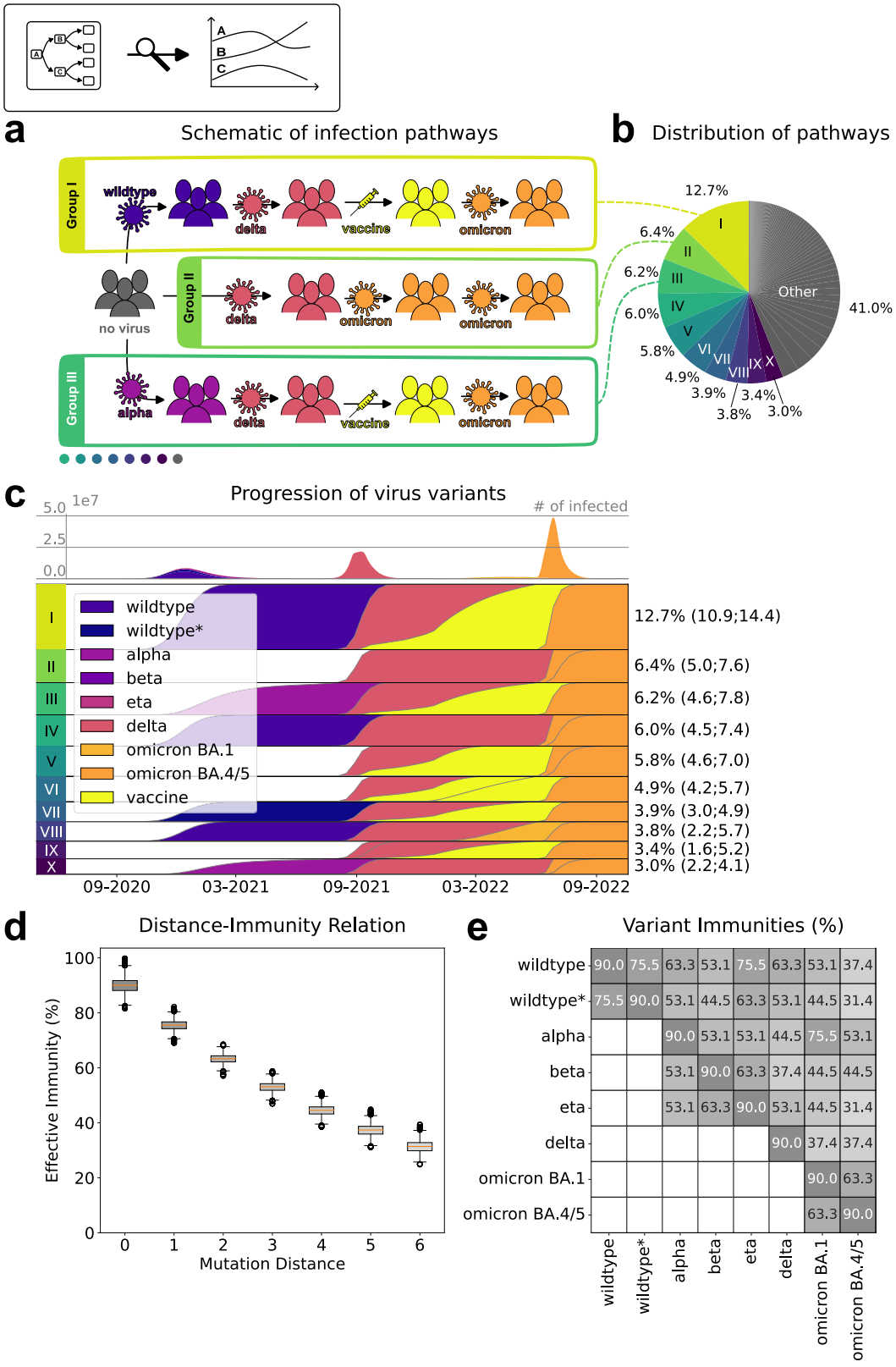
Overall, the multivariant model provided insights in the infection history by linking datasets collected for different groups of individuals at different time points. Based on this, it sheds light on the differential susceptibility to reinfection based on the genetic distances between variants.

Antibody-level model predicts high immunity and slow decline

The multivariant model enabled the assessment of the Anti-S antibody and variant data, yet, it is unable to fully exploit the comprehensive assessment of Anti-N and Anti-S antibody levels (Fig. 1a–e) available for a large fraction of our cohort. As this is necessary to assess waning immunity and the impact of vaccination rates, we decided to develop a tailored model for the analysis of these aspects.

We constructed an antibody-level model describing the dynamics of the Anti-N and Anti-S antibody levels. Following the analysis of the measurement data (Supplementary Information Fig. SN1), we implemented a discretization of both antibody levels in *low (negative)*, *medium* and *high*, which yielded a model with 9 state variables (Fig. 5a). Thresholds for these categories were inferred from the data (cf. antibody subsection of Results section and Supplementary Note 1 of Supplementary Information). Infections are assumed to result in increases of Anti-S and Anti-N antibody levels to the next higher category, while vaccinations are assumed to result only in an increase of Anti-S antibody levels to the next higher category. To account for the semi-quantitative nature of Anti-N measurements and the possibility of boosting Anti-N to *high* levels with a single infection, the model allows for a fraction of individuals in the Anti-N *low* category to directly transition to the *high* category. Antibody waning results in a shift to a lower category.

https://doi.org/10.1038/s41467-024-47556-2



To capture the dynamics of the antibody levels in Addis Ababa and Jimma, the antibody-level model used the available information about variants and vaccinations as inputs. The vaccination rate was calculated as monthly averaged rates based on the vaccination information provided by the participants of the antibody study, and the relative abundance of variants was computed by fitting Gaussian kernels to the data and using them as weights for the time-dependent effective transmission rate, i.e., the weighted sum of all variant transmission rates. The results of these computations can be seen in Fig. SN10 and Fig. SN11 of the Supplementary Information.

Additionally, the model incorporates seasonality, as described for the multivariant model. Furthermore, two immunity factors are introduced as multipliers of the transmission rate: one applied if either

Fig. 4 | Analysis of estimated variant-pathways and cross-immunities. **a** Illustration of three common pathways depicting the progression from a susceptible state to acquiring up to four different infections and/or vaccinations over time. **b** Proportions of variant-pathway-groups within the population, highlighting groups that constitute more than 3% of the total population. **c** Timeline of total number of infected people (first row) and time-resolved compositions of each group highlighting the portions of last variant recovered from or vaccination obtained by color (subsequent rows). **d**–**e** Estimated cross-immunity-levels, with 100% corresponding to a zero percent infection probability and 0% corresponds to infection risk as without previous infection. **d** Boxplot of estimated immunity-levels including sampled uncertainty. Immunity depicted with respect to MOIC mutation distance between newly encountered and previously encountered variants (-combinations) (Center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points, outliers). **e** Heatmap of cross-immunity levels between variants. Y-axis corresponding to previous and x-axis to new variant. Intensity of colors corresponds to strength of cross-immunity, with darker shades indicating higher levels of immunity. Empty cells indicating infection combinations excluded a priori from models based on the world wide variant wave chronology, e.g. a wildtype infection after recovery from delta. **c**-**e** Median and CIs obtained from $n = 6001$ samples after burn-in from Markov chain Monte Carlo. Source data are provided as a Source Data file.

of the antibody levels is in the *medium* category and a second applied on top of the first factor if either level is in the *high* category.

Unlike for the multivariate model, the distribution of variants is derived a priori from available data, and only their transmission rates and initial time of the overall disease dynamics are estimated. Incubation and recovery times are also estimated. For further information on the model setup, parameter details, and estimation results, we refer readers to the supplementary materials.

The antibody-level model possesses several unknown parameters, including the rates of antibody waning, the infection rates for different variants, and the fraction of infections, which directly result in a *high* Anti-N category. We estimate these parameters using Markov chain Monte Carlo sampling from the available data, which are the fraction of individuals in different categories and the national PCR test positivity rate. The parameter estimation provided a model which describes all these data well (Fig. 5b and c (left)). Indeed, credible intervals for parameter estimates (Supplementary Information Table SN8), state variables (Fig. 5b) and predictions (Fig. 5c) were mostly tight, indicating a low uncertainty of model predictions. In alignment with immune escape properties of later variants, we estimated higher valued infectiousness parameters for them, e.g. omicron BA.4/5 having 3.3 times the delta and 10.6 times the wildtype infectiousness. Relative infectiousness for all variants can be deduced from Supplementary Note Tables SN3 and SN6.

As for the multivariant model there is some discrepancy between national test positivity rate and model description (which might be caused by different regions in Ethiopia). Nevertheless, the antibody-level model provides an accurate description of the available antibody data, so that we used it to predict the current antibody levels, including observations of antibody levels until April 2022. We found that following the omicron wave, our model predicts a remarkable trend (Fig. 5b): up to 100% of the population is projected to fall into the high antibody category for both Anti-N and Anti-S antibodies. This prediction is subject to minimal uncertainties. Notably, the parameter estimation determined slow decay of both Anti-N and Anti-S antibody levels, leading to sustained high levels in the high antibody category until present times.

Given that the sequence of infections and vaccinations was predicted to yield high antibody levels, we explored the impact of vaccination rates. In addition to the actual reported vaccination rate, we considered a 5- and 10-times increased vaccination rate (Fig. 5c, middle), two levels, which could have been achieved using redistribution on the global scale. The artificial experiments indicated that increased vaccination rates would have led to a substantial reduction in infections during the delta wave. For the omicron wave, a reduced impact is predicted due to the higher transmission rate, but the number of hospitalizations could have been substantially lower with higher vaccination rates.

The second type of prediction involved retrospectively examining the impact of varying vaccination rates on the overall virus spread. By multiplying the actual vaccination rate by different factors larger than 1, we investigated how improved vaccination scenarios could have affected the course of the pandemic. Our analysis reveals compelling insights: a vaccination rate five times as high as the actual rate, equivalent to 11.2e7 vaccinated dosages instead of the actually observed 2.7e7, would have significantly mitigated the delta wave. Furthermore, higher vaccination rates of 5 or even 10 times the actual rate could have substantially reduced infections during the omicron wave, potentially halving or lowering it even further (Fig. 5c second and third subplot).

Overall, the predictions of the antibody-level model highlight the critical role of early vaccination in controlling the spread of the virus and provide valuable information for policymakers and public health officials. The results of our model offers evidence-based projections that shed light on the potential outcomes of different vaccination scenarios, emphasizing the importance of accelerated vaccination efforts early on in curbing the impact of viral variants, while implying a minor role of later vaccinations in already saturated natural immunity level scenarios.

Discussion

The course of the COVID-19 pandemic and current immunity status for many countries is still not sufficiently understood to inform decision-making about the effectiveness of past measures and strategies for future pandemics. This study provides data and model-based analysis to close some of the gaps for Ethiopia. By performing wide sampling before the omicron wave and quantifying antibody titres, we provide insights into the cumulative infection numbers, including the prevalence of reinfections. This suggests that by the end of the last sampling round in April 2022, already 55.1% of the inhabitants of Ethiopia recovered from two SARS-CoV-2 infections. Another 4.1% of the inhabitants of Ethiopia recovered from three SARS-CoV-2 infections.Comparing this to the roughly 470,000 officially confirmed case numbers at the end of April 2022 and the official WHO number of 500,000 cases by late spring 2023 (WHO Covid-19 Dashboard), it is clear that drastic underreporting regarding the number of SARS-CoV-2 infections has been and is still happening in Ethiopia.

Our broad longitudinal analysis of PCR-positive swabs complemented the information about antibody levels and provided an overview of disease-driving mutations. In Ethiopia, wildtype, alpha, delta and Omicron BA.4/5 were the most influential SAR-CoV-2 variants and appeared (except alpha) with a slight delay compared to the global appearance (Supplementary Information Fig. SN9). In relation with the Ethiopian variant survey of Sisay et al. our key findings are confirmed[9]: The importance of B.1.480 (wildtype* in our case) and non-concerning B.1 sublinages (wildtype in our case), the minor role of beta and the general timeframe and dominance of alpha, delta, and omicron waves are common discoveries. Since the observation period of Sisay et al. ends in February 2022, which is around the time when the statistical power of our sequencing data decreases substantially, future research about the precise transition between the omicron waves BA.1 and BA.4/5 could be worth exploring.

To fully exploit the large datasets, we developed two models in this study. The multivariant model provides, to the best of our knowledge, one of the most detailed descriptions of the dynamics of the SARS-CoV-2 pandemic for an African country and is unique as it

https://doi.org/10.1038/s41467-024-47556-2

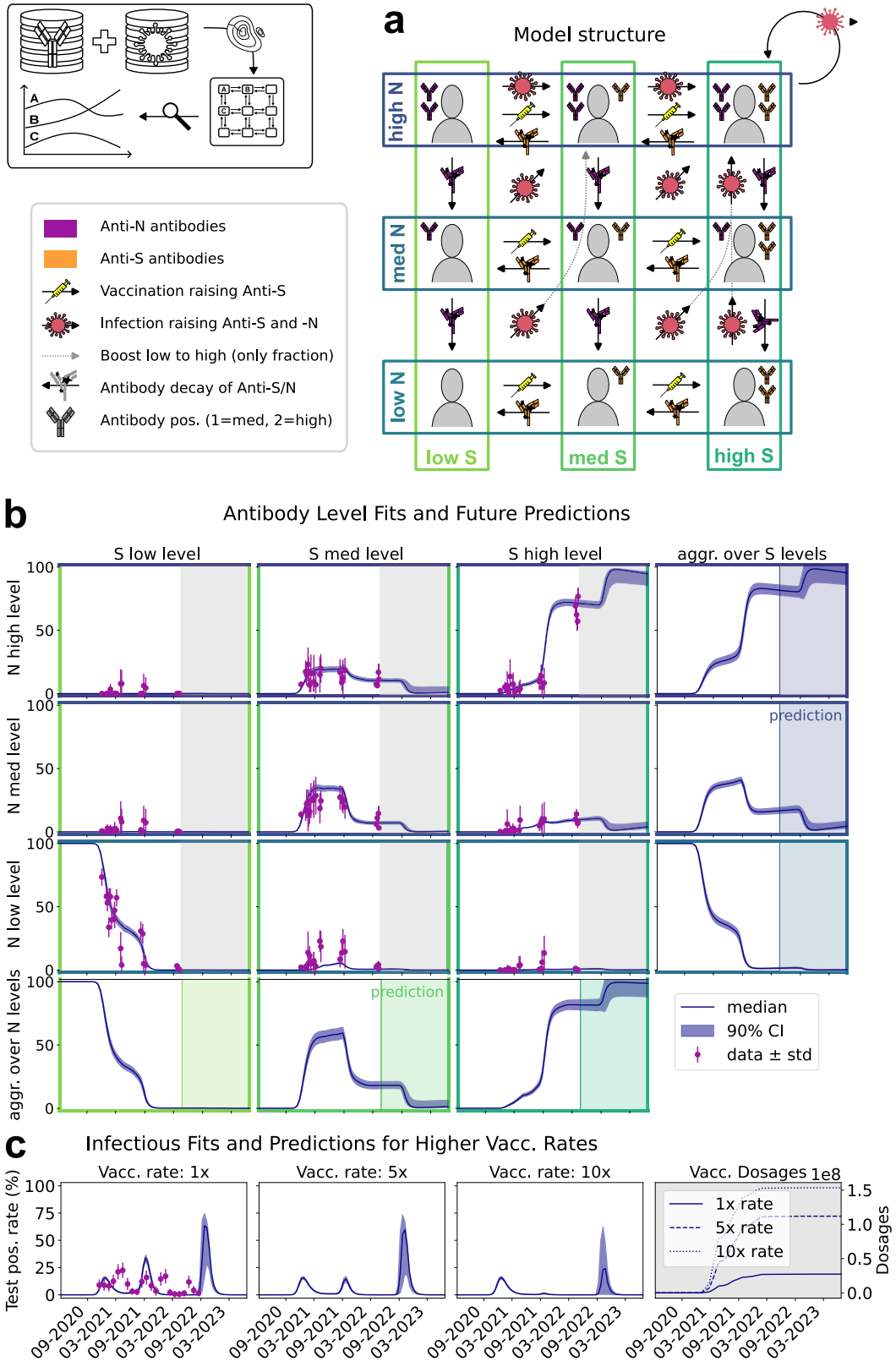


Fig. 5 | Modeling, fitting results and predictions of antibody-category model. **a** Model structure of different antibody category levels and transmission between them. Antibody levels against N or S, respectively, are indicated by colors. **b** The inner grid presents the fitting results to mean measurements (with standard deviations) along with 90% confidence bands derived from parameter sampling. The aggregated levels of Anti-S and Anti-N antibodies are displayed in the lower row and right column, respectively. The prediction phase, where no new antibody data was measured, is highlighted. **c** First plot showing the fit of simulated incidences to measured national test positivity rates (mean and standard deviation taken per month). Second and third plots illustrating predictions of test positivity rates under hypothetical scenarios with vaccination rates 5 and 10 times as high as the actual rate. Last plot showing how different vaccination rates translate to vaccinated dosages. **b, c** Prediction simulations performed on $n = 30{,}001$ parameter samples after burn-in from Markov chain Monte Carlo. Sample sizes of data points provided in Supplementary Note 3 (Table SN7). Source data are provided as a Source Data file.

allows for the description of multiple waves and the variant replacement dynamics. Many of the important studies on African countries presented so far focus on individual waves[19–21], a specific variant replacement event[22,23], or do not explicitly account for variants[24,25]. Here, we showed that our model provides a new way to assess pathways of infections and vaccinations as well as cross-immunity between variants with low prediction uncertainties. By integrating three complementary datasets: antibody, variant, and test positivity data, the model identified the four most dynamic driving variants and accurately mapped the timing of large-scale occurrences of second infections to the delta wave. Surprisingly, with the omicron variant, almost the entire population had a second infection, and third infections also became relevant. The investigation of cross-immunity revealed that a simple model based on the distance in MOIC is sufficient to describe the observed data. The model predicts cross-immunities ranging from 24.5% to 68.6% risk reduction.

The estimates and predictions provide an in-depth assessment of the situation in Ethiopia. On the high level, they also agree with other studies, including the meta-analysis by the COVID-19 Forecasting Team, which used Bayesian meta-regression to pool results of 65 studies from 19 different countries on protection against new variants by past infections with earlier variants[26]. For pooled protection against ancestral variants, which the COVID-19 Forecasting Team uses as a collective term for all variants which occurred earlier than the alpha variant, they obtained protection levels of 84.9% (72.8, 91.8). Comparing their result (95% CI) to our findings (with 90% CIs) of 90.0% (85.3, 94.9) of wildtype and wildtype*, which in our context corresponds to variants earlier than alpha variant, against themselves and 75.5% (72.2, 78.7) against each other, we see that our estimates lay well inside the study's CI (Fig. 4e, Supplementary Fig. 2). Pooled protection against the alpha variant is stated to be 90.0% (54.8, 98.4) while our values range from 53.1% (50.2, 56.0) to 90.0% (85.3, 94.9) (Fig. 4e, column on alpha variant), where our lowest median is only slightly below their CI's lower bound and the CIs overlap (Supplementary Fig. 2). Protection against beta is reported to be 85.7% (83.4, 87.7). Since the beta variant did not play a large role in Ethiopia according to our data, it is not surprising that this very tight interval is not represented by the values of our beta column in Fig. 4e. The eta variant was not explicitly investigated by the COVID-19 Forecasting Team. Delta induces reported protection of 82.0% (63.5, 91.9). Our model suggests lower protection values despite that wildtype, which is the main variant after which delta reinfections happened according to our model, has a median of 63.3% (60.6, 65.9), i.e. for delta the CIs are overlapping. For omicron BA.1, the COVID-19 Forecasting Team states protection levels of 45.3% (17.3, 76.1), which completely covers our values for previous infection with other variants. Only reinfection with BA.1, which does not play a role in our findings, is above this interval. For a meta-analysis on BA.4/5, there were insufficient publications available. They only cite one study[27] with protection levels of 76.2% (66.4, 83.1) for previous omicron BA.1 and 35.5% (12.1, 52.7), where the former is only slightly undercut and the latter slightly exceeded by our median values depicted in the last column of Fig. 4e. Overall, for the variant-variant combinations, which play a major role according to our model and are also part of the meta-study, the cross-immunities we obtained are mostly in accordance with the COVID-19 Forecasting Team's findings. The other variants must be treated more cautiously since either their minor role in our model makes it difficult to compare to the pooled data of the meta-study or the meta-study lacked sufficient statistical power to report on them.

The analysis based on the multivariant model was complemented using a tailored model for the description of antibody levels. The analysis of the available data using this model suggested that antibody decay is slow, in particular for Anti-S antibodies. This is in accordance with other research on SARS-CoV-2 antibody decay[28], although direct comparison of numbers is difficult due to the 2-dimensionality and 3-category setup of our model, tackling the issue of limited individuals participating in all rounds of data collection. Van Elslande et al. reported a median time to 50% seronegativity of 809.6 days in non-severe patients (resp. 985.9 days for severe cases) for Anti-S and 273.1 days in nonsevere patients (resp. 327.3 days for severe cases) for Anti-N[28]. The decay is assay-specific and thus, should be interpreted based on the test system used. We have investigated the decay in unpublished longitudinal cohorts in Munich using the same test system as this study (Ro-N-Ig and Ro-RBD-Ig-quant, for details see methods section) and see similarly slow decay of Anti-N and even slower decay of Anti-S signals. In accordance with the results, our antibody model indicated that, particularly with respect to the S-protein, antibody levels remain in the *high* category in the population to date, suggesting that current vaccinations may have a negligible effect. This is based on the general population, and thus does not take into account additional needs of vulnerable groups which might still benefit from vaccination in this setting of recurrent infection waves. Furthermore, by simulating higher vaccination rates retrospectively, we concluded that it would have been possible to substantially mitigate the delta and omicron waves with more administered vaccines. For the delta wave this is strongly supported by our healthcare worker antibody data, where in August 2021 most of the *high* antibody levels were caused by vaccination in comparison to community members with almost no vaccination, but similarly *high-level* percentages (Fig. 1d, Supplementary Fig. 3d). On the other hand, for omicron we have high uncertainties in our predictions (Fig. 5c). Taking into account the high immune escape property of omicron we would probably still have seen a substantial wave, nevertheless with a notably smaller peak. Moreover, from then on most of the population was exposed multiple times and thus benefits of the titres are less pronounced now.

It is important to approach these findings with caution, since we assessed total levels of antibodies, not neutralizing levels, and the relationship between overall antibody levels and reinfection risk is still an area of ongoing research. There is literature confirming that relative reinfection risk after first infection is around the median 32% that our multivariant model estimated. For example, Iversen et al. present 35% relative risk after first infection of Danish healthcare workers[29]. Transfer of protection data from the literature to Ethiopia is complicated, as the conditions of most studies in the field are vastly different. Protection varies considerably depending on the width of the pre-existing immune response and the time between last exposure and the exposure in question. The magnitude of the measured antibody levels also varies depending on the specificity profile of the antibodies and antigens used in the tests. With larger differences in antigenic structure, cross-protection decreases and variation in the serology results increases.

We focused on analyzing data from community members to investigate the antibody progression associated with SARS-CoV-2 infection. Virus variant-specific information was available for isolates from the clinics also derived primarily from community members and not specifically for healthcare workers. A detailed analysis of the antibody progression among healthcare workers can be found in Supplementary Fig. 3.

The study presented here provides several new insights, but also has weaknesses. On the data collection side, the low number of sequenced swabs after the end of 2021 is problematic. We thus accounted for inhomogeneous sampling in the statistical analysis and the parameter estimation. The models we propose here are based on antibody and variant data from Addis Ababa and Jimma, as well as nation-wide test-positivity rates. While the sampling regions in Addis Ababa and Jimma cover areas of different population density and should prove a broad picture, they might not be fully representative for the spread of SARS-CoV-2 in Ethiopia. An indication for this is that the nation-wide test-positivity rate increases in April 2021 and January 2022, while the antibody data do not show substantial changes at these

time points or briefly afterwards. Hence, the use of the combined dataset for the assessment of Ethiopia is an extrapolation. Moreover, (i) the description of cross-immunity factors as a function solely depending on MOIC neglects that other mutations might also affect immune escape potential, (ii) the dependency of cross-immunity after infections with different variants on the union of mutations from previous variants might overemphasize later variants (since secondary infections are assumed to mainly recall cross-reactive antibodies). Yet, these simplifications were important to ensure computational feasibility and balance model complexity and statistical power in the data. A consideration of all mutations would have increased the number of model parameters by a factor of 9.5 and the dataset would have been insufficient to inform them. Despite its limitations, this study provides an unprecedented insight into the dynamics of COVID-19 infections over time and the impact of the variants in Ethiopia. The findings have valuable implications for current and future research and policy-making, enabling a better understanding of the actual situation and offering potential directions for vaccination policies.

To conclude the dynamics of the SARS-CoV-2 variants in Ethiopia between 2020 and 2022 had similar trends as those observed globally. However, our five rounds of seroepidemiological survey in Addis Ababa and Jimma between August 2020 and April 2022, revealed that in our study group over 96% were exposed at least once to the virus by the last round of our survey. This figure is much higher than in other nation-wide reports. Combining longitudinal serology, viral sequencing data, national test positivity rates, and mathematical modeling, we conclude that most Ethiopians have had multiple exposures to SARS-CoV-2, leading to high antibody titres with slow decay characteristics. Due to recurrent infections with different variants and vaccination in many individuals in Ethiopia, we expect a strong hybrid immunity to date.

The models developed based on the antibody and virus variant dynamics show that earlier and more widespread vaccination of the population would have reduced the overall number of infections considerably. However, the general population has now undergone multiple infections as detected by serology and most likely will not benefit much from further vaccinations, especially if the vaccine still harbors the wild type receptor binding domain sequences. Due to persistent circulation of the virus with obvious underreporting, the main focus for preventive actions should be focused on the most vulnerable groups of the population.

Methods

Ethics

In this study, samples were collected as a follow-up to our previously published work[1].

In brief, we conducted a follow-up investigation on antibody prevalence at two centers in Ethiopia: Jimma Medical Center [JMC] in Jimma and St Paul's Hospital Millennium Medical College in Addis Ababa. The research was approved by the Institutional Review Boards of Jimma University Institute of Health (IHRPGD/978/2020 and IHRPGD/361/2021) and St Paul's Hospital Millennium Medical College (PM23/239/2020 and PM23/003/2020) as well as Ludwig Maximilian University of Munich (21-0293). Further approval from Addis Ababa and Oromia Regional Health Bureaus was also obtained (BEFO/KBTFU/1-16/488). Written informed consent in local languages was obtained prior to admission to the study. For participants unable to read or write, an impartial witness was involved and fingerprints were obtained for consent. Preliminary results were presented to the Ethiopian Public Health Institute, Federal Ministry of Health of Ethiopia, and Ethiopian Medical Association.

Antibody data acquisition

Community members and healthcare workers were recruited for the serology study based on convenience sampling. Hospital workers – including clinical staff, medical interns, cleaners, guards, food handlers, and administrative personnel – were recruited at two hospitals, the St Paul's Hospital in Addis Ababa and the Jimma Medical Center in Jimma. In Addis Ababa, community members from Addis Ketema and Yeka subcities were recruited. In Jimma, no specific region was chosen and rural participants were recruited around the Jimma Zone. Sample sizes were initially calculated in July, 2020, when not much baseline data was available and later became flexible as more data became available. Moreover, as the rate of dropout was more than 30% (our initial expectation), we recruited more participants to compensate for the dropouts (c.f. Supplementary Fig. 1 for detailed studyflow). One participant per household was sampled to avoid any clustering effects and households were selected randomly in a way that avoided frequent interaction from the next candidate household to prevent cross-contamination. Overall the median age was 30 with 90% percentile (20,60) and 55.6% of participants, which provided information about sex were female (for round and site-specific demographics see Table 1 and Supplementary Table 1). All participants of the first 3 rounds were enrolled before the introduction of COVID-19 vaccines in Ethiopia. In later rounds participants provided their vaccination status and dates through a questionnaire. For more details see in-depth description in Gudina et al.[1].

In total, 3 ml of venous blood was collected in standard serum tubes. After full coagulation at room temperature, serum was harvested by centrifugation and stored at −20 °C on the same day as sampling. The Roche Elecsys® anti-SARS-CoV-2 [Ro-N-Ig] and the Roche Elecsys® anti-SARS-CoV-2 S [Ro-RBD-Ig-quant] were used for serologic analysis. Both assays are double-antigen sandwich assays, detecting antibodies of all subclasses against SARS-CoV-2. Measurements were performed on a Cobas e801 analytical unit (Roche Diagnostics, Basel, Switzerland) in Munich, Germany, or a Cobas e601 unit (Roche Diagnostics, Basel, Switzerland) in Jimma and Addis Ababa, Ethiopia, using electrochemiluminescence (ELECSYS) technology.

The Ro-RBD-Ig-quant assay uses a truncated S1 protein as an antigen and is a quantitative assay validated for use with human serum and plasma. It is linear between 0.4 and 250 Units (U) per ml, which are equivalent to the standardized (WHO publication WHO/BS.2020.2403) BAU (Binding Antibody Units) according to the manufacturer's manual. Values above 250 U/ml were diluted in 10-fold until the linear range was reached according to the manufacturer's procedures. Values in this study were measured within the linear ranges and back-calculated depending on the dilution as appropriate.

The Ro-N-Ig assay is a qualitative assay similar to Ro-RBD-Ig-quant, but using nucleocapsid as an antigen. The results are given as cut off index (COI), and only a cutoff for positivity is provided by the manufacturer. A linear range is not officially established. We use the raw COI values in a semi-quantitative manner, as we have observed a good dynamic range and excellent repeatability of the values. Anti-N measurements were not diluted, so can be outside the linear range in this work.

Variant data acquisition

A total of 1873 SARS-CoV-2 RT-PCR positive swabs were collected in Jimma and Addis Ababa, Ethiopia between October 2020 and July 2022. Sample dates were not always available as an exact date, but rather month and year only. Therefore, the midpoint of the respective sampling month was used for all samples analyzed. The swabs were collected from individuals presenting with COVID-19-related symptoms, contacts of confirmed COVID-19 cases, and high-risk populations such as healthcare workers. The specimens were collected at Jimma Medical Center and St. Paul's Hospital.

Jimma Medical Center in Jimma Town and St. Paul Hospital in Addis Ababa are among the major COVID-19 testing and treatment sites in Ethiopia. Jimma COVID-19 center serves as the only COVID-19 diagnostic facility in southwest Ethiopia, home to about 20 million

inhabitants. It is the only facility with intensive care for severe COVID-19 cases in the region. St. Paul Hospital in Addis Ababa is a public tertiary referral hospital serving as a COVID-19 diagnostic and treatment center for Addis Ababa and surrounding areas.

All RT-PCR-positive specimens were stored at −80 °C at these two sites during the study period. Specimens in poor storage conditions and those without proper documentation of data collection dates were excluded. The stored samples were transported on dry ice to Munich in Germany. There, whole nucleic acid extraction was performed using the tanbead maelstrom 4800 instrument (TANBead, Taiwan) and the TANBead Optipure Viral Auto Tube / Plate extraction kits (TANBead, Taiwan). cDNA of the extracts was generated using the LunaScript one step RT (New England Biolabs).

Following the ARTIC network nCoV-2019 sequencing protocol v2[30], amplicons spanning the whole SARS-CoV2 genome were amplified from the cDNA samples. The resulting products were pooled, tagmented with NexteraXT library prep kit (Illumina, San Diego, USA), barcoded, and sequenced on an Illumina NextSeq 2000. For each sample, the sequenced reads were demultiplexed and mapped to the SARS-CoV-2 reference genome (NC 045512.2) with bwa-mem[31]. The consensus sequences were obtained from the sequenced amplicons using the iVar package[32]. Briefly, the package trims the primer sequences from the mapped reads and filters them by a base quality >20 and minimal read length of 30 nt. Pileup files are generated from the mapped reads which are used to assemble the consensus sequence. The consensus sequence was assigned to SARS-CoV-2 lineages using the Pangolin tool[33].

Analysis of antibody data

To ensure a broad analysis, we merged the data collected for community members in Addis Ababa and Jimma for each round, as the timing of the sampling campaigns overlapped significantly. This allowed us to combine the data effectively and to capture a more comprehensive picture of the antibody dynamics in these communities.

To facilitate meaningful analysis while preserving the relative order of magnitude and accounting for zero measurements, we transformed the antibody measurements using the shifted logarithm base 10 function ($\log_{10}(x+1)$). This transformation enabled us to easily analyze the data across different scales while still maintaining the interpretation of zero as the absence of detectable antibodies.

For categorizing the antibody levels, we considered measurements for each antibody type independently, disregarding the round in which they were obtained. Anti-N values and measurements below the predefined cutoff were excluded from the analysis. We performed k-means clustering with two means, i.e. $k = 2$, on the remaining samples to assign the measurements into distinct antibody level categories.

Smooth changes in antibody levels over time were visualized using a monotonic spline-fitting approach. This allowed us to capture the overall trend and highlight gradual variations in the antibody responses.

To ensure an adequate number of data points for model fitting while remaining reasonable errors for the analysis, we performed k-means clustering ($k = 2$) on the dates of each round. Subsequently, we split each round into two subgroups based on the clustering results and aggregated the antibody responses within these subgroups. Additionally, to estimate high-confidence intervals for error analysis, we fitted a multinomial model to the distribution of the three antibody categories.

To estimate vaccination rates in our study, we employed a fitting approach using monotonic splines applied to the vaccination information provided by the participants, allowing us to capture the temporal trends and variations in vaccination rates accurately. For comprehensive details on the specific methodologies and results of the vaccination rate estimation, we refer readers to the supplementary materials.

For more detailed information and results, we encourage readers to refer to the corresponding sections in our manuscript.

Analysis of variant data

Whole genome sequencing and subsequent analysis utilized Nextstrain's[34] Augur software, coupled with Auspice for phylogenetic analysis and visualization. To classify the sequenced genomes, we employed pango lineages[33] and grouped them based on shared mutations of interest or concern (MOIC) on the S1 protein according to outbreak.info[12,13]. To quantify the genetic distances between these variant groups, we utilized the Hamming distance, a metric often used to measure distance in gene alignment[14–16]. Here we calculate the distance only based on different MOIC and not all mutations to grasp only the major immune escape changing differences. For our models below we allow for additional behavioral differences independently of this distance. To capture the temporal dynamics of variant prevalence, we organized the samples according to the month of collection and calculated the fractions of each variant. For a smooth visualization of these trends, we applied monotonic spline fitting to generate smoothed curves. To estimate the errors for later parameter estimation, we utilized a multinomial model and fitted it to the monthly variant distributions. To obtain an input function of variant distribution for the antibody level category model while maintaining a reasonable level of complexity, we aggregated the samples over two-month intervals before applying monotonic spline fitting. These procedures allowed us to effectively characterize the variant dynamics and obtain essential inputs for subsequent modeling analyses.

Modeling

The model-based analysis was performed using compartment models. Utilizing the SEIR (susceptible, exposed, infectious, and recovered) framework, which has been shown to be reliable for modeling the spread of Covid-19[1,35], we aimed to analyze and predict the dynamics of the pandemic.

For the multivariant model we constructed pathways, i.e., chains of SEIR strands, allowing up to four consecutive infections or vaccinations. Pathways which deviated from the chronological order of variant appearances worldwide were excluded. Furthermore, the model only allows for a third infection with the two omicron variants and a fourth infection exclusively by omicron BA.4/5 to account for the reported inter-infection intervals. We allow for different transmission rates for each variant—thereby implicitly considering all mutations – and model their cross-immunity as a function of difference in MOIC. Rates for first, second and third vaccination were estimated a priori as splines from the vaccination information of the antibody study participants and implemented as time-dependent functions into the model.

The antibody-level model does not trace pathways of variants and infections, but categories of antibody levels for Anti-S and Anti-N. Here the SEIR strands are connecting the categories allowing for a boost in antibody levels by infection and recovery. For this model the vaccination is calculated a priori as an average vaccination rate and implemented as a time-dependent function into the model. Moreover, we made the assumptions that people with already high Anti-S levels do not get vaccinated anymore, i.e., the amount of people still applying for vaccination after two infections or vaccinations is negligible. Because of the non-pathway nature of this model we also fitted the variant distribution a priori and used this fit as weights for a sum over the variants' transmission rates to obtain an effective transmission rate. The exact formula for this can be found in Supplementary Note 3 of the Supplementary Information.

The models were encoded using the Systems Biology Markup Language (SBML)[36] and simulated via the software toolbox AMICI[37]. More comprehensive details regarding the modeling methodology are provided in the model subsections of Supplementary Notes 2 and 3 of the Supplementary Information.

Parameter estimation

To estimate the model parameters, we adopted a Bayesian approach, integrating categorial antibody data and sequenced variant information, along with national test positivity rates and previous knowledge derived from the literature regarding disease progression rates. The model parameter inference was performed using an adaptive Metropolis-Hastings algorithm from a starting point estimated with frequentistic, gradient-based optimization, both expertly implemented in the Python Parameter Estimation Toolbox (pyPESTO)[38]. In order to capture the temporal dynamics of the antibody levels, we split each antibody round into early and late phases using the k-means clustering technique. The resulting samples from the posterior distribution were post-processed, e.g., by removing the burn-in, and convergence was assessed visually and using the Geweke test. The samples were then utilized to derive predictions and associated credible intervals (CIs), providing valuable insights into the dynamics of the pandemic. The parameter estimation problems were formulated using the Parameter Estimation table (PEtab)[39] standard. More information on the parameter estimation setup and results can be found in the corresponding subsections of Supplementary Notes 2 and 3.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The models and population average data are available at Zenodo [https://doi.org/10.5281/zenodo.10871139]. The variant sequences are published in the Sequence Read Archive[40] under project number PRJNA1017685. Individual level data will be made available to other researchers in a reasonable timeframe upon qualified request to the corresponding authors AK and AW, due to limitations of data sharing in the ethics statements. Source data are provided with this paper.

Code availability

The code for model creation,data aggregation and figure plotting is available at Zenodo [https://doi.org/10.5281/zenodo.10871139].

References

1. Gudina, E. K. et al. Seroepidemiology and model-based prediction of SARS-CoV-2 in Ethiopia: longitudinal cohort study among frontline hospital workers and communities. *Lancet Glob. Health* **9**, e1517–e1527 (2021).
2. Mulu, A. et al. The challenges of COVID-19 testing in Africa: the Ethiopian experience. *Pan Afr. Med. J.* **38**, 6 (2021).
3. Lamptey, E., Senkyire, E. K., Benita, D. A. & Boakye, E. O. COVID-19 vaccines development in Africa: a review of current situation and existing challenges of vaccine production. *Clin. Exp. Vaccin. Res.* **11**, 82–88 (2022).
4. Sahile, A. T., Gizaw, G. D., Mgutshini, T., Gebremariam, Z. M. & Bekele, G. E. COVID-19 Vaccine Acceptance Level in Ethiopia: A Systematic Review and Meta-Analysis. *Can. J. Infect. Dis. Med. Microbiol.* **2022**, 2313367 (2022).
5. Tonen-Wolyec, S., Mbumba Lupaka, D.-M., Batina-Agasa, S., Mbopi Keou, F.-X. & Bélec, L. Review of authorship for COVID-19 research conducted during the 2020 first-wave epidemic in Africa reveals emergence of promising African biomedical research and persisting asymmetry of international collaborations. *Trop. Med. Int. Health* **27**, 137–148 (2022).
6. Abagero, A. et al. A Review of COVID-19 Response Challenges in Ethiopia. *Int. J. Environ. Res. Public Health* **19**, 11070 (2022).
7. Gelanew, T. et al. High seroprevalence of anti-SARS-CoV-2 antibodies among Ethiopian healthcare workers. *BMC Infect. Dis.* **22**, 261 (2022).
8. Abdella, S. et al. Prevalence of SARS-CoV-2 in urban and rural Ethiopia: Randomized household serosurveys reveal level of spread during the first wave of the pandemic. *EClinicalMedicine* **35**, 100880 (2021).
9. Sisay, A. et al. Molecular Epidemiology and Diversity of SARS-CoV-2 in Ethiopia, 2020-2022. *Genes* **14**, 705 (2023).
10. Olbrich, L. et al. Head-to-head evaluation of seven different seroassays including direct viral neutralisation in a representative cohort for SARS-CoV-2. *J. Gen. Virol.* **102**, 001653 (2021).
11. Le Gleut, R. et al. The representative COVID-19 cohort Munich (KoCo19): from the beginning of the pandemic to the Delta virus variant. *BMC Infect. Dis.* **23**, 1–15 (2023).
12. Gangavarapu, K. et al. Outbreak.info genomic reports: scalable and dynamic surveillance of SARS-CoV-2 variants and mutations. *Nat. Methods* **20**, 512–522 (2023).
13. Tsueng, G. et al. Outbreak.info Research Library: a standardized, searchable platform to discover and explore COVID-19 resources. *Nat. Methods* **20**, 536–540 (2023).
14. Mohammadi-Kambs, M., Hölz, K., Somoza, M. M. & Ott, A. Hamming Distance as a Concept in DNA Molecular Recognition. *ACS Omega* **2**, 1302–1308 (2017).
15. Kindhi, B. A., Hendrawan, M. A., Purwitasari, D., Sardjono, T. A. & Purnomo, M. H. Distance-based pattern matching of DNA sequences for evaluating primary mutation. in *2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)* 310–314. https://doi.org/10.1109/ICITISEE.2017.8285518 (2017).
16. Chen, Z., Bancej, C., Lee, L. & Champredon, D. Antigenic drift and epidemiological severity of seasonal influenza in Canada. *Sci. Rep.* **12**, 15625 (2022).
17. Ethiopia launches a COVID-19 vaccination campaign targeting the 12 years and above population. *WHO | Regional Office for Africa* https://www.afro.who.int/news/ethiopia-launches-covid-19-vaccination-campaign-targeting-12-years-and-above-population.
18. Ma, K. C. et al. Trends in Laboratory-Confirmed SARS-CoV-2 Reinfections and Associated Hospitalizations and Deaths Among Adults Aged ≥18 Years – 18 U.S. Jurisdictions, September 2021-December 2022. *MMWR Morb. Mortal. Wkly. Rep.* **72**, 683–689 (2023).
19. Habenom, H., Aychluh, M., Suthar, D. L., Al-Mdallal, Q. & Purohit, S. D. Modeling and analysis on the transmission of covid-19 Pandemic in Ethiopia. *Alex. Eng. J.* **61**, 5323–5342 (2022).
20. Nkwayep, C. H., Bowong, S., Tsanou, B., Alaoui, M. A. A. & Kurths, J. Mathematical modeling of COVID-19 pandemic in the context of sub-Saharan Africa: a short-term forecasting in Cameroon and Gabon. *Math. Med. Biol.* **39**, 1–48 (2022).
21. Akuka, P. N. A., Seidu, B. & Bornaa, C. S. Mathematical Analysis of COVID-19 Transmission Dynamics Model in Ghana with Double-Dose Vaccination and Quarantine. *Comput. Math. Methods Med.* **2022**, 7493087 (2022).
22. Khan, M. A. & Atangana, A. Mathematical modeling and analysis of COVID-19: A study of new variant Omicron. *Phys. A* **599**, 127452 (2022).
23. Li, X.-P. et al. Assessing the potential impact of COVID-19 Omicron variant: Insight through a fractional piecewise model. *Results Phys.* **38**, 105652 (2022).
24. Wangari, I. M. et al. Mathematical Modelling of COVID-19 Transmission in Kenya: A Model with Reinfection Transmission Mechanism. *Comput. Math. Methods Med.* **2021**, 5384481 (2021).
25. Oke, A. S., Bada, O. I., Rasaq, G. & Adodo, V. Mathematical analysis of the dynamics of COVID-19 in Africa under the influence of asymptomatic cases and re-infection. *Math. Methods Appl. Sci.* **45**, 137–149 (2022).
26. Stein, C. et al. Past SARS-CoV-2 infection protection against re-infection: a systematic review and meta-analysis. *Lancet* **401**, 833–842 (2023).

27. Altarawneh, H. N. et al. Protection against the Omicron Variant from Previous SARS-CoV-2 Infection. *N. Engl. J. Med.* **386**, 1288–1290 (2022).
28. Van Elslande, J. et al. Lower persistence of anti-nucleocapsid compared to anti-spike antibodies up to one year after SARS-CoV-2 infection. *Diagn. Microbiol. Infect. Dis.* **103**, 115659 (2022).
29. Iversen, K. et al. Seroprevalence of SARS-CoV-2 antibodies and reduced risk of reinfection through 6 months: a Danish observational cohort study of 44 000 healthcare workers. *Clin. Microbiol. Infect.* **28**, 710–717 (2022).
30. Quick, J. nCoV-2019 sequencing protocol v2 (GunIt) v2 https://doi.org/10.17504/protocols.io.bdp7i5rn (2020).
31. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013).
32. Grubaugh, N. D. et al. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biol.* **20**, 8 (2019).
33. O'Toole, Á. et al. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evol.* **7**, veab064 (2021).
34. Hadfield, J. et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).
35. Raimúndez, E. et al. COVID-19 outbreak in Wuhan demonstrates the limitations of publicly available case numbers for epidemiological modeling. *Epidemics* **34**, 100439 (2021).
36. Hucka, M. et al. The Systems Biology Markup Language (SBML): Language Specification for Level 3 Version 2 Core Release 2. *J. Integr. Bioinform.* **16**, 20190021 (2019).
37. Fröhlich, F. et al. AMICI: high-performance sensitivity analysis for large ordinary differential equation models. *Bioinformatics* **37**, 3676–3677 (2021).
38. Schälte, Y. et al. pyPESTO: A modular and scalable tool for parameter estimation for dynamic models. *Bioinformatics* **39**, btad711 (2023).
39. Schmiester, L. et al. PEtab-Interoperable specification of parameter estimation problems in systems biology. *PLoS Comput. Biol.* **17**, e1008646 (2021).
40. Leinonen, R., Sugawara, H. & Shumway, M., International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic Acids Res* **39**, D19–D21 (2011).

Acknowledgements

We are grateful for research funding provided by the Bavarian State Ministry of Sciences, Research and the Arts (Bayerisches Staatsministerium, F.4-V0122.4/3/20 (AK, AW)); the Germany Ministry of Education and Research (MoKoCo19; 01KI20271 (JH, AW) and FitMultiCell; 031L0159C (SM, JH) and INSIDe; 031L0297A (SM, JH, AW) and GEN-Immune; 031L0292F (SM, JH)); the EU Horizon 2020 programme (ORCHESTRA; 101016167 (JH, AW)); and Volkswagenstiftung (E2; 99 450 (SM, JH)). This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) via project funding (SEPAN; 7376/3-1 (JH)), under Germany's Excellence Strategy (project IDs 390685813 - EXC 2047 (JH) and 390873048 - EXC 2151 (JH)), and by the University of Bonn via the Schlegel professorship to JH. This study was partially funded by the Free State of Bavaria under the FORCOVID (MM) and BayVOC (MM) research initiatives. We thank participants, study teams, Jimma Medical Center, Oromia Regional Health Bureau, St Paul's Hospital Millennium Medical College, and Addis Ababa Health Bureau for the support provided during data collection.

Author contributions

E.K.G., S.A., A.K., J.H., M.H. and A.W. conceived of and designed the study. E.K.G., S.A., W.A., Gize, S.S.B., B.F., D.B., M.G., E.G., D.Y., A.Z. participated in the data and sample collection. S.A., Gize, M.B., R.R.A., I.P., M.G., A.W. performed serologic analysis. M.M., Graf, S.K., and H.B. performed cDNA synthesis and sequencing. K.E., R.K., J.H. and S.M. summarized, cleaned, and analyzed the data. S.M. and J.H. did the modeling and parameter estimation. E.K.G., S.A., A.K., J.H., M.H., M.M., A.W. interpreted the results. S.M., E.K.G., S.A., J.H., A.K., and A.W. drafted the manuscript. All authors contributed to the writing of the final version of the manuscript. S.M., E.K.G., S.A., K.E., and A.W. have accessed and verified the data; all authors accepted responsibility for the decision to submit for publication.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare the following competing interests: The medical center of the LMU received reagents and an analyzer from Roche with reduced rates for other studies regarding SARS-CoV-2 serology. MH and AW received different consultancy contracts and support for studies regarding SARS-CoV-2 serology, independent of this project. This did, however, not influence the interpretation of the data, or the data reported. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-47556-2.

Correspondence and requests for materials should be addressed to Jan Hasenauer, Arne Kroidl or Andreas Wieser.

Peer review information *Nature Communications* thanks Christian Bottomley, James San and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

[1]Life and Medical Sciences (LIMES), University of Bonn, Bonn, Germany. [2]Saint Paul's Hospital Millennium Medical College, Addis Ababa, Ethiopia. [3]Jimma University Clinical Trial Unit, Jimma University Institute of Health, Jimma, Ethiopia. [4]CIH LMU Center for International Health, LMU Munich, Munich, Germany. [5]Max von Pettenkofer Institute and Gene Center, Virology, National Reference Center for Retroviruses, LMU Munich, Munich, Germany. [6]German Center for Infection Research (DZIF), partner site Munich, Munich, Germany. [7]Laboratory for Functional Genome Analysis, Gene Center, LMU Munich, Munich, Germany. [8]Division of Infectious Diseases and Tropical Medicine, LMU University Hospital, LMU Munich, Munich, Germany. [9]Institute for Medical Information Processing, Biometry and Epidemiology (IBE), Faculty of Medicine, LMU Munich, Munich, Germany. [10]Immunology, Infection and Pandemic Research IIP, Fraunhofer ITMP, Munich, Germany. [11]Unit Global Health, Helmholtz Zentrum München—German Research Center for Environmental Health, Neuherberg, Germany. [12]Institute of Computational Biology, Helmholtz Zentrum München—German Research Center for Environmental Health, Neuherberg, Germany. [13]Center for Mathematics, Technische Universität München, Garching, Germany. [14]Faculty of Medicine, Max Von Pettenkofer Institute, LMU Munich, Munich, Germany. [15]These authors contributed equally: Simon Merkt, Solomon Ali, Esayas Kebede Gudina. [16]These authors jointly supervised this work: Jan Hasenauer, Arne Kroidl, Andreas Wieser. ✉e-mail: jan.hasenauer@uni-bonn.de; akroidl@lrz.uni-muenchen.de; andreas.wieser@lmu.de

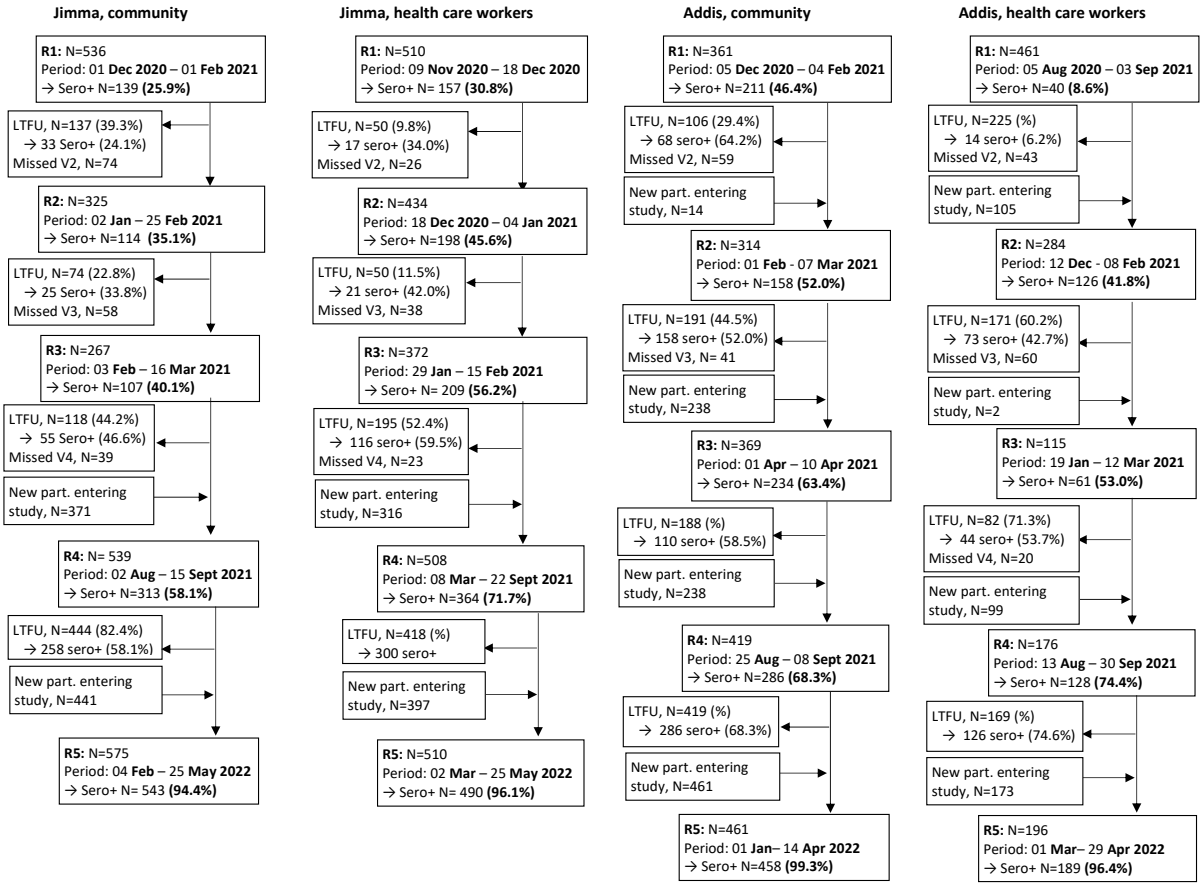
Supplementary Information to long-term monitoring of SARS-CoV-2 seroprevalence and variants in Ethiopia provides prediction for immunity and cross-immunity

Simon Merkt[*1], Solomon Ali[*2], Esayas Kebede Gudina[*3], Wondimagegn Adissu[3], Addisu Gize[2,4], Maximilian Muenchhoff[5,6], Alexander Graf[7], Stefan Krebs[7], Kira Elsbernd[8,9], Rebecca Kisch[8], Sisay Sirgu[2], Bereket Fantahun[2], Delayehu Bekele[2], Raquel Rubio-Acero[8], Mulatu Gashaw[3], Eyob Girmai[3], Daniel Yilma[3], Ahmed Zeynudin[3], Ivana Paunovic[8,10], Michael Hoelscher[6,8,10,11], Helmut Blum[7], Jan Hasenauer[+1,12,13], Arne Kroidl[+6,8], and Andreas Wieser[+,6,8,10,14]

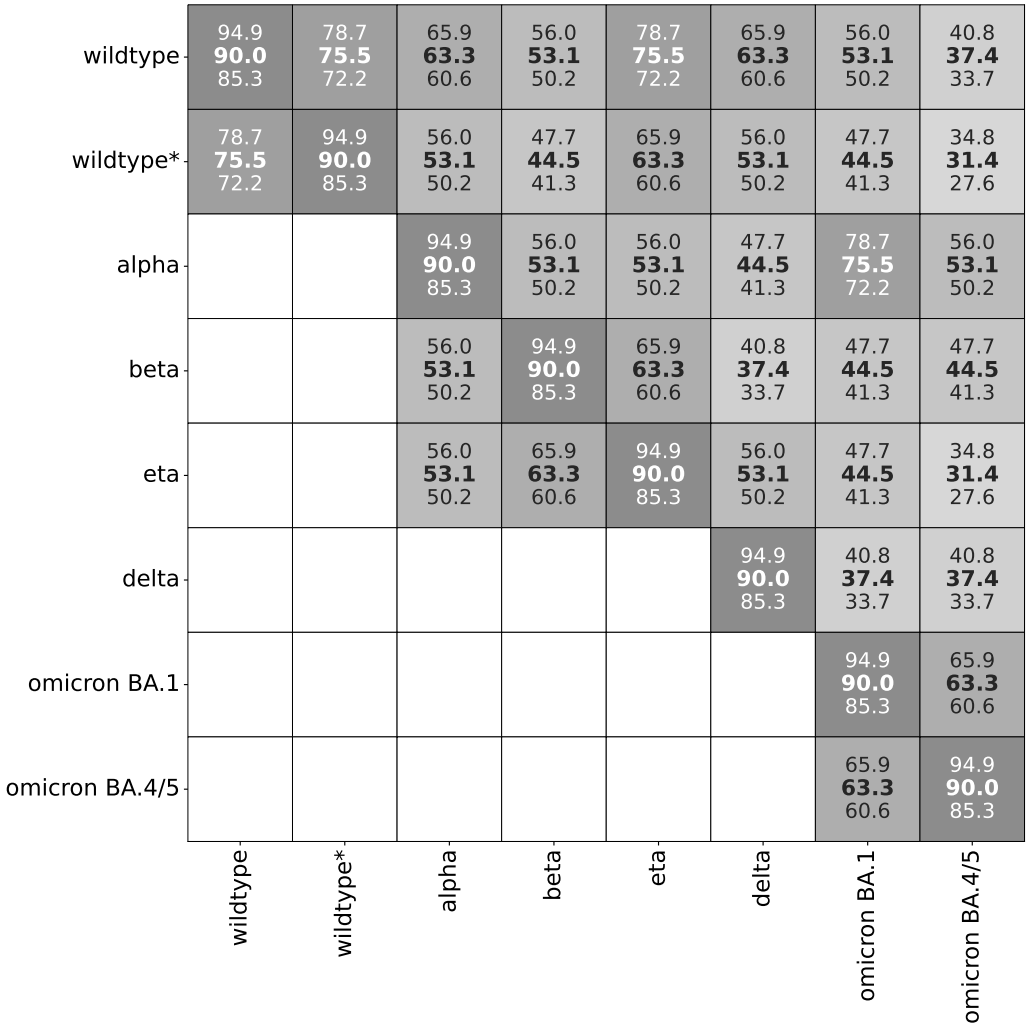
[1]Life and Medical Sciences (LIMES), University of Bonn, Bonn, Germany
[2]Saint Paul's Hospital Millennium Medical College, Addis Ababa, Ethiopia
[3]Jimma University Clinical Trial Unit, Jimma University Institute of Health, Jimma, Ethiopia
[4]CIH[LMU] Center for International Health, LMU Munich, Munich, Germany
[5]Max von Pettenkofer Institute and Gene Center, Virology, National Reference Center for Retroviruses, LMU Munich, Munich, Germany
[6]German Center for Infection Research (DZIF), partner site Munich, Munich, Germany
[7]Laboratory for Functional Genome Analysis, Gene Center, LMU Munich, Munich, Germany
[8]Division of Infectious Diseases and Tropical Medicine, Medical Center LMU Munich, Munich, Germany
[9]Institute for Medical Information Processing, Biometry, and Epidemiology, LMU Munich, Munich, Germany
[10]Immunology, Infection and Pandemic Research IIP, Fraunhofer ITMP, Munich, Germany
[11]Unit Global Health, Helmholtz Zentrum München—German Research Center for Environmental Health, Neuherberg, Germany
[12]Institute of Computational Biology, Helmholtz Zentrum München—German Research Center for Environmental Health, Neuherberg, Germany
[13]Center for Mathematics, Technische Universität München, Garching, Germany
[14]Faculty of Medicine, Max Von Pettenkofer Institute, LMU Munich, Munich, Germany
[*]These authors contributed equally
[+]Corresponding authors: jan.hasenauer@uni-bonn.de, akroidl@lrz.uni-muenchen.de, andreas.wieser@lmu.de

Contents

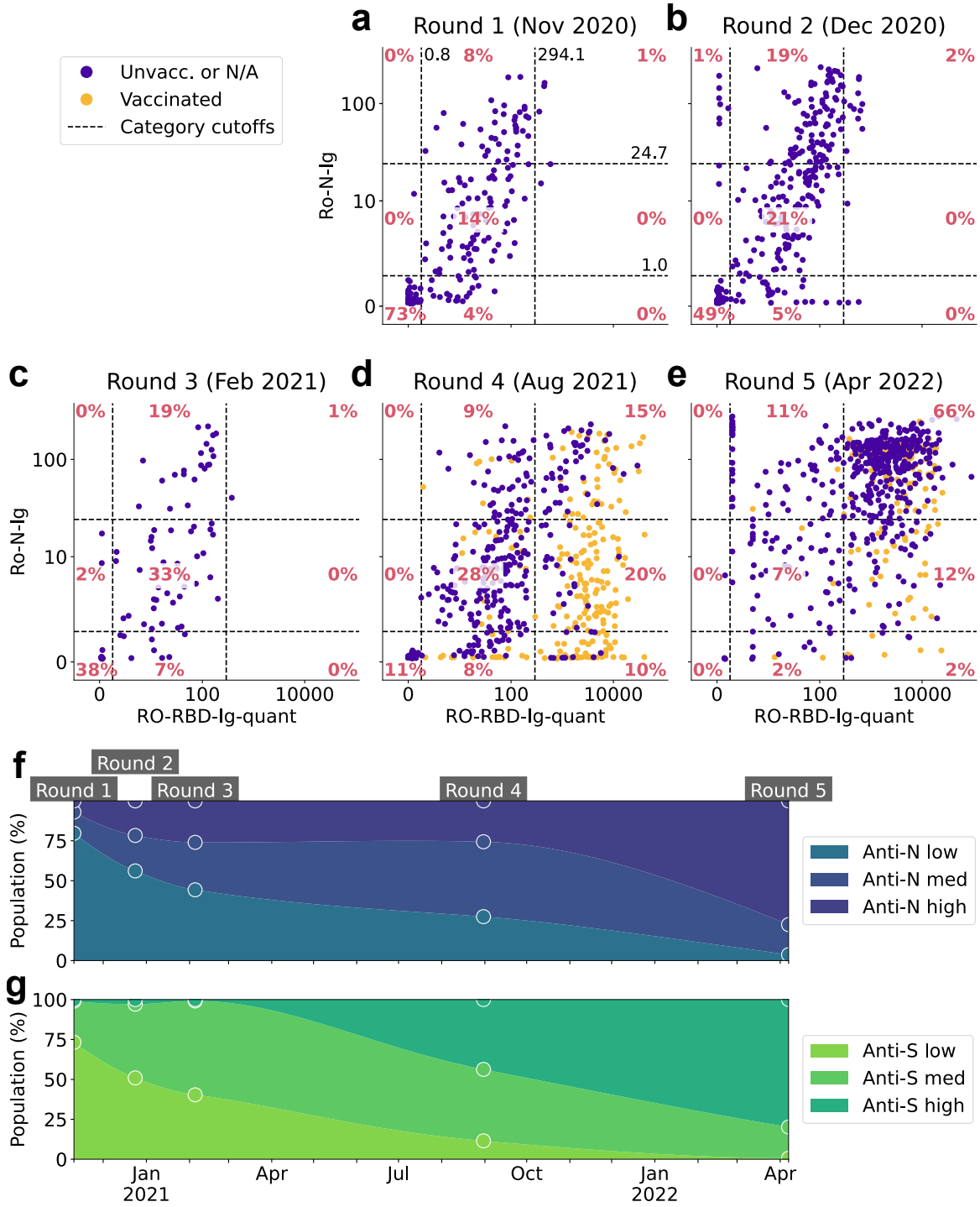
1 Supplementary Figures



Supplementary Figure 1. Study flow and point prevalence for SARS CoV-2 seropositivity in healthcare workers and community members recruited in Jimma and Addis Ababa including long-term follow-up (LTFU) numbers and percentages.



Supplementary Figure 2. Heatmap of median (bold) cross-immunity levels between variants including 90% CIs (n=6001 samples after burn-in from Markov chain Monte Carlo). Source data are provided as a Source Data file.



Supplementary Figure 3. Ro-N-Ig and Ro-RBD-Ig-quant measurements of five rounds of convenience sampled healthcare workers. a–e Scatterplots displaying the relationship between levels of N- and S-specific antibodies (y-axis, resp. x-axis) across five rounds of measurement. Known vaccination status of each participant indicated by colors, cutoff levels indicated by dashed lines and percentages of people per category annotated in red. **f–g** Evolution of antibody levels over time between Fall of 2020 and April 2022. Times of sample acquisition are highlighted as circles. Source data are provided as a Source Data file.

2 Supplementary Table

Supplementary Table 1. Demographic characteristics of healthcare workers study participants. Age denoted as median and 90% quantiles. Round 1-3 (R1-R3) are the previous study of Gudina et al 2021.

	Jimma Medical Center					St Paul's Hospital				
	R1 (Nov 20)	R2 (Dec 20)	R3 (Feb 21)	R4 (Aug 21)	R5 (Apr 22)	R1 (Aug 20)	R2 (Dec 20)	R3 (Feb 21)	R4 (Sep 21)	R5 (Apr 22)
Participants	510	434	372	508	510	461	284	116	176	196
Age	26 (22, 39)	26 (23, 41)	26 (23, 39)	28 (21, 39)	29 (23, 50)	28 (22, 42)	28 (20, 42)	26 (20, 42)	26 (21, 42)	30 (23, 40)
Sex										
Female	271 (53.1%)	231 (53.2%)	199 (53.5%)	273 (53.7%)	68 (13.3%)	236 (51.2%)	103 (36.3%)	44 (37.9%)	92 (52.3%)	4 (2.0%)
Male	239 (46.9%)	203 (46.8%)	173 (46.5%)	233 (45.9%)	45 (8.8%)	222 (48.2%)	76 (26.8%)	30 (25.9%)	56 (31.8%)	4 (2.0%)
Missing	0 (0.0%)	0 (0.0%)	0 (0.0%)	2 (0.4%)	397 (77.8%)	3 (0.7%)	105 (37.0%)	42 (36.2%)	28 (15.9%)	188 (95.9%)
Anti-N positive	157 (30.8%)	198 (45.6%)	209 (56.2%)	364 (71.7%)	490 (96.1%)	40 (8.7%)	112 (39.4%)	60 (51.7%)	128 (72.7%)	189 (96.4%)
Vaccinated	0 (0.0%)	0 (0.0%)	0 (0.0%)	217 (42.7%)	149 (29.2%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	71 (40.3%)	5 (2.6%)

3 Supplementary Note 1: Analysis of Antibody and Variant Data

Clustering of Antibody Data

The distribution for the Anti-S antibody levels in Figure SN1 has three distinct peaks: one peak close to zero, one peak at 2 and one peak at 3.5. Comparing to the distributions with reactivity of Anti-N and with the vaccination information one can derive that the first of those two peaks corresponds to one infection or vaccination and the second to two or more infections or vaccinations.

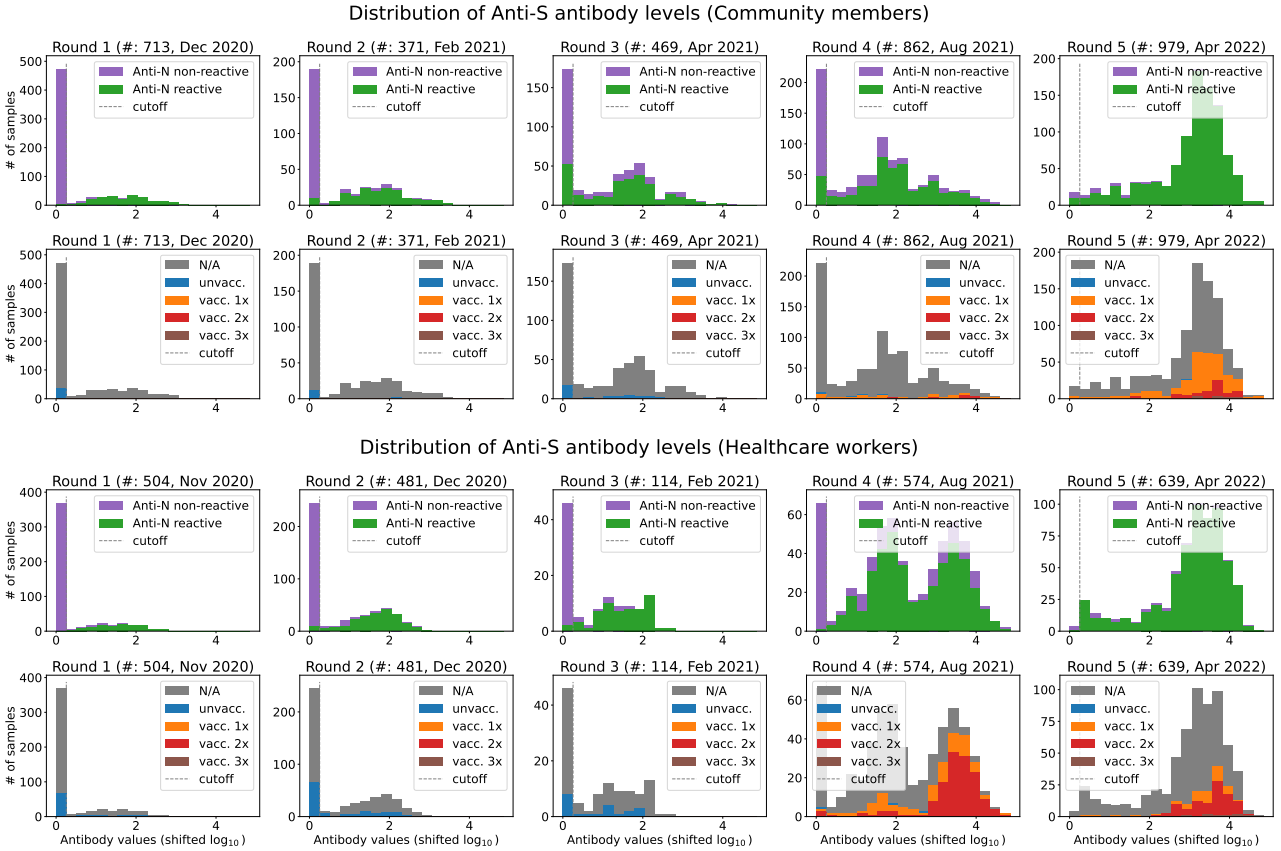


Figure SN1. Distributions of Anti-S antibody levels for healthcare workers and community members. Reactivity of Anti-N and vaccination status highlighted by colors (first resp. second row for each group). Source data are provided as a Source Data file.

As we use a compartment model for the subsequent analysis, we decided to categorize the antibody measurements based on the observed groups. Therefore, we combined all community measurements, respectively healthcare worker measurements, from different sites and rounds. First, we individually processed N and S measurements, excluding NaN values and measurements below the detection threshold. We utilized scikit-learn's k-means clustering implementation to categorize the remaining data points above the threshold into two distinct groups[1]. We chose clustering the antibody datasets separately, i.e. 1-dimensional clustering, motivated by the bi-modal distributions we observed in the histograms for Anti-S. Moreover, the separate clustering of the Anti-N or Anti-S data provides: (i) a slightly higher statistical power, since for some study participants only one the antibody tests, Anti-N or Anti-S, was successful; and (ii) clear cutoff values for aggregated Anti-S measurements (e.g. by using midpoint of the two

resulting groups' centers), which is necessary for the multivariant model. The performance of the k-means clustering can be observed in Figure SN2.

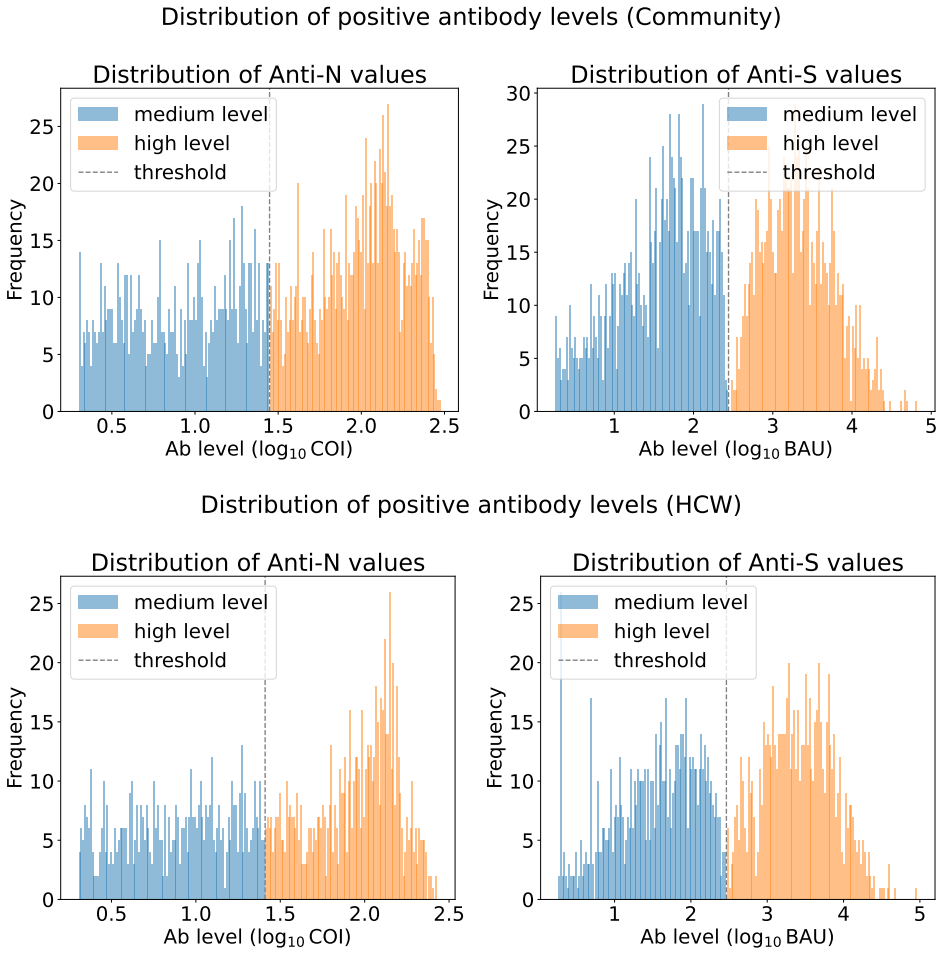


Figure SN2. Distribution of positive antibody measurements for community members and healthcare workers (HCW). Thresholds computed with k-means clustering are highlighted. Source data are provided as a Source Data file.

To visualize the data, the three groups (below the detection limit, above the limit but below the category separation, and above the category separation) were aggregated for each round. We employed the monotonic cubic spline interpolation of the scipy[2] package to interpolate the resulting values (Figure 1 of the main manuscript).

There was no vaccine publicly available in Ethiopia until after Round 3[3]. Because of this information about general vaccine availability in combination with our previous observation that vaccinated individuals are more likely to answer questions on the vaccination status on the questionnaire than unvaccinated individuals, we considered individuals without an answer ("N/A") as "unvaccinated" for modelling. This is also supported by official nation wide numbers of people with at least one dose of vaccine, provided by Our World in Data (ourworldindata.org) and depicted in Figure SN3. Moreover, we treat the effect of vaccine and infection on Anti-S levels analogously. This is based on the comparison of the observed antibody levels for healthcare workers (Supplementary Figure 1) and community members (Figure 1). There from Round 3 to Round 4 for healthcare workers a clear shift from medium

Anti-S to high Anti-S is observed in response to vaccination, but community members reach the same levels by infections alone.

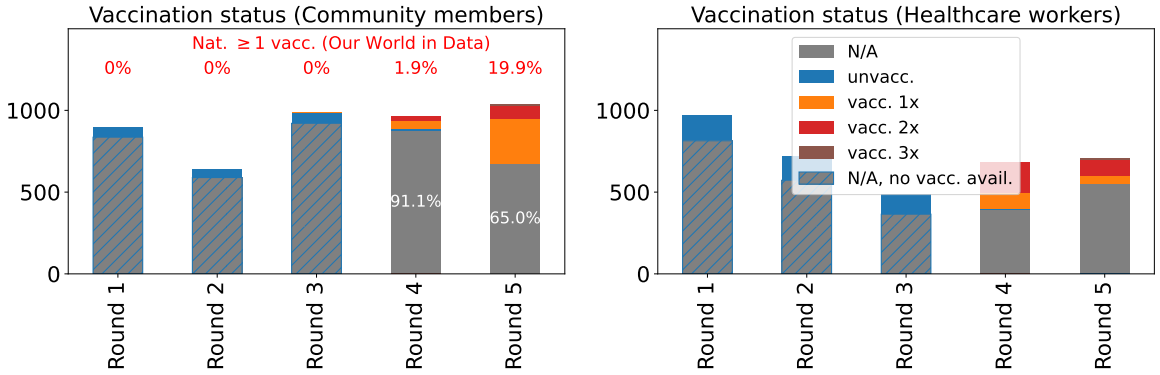


Figure SN3. Histograms of distributions of vaccination information from study participants at each round. "N/A" responses before public availability of vaccine in Ethiopia are highlighted by hatching. For community members official, national vaccination numbers (provided by Our World in Data) are indicated in red above each round and percentages from our data set of "N/A" responses after public availability of vaccines in Ethiopia are displayed inside of the corresponding bars. Source data are provided as a Source Data file.

Sequencing Result of Variant Data

For the sequencing data, we merged the data sets from Addis Ababa and Jimma sites and removed entries, where sequencing failed. The observed Pango[4] lineages were assessed for Mutations of Interest or Concern (MOIC) by referencing the outbreak.info[5,6] database. Based on these mutations, the lineages were grouped and groups lacking sufficient statistical power, i.e. sample size below 3, were dropped from the data set. The complete list of observed lineages and their mutations is provided in Table SN1. The samples were then aggregated by the month of collection and interpolated using scipy's monotonic cubic spline interpolation for visualization purposes (Figure 2a of main part).

Table SN1. Variants detected by sequencing.

Pango lineage	Samples	MOIC	Grouped lineage
A	4	-	wildtype
A.24	1	-	wildtype
A.29	2	N501Y	*dropped*
AY.120	14	L452R, P681R	delta
AY.127.1	1	L452R, P681R	delta
AY.20	9	L452R, P681R	delta
AY.26	1	L452R, P681R	delta
AY.32	8	L452R, P681R	delta
AY.4	4	L452R, P681R	delta
AY.43	5	L452R, P681R	delta

continued on next page

Table SN1, continued

Pango lineage	Samples	MOIC	Grouped lineage
AY.44	7	L452R, P681R	delta
AY.45	1	L452R, P681R	delta
AY.46	1	L452R, P681R	delta
AY.65	4	L452R, P681R	delta
AY.83	1	L452R, P681R	delta
AY.85	1	L452R, P681R	delta
AY.95	1	L452R, P681R	delta
B.1	56	-	wildtype
B.1.1	6	-	wildtype
B.1.1.7	182	N501Y, P681H	alpha
B.1.117	1	-	wildtype
B.1.160	2	-	wildtype
B.1.177.73	1	-	wildtype
B.1.178	3	-	wildtype
B.1.351	11	N501Y, E484K, K417N	beta
B.1.351.5	1	N501Y, E484K, K417N, L18F	*dropped*
B.1.36.17	2	-	wildtype
B.1.36.19	1	-	wildtype
B.1.395	1	-	wildtype
B.1.402	1	-	wildtype
B.1.480	45	N439K	wildtype*
B.1.525	11	E484K	eta
B.1.558	1	-	wildtype
B.1.576	1	-	wildtype
B.1.617.2	55	L452R, P681R	delta
BA.1	24	S477N, N501Y, P681H	omicron BA.1
BA.1.1	57	S477N, N501Y, P681H	omicron BA.1
BA.1.14	1	S477N, N501Y, P681H	omicron BA.1
BA.1.17	14	S477N, N501Y, P681H	omicron BA.1
BA.1.18	1	S477N, N501Y, P681H	omicron BA.1
BA.1.9	2	S477N, N501Y, P681H	omicron BA.1
BA.2	1	S477N, N501Y, K417N, P681H	*dropped*
BA.2.10	1	S477N, N501Y, K417N, P681H	*dropped*
BA.4	1	L452R, S477N, N501Y, K417N, P681H	omicron BA.4/5
BA.4.1	20	L452R, S477N, N501Y, K417N, P681H	omicron BA.4/5
BA.4.1.1	1	L452R, S477N, N501Y, K417N, P681H	omicron BA.4/5
BA.5.2	1	L452R, S477N, N501Y, K417N, P681H	omicron BA.4/5
BF.2	1	L452R, S477N, N501Y, K417N, P681H	omicron BA.4/5
Q.1	2	N501Y, P681H	alpha
Q.4	1	P681R, N501Y, P681H	*dropped*

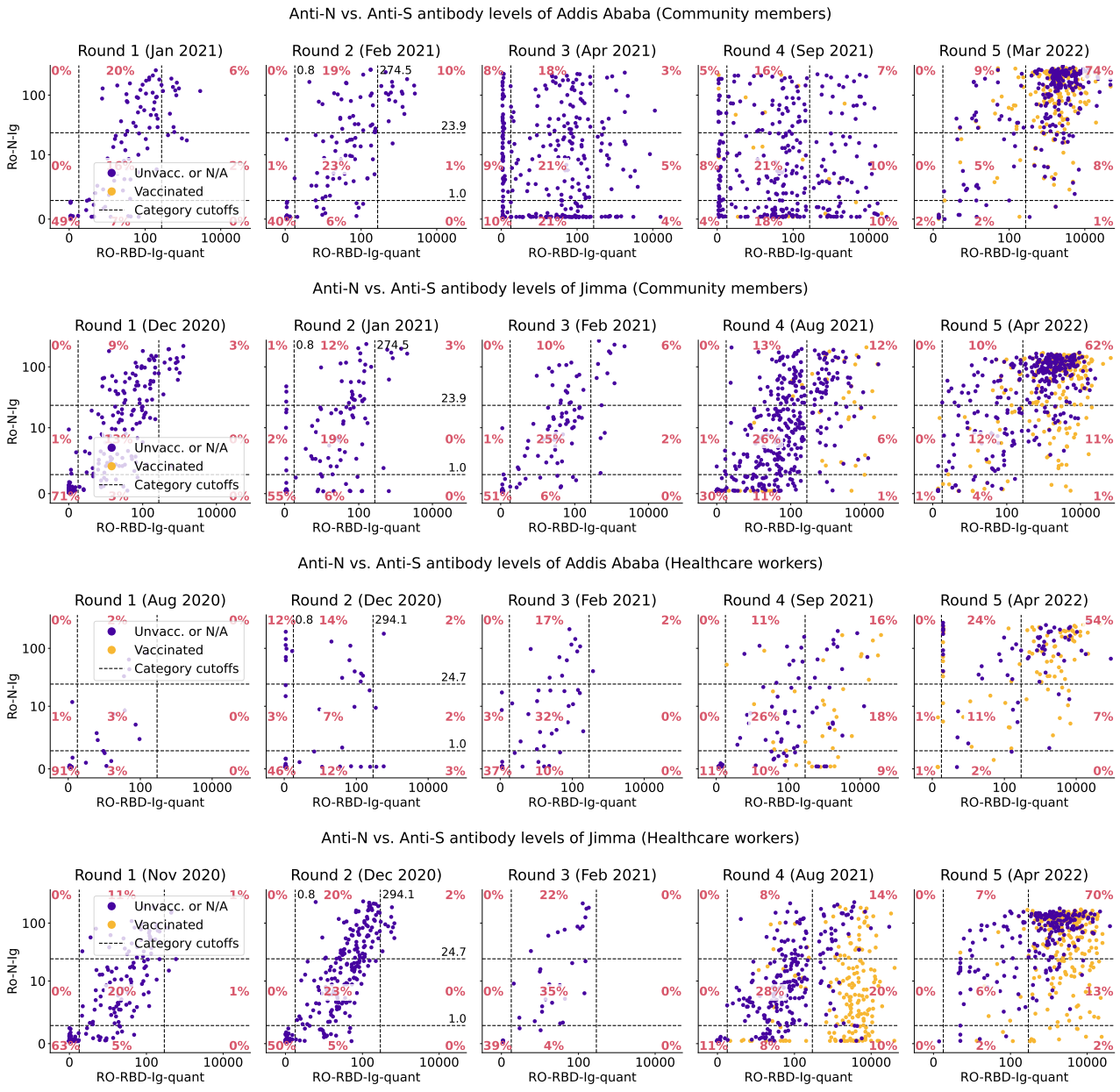


Figure SN4. Antibody data of community members and healthcare workers by site of collection. Source data are provided as a Source Data file.

4 Supplementary Note 2: Multivariant Model

The multivariant model is encoded in the SBML[7] format, integrated with the parameter estimation problem in the PEtab[8] format and made available at Zenodo[9]. In the following, we provide a compact mathematical description, while for additional details we refer to the SBML file and the code.

Model Equations

We utilize the SEIR (susceptible, exposed, infectious, and recovered) framework as basis for our model structure. Assuming a maximum number of 4 infections all combinations of our 8 variants would lead to a system of $8^4 = 4096$ pathways. Hence in order to obtain a computationally feasible system while still retaining realism we exclude pathways which deviated from the chronological order of variant appearances worldwide. We define by P_i the set of potential reinfections after infection with variant i, described by Table SN2 where vaccination is treated as previous infection with the wildtype variant. Furthermore to account for the reported inter-infection intervals we assume third infections before omicron played a negligible role and allow a fourth infection only for omicron BA.4/5, i.e. P_i collapses to $\{7,8\}$ resp. $\{8\}$. For $i = 1,\dots,8$ representing the variant index, where these numbers correspond to columns in Table SN2, we have the following equations for first infection or vaccination

$$
\begin{aligned}
\dot{S} &= -\frac{\beta_i \hat{I}_i S}{N} - v_1 S && S(0) = 120.3\text{e}6 \\
\dot{E}_i &= \frac{\beta_i \hat{I}_i S}{N} - \kappa E_i && E_i(0) = 0 \\
\dot{I}_i &= \kappa E_i - \gamma I_i && I_i(t_{0i}) = 1 \\
\dot{R}_i &= \gamma I_i - \sum_{j \in P_i} \frac{\beta_{ij} \hat{I}_j R_i}{N} - v_1 R_i && R_i(0) = 0 \\
\dot{R}_v &= v_1 S - \sum_{j=1,\dots,8} \frac{\beta_j \hat{I}_j R_v}{N} - v_2 R_v && R_v(0) = 0,
\end{aligned}
$$

where $\hat{I}_i$ is the sum of all currently infected with variant i, N the sum of all state variables, t_{0i} the entrance date of variant i and v_k denote the k-th vaccination rates.

Table SN2. Boolean table of possible reinfections, where 1 means reinfection in model possible and 0 means no reinfection allowed. Rows represent variants of previous infection and columns the variants of reinfection.

	wildtype	wildtype*	alpha	beta	eta	delta	omicron BA.1	omicron BA.4/5
wildtype	1	1	1	1	1	1	1	1
wildtype*	1	1	1	1	1	1	1	1
alpha	0	0	1	1	1	1	1	1
beta	0	0	1	1	1	1	1	1
eta	0	0	1	1	1	1	1	1
delta	0	0	0	0	0	1	1	1
omicron BA.1	0	0	0	0	0	0	1	1
omicron BA.4/5	0	0	0	0	0	0	1	1

The second infections and vaccinations for $i = 1,\dots,8,v$ (numbers for infections, v for vaccination) and $j = 1,\dots,8$ are described by

$$\begin{aligned}
\dot{E}_{ij} &= \frac{\beta_{ij}\hat{I}_j R_i}{N} - \kappa E_{ij} && E_{ij}(0) = 0 \\
\dot{I}_{ij} &= \kappa E_{ij} - \gamma I_{ij} && I_{ij}(0) = 0 \\
\dot{R}_{ij} &= \gamma I_{ij} - \sum_{k=7,8} \frac{\beta_{ijk}\hat{I}_k R_{ij}}{N} - \nu_{n(i,j)+1} R_{ij} && R_{ij}(0) = 0 \\
\dot{R}_{iv} &= \nu_{n(i,v)} R_i - \sum_{k=7,8} \frac{\beta_{ivk}\hat{I}_k R_{iv}}{N} - \nu_{n(i,v)+1} R_{iv} && R_{iv}(0) = 0,
\end{aligned}$$

where $n(\mathbf{Idx}) := \#\{v \in \mathbf{Idx}\}$.

For $i, j = 1,\dots,8,v$ and $k = 7,8$ we obtain the third infection or vaccination equations

$$\begin{aligned}
\dot{E}_{ijk} &= \frac{\beta_{ijk}\hat{I}_k R_{ij}}{N} - \kappa E_{ijk} && E_{ijk}(0) = 0 \\
\dot{I}_{ijk} &= \kappa E_{ijk} - \gamma I_{ijk} && I_{ijk}(0) = 0 \\
\dot{R}_{ijk} &= \gamma I_{ijk} - \frac{\beta_{ijk8}\hat{I}_8 R_{ijk}}{N} - \nu_{n(i,j,k)+1} R_{ijk} && R_{ijk}(0) = 0 \\
\dot{R}_{ijv} &= \nu_{n(i,j,v)} R_{ij} - \frac{\beta_{ijv8}\hat{I}_8 R_{ijv}}{N} - \nu_{n(i,j,v)+1} R_{ijv} && R_{ijv}(0) = 0,
\end{aligned}$$

where $\nu_4 = 0$.

And finally for the fourth infection we have the equations for $i, j = 1,\dots,8,v$ and $k = 7,8,v$

$$\begin{aligned}
\dot{E}_{ijk8} &= \frac{\beta_{ijk8}\hat{I}_8 R_{ijk}}{N} - \kappa E_{ijk8} && E_{ijk8}(0) = 0 \\
\dot{I}_{ijk8} &= \kappa E_{ijk8} - \gamma I_{ijk8} && I_{ijk8}(0) = 0 \\
\dot{R}_{ijk8} &= \gamma I_{ijk8} && R_{ijk8}(0) = 0.
\end{aligned}$$

The effective infection rates $\beta_{\mathbf{Idx}}$ are split into three parts

$$\beta_{\mathbf{Idx}} = s_{\text{seas}} s_{\text{reinf}}(\mathbf{Idx}) \hat{\beta}_{\mathbf{Idx}[-1]},$$

the seasonality factor s_{seas}, the reinfection factor s_{reinf} and the transmission rate $\hat{\beta}_{\mathbf{Idx}[-1]}$ of the currently encountered variant $\mathbf{Idx}[-1]$, i.e. variant corresponding to last index entry of $\mathbf{Idx}$.

The seasonality is formulated as follows

$$s_{\text{seas}}(t) = (1 - s_{\text{frac}}) + s_{\text{frac}} \exp(\sin(2\pi t/365 + s_{\text{shift}}))/\exp(1),$$

where s_{frac} denotes the fraction of seasonality effect, i.e. it equals 1 if transmission rates are fully governed by as yearly cycle and it equals 0 if there is no seasonal effect. The sinus function introduces the periodicity which is

scaled to have a period of one year. Its peak is shifted by the parameter $seas_{\text{shift}}$ and the exponential function ensures positivity.

The reinfection factor depends on the previously encountered variants encoded in all but the last index entries $\mathbf{Idx}[:-1]$ and the currently encountered variant encoded in the last index entry $\mathbf{Idx}[-1]$ and is formulated as follows

$$s_{\text{reinf}}(\mathbf{Idx}) = \begin{cases} 1, & \text{if } |\mathbf{Idx}| = 1 \\ (1-s_0)(1-s)^{d(\mathbf{Idx}[:-1],\mathbf{Idx}[-1])}, & \text{otherwise.} \end{cases}$$

Here $d(x,y)$ is the Hamming distance between MOIC observed in variant y and MOIC observed in variant or combination of variants x. The case where the previous infection(s) x is only one variant, not multiple ones, is depicted in main paper's Figure 2e. The parameters s_0 and s encode the risk reduction for being encountered with the same variant as previously and the risk reduction for an infection with a variant with mutation distance 1 to the former infection's variant, respectively.

In order to incorporate prior knowledge about the variants transmission rates $\hat{\beta}_i$, which is often provided relative between different variants, they are defined as multiplicatives of a base transmission rate β_b or of other $\hat{\beta}_j$ as depicted in Table SN3.

Table SN3. Definition of transmission rates for different variants.

wildtype	wildtype*	alpha	beta	eta	delta	omicron BA.1	omicron BA.4/5
$\hat{\beta}_1 = \tilde{\beta}_1 \cdot \beta_b$	$\hat{\beta}_2 = \tilde{\beta}_2 \cdot \beta_b$	$\hat{\beta}_3 = \tilde{\beta}_3 \cdot \beta_b$	$\hat{\beta}_4 = \tilde{\beta}_4 \cdot \beta_b$	$\hat{\beta}_5 = \tilde{\beta}_5 \cdot \hat{\beta}_3$	$\hat{\beta}_6 = \tilde{\beta}_6 \cdot \hat{\beta}_3$	$\hat{\beta}_7 = \tilde{\beta}_7 \cdot \hat{\beta}_6$	$\hat{\beta}_8 = \tilde{\beta}_8 \cdot \hat{\beta}_7$

Data Integration

The initial time of our model $t = 0$ is set to be the 13^{th} of March 2020 as this was stated by national test positivity data as first Covid-19 case in Ethiopia.

In order to map the model to our data we define three types of observables: Anti-S antibody prevalence, variant distribution and national incidence numbers. Antibody prevalence is observed as levels of 1 infection or vaccination and 2 or more infections or vaccinations and hence, its observable functions are defined by

$$A_1 = \left(\frac{\sum_{|\mathbf{Idx}|=1} R_{\mathbf{Idx}} + \sum_{|\mathbf{Idx}|=2} (E_{\mathbf{Idx}} + I_{\mathbf{Idx}})}{N} \right)$$
$$A_2 = \left(\frac{\sum_{|\mathbf{Idx}|>1} R_{\mathbf{Idx}} + \sum_{|\mathbf{Idx}|>2} (E_{\mathbf{Idx}} + I_{\mathbf{Idx}})}{N} \right).$$

The variant observables are defined for $i = 1, \ldots, 8$ as

$$V_i = \frac{\hat{I}_i}{\sum_j \hat{I}_j}.$$

Finally the national test positivity rate is mapped to model simulations by

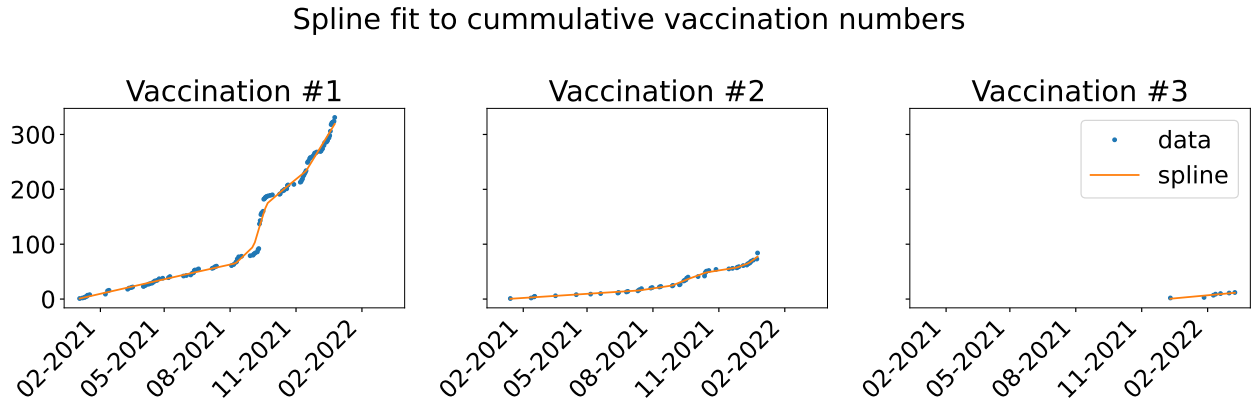


Figure SN5. Spline fits to cumulative counts of first, second and third vaccination information obtained from the antibody study's participants. Source data are provided as a Source Data file.

$$I_{\text{tpr}} = s_{\text{tpr}} \frac{\sum_j \hat{I}_j}{N},$$

where s_{tpr} will be estimated.

Measurement errors are assumed to be normally distributed for each time point and observable with standard deviations taken from the multinomial error estimation described below.

The three vaccination rates for first, second and third vaccination v_1, v_2 and v_3 are fit previously to the parameter estimation as part of the modeling, by fitting monotonic cubic splines to the antibody cohorts' vaccination information and incorporating those splines directly as time dependent functions into the model. The result of those fittings can be seen in Figure SN5.

For improved time resolution of the antibody data in the estimation process while remaining reasonable errors we split each round into two subrounds by performing k-means clustering on its sampling dates (Figure SN6). The antibody prevalence levels were clustered as described in Supplementary Note 3 and then aggregated within the subrounds.

Error estimates for antibody and variant data were obtained by fitting multinomial models for each data-type timepoint combination using pymc3[10]. Error estimates for the national test positivity rate were obtained by fitting binomial models using pymc3. The sample sizes used for these estimations are listed in Table SN4.

Parameter Estimation

There are a total of 24 model and observation parameters subject to estimation. They are listed in Table SN5 including prior information, boundaries, the maximum a-posteriori used as starting point of sampling (obtained by gradient based optimisation) and their sampling result.

For the base transmission rate we use as priors the Bayesian estimation results of the SEIR model of our previous study and priors for incubation and recovery times are taken from literature as established in before[11]. Also prior information about the increased transmission rates of variants are taken from literature, where available.

The parameter sampling for the multivariant model was performed with a sample size of 1.5e4. The first 9e3 samples

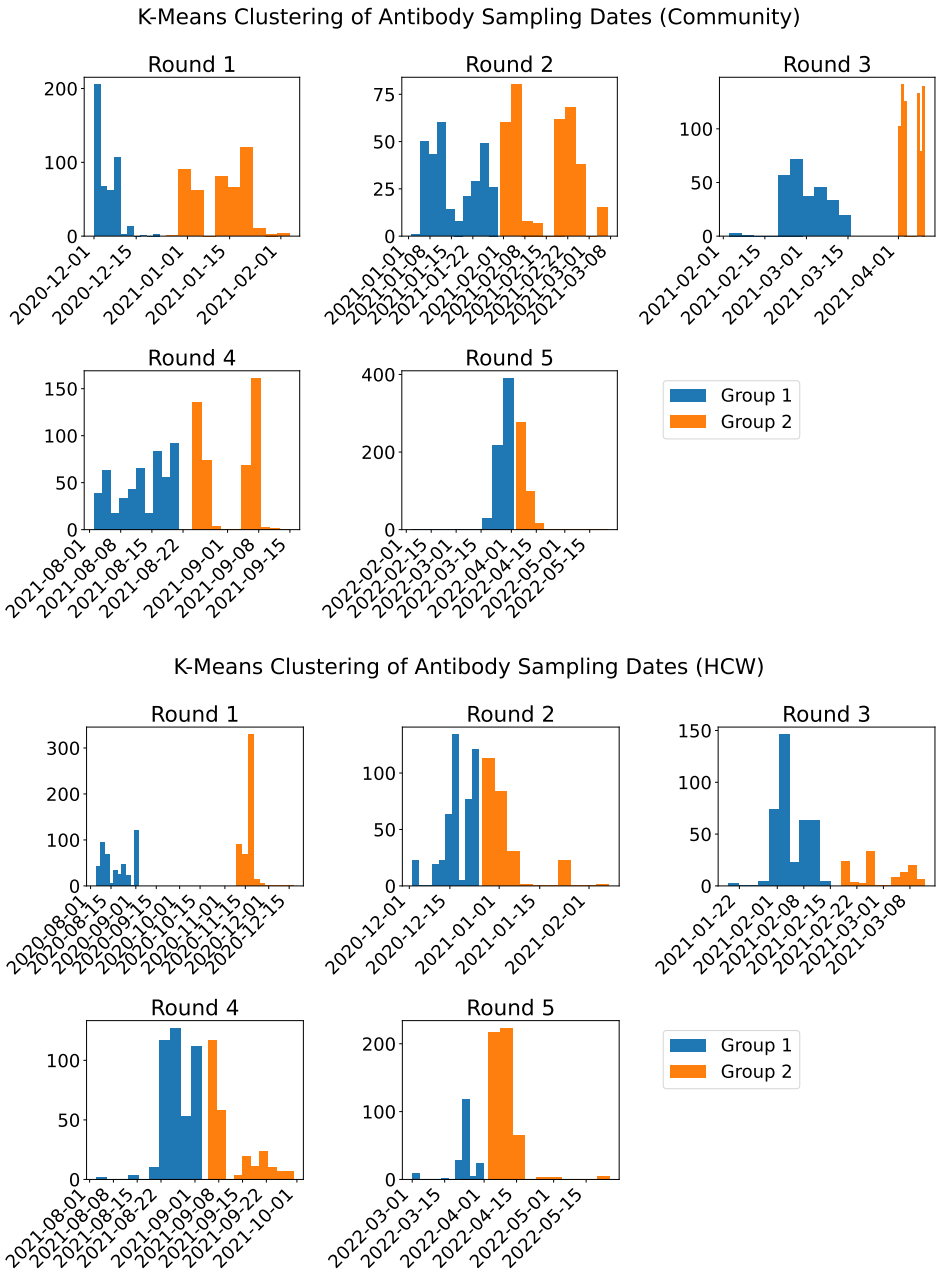


Figure SN6. Subgroups of antibody sampling rounds obtained by k-means clustering of sampling dates for community members and healthcare workers. Source data are provided as a Source Data file.

Table SN4. Sample sizes of aggregated measurements used for fitting the multivariant model. Listed per corresponding observable and total sample sizes per time point. Time depicted in days since 20$^{\text{th}}$ March 2020.

(a) Anti-S antibody levels

Time	A_1	A_2	neg.	$\sum$
266	105	12	335	452
295	44	6	69	119
301	44	6	63	113
310	36	9	22	67
313	27	3	44	74
320	30	2	56	88
328	38	12	33	83
346	45	11	52	108
347	43	7	37	87
359	27	2	40	69
385	79	11	50	140
391	95	26	30	151
512	137	39	84	260
524	135	60	80	275
530	85	30	29	144
543	98	57	28	183
741	38	194	8	240
747	78	291	8	377
754	58	106	1	165
757	30	164	0	194

(b) Variant distributions

Time	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	$\sum$
216	5	2	0	0	0	0	0	0	7
247	4	1	0	0	0	0	0	0	5
277	4	1	0	0	0	0	0	0	5
311	18	9	15	1	0	0	0	0	43
333	23	19	28	6	0	0	0	0	76
369	25	12	60	0	0	0	0	0	97
403	1	0	37	3	6	0	0	0	47
426	1	1	22	1	5	1	0	0	31
459	0	0	8	0	0	0	0	0	8
489	0	0	14	0	0	1	0	0	15
520	0	0	0	0	0	13	0	0	13
551	0	0	0	0	0	55	0	0	55
583	0	0	0	0	0	18	0	0	18
610	0	0	0	0	0	14	1	0	15
646	0	0	0	0	0	11	48	0	59
666	0	0	0	0	0	0	41	0	41
692	0	0	0	0	0	0	4	0	4
723	0	0	0	0	0	0	1	0	1
817	0	0	0	0	0	0	3	4	7
843	0	0	0	0	0	0	0	20	20

(c) National test positivity rates

Time	247	278	309	338	368	398	429	459	490	521	551	582	612	643	674	703	733	763	794	824	855	886
I_{tpr}	30	31	31	28	31	30	31	29	31	31	30	31	30	31	31	28	31	30	31	30	31	31

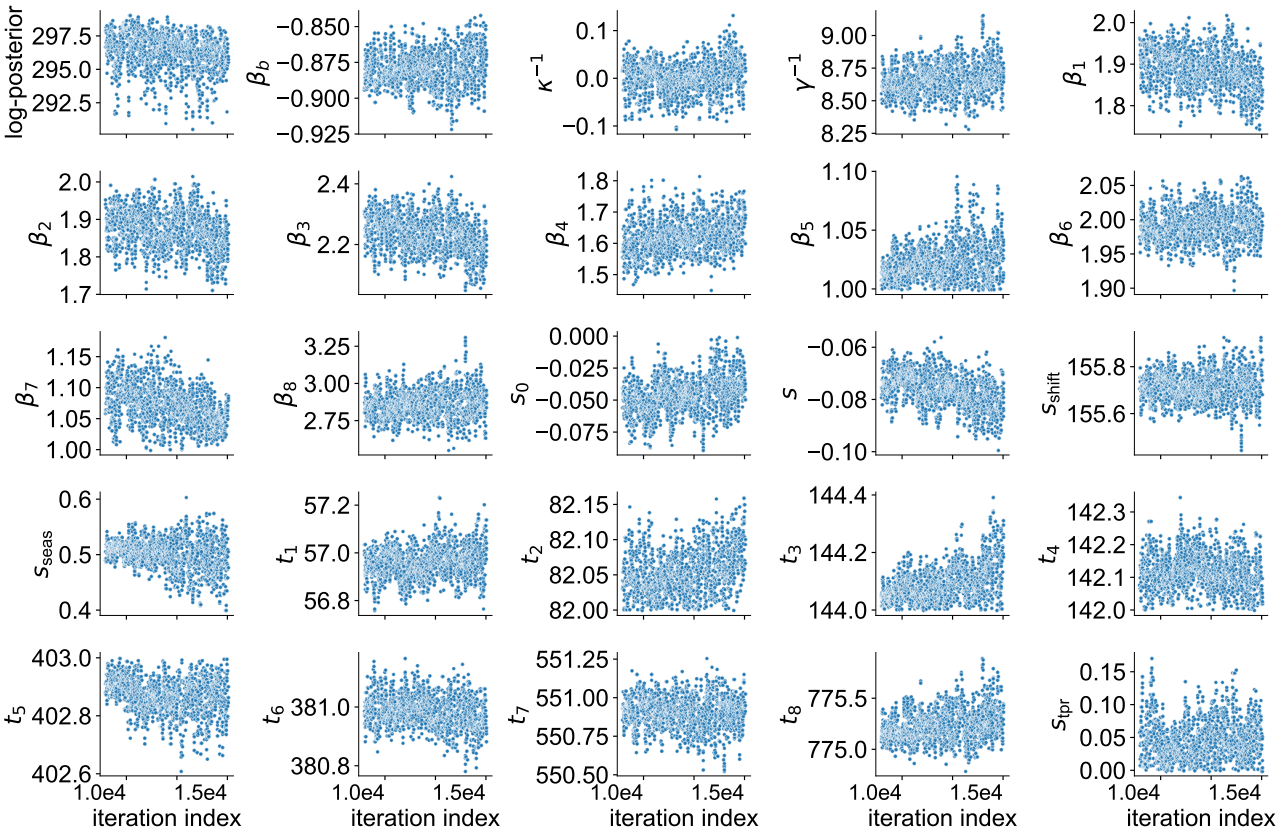


Figure SN7. Multivariant model's sampled log-posterior and parameter traces. Burn in phase is cut off. Source data are provided as a Source Data file.

were removed as burn-in. The remaining sample passed the Geweke convergence test as well as visual examination (Figure SN7). Parameter correlations and distributions are depicted in Figure SN8.

Model Analysis

The model estimates the entry time points of most variants substantially later than the first global appearance according to outbreak.info (Figure SN9). Global reporting and local estimation coincide only for wildtype*, alpha and beta, the latter of which does not play a large role in the overall dynamics observed and estimated by us.

Including the top ten infection-vaccination pathways depicted in main paper's Figure 4 the model estimated 68 pathways contributing more than 0.1 % (Table SN6). We calculated them by checking sizes of all, and in particular the recovered compartments, after simulating the model until $t = 1200$, where we encountered equilibrium due to the lack of new variants after omicron. Then we investigated the transitions inside the pathways (Figure 4c of main manuscript) by appropriately scaling and stack-plotting the time courses of all recovery states being part of the pathway. For example for $R_{1,2,3,4}$ this we would plot R_1, $R_{1,2}$, $R_{1,2,3}$ and $R_{1,2,3,4}$.

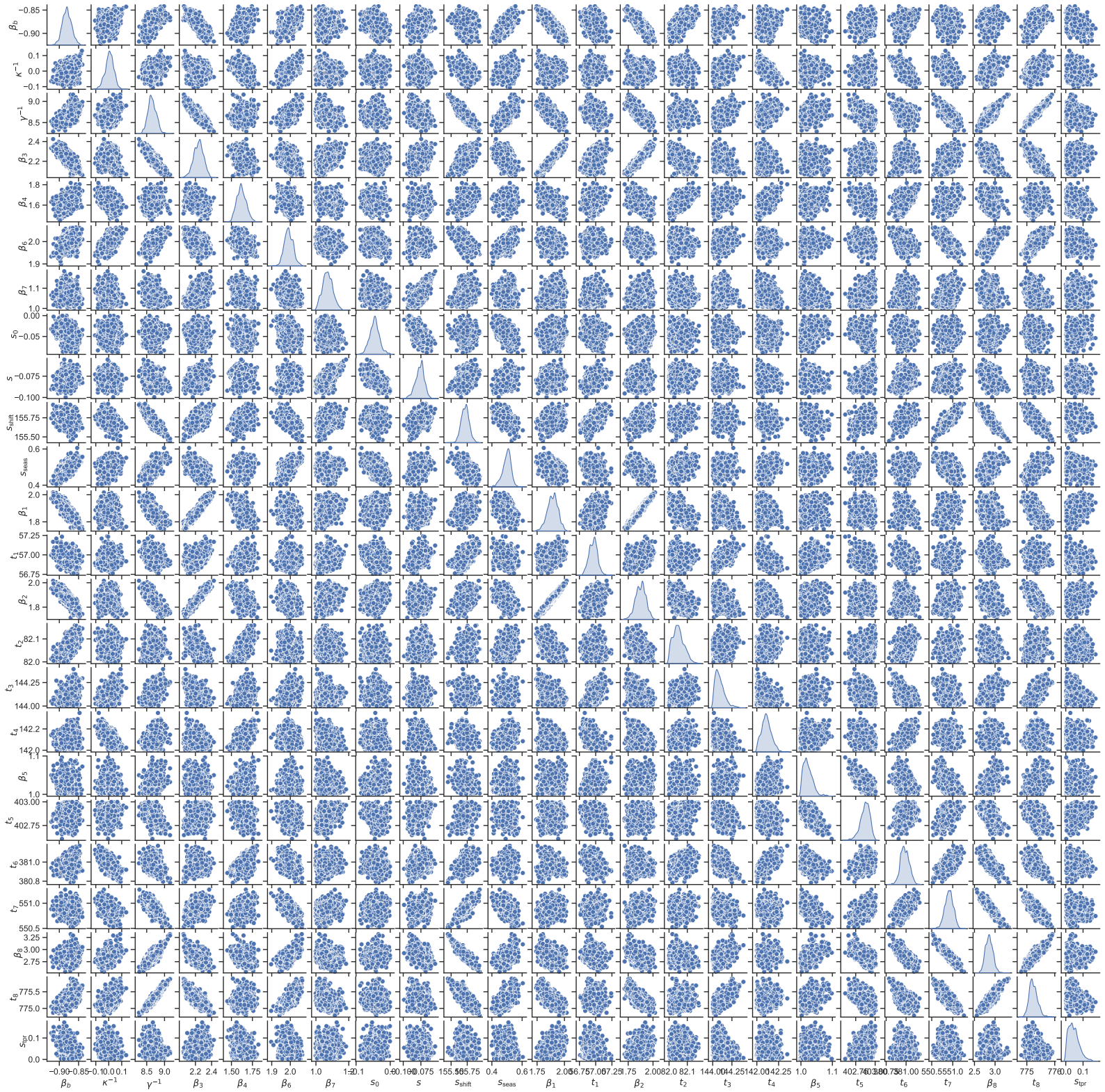


Figure SN8. Scatter plots and distributions of multivariant model's sampled parameters. Source data are provided as a Source Data file.

Table SN5. Parameters of the multivariant model.

Parameter	Sampling result - Median (CI 95%)	Parameter bounds	Scale used for sampling	Prior (in scale)	Maximum a-posteriori	Unit
β_b	0.13 (0.12, 0.14)	$[0.01, 1]$	$\log_{10}$	$\mathcal{N}(-1.10; 0.06)$[11]	0.13	day^{-1}
κ^{-1}	1.01 (0.94, 1.07)	$[0.1, 100]$	log	$\mathcal{N}(1.63; 0.50)$[12]	1.12	day(s)
γ^{-1}	8.64 (8.44, 8.89)	$[0.1, 100]$	linear	$\mathcal{N}(15.7; 6.7)$[13]	8.38	day(s)
β_1	1.89 (1.79, 1.98)	$[1.0, 10]$	linear	-	1.97	-
β_2	1.88 (1.77, 1.97)	$[1.0, 10]$	linear	-	1.96	-
β_3	2.24 (2.11, 2.35)	$[1.0, 10]$	linear	$\mathcal{N}(1.82; 0.22)$[14]	2.35	-
β_4	1.62 (1.52, 1.73)	$[1.0, 10]$	linear	$\mathcal{N}(1.50; 0.24)$[15]	1.50	-
β_5	1.02 (1.00, 1.06)	$[1.0, 10]$	linear	-	1.00	-
β_6	1.99 (1.95, 2.04)	$[1.0, 10]$	linear	$\mathcal{N}(1.99; 0.04)$[16]	2.00	-
β_7	1.07 (1.02, 1.13)	$[1.0, 10]$	linear	$\mathcal{N}(1.1; 0.05)$[17]	1.09	-
β_8	2.85 (2.69, 3.04)	$[1.0, 10]$	linear	-	2.70	-
s_0	0.90 (0.85, 0.97)	$[0.001, 1]$	$\log_{10}$	-	0.87	-
s	0.84 (0.81, 0.86)	$[0.001, 1]$	$\log_{10}$	-	0.85	-
s_{shift}	155.72 (155.60, 155.84)	$[0.0, 365]$	linear	-	155.80	day(s)
s_{frac}	0.50 (0.43, 0.55)	$[0.0, 1]$	linear	-	0.48	-
t_1	56.97 (56.83, 57.09)	$[1.0, 216]$	linear	-	57.00	day(s)
t_2	82.05 (82.00, 82.12)	$[82.0, 216]$	linear	-	82.00	day(s)
t_3	144.08 (144.01, 144.26)	$[144.0, 311]$	linear	-	144.00	day(s)
t_4	142.11 (142.01, 142.23)	$[142.0, 311]$	linear	-	142.00	day(s)
t_5	402.88 (402.73, 402.97)	$[1.0, 403]$	linear	-	403.00	day(s)
t_6	380.98 (380.88, 381.09)	$[323.0, 426]$	linear	-	381.00	day(s)
t_7	550.91 (550.71, 551.10)	$[508.0, 610]$	linear	-	551.00	day(s)
t_8	775.22 (774.97, 775.55)	$[560.0, 817]$	linear	-	775.00	day(s)
s_{tpr}	1.10 (1.01, 1.32)	$[1.0, 10]$	$\log_{10}$	-	1.01	-

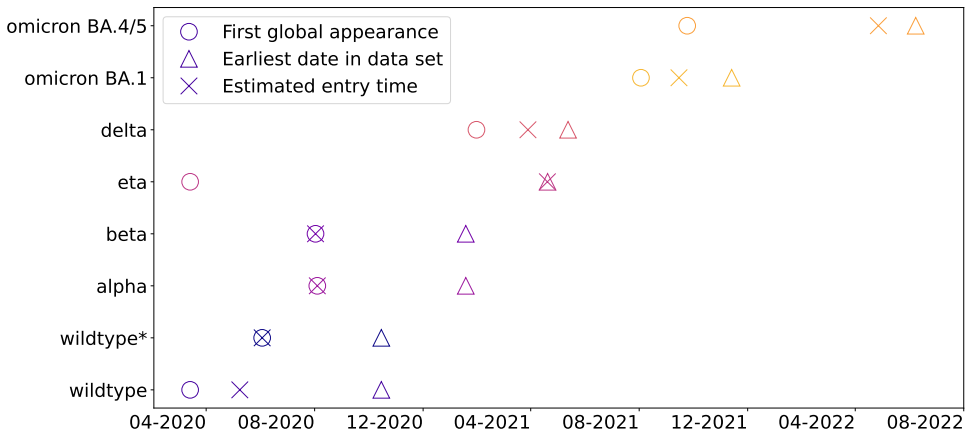


Figure SN9. Estimated entry times of variants. First global appearance and earliest date in sequenced data set included for comparison. Source data are provided as a Source Data file.

Table SN6. Pathways of the multivariant model which account for more than 0.1%

Rank	Pathway	Median - 90% CI
1	wildtype - delta - vaccine - omicron BA.4/5	12.7% (10.9%,14.4%)
2	delta - omicron BA.4/5 - omicron BA.4/5	6.4% (5.0%,7.6%)
3	alpha - delta - vaccine - omicron BA.4/5	6.2% (4.6%,7.8%)
4	wildtype - delta - omicron BA.4/5 - omicron BA.4/5	6.0% (4.5%,7.4%)
5	delta - vaccine - omicron BA.4/5 - omicron BA.4/5	5.8% (4.6%,7.0%)
6	delta - vaccine - vaccine - omicron BA.4/5	4.9% (4.2%,5.7%)
7	wildtype* - delta - vaccine - omicron BA.4/5	3.9% (3.0%,4.9%)
8	wildtype - delta - omicron BA.1 - omicron BA.4/5	3.8% (2.2%,5.7%)
9	delta - delta - vaccine - omicron BA.4/5	3.4% (1.6%,5.2%)
10	alpha - delta - omicron BA.4/5 - omicron BA.4/5	3.0% (2.2%,4.1%)
11	delta - omicron BA.1 - omicron BA.4/5	3.0% (1.8%,4.3%)
12	wildtype - vaccine - vaccine - omicron BA.4/5	2.9% (2.5%,3.4%)
13	delta - vaccine - omicron BA.1 - omicron BA.4/5	2.5% (1.5%,3.7%)
14	wildtype* - delta - omicron BA.4/5 - omicron BA.4/5	2.0% (1.4%,2.5%)
15	wildtype - alpha - vaccine - omicron BA.4/5	1.9% (1.5%,2.5%)
16	wildtype - vaccine - omicron BA.4/5 - omicron BA.4/5	1.8% (1.1%,2.3%)
17	delta - delta - omicron BA.4/5 - omicron BA.4/5	1.7% (0.7%,2.9%)
18	alpha - delta - omicron BA.1 - omicron BA.4/5	1.4% (0.8%,2.3%)
19	wildtype - vaccine - omicron BA.4/5	1.3% (0.8%,2.1%)
20	wildtype* - delta - omicron BA.1 - omicron BA.4/5	1.3% (0.7%,2.0%)
21	wildtype - omicron BA.4/5 - omicron BA.4/5	1.3% (0.8%,1.6%)
22	wildtype - delta - omicron BA.4/5	1.1% (0.9%,1.3%)
23	vaccine - delta - vaccine - omicron BA.4/5	1.1% (0.9%,1.2%)
24	delta - vaccine - omicron BA.4/5	1.1% (0.8%,1.5%)
25	delta - delta - omicron BA.1 - omicron BA.4/5	1.0% (0.5%,1.6%)
26	wildtype - vaccine - omicron BA.1 - omicron BA.4/5	1.0% (0.5%,1.6%)
27	alpha - vaccine - vaccine - omicron BA.4/5	0.8% (0.6%,1.0%)
28	delta - omicron BA.1 - vaccine - omicron BA.4/5	0.8% (0.4%,1.3%)
29	wildtype* - vaccine - vaccine - omicron BA.4/5	0.7% (0.5%,0.9%)
30	wildtype - omicron BA.1 - omicron BA.4/5	0.7% (0.4%,1.1%)
31	wildtype* - alpha - vaccine - omicron BA.4/5	0.7% (0.5%,0.9%)
32	vaccine - delta - omicron BA.4/5 - omicron BA.4/5	0.6% (0.5%,0.8%)
33	alpha - delta - omicron BA.4/5	0.6% (0.5%,0.9%)
34	wildtype - alpha - omicron BA.4/5 - omicron BA.4/5	0.6% (0.3%,0.9%)
35	wildtype - wildtype - vaccine - omicron BA.4/5	0.6% (0.3%,1.0%)
36	delta - omicron BA.4/5 - vaccine	0.6% (0.4%,0.8%)
37	wildtype* - vaccine - omicron BA.4/5 - omicron BA.4/5	0.6% (0.4%,0.7%)
38	alpha - vaccine - omicron BA.4/5 - omicron BA.4/5	0.5% (0.3%,0.7%)
39	wildtype - omicron BA.4/5 - vaccine	0.5% (0.3%,0.8%)
40	wildtype - wildtype* - vaccine - omicron BA.4/5	0.5% (0.3%,0.6%)

continued on next page

Table SN6, continued

Rank	Pathway	Median - 90% CI
41	wildtype - alpha - omicron BA.4/5	0.5% (0.3%,0.8%)
42	vaccine - delta - omicron BA.1 - omicron BA.4/5	0.4% (0.2%,0.6%)
43	wildtype* - wildtype - vaccine - omicron BA.4/5	0.4% (0.3%,0.5%)
44	alpha - vaccine - omicron BA.4/5	0.4% (0.2%,0.7%)
45	delta - omicron BA.1 - omicron BA.1	0.3% (0.1%,0.8%)
46	wildtype* - omicron BA.4/5 - omicron BA.4/5	0.3% (0.2%,0.4%)
47	delta - delta - omicron BA.4/5	0.3% (0.2%,0.4%)
48	wildtype* - alpha - omicron BA.4/5 - omicron BA.4/5	0.3% (0.2%,0.4%)
49	alpha - omicron BA.4/5 - omicron BA.4/5	0.3% (0.2%,0.4%)
50	wildtype* - vaccine - omicron BA.1 - omicron BA.4/5	0.3% (0.1%,0.4%)
51	vaccine - vaccine - omicron BA.4/5 - omicron BA.4/5	0.2% (0.2%,0.3%)
52	vaccine - omicron BA.4/5 - omicron BA.4/5	0.2% (0.2%,0.3%)
53	wildtype - alpha - omicron BA.1 - omicron BA.4/5	0.2% (0.1%,0.3%)
54	wildtype - wildtype* - omicron BA.4/5 - omicron BA.4/5	0.2% (0.1%,0.2%)
55	wildtype* - delta - omicron BA.4/5	0.2% (0.1%,0.2%)
56	vaccine - vaccine - omicron BA.4/5	0.2% (0.1%,0.3%)
57	wildtype - wildtype - omicron BA.4/5 - omicron BA.4/5	0.2% (0.0%,0.3%)
58	wildtype - omicron BA.1 - vaccine - omicron BA.4/5	0.2% (0.1%,0.3%)
59	wildtype* - wildtype - omicron BA.4/5 - omicron BA.4/5	0.2% (0.1%,0.2%)
60	wildtype* - omicron BA.1 - omicron BA.4/5	0.1% (0.1%,0.2%)
61	alpha - vaccine - omicron BA.1 - omicron BA.4/5	0.1% (0.1%,0.2%)
62	delta - omicron BA.4/5 - vaccine - omicron BA.4/5	0.1% (0.1%,0.2%)
63	alpha - alpha - vaccine - omicron BA.4/5	0.1% (0.1%,0.2%)
64	wildtype - wildtype - omicron BA.4/5	0.1% (0.1%,0.2%)
65	vaccine - delta - omicron BA.4/5	0.1% (0.1%,0.1%)
66	alpha - omicron BA.4/5 - vaccine	0.1% (0.1%,0.2%)
67	vaccine - omicron BA.1 - omicron BA.4/5	0.1% (0.1%,0.2%)
68	wildtype - wildtype - omicron BA.1 - omicron BA.4/5	0.1% (0.0%,0.2%)

Alternative Model Formulations

Initially we considered three potential model extensions: (i) Describing cross-immunities independently of MOIC. (ii) Allowing all pathways between variants. (iii) No grouping of variants.

In the end all of these formulations proved impractial. For (i) we would have to model individual parameters for each combination of past infections and new infections. Even with the other simplifications of the model still in place this leads to a total of 205 immune escape factors instead of the two we have in the current model. For such a high dimensional parameter estimation the dataset would have been insufficient to inform. (ii) would result in a model with 12289 different states being computationally infeasible. Extension (iii) implies 50 different variants instead of the current 8 lineages. Even if we disregard the low statistical power we have for some of these single sublineages, we would still end up with more than 10000 different model states and five times as many parameters

as in our current model, make this computationally and with respect to the information in our data set infeasible.

5 Supplementary Note 3: Antibody-level Model

The antibody-level model is encoded in the SBML format, integrated with the parameter estimation problem in the PEtab format and made available at Zenodo[9]. In the following, we provide a compact mathematical description, while for additional details we refer to the SBML file and the code.

Model Equations

The antibody-level model described the distribution of individuals with a certain combination of Anti-S and Anti-N antibody levels. For each antibody, we consider three discrete catgories, with index i =0 (low), 1 (medium), 2 (high) being used for Anti-S antibody categories and index j =0 (low), 1 (medium), 2 (high) being used for Anti-N antibody categories. The distribution changes over time due to infection as well as vaccination and antibody decay. Defining χ_{bool} as the indicator function, i.e. $\chi_{\text{true}} = 1$ and $\chi_{\text{false}} = 0$, we modelled the time evolution of A_{ij}, i.e. individuals with Anti-S antibody levels in category i and Anti-N antibody levels in category j, as

$$
\begin{aligned}
\dot{A}_{ij} =& -\frac{\beta_{ij}\hat{I}A_{ij}}{N} - \nu A_{ij}\chi_{i\leq 1} \\
&+ \gamma\left(I_{i,j-1}\chi_{i=2} + I_{i-1,j}\chi_{j=2} + I_{i,j}\chi_{i=2}\chi_{j=2} + (1-\theta^{\chi_{i=1}})I_{i-1,j-1} + \theta^{\chi_{i=1}}I_{i-2,j-1}\right)\chi_{i\geq 1}\chi_{j\geq 1} \\
&+ \delta_N A_{i+1,j}\chi_{i\leq 1} + \delta_S A_{i,j+1}\chi_{j\leq 1} + \delta_{\text{SN}} A_{i+1,j+1}\chi_{i\leq 1}\chi_{j\leq 1} \\
&+ \nu A_{i-1,j}\chi_{i\geq 1} \\
&- (\delta_N\chi_{i\geq 1} + \delta_S\chi_{j\geq 1} + \delta_{\text{SN}}\chi_{i\geq 1}\chi_{j\geq 1})A_{ij} \\
\dot{E}_{ij} =& \frac{\beta_{ij}\hat{I}A_{ij}}{N} - \kappa E_{ij} \\
\dot{I}_{ij} =& \kappa E_{ij} - \gamma I_{ij},
\end{aligned}
$$

with initial conditions

$$
\begin{aligned}
A_{ij}(0) &= \begin{cases} 120.3\text{e}6 & \text{if } i=j=0 \\ 0 & \text{otherwise} \end{cases} \\
E_{ij}(0) &= 0 \\
I_{ij}(t_0) &= \begin{cases} 1 & \text{if } i=j=0 \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}
$$

Here, $\hat{I}$ is the sum of all infected and N is the sum of all state variables. The fraction θ of getting boosted Anti-N levels after recovery as well as the decays δ will be estimated. Moreover β_{ij} is structured as

$$
\beta_{ij} = (1-s_1)^{\chi_{i\geq 1 \text{ or } j\geq 1}}(1-s_2)^{\chi_{i=2 \text{ or } j=2}}\sum_{k=1}^{8}\alpha_k\hat{\beta}_k,
$$

where the immunity factors s_1 and s_2 are obtained via estimation. The α_i are the normalized Gaussian fits to the variant distributions, described above and shown in Figure SN11. β_i are the variants' transmission rates again defined as multiplicatives of each other as for the multivariant model depicted in Table (SN3). Without loss of generality here we assume that $\beta_1 = \beta_b$.

Table SN7. Sample sizes for aggregated 2-dimensional antibody measurements corresponding to the observables $\tilde{A}_{ij}$ and total sample sizes per time point. Time depicted in days since 20$^{\text{th}}$ March 2020.

Time	$\tilde{A}_{00}$	$\tilde{A}_{01}$	$\tilde{A}_{02}$	$\tilde{A}_{10}$	$\tilde{A}_{11}$	$\tilde{A}_{12}$	$\tilde{A}_{20}$	$\tilde{A}_{21}$	$\tilde{A}_{22}$	$\sum$
266	332	10	0	3	62	1	0	33	11	452
295	67	3	0	0	19	0	0	20	6	115
301	60	2	0	3	25	1	0	17	5	113
310	22	9	0	0	11	4	0	15	4	65
313	43	4	0	1	10	0	0	13	3	74
320	51	10	1	2	14	0	3	6	1	88
328	32	4	0	0	20	1	0	13	11	81
346	51	8	0	1	27	3	0	10	8	108
347	33	5	0	2	22	0	0	13	6	81
359	39	2	0	1	20	1	0	5	1	69
385	24	32	5	15	26	2	11	21	4	140
391	6	28	6	12	37	14	12	30	6	151
512	80	22	1	4	71	13	0	44	25	260
524	79	38	3	1	72	19	0	25	38	275
530	7	33	9	13	34	14	9	18	7	144
543	7	26	24	13	35	18	8	33	15	179
741	8	5	3	0	15	24	0	18	167	240
747	6	13	3	2	41	53	0	24	235	377
754	1	7	0	0	24	11	0	27	95	165
757	0	2	0	0	6	15	0	22	149	194

Data Integration

Initial time of the model $t = 0$ is set to be the 13^{th} of March 2020 as for the multivariant model.

The observables mapping the antibody-level model to data are

$$\tilde{A}_{ij} = \frac{A_{ij} + E_{ij} + I_{ij}}{N}$$

and the national test positivity data is mapped with a scaling as before for the multivariant model.

Measurement errors are assumed to be normally distributed and obtained by multinomial error modelling as described above. The sample sizes used for the error estimates of the nine antibody categories are listed in Table SN7.

The antibody rounds are split into subgroups by sampling dates and clustered into categories as before. Moreover errors of all data types for estimation are again obtained by multinomial, resp. binomial, models.

The vaccination rate ν is implemented as a piecewise linear function which is calculated by monthly averaging the vaccination information of the antibody sampling cohort a priori to the parameter estimation. The results of this can be seen in Figure SN10. Note in the equations of the model we made the assumptions that people with already high Anti-S levels do not get vaccinated anymore, i.e. the amount of people still applying for vaccination after two infections or vaccinations is negligible.

For the antibody-level model the variant data is directly incorporated as a time-dependent function. First, the variant data is aggregated into 2-month bins, and Gaussian kernels are fit to the distributions using scipy's "minimize"

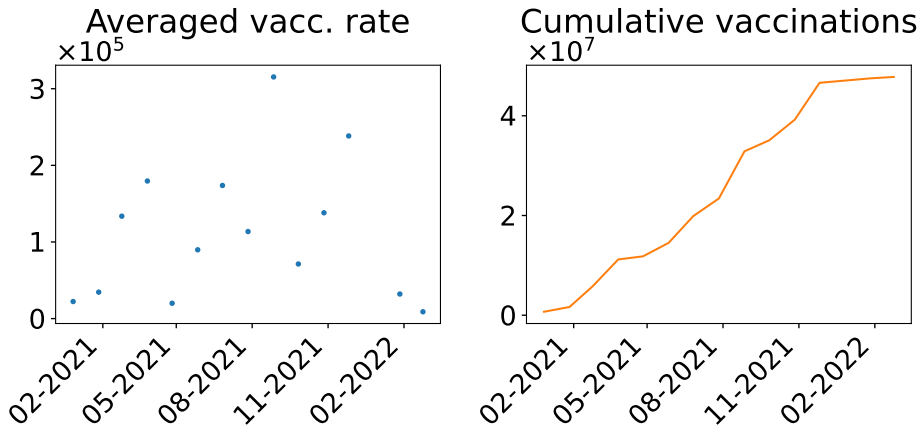


Figure SN10. Monthly averaged vaccination rates and cumulative vaccinations of antibody study's cohort. Source data are provided as a Source Data file.

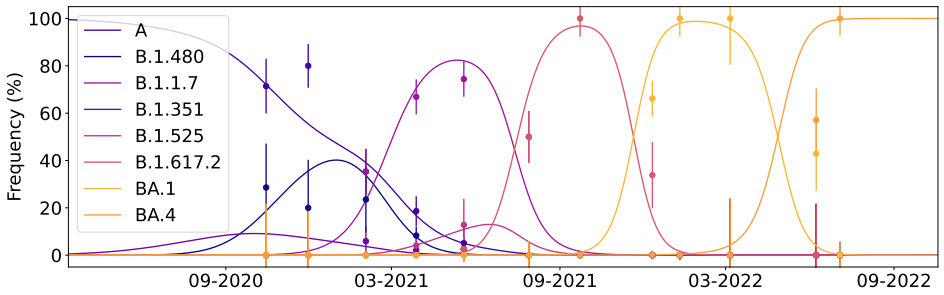


Figure SN11. Fits of normalized Gauss kernels to mean variant data. Variant data depicted as mean -/+ standard deviations. Sample sizes listed in Table SN4(b). Source data are provided as a Source Data file.

function. Finally, the distributions are normalized so that they sum up to 1. The result of this fitting process is illustrated in Figure SN11.

Parameter Estimation

There are a total of 20 model and observation parameters subject to estimation. They are listed in Table SN8 including prior information, boundaries, the maximum a-posteriori used as starting point of sampling (obtained by gradient based optimisation) and their sampling result.

The parameter sampling for the multivariant model was performed with a sample size of 1e5. The first 7e4 samples were removed as burn-in. The remaining sample passed the Geweke convergence test as well as visual examination (Figure SN12). Parameter correlations and distributions are depicted in Figure SN13.

Table SN8. Parameters of the antibody-level model.

Parameter	Sampling result - Median (CI 95%)	Parameter bounds	Scale used for sampling	Prior (in scale)	Maximum a-posteriori	Unit
κ^{-1}	1.1 (0.868, 1.35)	$[0.01, 100]$	log	$\mathcal{N}(1.63; 0.50)$[12]	1.22	day(s)
γ^{-1}	18.7 (18.4, 19)	$[0.01, 100]$	linear	$\mathcal{N}(15.7; 6.7)$[13]	18.8	day(s)
β_1	0.152 (0.145, 0.159)	$[0.01, 1]$	$\log_{10}$	$\mathcal{N}(-1.10; 0.06)$[11]	0.153	day^{-1}
β_2	9.93 (9.8, 10)	$[1, 10]$	linear	-	9.99	-
β_3	1.67 (1.45, 1.92)	$[1, 10]$	linear	$\mathcal{N}(1.82; 0.22)$[14]	1.67	-
β_4	1.5 (1.21, 1.79)	$[1, 10]$	linear	$\mathcal{N}(1.50; 0.24)$[15]	1.54	-
β_5	1.4 (1.03, 2.03)	$[1, 10]$	linear	-	1.2	-
β_6	1.99 (1.92, 2.05)	$[1, 10]$	linear	$\mathcal{N}(1.99; 0.04)$[16]	1.99	-
β_7	1.09 (1.02, 1.17)	$[1, 10]$	linear	$\mathcal{N}(1.1; 0.05)$[17]	1.1	-
β_8	2.77 (2.45, 3)	$[1, 10]$	linear	-	2.89	-
δ_N	0.000135 (6.18e-05, 0.000245)	$[1e\text{-}05, 0.01]$	$\log_{10}$	-	0.00019	day^{-1}
δ_S	2.29e-05 (1.09e-05, 6.15e-05)	$[1e\text{-}05, 0.01]$	$\log_{10}$	-	1.94e-05	day^{-1}
δ_{SN}	1.29e-05 (1.02e-05, 1.96e-05)	$[1e\text{-}05, 0.01]$	$\log_{10}$	-	1.58e-05	day^{-1}
t_0	91.4 (76.1, 106)	$[1, 250]$	$\log_{10}$	-	100	day(s)
s_{frac}	0.995 (0.984, 1)	$[0, 1]$	linear	-	0.999	-
s_{shift}	213 (213, 213)	$[0, 365]$	linear	-	213	day(s)
θ	0.342 (0.296, 0.386)	$[0.001, 1]$	$\log_{10}$	-	0.351	-
s_1	0.736 (0.677, 0.791)	$[0.001, 1]$	$\log_{10}$	-	0.737	-
s_2	0.635 (0.523, 0.753)	$[0.001, 1]$	$\log_{10}$	-	0.629	-
s_{tpr}	1.03 (1, 1.09)	$[1, 10]$	$\log_{10}$	-	1	-

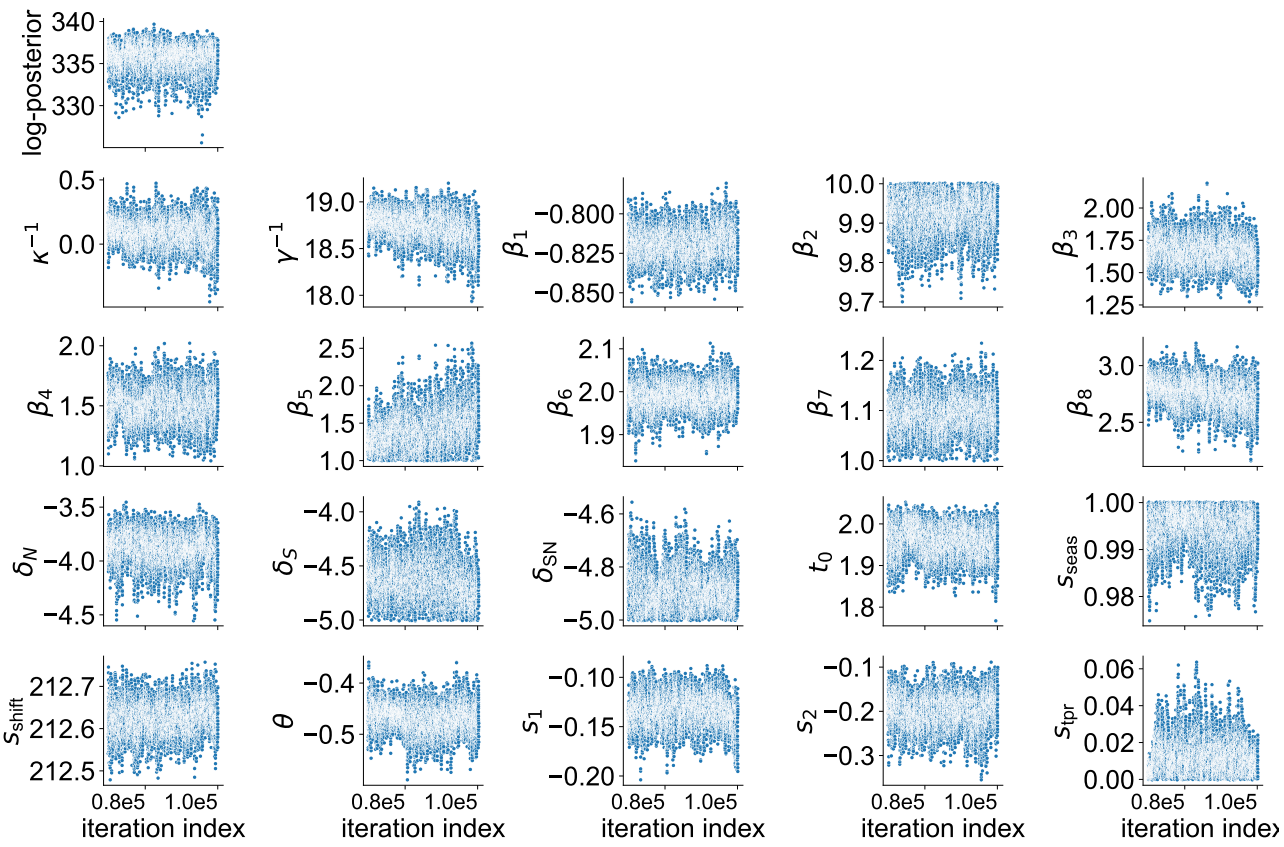


Figure SN12. Antibody-level model's sampled log-posterior and parameter traces. Burn in phase is cut off. Source data are provided as a Source Data file.

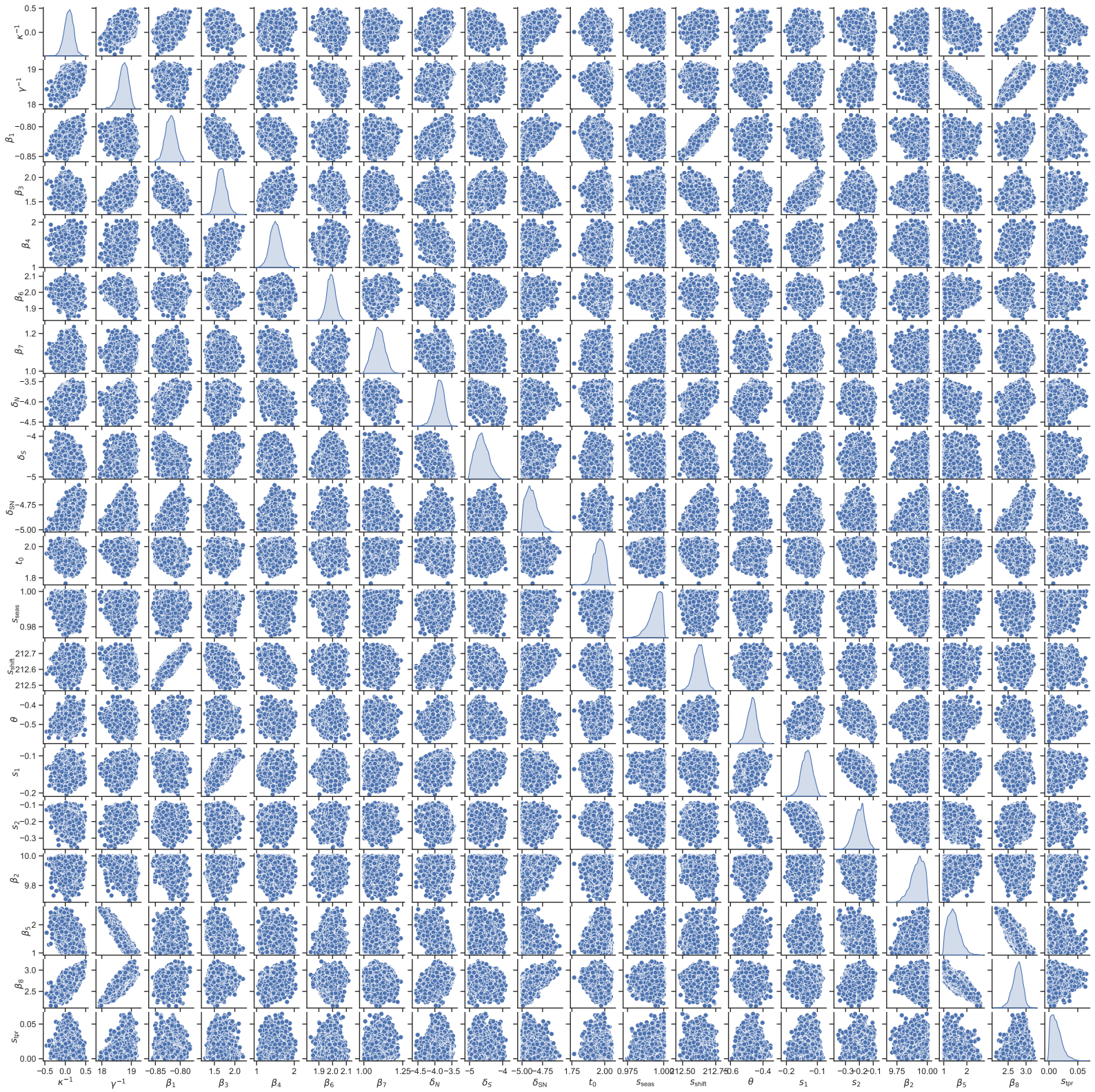


Figure SN13. Scatter plots and distributions of antibody-level model's sampled parameters. Source data are provided as a Source Data file.

Supplementary References

1. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python . *J. Mach. Learn. Res.* **12**, 2825–2830, DOI: 10.48550/arXiv.1201.0490 (2011).
2. Virtanen, P. *et al.* SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **17**, 261–272, DOI: 10.1038/s41592-019-0686-2 (2020).
3. WHO. Ethiopia launches a covid-19 vaccination campaign targeting the 12 years and above population. https://www.afro.who.int/news/ethiopia-launches-covid-19-vaccination-campaign-targeting-12-years-and-above-population (2021). Accessed: 2023-11-26.
4. O'Toole, Á. *et al.* Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evol* **7**, veab064, DOI: 10.1093/ve/veab064 (2021).
5. Gangavarapu, K. *et al.* Outbreak.info genomic reports: scalable and dynamic surveillance of sars-cov-2 variants and mutations. *Nat. Methods* **20**, 512–522, DOI: 10.1038/s41592-023-01769-3 (2023).
6. Tsueng, G. *et al.* Outbreak.info research library: a standardized, searchable platform to discover and explore covid-19 resources. *Nat. Methods* **20**, 536–540, DOI: 10.1038/s41592-023-01770-w (2023).
7. Hucka, M. *et al.* The systems biology markup language (sbml): Language specification for level 3 version 1 core. *J. Integr. Bioinforma.* **12**, 382–549, DOI: 10.1515/jib-2015-266 (2015).
8. Schmiester, L. *et al.* Petab—interoperable specification of parameter estimation problems in systems biology. *PLOS Comput. Biol.* **17**, 1–10, DOI: 10.1371/journal.pcbi.1008646 (2021).
9. Merkt, S. *et al.* Supplementary files to Long-term monitoring of SARS-CoV-2 seroprevalence and variants in Ethiopia provides prediction for immunity and cross- immunity, DOI: 10.5281/zenodo.10871139 (2024).
10. Salvatier, J., Wiecki, T. V. & Fonnesbeck, C. Probabilistic programming in python using PyMC3. *PeerJ Comput. Sci.* **2**, e55, DOI: 10.7717/peerj-cs.55 (2016).
11. Gudina, E. K. *et al.* Seroepidemiology and model-based prediction of SARS-CoV-2 in ethiopia: longitudinal cohort study among front-line hospital workers and communities. *Lancet Glob Heal.* **9**, e1517–e1527, DOI: 10.1016/S2214-109X(21)00386-7 (2021).
12. Fang, Z. *et al.* Comparisons of viral shedding time of SARS-CoV-2 of different samples in ICU and non-ICU patients. *J. Infect.* **81**, 147–178, DOI: 10.1016/j.jinf.2020.03.013 (2020).
13. McAloon, C. *et al.* Incubation period of COVID-19: a rapid systematic review and meta-analysis of observational research. *BMJ Open* **10**, e039652, DOI: 10.1136/bmjopen-2020-039652 (2020).
14. Davies, N. G. *et al.* Estimated transmissibility and impact of SARS-CoV-2 lineage b.1.1.7 in england. *Science* **372**, eabg3055, DOI: 10.1126/science.abg3055 (2021).
15. Pearson, C. A. *et al.* Estimates of severity and transmissibility of novel SARS-CoV-2 variant 501Y.V2 in south africa. https://cmmid.github.io/topics/covid19/sa-novel-variant.html (2021). Accessed: 2023-7-26.
16. Agency, U. H. S. Investigation of SARS-CoV-2 variants: technical briefing 12. https://www.gov.uk/government/publications/investigation-of-sars-cov-2-variants-technical-briefings (2021). Accessed: 2023-7-26.
17. Lyngse, F. P. *et al.* Household transmission of the SARS-CoV-2 omicron variant in denmark. *Nat. Commun.* **13**, 5573, DOI: 10.1038/s41467-022-33498-0 (2022).

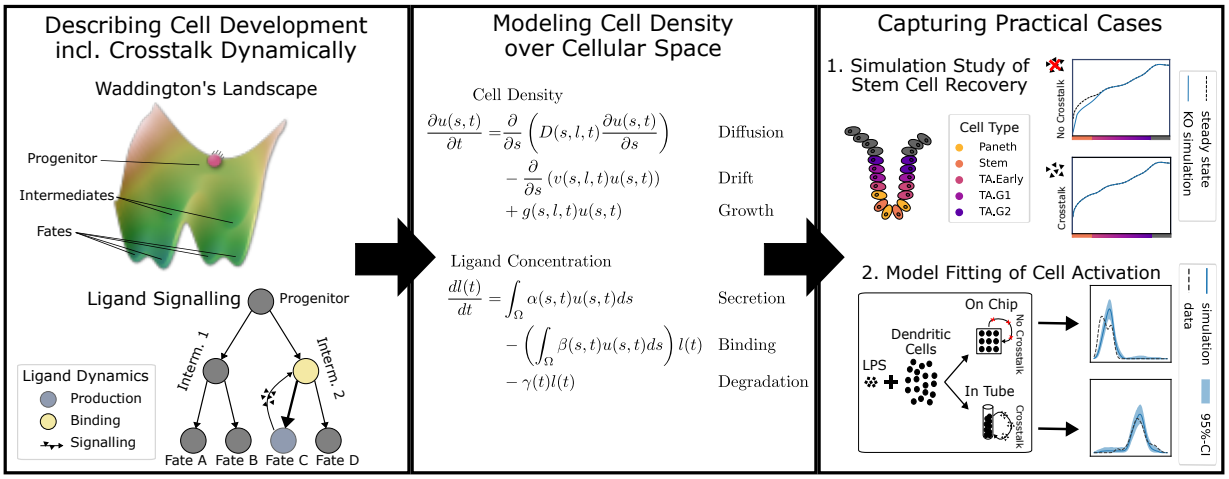
C A dynamic model for Waddington's landscape accounting for cell-to-cell communication

This preprint of this paper is reprinted as part of this thesis. The reference for the original preprint is as follows:
Merkt, S. *et al. A Dynamic Model for Waddington's Landscape Accounting for Cell-to-Cell Communication* Preprint available on *SSRN*. 2024. doi:10.2139/ssrn.5051345

Graphical Abstract

A dynamic model for Waddington's landscape accounting for cell-to-cell communication

Simon Merkt, Lara Fuhrmann, Erika Dudkin, Andreas Schlitzer, Barbara Niethammer, Jan Hasenauer



Highlights

A dynamic model for Waddington's landscape accounting for cell-to-cell communication

Simon Merkt, Lara Fuhrmann, Erika Dudkin, Andreas Schlitzer, Barbara Niethammer, Jan Hasenauer

- PDE-ODE system allows dynamic population-level modeling of communicating cells.
- Existence and uniqueness proof for model solutions provides theoretical foundation.
- Simulation study of stem cell recovery matches empirical data.
- Model fits single-cell data from LPS-stimulated dendritic cells.
- Findings showcase the importance of ligand-mediated cell-to-cell communication.

A dynamic model for Waddington's landscape accounting for cell-to-cell communication

Simon Merkt[a,1], Lara Fuhrmann[b,c,1], Erika Dudkin[d], Andreas Schlitzer[a], Barbara Niethammer[e], Jan Hasenauer[a,f,*]

[a]*Life and Medical Sciences (LIMES), University of Bonn, Carl-Troll-Str. 31, Bonn, 53115, North Rhine-Westphalia, Germany*

[b]*Department of Biosystems Science and Engineering, ETH Zurich, Schanzenstr. 44, Basel, 4058, Basel-Stadt, Switzerland*

[c]*SIB Swiss Institute of Bioinformatics, Elisabethenstr. 43, Basel, 4051, Basel-Stadt, Switzerland*

[d]*Data Science, Software Competence Center Hagenberg (SCCH), Softwarepark 21, Hagenberg im Mühlkreis, 4232, Upper Austria, Austria*

[e]*Institute for Applied Mathematics, University of Bonn, Endenicher Allee 60, Bonn, 53115, North Rhine-Westphalia, Germany*

[f]*Institute of Computational Biology, Helmholtz Zentrum München – German Research Center for Environmental Health, Ingolstädter Landstr. 1, Neuherberg, 85764, Bavaria, Germany*

Abstract

Waddington's landscape provides a conceptual model for developmental processes. It is the basis of various mathematical models describing cell maturation and development at cell and population levels. Yet, these mathematical models mostly disregard cell-to-cell communication, an essential process that modulates cellular decision-making and population dynamics.

In this study, we provide a dynamical model for cell maturation and development which can be seen as an extension of Waddington's landscape. The coupled system of partial and ordinary differential equations describes cell density along the cell state together with ligand concentrations. Cell-state-dependent ligand production determines ligand availability, which controls population-level processes. We provide proof of the existence and uniqueness of solutions for our coupled differential equation system and demonstrate the model's validity by analyzing single-cell transcriptomics data. Our results show that cell-to-cell communication is essential for accurately depicting biological recovery processes, such as the regeneration of stem cells in the intestine's crypt and the response of immune cells upon LSP stimulation.

Our findings underscore the importance of incorporating cell-to-cell communication into mathematical models of biological development. By doing so, we unlock the potential for

*Corresponding Author: jan.hasenauer@uni-bonn.de

[1]These authors contributed equally

deeper insights into complex processes such as tissue regeneration and immune responses, offering new avenues for understanding and predicting the dynamics of biological recovery and cell activation.

Keywords:
Population dynamics, Parameter estimation, Waddington's Landscape, Cell-to-cell communication, Stem Cell Regeneration, Dendritic Cell Activation, Single-Cell Transcriptomics Data

---

1. Introduction

Cell-to-cell communication is a fundamental biological process essential for a variety of cell- and tissue-level functions, including cell proliferation and differentiation along with development processes and cell activation [1, 2, 3, 4, 5]. Cells communicate through cell-to-cell contacts or via secretion, binding, and uptake of biochemical substances. Ligand-receptor interactions are crucial since they initiate intracellular signaling cascades, e.g., via protein-protein interaction networks. Indeed, cell-to-cell communication can substantially impact gene expression and may even be inferred from it [6].

Given the importance of cell-to-cell communication, it is surprising that many core concepts conceptualizing cellular development do not account for it. An important example is Waddington's landscape [7], which conceptualizes cell maturation and differentiation and links concepts in systems theory [8]. The perspective of cell maturation and differentiation as dynamical processes in an energy landscape allows for the mathematical modeling of cell population dynamics, particularly benefiting the analysis of single-cell transcriptomics data. Differential equations of various forms can describe the dynamics. Ordinary differential equation (ODE) systems take gene expressions as state variables and capture the genetic interdependencies by differential equations [9, 10]. Randomness inherently involved in biological processes can be accounted for by extending ODEs to stochastic differential equations [11] or through a mean-field approach by using partial differential equations (PDEs) [12, 13, 14]. PDEs describe evolving cell densities over the molecular space or a dimensionally reduced version, and stochasticity is integrated via a diffusion component. Moreover, there are also PDE models disregarding diffusion and employing optimal transport frameworks, where the core assumption is that cell distributions will choose an optimal path for changing between timepoints [15, 16].

In contrast to these dynamical models, which lack cell-to-cell communication, several statistical methodologies specifically aiming at analyzing cell-to-cell interactions using single-cell transcriptomic data were introduced recently [17]. They identify genetic sequences corresponding to ligand and receptor proteins and use predefined interaction databases to map them together. Cells with a high abundance of a specific ligand sequence are then connected to cells that express compatible receptors' RNA, which leads to an extensive cell-to-cell interaction network. Some approaches focus on directly pairing ligands to receptors [18], while others also take subunits of ligand-receptor complexes into account [19]. Moreover, there are methods additionally investigating gene regulatory effects of communication on interacting

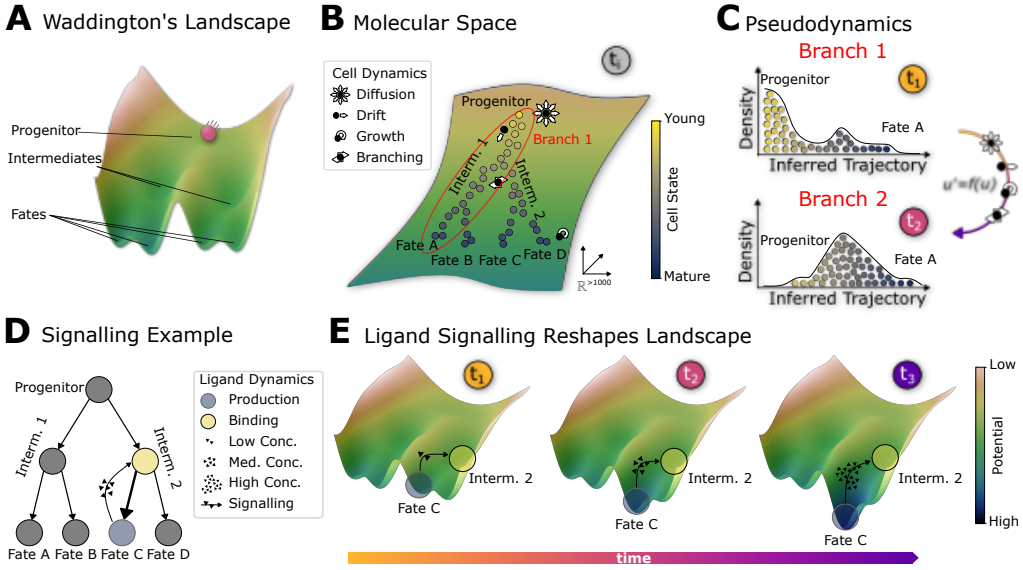


Figure 1: **Conceptualization of Waddington's Landscape, Pseudodynamics, and ligand signaling.** (A) Waddington's Landscape: Understanding cell state changes like a ball rolling down a hill from progenitor state through intermediate stages to different cell fates. (B) Cell measurements on low dimensional submanifold of molecular space. Dynamics on this manifold governed by diffusion, drift, growth, and branching. (C) Pseudodynamics modeling dynamics of cell densities on inferred trajectories through multiple snapshot measurements. (D) Schematic of a signaling example, where cells at one fate signal to cells at an intermediate branching state that they should develop into this fate at higher rates. (E) Potential landscape reshaped over time by ligand signaling, e.g., from one fate to an intermediate state.

cells [20]. These statistical methods, on their own, consider only static distinct time points without incorporating any dynamic component. However, recent approaches try integrating these statistical methods into dynamic models. Sha et al. [16] use a reversible dimension reduction, compute unobserved time points with a dynamic model, and apply ligand-receptor analysis on those after recomputing the unreduced representation. This approach enables the investigation of unobserved time points but, as the underlying receptor-ligand pairing methods, remains susceptible to database biases inherent in the ligand-receptor pairings [17].

To date, we are unaware of any modeling approaches that directly incorporate cell-to-cell communication into a dynamic model and could address the abovementioned limitations. We propose a dynamical model that simultaneously describes cell distributions over cell states and concentrations of signaling ligands. To do so, we build on the idea of Fischer et al. [12], who denoted their reaction-diffusion-drift equation as pseudodynamics and facilitated the assessment of cell-state dependent differentiation and proliferation rates. We formulate a more general version of their model by integrating it into a combined PDE and ODE system. This extended model describes the concentrations of both cells and ligands, thereby capturing the dynamic nature of developmental potential across a time period rather than at a single instance (Figure 1). After laying out the precise formulation of our differential equations, we provide proof for the existence and uniqueness of solutions to this system. Then, through a simulation study utilizing parameters derived from existing literature, we demonstrate that incorporating the communication between cells is crucial for accurately depicting biological

recovery processes, such as the regeneration of stem cells in the intestine's crypt (data from [21]). Furthermore, using a dataset from Shalek et al. [22], we illustrate that the dynamics of dendritic cell activation can be deduced from measurements of communicating and non-communicating cells, even without direct data on ligand concentrations. Overall this paper showcases how cell-to-cell communication can be implemented in dynamic mathematical models of biological development.

2. Mathematical Model

This study considers a population of cells that communicate via ligands. The state of a cell is denoted by $s(t) \in \Omega$, where $\Omega \subset \mathbb{R}^{n_s}$ is bounded, and the ligand concentration by $l(t) \in \mathbb{R}^{n_l}_{\geq 0}$. Cells can undergo multiple processes:

1. **Cell development:** A cell can change its state in response to intra- and extracellular processes. The dynamics of the cell state are governed by a stochastic differential equation $ds = v(s,l,t)dt + D^{1/2}(s,l,t)dB_t$, with drift term $v(s,l,t)$ and diffusion term $D^{1/2}(s,l,t)$. The drift term is considered to be related to the developmental potential W defined by Waddington's landscape (Figure 1A) as $\frac{\partial}{\partial s}W = -v$.
2. **Cell division and death:** A cell can divide and die. The effective proliferation rate, which incorporates cell division and death, is denoted by g. We assume in the following that cells in state s divide in two daughter cells which are also in state s.

In the limit of large cell numbers, we obtain a population balance model for the state- and time-dependent cell number density function $u(s,t) \in \mathbb{R}_{\geq 0}$:

$$\frac{\partial u(s,t)}{\partial t} = \frac{\partial}{\partial s}\left(D(s,l,t)\frac{\partial u(s,t)}{\partial s}\right) - \frac{\partial}{\partial s}\left(v(s,l,t)u(s,t)\right) + g(s,l,t)u(s,t) \tag{1}$$

with initial condition

$$u(s,0) = u_0(s) \geq 0 \quad \forall s \in \Omega \tag{2}$$

and no-flux boundary conditions, i.e., drift and diffusion cancel out on the boundary:

$$\left(D(s,l,t)\frac{\partial}{\partial s}u(s,t) - v(s,l,t)u(s,t)\right)\Big|_{s\in\partial\Omega} = 0 \quad \forall t \in [0,T]. \tag{3}$$

The total number of cells at time t is given by the integral u over the open set of possible cell states Ω. We assume that the range of cell states is finite and that no-flux conditions hold at the boundary $\partial\Omega$. The cell state s is defined based on essential characteristics of cells, which can be transcript expression and protein abundance of marker genes as measured by single-cell RNA-sequencing [23, 24] or flow cytometry [25, 26]. Yet, low dimensional cell state specifications can also be used, e.g., based on diffusion maps [27] as in [12]. This exploits that single-cell data often lie on or are close to low-dimensional manifolds (Figure 1B-C).

We consider that cells communicate using a set of ligands which participate in three classes of processes:

1. **Ligand secretion:** Ligands are secreted by cells in state s at rate $\alpha(s,t)$.

2. **Ligand binding:** Ligands bind to cells in state s expressing the receptors, yielding a cell-state specific binding rate $\beta(s,t)l$.

3. **Ligand degradation:** Ligands are naturally degraded at rate $\gamma(t)l$.

We assume the ligands have a high diffusion coefficient or small distances between cells. Accordingly, all cells are exposed to the same ligand concentration, and spatial effects on ligand concentrations can be disregarded. The governing equation for the extracellular ligand concentration l is

$$\frac{dl(t)}{dt} = \int_\Omega \alpha(s,t)u(s,t)ds - \left(\int_\Omega \beta(s,t)u(s,t)ds\right) l(t) - \gamma(t)l(t) \tag{4}$$

with initial condition

$$l(0) = l_0 \geq 0. \tag{5}$$

Following the previous formulations, we model the overall dynamics of a developing population of cells that communicate via ligands by the collection of Equations (1)-(5). These equations form a coupled ODE-PDE system. The equations are coupled via the process parameters, source, and sink terms. Cell drift v, diffusion D, and proliferation g can depend on the cell state s, ligand concentrations l, and time t. Ligand secretion α and binding β can depend on cell state s and time t, and ligand degradation γ can depend on time t. Furthermore, all process parameters can depend on experimental conditions, which we omitted in the expressions to ensure readability. The dependencies are considered to be smooth, using, e.g., Hill functions, Gaussian kernels, or splines. For example, mature cells in certain states might produce ligands to signal cells in intermediate states, where cells binding the ligand are more inclined to develop towards this fate (Figure 1D-E).

3. Analysis of Mathematical Model

The coupled ODE-PDE system (1)-(5) has not yet been used to study cell population dynamics. To assess its mathematical validity, we analyze the global existence and uniqueness of its solution by applying the approach presented in [28] and [29].

Therefore, we assume regularity of the model coefficients:

Condition 1. *For the ligand dynamics we assume that* $\alpha, \beta \in C([0,T], L^\infty(\Omega)$ *are positive a.e. and that* $\gamma(t) \in C([0,T], \mathbb{R}_{>0})$.

Condition 2. *For the cell population dynamics, we assume:*

(i) $D(s,l,t), v(s,l,t), g(s,l,t) \in L^\infty(\Omega \times [0,\infty] \times [0,T])$.

(ii) $D(s,l,t) \geq \eta > 0$ *for all* $(s,l,t) \in \Omega \times \mathbb{R}_{\geq 0} \times [0,T]$.

(iii) $u_0(s) \in L^2(\Omega)$

(iv) $D(s,l,t), v(s,l,t)$ and $g(s,l,t)$ are globally Lipschitz in $l \in [0,\infty)$.

Given these regularity conditions, the following theorem holds:

Theorem 1 (Existence of unique weak solution)**.** *If Conditions 1 and 2 hold, then the ODE-PDE system* (1)-(5) *possesses a unique weak solution* (u,l) *satisfying*

$$\begin{aligned} u &\in L^2([0,T],H^1(\Omega)) \cap C([0,T],L^2(\Omega)),\\ \partial_s u &\in L^2([0,T],H^{-1}(\Omega)),\\ l &\in C([0,T]). \end{aligned}$$

To prove Theorem 1, we will show that (I) the ODE for the ligand dynamics has a unique solution l^u for a fixed u, (II) the PDE for the cell population density has a unique solution $\hat{u}$ for a fixed l^u, and (III) $\mathcal{B} : u \mapsto l^u \mapsto \hat{u} = \mathcal{B}(u)$ is a contraction. In this case, Banach's fixed point theorem can be applied and ensures existence and uniqueness [28].

For the subsequent analysis, we consider the closed set

$$\begin{aligned} X := \{u \in C([0,T],L^2(\Omega)) \cap L^2([0,T],H^1(\Omega))\\ \text{s.t. } u \geq 0 \text{ and } \max_{t\in[0,T]} ||u(t)||_{L^2(\Omega)} \leq C_X ||u_0||_{L^2(\Omega)}\} \end{aligned}$$

with the norm $||u||_X := \max_{t\in[0,T]} ||u(\cdot,t)||_{L^2(\Omega)}$ and $C_X = C(\Omega, D, v, g, T)$.

We start by verifying existence and uniqueness locally, i.e., on the time interval $[0,T_0]$ for a sufficiently small $T_0 \leq T$:

I. The ligand dynamics are governed by a 1st order linear ODE with time-varying coefficients. Condition 1, i.e. $\alpha, \beta \in C([0,T],L^\infty(\Omega))$ and $\gamma \in C([0,T],\mathbb{R}_{>0})$, ensures that the solution $l^u(t)$ to (4) + (5) for a fixed $u \in X$ is continuous in t and therefore bounded on the closed time interval $[0,T_0]$. The proof is provided in Appendix A, Lemma 1.

II. The cell population dynamics are shaped by a 2nd-order parabolic PDE with Robin boundary conditions. Condition 2 (i)-(iv) and the application of Galerkin Approximations and Energy Estimates as in Chapter 7 in Evans' book [29], provide that the PDE (1)-(3) for fixed $l^u \in C([0,T])$ possesses a unique weak solution $\hat{u} \in L^2([0,T],H^1(\Omega))$ with $\hat{u}' \in L^2([0,T],H^{-1}(\Omega))$. Moreover, $\hat{u} \in C([0,T],L^2(\Omega))$ and the following estimate holds for $\hat{u}$:

$$\max_{0\leq t\leq T} ||\hat{u}(t)||_{L^2(\Omega)} + ||\hat{u}||_{L^2([0,T],H^1(\Omega))} + ||\hat{u}'||_{L^2([0,T],H^{-1}(\Omega))} \leq C||u_0||_{L^2(\Omega)},$$

with $C = C(T,\Omega,D,v,g)$. Additionally, if $\hat{u}(s,0) = u_0(s) \geq 0$, then $\hat{u}(s,t) \geq 0$ for all $t \in [0,T]$, i.e. non-negativity is preserved. The proof details are provided in Appendix A, Lemma 2.

III. Given that the above holds, we can make use of Banach's fixed point theorem, which implies that there exists a unique fixed point $\mathcal{B}(u) = \hat{u} = u$ if $\mathcal{B}$ is a contraction.

To be precise, there has to exist a sufficiently small $T_0 > 0$ and $\rho \in (0,1)$ such that $\mathcal{B} : X_0 \to X_0$ is a contraction, i.e.

$$||\mathcal{B}(u_1) - \mathcal{B}(u_2)||_X \leq \rho ||u_1 - u_2||_X \ \forall u_1, u_2 \in X_0,$$

where X_0 is the space X reduced to the time interval $[0, T_0]$ instead of $[0, T]$.

To prove the contraction property, we use that there exists a constant $K(T) > 0$ such that

$$||\hat{u}_1 - \hat{u}_2||_X \leq K(T)||u_1 - u_2||_X^2. \tag{6}$$

where for $T \to 0$, we have $K(T) \to 0$. Since the construction of K is rather technical, we provide it in Appendix A.3. Since T is chosen arbitrarily we can conclude that:

$$\exists T_0 > 0 \text{ such that } K(T_0) < 1 \Rightarrow \mathcal{B} \text{ contraction },$$

where $T_0 = T_0(|l_0|, ||u_0||_{L^2(\Omega)}, \Omega, D, v, g)$. Since $\mathcal{B}$ is a contraction on X_0, we can apply Banach's fixed point theorem for closed subsets to $\mathcal{B}$ and obtain

$$\exists! u \in X_0 : \mathcal{B}(u) = u.$$

This completes the proof of the local existence of a unique solution (u, l).

To generalize the result for the local existence of a unique solution (u, l) on the time interval $[0, T_0]$ to any bounded time interval $[0, T]$ for any $T > 0$, we define new initial conditions: $u_0(s) = u(s, T_0)$ and $l_0 = l(T_0)$. As the results (I)-(III) presented above hold for any initial conditions which are non-negative (and L^2 for u_0), it follows the existence of a time T_1 such that the problem (1)-(5) with the new initial conditions possesses a unique local solution on $[T_0, T_1]$. We can extend the solution to the bigger interval $[0, T_1]$. This strategy can be applied repeatedly, since l is bounded on a bounded interval and for any $T > 0$, there exists a constant $C = C(D, v, g)$ such that $||u(., t)||_{L^2(\Omega)}^2 \leq e^{CT}||u_0||_{L^2(\Omega)}^2$. The proof details are provided in Appendix A.4, Lemma 5. Therefore, we can step-wise extend the existence interval of the local solution to any interval $[0, T]$, which proves the global existence of a unique solution for any $T > 0$.

4. Numerical simulation and model-based data analysis

The coupled ODE-PDE system (1)-(5) describes cell population dynamics based on the properties of cells in different cell states and ligand characteristics. To study the process dynamics, we introduce a numerical simulation method. Furthermore, we formulate the mathematical problem for assessing cell and ligand properties based on experimental data and introduce parameter estimation and uncertainty analysis methods. To ensure reusability, we made the code for simulations, data integration, and models available at Zenodo [30].

4.1. Numerical simulation

To study the dynamics of the coupled ODE-PDE system (1)-(5), we use numerical simulation based on the finite volume method (FVM) [31, 32]. This method divides the cell state into a finite number of control volumes and then approximates the integral form of the conservation laws over these control volumes. It is based on the fundamental theorem of calculus and focuses on the fluxes of conserved quantities across the boundaries of the control volumes. These fluxes are calculated using approximate solutions at the interfaces between adjacent volumes, removing spatial derivatives. The FVM was chosen over the finite element method to ensure mass conservation and avoid population growth and shrinkage as a numerical artifact.

We will briefly sketch the discretization scheme for the case of a cell state reduction to a one-dimensional space, which is usually done for computational feasibility e.g. by trajectory inference, and a non-branching cell lineage to keep notation clear. This means we have $\Omega = [0, s_{\max}]$ and discretize it by choosing $n_b + 1$ equal-distant grid points

$$s_{\min} = s_{\frac{1}{2}} < ... < s_{i-\frac{1}{2}} < s_{i+\frac{1}{2}} < ... < s_{n_b+\frac{1}{2}} = s_{\max}, \tag{7}$$

which divide the cell state space into n_b control volumes, i.e. intervals, $[s_{i-\frac{1}{2}}, s_{i+\frac{1}{2}}]$ with $i \in \{1, ..., n_b\}$ of length $h = s_{i+\frac{1}{2}} - s_{i-\frac{1}{2}} = \frac{s_{\max}-s_{\min}}{n_b}$.

The grid points on the left edge of the i-th control volume are denoted by $s_{i-\frac{1}{2}} = s_{\min}+(i-1)h$ and the grid points on the right edge of the interval by $s_{i+\frac{1}{2}} = s_{\min} + ih$. The centers of these control volumes are given by:

$$s_i = \frac{s_{i-\frac{1}{2}} + s_{i+\frac{1}{2}}}{2} = s_{\min} + (i - \frac{1}{2})h, \qquad \text{for } i = 1, ..., n_b.$$

For each center point, the average density over its control volume is given by:

$$u_i(t) = \frac{1}{h} \int_{s_{i-\frac{1}{2}}}^{s_{i+\frac{1}{2}}} u(s, t)ds. \tag{8}$$

The difference between the actual value at a center point s_i, $u(s_i, t)$, and the average over the interval, u_i, is $O(h^2)$ [33]. Applying the finite volume method by integrating over each control volume $[s_{i-\frac{1}{2}}, s_{i+\frac{1}{2}}]$ and integrating by parts results in a space-discretization of (1). Accordingly, we approximate the integrals of the ligand ODE by sums. Hence, the coupled

PDE-ODE system (1)-(5) reduces to a system of coupled ODEs:

$$\frac{du_1}{dt} = -\frac{1}{h^2}\left(D_{1+\frac{1}{2}}\left(u_1 - u_2\right)\right) - \frac{1}{2h}v_{1+\frac{1}{2}}\left(u_1 + u_2\right) + g_1 u_1$$

$$\begin{aligned}\frac{du_i}{dt} = &-\frac{1}{h^2}\left(D_{i-\frac{1}{2}}\left(u_{i-1} - u_i\right) - D_{i+\frac{1}{2}}\left(u_i - u_{i+1}\right)\right) \\ &+ \frac{1}{2h}\left(v_{i-\frac{1}{2}}\left(u_{i-1} + u_i\right) - v_{i+\frac{1}{2}}\left(u_i + u_{i+1}\right)\right) + g_i u_i, \quad i = 2, \ldots, n_b - 1\end{aligned}$$

$$\begin{aligned}\frac{du_{n_b}}{dt} =& \frac{1}{h^2}\left(D_{n_b-\frac{1}{2}}\left(u_{n_b-1} - u_{n_b}\right)\right) + \frac{1}{2h}v_{n_b}\left(u_{n_b-1} + u_{n_b}\right) + g_{n_b}u_{n_b} \\ \frac{dl}{dt} =& \sum_{i=1}^{n} u_i\alpha_i - \left(\sum_{i=1}^{n} u_i\beta_i\right) l - \gamma l,\end{aligned}$$

with $l = l(t)$ and $\gamma = \gamma(t)$. Moreover, $\alpha_i = \alpha(s_i, t)$, $\beta_i = \beta(s_i, t)$, $D_{i\pm\frac{1}{2}} = D(s_{i\pm\frac{1}{2}}, l, t)$ and $v_{i\pm\frac{1}{2}} = v(s_{i\pm\frac{1}{2}}, l, t)$ for $i = 0, ...n_b$, and $g_i = g(s_i, l, t)$ for $i = 1, ..., n_b$.

We developed two Python implementations for numerical simulation of (1)-(5) for given initial conditions and parameters. The *educational implementation* has been designed for testing and avoids using advanced numerical simulation methods. The discretization is constructed explicitly, and the numerical simulation of the system provided by the ODE and the discretized PDE is performed using SUNDIALS CVODE [34, 35] via the simulation toolbox AMICI [36]. The *computationally efficient implementation* has been designed to enable parameter estimation and uncertainty analysis. This implementation builds on the PDE solver package FiPy [37], which handles the spatial discretization described above and the time-stepping. For numerical integration over time it employs the approximation

$$\int_{s_{i-\frac{1}{2}}}^{s_{i+\frac{1}{2}}} \partial_t u(s, t + \Delta t) ds \approx \frac{(u_i(t + \Delta t) - u_i(t))h}{\Delta t}.$$

Setting $t_{\hat{k}} = \hat{k}\Delta t$ for a user chosen time step size Δt and defining $u_{i\hat{k}} = u_i(t_{\hat{k}})$, we obtain a system of linear equations. For solving this system, FiPy offers a variety of linear solvers, out of which we employed the default: scipy's linear LU solver[38]. The implementation using FiPy requires reduced user input and allows for a straightforward adaptation.

We assessed the accuracy and efficiency of the two implementations for numerical simulation on a 2-dimensional PDE test case with constant ligand concentration. We compared results for different spatial discretization and time step sizes, which ensured a high quality of the numerical solution.

4.2. Parameter estimation

We propose a model-based data analysis approach to determine cell and ligand properties based on experimental data. This approach constructs parameters and initial condition of

(1)-(5) from measured cell population density y_{kj}, which are collected at time points t_k, $k = 1, \dots, n_t$ and experimental conditions c_j, $j = 1, \dots, n_c$. The resulting model comprehensively describes the available experimental data, and the parameters offer insights into the underlying biological processes.

To ensure that the inverse problem of determining parameters and initial conditions is computationally feasible, we employ parametric functions for drift v, diffusion D, proliferation g, and initial value u_0, as well as ligand secretion α, binding β, degradation γ, and initial value l_0. We consider constants, Hill functions, splines, and combinations thereof. The unknown parameters of the parametric functions are denoted by θ_v, θ_D, θ_g, θ_{u_0}, θ_α, θ_β, θ_γ, and θ_{l_0}, and collectively as $\theta = (\theta_v, \theta_D, \theta_g, \theta_{u_0}, \theta_\alpha, \theta_\beta, \theta_\gamma, \theta_{l_0})$. The vector of unknown parameters is real-valued and constrained to the set Θ, i.e. $\theta \in \Theta \subset \mathbb{R}^{n_\theta}$.

The maximum likelihood estimate of the parameter vector, θ^{ml}, is obtained by maximizing the likelihood of observing the data given the model. Assuming independence of the measurements for different time points and conditions, the optimization problem is formulated as

$$\theta^{\mathrm{ml}} = \arg\max_{\theta \in \Theta} \prod_{k=1}^{n_t} \prod_{j=1}^{n_c} p(y_{kj} | \phi_u(t_k, \cdot; c_j, \theta)),$$

with $\phi_u(\cdot, t; c, \theta))$ denoting the solution operator for the population density u in (1)-(5) for condition c and parameters θ.

The formulation of the likelihood function depends on the measurement technique and the subsequent data processing. Here, we assume that the states of m_{kj} cells are assessed using single-cell measurement technology followed by dimension reduction to a locally one-dimensional manifold. For parameter estimation, histograms are constructed from the single-cell measurements. This yields a vector of counts $y_{kj} \in \mathbb{N}_0^{n_b}$, with the i-th entry, y_{kji}, indicating the number of cells in bin i, i.e. with $s \in (s_{i-\frac{1}{2}}, s_{i+\frac{1}{2}}]$. For the case of a non-branching cell state ($s \in \Omega = (s_{\min}, s_{\max})$), the bin intervals are given by $s_{\min} = s_{\frac{1}{2}} < \dots < s_{i-\frac{1}{2}} < s_{i+\frac{1}{2}} < \dots < s_{n_b+\frac{1}{2}} = s_{\max}$ as in the FVM discretization (7). Assuming unbiased cell sampling, the probability of picking a cell with state in bin i is given by the fraction of the cells in $(s_{i-\frac{1}{2}}, s_{i+\frac{1}{2}}]$, which is

$$f_i(c, \theta) = \frac{\int_{s_{i-\frac{1}{2}}}^{s_{i+\frac{1}{2}}} u(s, t; c, \theta) ds}{\int_{s_{\min}}^{s_{\max}} u(s, t; c, \theta) ds}. \tag{9}$$

The likelihood for these types of population-level measurements is given by the multinomial probability mass function:

$$p(y_{kj} | u(t_k, \cdot; c_j, \theta)) = \frac{m_{kj}!}{\prod_{i=1}^{n_b} y_{kji}!} \prod_{i=1}^{n_b} f_i^{y_{kji}}(u(t_k, \cdot; c_j, \theta)).$$

In this study, we determine the maximum likelihood estimate by minimizing the negative log-likelihood function. For the considered likelihood function, this minimization problem is

given by

$$\theta^{\mathrm{ml}} = \arg\min_{\theta\in\Theta} \left(-\sum_{k=1}^{n_t}\sum_{j=1}^{n_c}\sum_{i=1}^{n_b} y_{kji} \log f_i(\phi_u(t_k, \cdot; c_j, \theta)) + \mathrm{const.} \right).$$

We follow a discretize-optimize strategy, assuming that the numerical simulation algorithm provides an accurate solution for the coupled ODE-PDE system. The time stepping $t_{\hat{k}}$ for the numerical solution is chosen in such a way that the measurement time points t_k are covered by $t_{\hat{k}}$ and no additional interpolation is required. The fraction f is directly computed from the finite volume approximation of $u(s, t; c, \theta)$, using the fact that plugging (8) into (9) yields

$$f_i(c, \theta) = \frac{u_i(t; c, \theta)}{\sum_{j=1}^{n_b} u_j(t; c, \theta)}.$$

We implement the parameter estimation using the pyPESTO framework [39], which offers a broad spectrum of local and global optimization methods. Following comprehensive testing, we decided on a multi-start local optimization using the gradient-based interior point algorithm IPOPT [40], where the gradient is computed using finite differences.

To assess the reproducibility of the parameter optimization, we evaluate its convergence with the waterfall plots of the optimization results. We compare the measurement data with the distribution of measurements expected for the estimate to evaluate the quality-of-fit of the maximum likelihood estimate. Therefore, we calculate for each time point and experimental condition Gaussian kernel density estimates based on the states of the experimentally observed cells and a distribution of Gaussian kernel density estimates for samples from the population density $u(t_k, \cdot; c_j, \theta^{\mathrm{ml}})$. In particular if the number of observed cells m_{kj} is small, kernel density estimates for individual samples from $u(t_k, \cdot; c_j, \theta^{\mathrm{ml}})$ can differ substantially from $u(t_k, \cdot; c_j, \theta^{\mathrm{ml}})$, which is important to consider this in the evaluation the quality-of-fit.

For models which provide an accurate description of experimental data, an assessment of parameter uncertainties is meaningful. We investigate this via the ensemble method, selecting the top K results from a multistart optimization with N starts, where $K \ll N$, as representatives of this set [41]. Then we can asses how tight the uncertainty set is around $\theta^{\mathrm{ml}}_{\bar{k}}$ for each dimension $\bar{k}$, i.e., how certain we are about this parameter. Given the high computational demand of solving our system, we can restrict the computational load to what has already been computed for the parameter estimation while efficiently obtaining results on parameter uncertainty.

5. Results

The numerical simulation and parameter estimation methods introduced in the previous section should facilitate the study of a broad range of biological processes. To assess this, we model cell population dynamics in the intestinal crypt and characterize the properties of dendritic cells upon activation.

5.1. Modeling cell population dynamics in intestinal crypts

To evaluate the applicability of the proposed modeling and simulation framework, we study cell proliferation and differentiation in intestinal crypts (Figure 2A). Here, we aim to determine whether our framework can effectively capture cell development and the significance of ligands in shaping this process. Therefore, we develop a model and study the impact of an established knockout experiment on the differentiation process.

Biological background. The epithelial cells of the crypt in the small intestine exhibit a rapid turnover, being renewed every 4-5 days [42]. Stem cells located at the base of the crypt primarily proliferate asymmetrically, resulting in one stem cell and one transit amplifying (TA) daughter cell. TA cells undergo rapid proliferation and migrate from the crypt base outward. During migration, these cells differentiate and undergo cell cycle arrest in the upper part of the crypt. As they exit the crypt, they mature into various cell types and migrate to the tips of the villi. Paneth cells are the exception to this upward migration; during maturation, they move down to the crypt base and reside there alongside stem cells [43, 42, 44, 45].

Communication between Paneth cells, stem cells, and TA cells is crucial for maintaining crypt homeostasis [46, 47, 48]. The canonical Wnt3 pathway is a key signal promoting the maintenance and proliferation of stem and TA cells in the lower crypt [42, 49, 50]. Indeed, there are strong indications that upon loss or damage of the stem cells, Wnt3 activates a dedifferentiation process whereby first- and second-generation TA cells regain stem cell properties and functionalities.

Model formulation. To develop a model for cell population dynamics in the intestinal crypt, we assess the structure of the cell state space using published single-cell RNA sequencing data for epithelial cells of the mouse small intestine [21]. Using the scanpy framework, specifically partition-based graph abstraction (PAGA) and diffusion pseudotime [51, 52, 53], we identify two differentiation trajectories within the epithelial dataset (Figure 2B-C) for the first two diffusion map components. Stem cells differentiate into early-generation TA cells, which either develop into Paneth cells or further differentiate into TA cells in subsequent cell cycle phases. Upon reaching cell cycle arrest, TA cells commit to mature into specific cell types. We represent cell state space as the union of two one-dimensional sets (Figure 2D). The main branch, with stem and TA cells, is mapped to $\Omega_1 = [0, 1]$ and the side branch with Paneth cells to $\Omega_2 = [0.3, 1]$. The number density on these two line segments is denoted by $u_1(s_1, t)$ and $u_2(s_2, t)$, and cells are allowed to switch in the branching region $s_1, s_2 \in [0.3, 0.4]$, reflecting that cell fate is decided at early TA cell stages (Figure 2D). This yields for all $t \in [0, T]$ the

population model

$$\begin{aligned}
\forall s_1 \in \Omega_1: \quad \partial_t u_1(s_1,t) &= \partial_{s_1}\left(D_1(s_1)\partial_{s_1}u_1(s_1,t)\right) \\
&\quad - \partial_{s_1}\left(v_1(s_1,l)u_1(s_1,t)\right) + g_1(s_1)u_1(s_1,t) \\
&\quad - \mathbb{1}_{[0.3,0.4]}(s_1)g_{12}u_1(s_1,t) \\
\forall s_2 \in \Omega_2: \quad \partial_t u_2(s_2,t) &= \partial_{s_2}\left(D_2(s_2)\partial_{s_2}u_2(s_2,t)\right) \\
&\quad - \partial_{s_2}\left(v_2(s_2)u_2(s_2,t)\right) + g_2(s_2)u_2(s_2,t) \\
&\quad + \mathbb{1}_{[0.3,0.4]}(s_2)g_{12}u_1(s_2,t)
\end{aligned}$$

where $\mathbb{1}_{[0.3,0.4]}$ is 1 on the interval $[0.3, 0.4]$ and 0 elsewhere. For initial and boundary conditions, we obtain

$$\begin{aligned}
u_1(s_1,0) &= u_{1,0}(s_1), \qquad \forall s_1 \in \Omega_1 \\
u_2(s_2,0) &= u_{2,0}(s_2), \qquad \forall s_2 \in \Omega_2
\end{aligned}$$

and

$$\begin{aligned}
\left(D_1(s_1)\partial_{s_1}u_1(s_1,t) - v_1(s_1,l)u_1(s_1,t)\right)\big|_{s_1\in\{0,1\}} &= 0 \\
\left(D_2(s_2)\partial_{s_2}u_2(s_2,t) - v_2(s_1,l)u_2(s_2,t)\right)\big|_{s_2\in\{0.3,1\}} &= 0
\end{aligned}$$

We account for ligand Wnt3 in the model due to its above-described importance and the observed in-homogeneous expression (Supplementary Figure S1B). We find Wnt3 to be highly expressed in Paneth cells, while the Wnt3 receptors Fzd-7 and Lrp5-receptor are highly expressed in stem cells and expressed at intermediate levels for TA cells (Supplementary Figure S1). This yields the ligand model

$$\frac{dl(t)}{dt} = \int_{\Omega_2} \alpha(s)u_2(s,t)ds - \left(\int_{\Omega_1} \beta(s)u_1(s,t)ds\right) l(t) - \gamma l(t),$$

with Hill functions

$$\alpha(s) = \alpha_{max}\frac{K_\alpha^{n_\alpha}}{K_\alpha^{n_\alpha} + s^{n_\alpha}} \qquad \text{and} \qquad \beta(s) = \beta_{max}\frac{K_\beta^{n_\beta}}{K_\beta^{n_\beta} + s^{n_\beta}},$$

with inflection points K_α and K_β and Hill coefficients n_α and n_β (Figure 3A). Following the observation that high levels of Wnt3 inhibit differentiation of cells in the lower crypt, we account for a dependence of the drift in the main branch on (a) ligand concentration and (b) cell state,

$$v_1(s_1,t,l) = \Bigg(1 + \underbrace{k\left(\kappa\frac{K_{v,l}^{n_{v,l}}}{K_{v,l}^{n_{v,l}} + l(t)^{n_{v,l}}} - 1\right)}_{=:v_{\text{ligand}}} \underbrace{\left(\frac{K_{v,s}^{n_{v,s}}}{K_{v,s}^{n_{v,s}} + s_1^{n_{v,s}}}\right)}_{=:v_{\text{state}}}\Bigg)\tilde{v}_1(s_1),$$

with inflection points $K_{v,l}$ and $K_{v,s}$, Hill coefficients $n_{v,l}$ and $n_{v,s}$, scaling parameters $k > 0$ and $\kappa > 1$), and baseline drift $\tilde{v}_1(s_1)$. The baseline growth, drift, and diffusion coefficients are chosen to reflect previous literature. The growth rates are chosen based on knowledge

about proliferation rates and cell survival times [42, 54, 49, 44], and are negative for terminal cells to reflect their exit out of the crypt. The drift was chosen to match reported transition times [42], and is high close to the stem cell state and lower for differentiated cells. The diffusion– on which information is scarce– is set to low values overall, despite the regime for Paneth cells, which appears highly variable. To ensure smoothness and flexibility, we use natural cubic splines (ncs):

$$\begin{aligned} g_i(s) &= \text{ncs}(s; \hat{g}_{i,0}, \ldots, \hat{g}_{i,n_i-1}), \\ \tilde{v}_i(s) &= \exp(\text{ncs}(s; \hat{v}_{i,0}, \ldots, \hat{v}_{i,n_i-1})), \\ D_i(s) &= \exp(\text{ncs}(s; \hat{D}_{i,0}, \ldots, \hat{D}_{i,n_i-1})), \end{aligned}$$

with $n_1 = 10$ and $n_2 = 7$ grid points (Figure 3B). We ensure positivity for the drift and diffusion by exponentiation of the natural cubic splines.

Model simulation and testing. The model formulation provides a high-level description of cell proliferation, differentiation, and communication in the intestinal crypt. Yet, while some parameters can be informed based on the literature, others remain unknown. We choose the unknown parameters and initial distribution to retain essential biological properties, most notably the cell type distribution [55, 45, 56, 57, 44]. The numerical simulation of the model for the selected parameters (Supplementary Table S1) reveals a realistic distribution in steady state (Figure 4A).

Following the positive evaluation, we want to determine if the proposed model captures the results of previous studies. In particular, we assess if the cell state distribution is stable and reverts to the previously observed distribution. Therefore, we perform an in-silico knockout of the stem cell compartment. We find that the simulations capture the experimentally reported replenishment of the stem cell compartment from the TA cells [47, 46, 48]. Indeed, the perturbation is suppressed and the system returns to the original steady state (Figure 4B). To assess if the model correctly describes the importance of the Wnt3-mediated feedback in the process [48], we perform a second in-silico experiment in which the effect of Wnt3 on the drift of stem and TA cells is disregarded (Figure 4B). In this case, the system does not return to the original steady state, but the stem cell fraction remains low. Moreover, we can see that the potential to develop towards later TA cells is decreased during peaking ligand concentrations before returning back to the original shape (Figure 4C).

In summary, the study shows that the model can describe cell population dynamics in the intestinal crypt. The model formulation can be easily informed using prior knowledge. The intuitive formulation even allows for directly extracting parameters from the available literature. Importantly, other continuum-based approaches, such as the work by [12], would not have been able to describe cell-to-cell communication.

5.2. Inference of LPS-induced Dendritic Cell Activation

To assess the reconstruction of cell-to-cell communication using the proposed modeling and inference framework, we study the activation of dendritic cells with lipopolysaccharide (LPS).

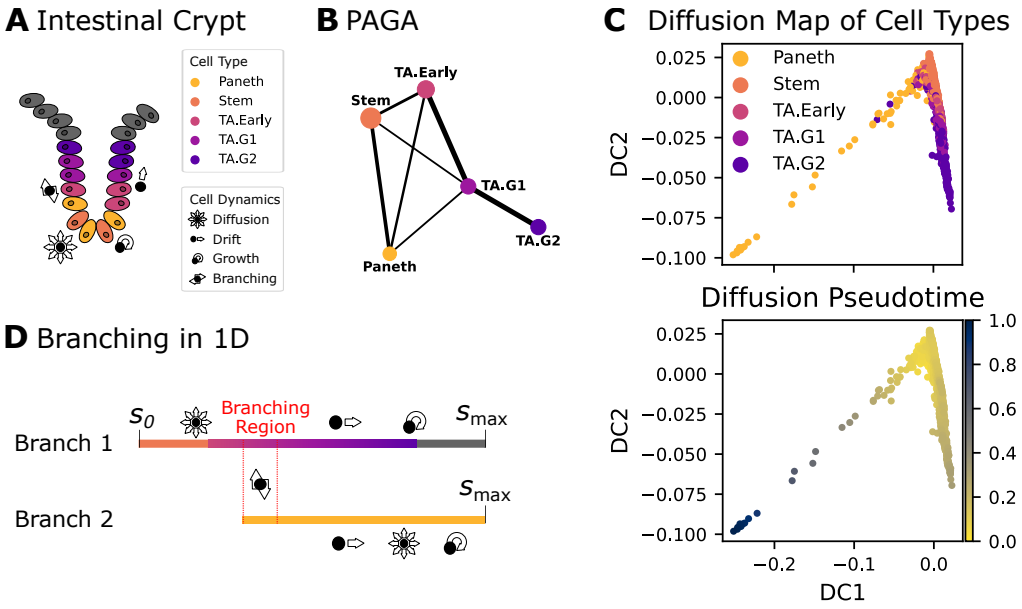


Figure 2: **Analysis of cells of the intestinal crypt.** (A) Visualization of intestinal crypt. Cell types highlighted by color. (B–C) Analysis of cells measured in the intestinal crypt using (B) PAGA and (C) diffusion map and diffusion pseudotime (D) Schematic of branching implementation for one-dimensional cell state space. Branching region highlighted in red.

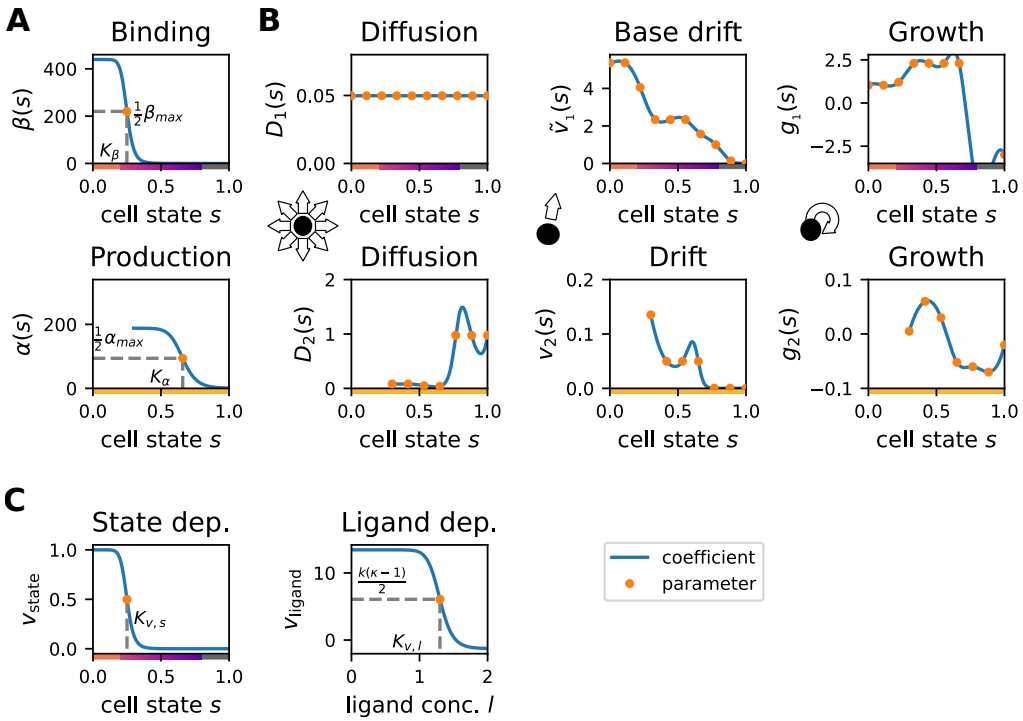


Figure 3: **Parameters chosen for steady state of intestinal crypt cells.** (A) Ligand coefficients binding and production rate. (B) Basic splines for cell dynamics parameters diffusion, drift, and growth on main branch (first row) and side branch (second row). (C) Dependency terms for ligand effect on drift.

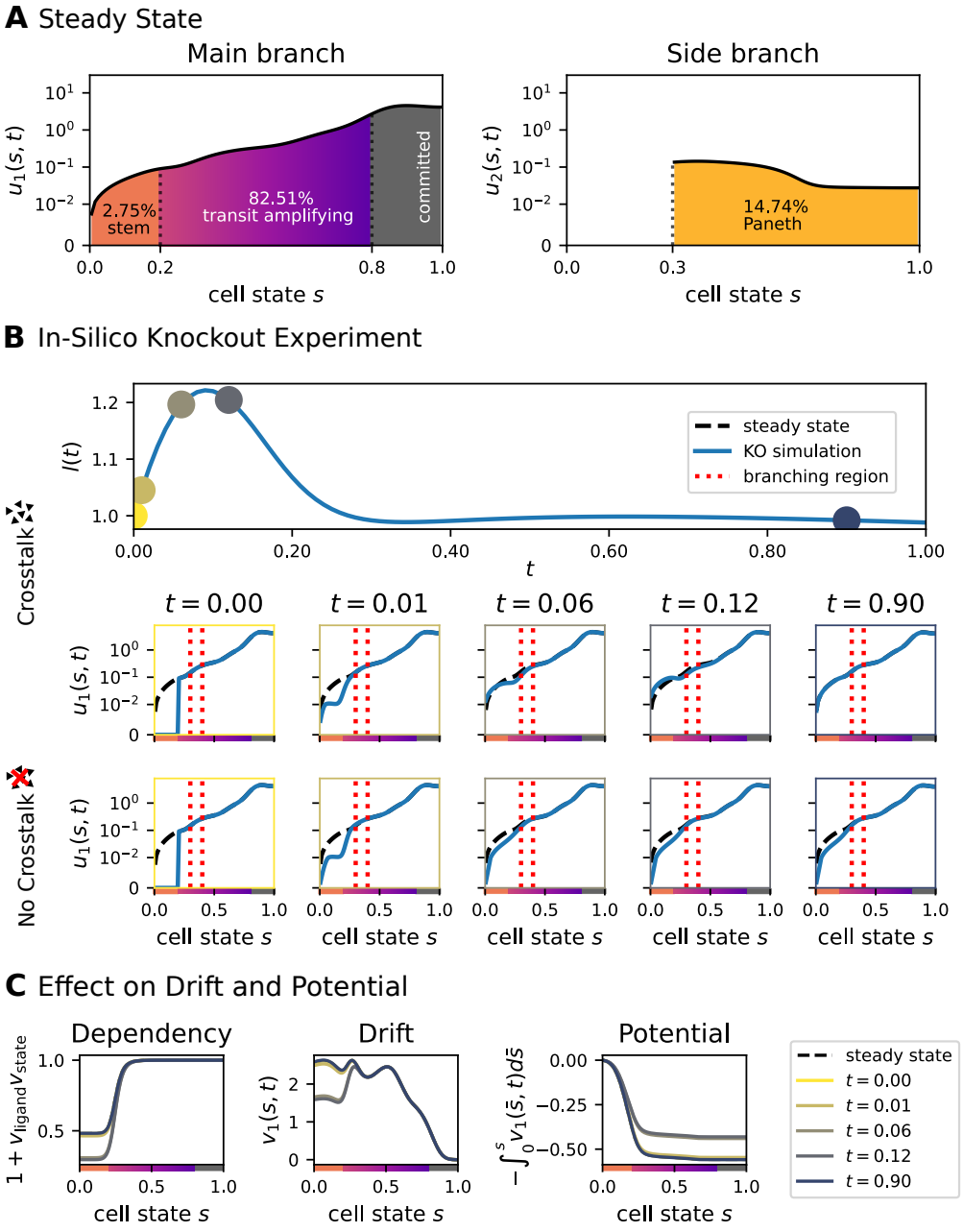


Figure 4: **Recovery of steady state after in-silico knockout of stem cells.** (A) Steady state on main branch and side branch. (B) Simulation result after knockout of stem cells. First row depicts result with ligand feedback, second row depicts results without feedback, and third row depicts ligand concentrations over time in feedback scenario. Dashed lines indicate a steady state before stem cell removal. (A–B) y-axis of cell densities provided in symlog scale with linearity threshold 10^{-2} (C) Product of the drift dependency factors and drift for multiple values of l.

Shalek et al. [22] investigated this process under two conditions: (a) an *in-tube* setup in which cells can communicate and (b) an *on-chip* setup in which cells are unable to communicate (Figure 5A–B). Here, we assess the impact of cell-to-cell communication (and its absence) on cell differentiation by inferring drift and diffusion rates. Therefore, we formulate a model and infer its parameters from the available experimental data.

Biological Background. LPS is a large molecule found in Gram-negative bacteria's outer membrane, which induces activation of various immune cells through ligation of Toll-like receptor 4 and CD14. Dendritic cells' response depends among other molecules on interferon-beta (IFN-β) [22]. Indeed, after the initial exposure, a subset of dendritic cells is activated and then communicates with the remaining cells, prompting them to change their cell state.

Model formulation. To develop a model for dendritic cell activation by LPS, we analyze the published single-cell RNA sequencing data by Shalek et al. [22]. Integrating the *on-chip* and *in-tube* data using Scanorama [58] within the scanpy framework revealed distinct clustering patterns. Measurements of *on-chip* cells 4 hours after LPS stimulation clustered closely with *in-tube* cells taken before and 1 and 2 hours post-stimulation, forming a cluster interpreted as inactivated cells. Conversely, *in-tube* cells measured at 4 and 6 hours post-stimulation formed a separate cluster, likely representing activated dendritic cells (Figure 5C).

To capture the variability, we introduce a one-dimensional cell-state, $s \in \Omega = [0, 1]$, using the trajectories inferred by diffusion pseudotime (Figure 5C), with inactivated cells located at low values of s and activated cells located at high values of s, yielding the model

$$\begin{aligned}\frac{\partial}{\partial t}u(s,t) &= \frac{\partial}{\partial s}\left(D(s,l)\frac{\partial}{\partial s}u(s,t)\right) - \frac{\partial}{\partial s}\left(v(s,l)u(s,t)\right)\\ \frac{dl(t)}{dt} &= \int_0^1 \alpha(s)u(s,t)ds - \left(\int_0^1 \beta(s)u(s,t)ds\right)l(t) - \gamma l(t),\end{aligned}$$

with non-flux boundary conditions and initial conditions $u(s, 0)$ and $l(0)$. Activated cells are assumed to influence inactive cells via a ligand, yielding

$$\alpha(s) = \alpha_{\max}\frac{s^{n_\alpha}}{K_\alpha^{n_\alpha} + s^{n_\alpha}} \qquad \text{and} \qquad \beta(s) = \beta_{\max}\frac{K_\beta^{n_\beta}}{K_\beta^{n_\beta} + s^{n_\beta}}$$

with inflection points $K_{v,s}$, K_α and K_β, Hill coefficients n_α and n_β, and maximal values $\alpha_{\max}$ and $\beta_{\max}$. The functional form of α and β is chosen in a way to reflect that a subset of early activated cells communicates to the inactivated cells, i.e., cells of higher cell state s express IFN-β and cells with lower cell state s bind it. Drift and diffusion are assumed to depend on the cell state and the ligand concentration. The baseline drift and diffusion rates are unknown and modeled using exponentials of cubic splines, $\tilde{v}(s)$ and $\tilde{D}(s)$. For the ligand dependence, we assume a functional form that captures the observation that cell activation,

i.e., drift and diffusion rates, at low cell states increase with the ligand concentration, yielding

$$v(s,l) = \tilde{v}(s) + v_{\max} \underbrace{\frac{K_{v,s}^{n_{v,s}}}{K_{v,l}^{n_{v,s}} + s^{n_{v,s}}}}_{=:v_{\text{state}}} \underbrace{\frac{l^{n_{v,l}}}{K_{v,l}^{n_{v,l}} + l^{n_{v,l}}}}_{=:v_{\text{ligand}}},$$

$$D(s,l) = \tilde{D}(s) + D_{\max} \underbrace{\frac{K_{D,s}^{n_{D,s}}}{K_{Ds}^{n_{D,s}} + s^{n_{D,s}}}}_{=:D_{\text{state}}} \underbrace{\frac{l^{n_{D,l}}}{K_{D,l}^{n_{D,l}} + l^{n_{D,l}}}}_{=:D_{\text{ligand}}},$$

with inflection points $K_{v,s}$, $K_{v,l}$, $K_{D,s}$, and $K_{D,l}$, Hill coefficients $n_{v,s}$, $n_{v,l}$, $n_{D,s}$, and $n_{D,l}$, and maximum effect size $v_{\max}$ and $D_{\max}$. We assume that there is an effect on diffusion and drift for all cells to which a ligand binds, i.e., $K_{D,s} = K_{v,s} = K_\beta$ and $n_{D,s} = n_{v,s} = n_\beta$. Cell proliferation and death are disregarded due to the short duration of the experiment in comparison to the lifespan of dendritic cells [59], yielding $g(s,t) = 0$. The baseline drift and diffusion coefficients are again described using natural cubic splines (ncs):

$$\tilde{v}(s) = 10^{\text{ncs}(s;\hat{v}_0,\ldots,\hat{v}_{n_i-1})},$$
$$\tilde{D}(s) = 10^{\text{ncs}(s;\hat{D}_0,\ldots,\hat{D}_{n_i-1})},$$

with 10 equally spaced grid points. For increased numeric stability we scaled the data to $[0, 0.9]$ and set the drift continuously to 0 at $s = 1$ with a cubic Hermite spline on $[0.9, 1]$.

The formulated model allows for the communication of cells via the ligand, thereby capturing the *in-tube* setup. To model the *on-chip* setup, we set $D_{\max} = v_{\max} = 0$, which implies the lack of any communication effect. For both setups, the initial population densities $u(s,0)$ are set to the kernel density estimates of the cell states obtained from the respective single-cell data, and the initial ligand concentration $l(0)$ is set to zero. The latter is plausible, as in the experiments, the medium is replaced before the start of the experiment.

Calibration. To infer the drift, diffusion, and growth rates from the observed data, we employ the parameter estimation procedure outlined in Section 4.2. We perform 386 local optimization runs. The Supplementary Table S2 and the Supplementary Figure S2A provide details on the parameter constraints and related properties.

The assessment of the estimation results reveal that the best 10 optimization runs achieve similar objective function values (Supplementary Figure S2B). Furthermore, the parameterized model agrees well with the data (Figure 6). The measured distribution is mainly contained in the confidence interval of the multinomial distribution obtained for the maximum likelihood estimate.

The parameterized model captures the dynamics observed by Shalek et al. [22]. Cells *in tube* remain for $t = 1h$ and $2h$ mostly inactive, corresponding to cell states $s < 0.5$. Only a small subset of cells was activated. The activated cells secret ligand, resulting in a steady increase of its concentration and the activation of most remaining cells at time $t = 4h$ and $6h$. Cells *on chip* show hardly any change over time. While the initial distribution is similar,

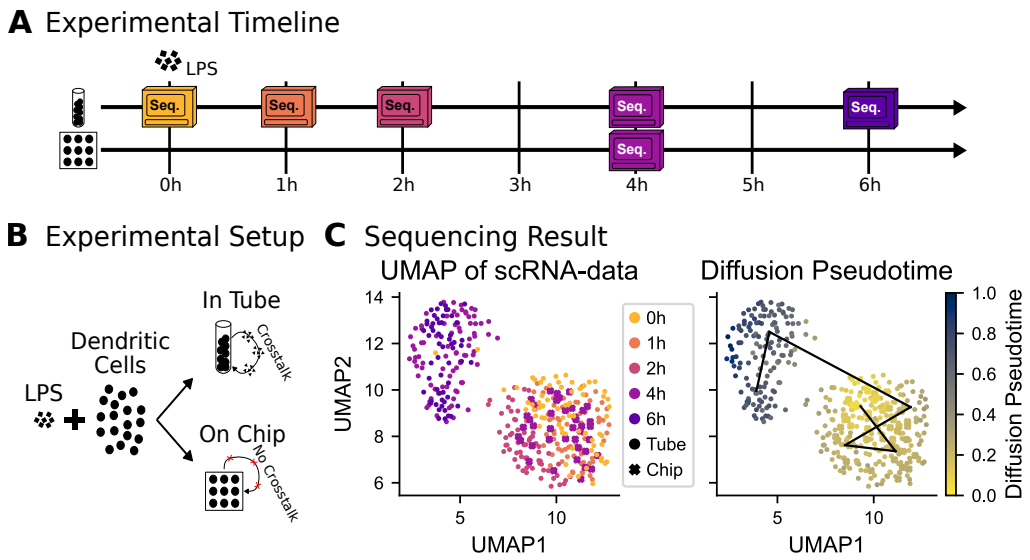


Figure 5: **Immune response data of dendritic cells to LPS-stimulation.** (A) Timeline and (B) experimental setup of data used from [22]. (C) Analysis of cells measured in the intestinal crypt using UMAP and diffusion pseudotime.

the absence of cell-to-cell communications limits the activation of the cell population by the few early responding cells. This is in agreement with the results of the UMAP visualization, in which measurements for cells *on chip* at $4h$ blended in with the measurements for cells *in tube* at $1h$ and $2h$, while measurements for cells *in tube* at $4h$ and $6h$ formed a separate cluster.

The assessment of the model parameters reveals that ligand production is estimated to occur only in fully activated cells. At the same time, binding plays a role in the lower half of the state space (Figure 7A). Additionally, we can see that in the absence of ligand, drift, and diffusion are zero besides for the intervals $s = [0, 0.2]$ and $s = [0.5, 0.6]$ (Figure 7B–C). Moreover, diffusion has a third, smaller peak in $s = [0.8, 0.9]$, which might serve as a stabilizing factor. Compared to the baseline drift $\tilde{v}$, the maximum ligand effect size on the drift $v_{\max}$ seems relatively small, and hence over time, an increase in drift is barely visible (Figure 7B). In contrast, baseline diffusion $\tilde{D}$ is in the same order of magnitude as $D_{\max}$ and over time, an increase of the diffusion in the affected state space regions is clearly visible (Figure 7C). In the context of Waddington's landscape, this suggests that instead of altering the landscape itself, increasing ligand concentration enhances the random movement of cells on the initial plateau, making it more likely for them to cross the edge and descend into lower potential states (c.f. time invariance of developmental potential in the third row of Figure 7B).

In summary, the analysis indicates that the framework allows for integrating different datasets and estimating condition-specific models. Furthermore, model calibration provides estimates for rates, enabling an in-depth analysis of the process dynamics. Importantly, the statistical framework allows for a coherent assessment of the model fit.

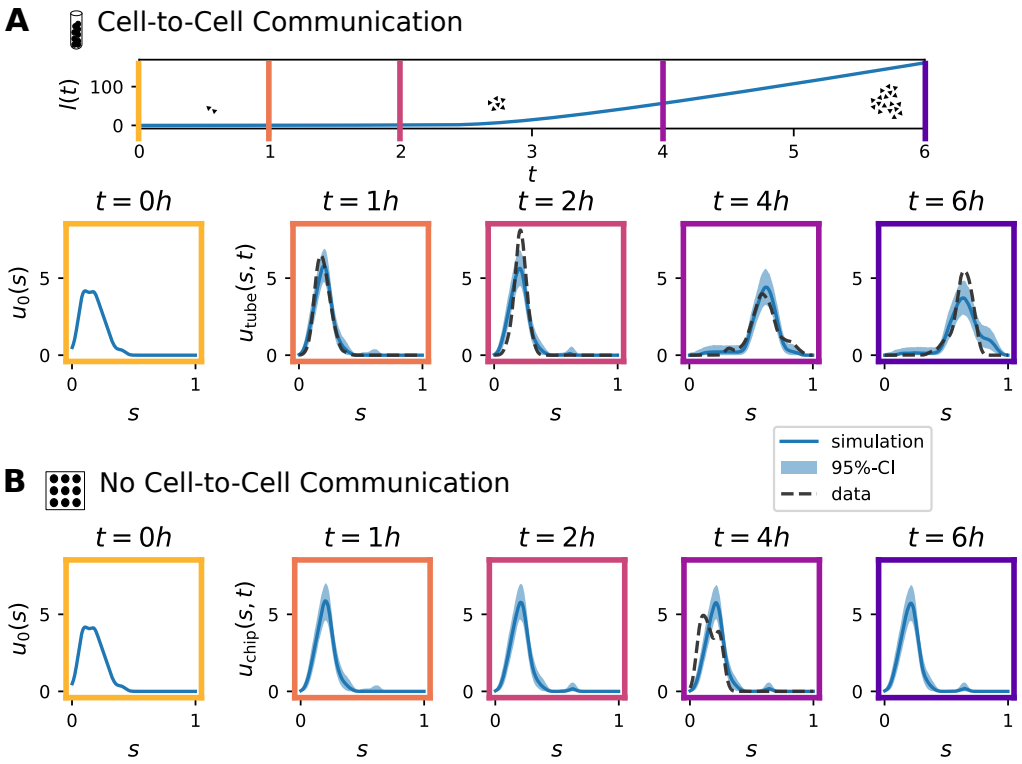


Figure 6: **Fitting results of dendritic cells after LPS-stimulation.** Fitting results of model to data. First row depicts ligand concentrations over time in feedback scenario, second row depicts results with feedback, and third row depicts results without ligand feedback. Dashed lines indicate measured data and measurement time points highlighted by color.

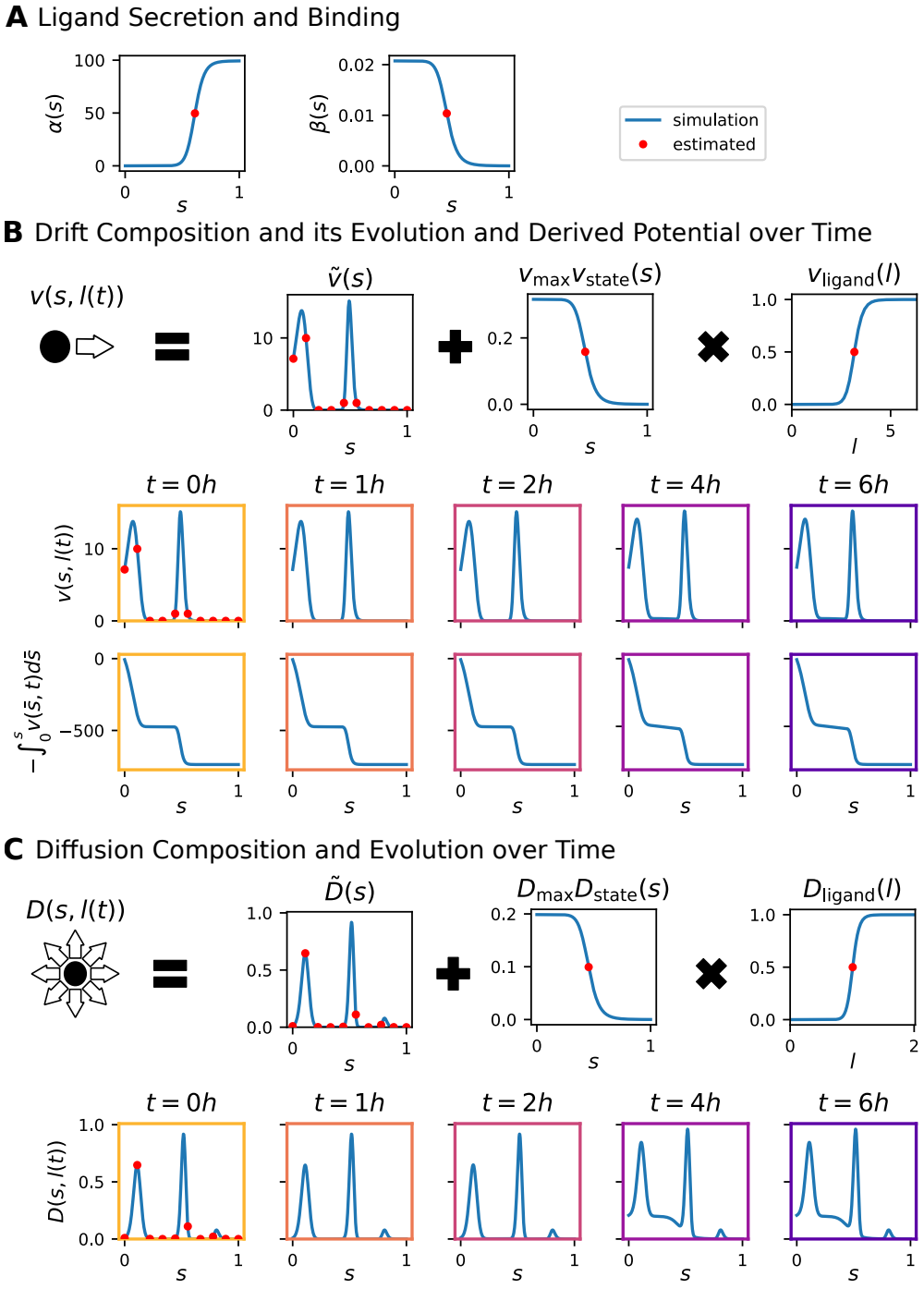


Figure 7: **Dynamic coefficients estimated from LPS-stimulated dendritic cells.** (A) Estimated ligand secretion α and binding β. Estimated Hill coefficients (x-value) and half-maximums (y-value) highlighted in red. (B–C) Estimated drift v and potential derived from it (B) and diffusion D (C) as composition of baseline, state- and ligand-dependency terms and their evolution over time. Estimated spline knot points and inflection points in red.

6. Discussion

Cell-to-cell communication is an essential biological process that modulates the dynamics and equilibria of cell populations. Here, we propose an extension of the established "pseudodynamics" framework by Fischer et al. [12], which offers significant advancements in understanding cell population dynamics. The proposed PDE-ODE system captures the temporal evolution of developmental processes by incorporating cell state distributions and ligand concentrations.

The flexibility of our model is demonstrated by its applicability to various biological scenarios, from tissue regeneration to immune cell activation. We achieved precise simulations that align closely with experimental data by parameterizing the state-dependent coefficients and employing advanced numerical methods. Using finite volume methods and tools like AMICI and FiPy facilitated robust numerical implementation, ensuring the model's reliability across different applications. The successful fitting of both whole activation processes and specific marker gene expressions further highlights the model's versatility.

Our in-silico knockout experiment of the intestinal crypt highlighted the critical role of cell-to-cell communication in tissue regeneration. The model demonstrates that the feedback mechanisms, particularly the interaction between Paneth cells and stem cells via Wnt3 signaling, are essential for the crypt's recovery. The ability of early TA cells to dedifferentiate in response to increased Wnt3 levels underscores the robustness of the crypt's regenerative processes. The absence of such feedback results in incomplete recovery, aligning with empirical observations that by inhibiting Wnt signaling, one compromises crypt integrity. Future improvements for similar applications in a model calibration scenario could be achieved by further investigating diffusion's dominant role in the recovery process. However, as a proof of concept simulation study, it validates our model's potential to simulate complex biological phenomena.

The application of our model to LPS-stimulated dendritic cells, using single-cell transcriptomics data from Shalek et al., further underscores the importance of intercellular communication. Our model effectively captures the distinct responses observed between cells cultured in tubes (with communication) and on-chip (without communication). The ligand concentration dynamics provides insights into the activation process, where the presence of communication allows for the proper activation of dendritic cells. Furthermore, we are able to fit the model to a representation of the whole cell activation through diffusion pseudotime, as well as to the markers for the core antiviral response. This suggests that ligand-mediated signaling is a critical component of the immune response, facilitating a coordinated and efficient activation of immune cells.

Our findings emphasize the necessity of incorporating cell-to-cell communication into mathematical models of biological development. The accuracy in simulating cell differentiation and activation dynamics offers valuable insights for developmental biology and immunology. Future work should focus on extending the model to include additional signaling pathways and interactions, potentially involving multiple ligands and receptor types.

A limitation of the proposed approach is that the available data might be too scarce to

identify the ligand responsible for signaling. Since ligand concentrations are usually not measured directly, a well-informed parameter estimation needs data on a communication scenario and a scenario with inhibited communication (through knockout or physically via chips). Moreover, to fully use the potential of our model and investigate temporal dynamics and not only steady states, snapshot data of at least two time points are required. This broad experimental setup is still uncommon in already published data sets.

In conclusion, integrating cell-to-cell communication into cell population models represents a significant step forward in capturing the dynamic nature of biological processes. Our extended PDE-ODE model provides a powerful tool for exploring the complexities of cell development, offering a deeper understanding of the mechanisms underlying tissue regeneration and immune activation.

CRediT authorship contribution statement

S.M.: Conceptualization, Formulation of mathematical model, Formal analysis of existence and uniqueness, Implementation of numerical simulation, Implementation of parameter estimation, Preparation of figures, Writing – original draft (Introduction, Discussion, Model Fitting), Writing – review and editing; **L.F.**: Formulation of mathematical model, Formal analysis of existence and uniqueness, Implementation of numerical simulation, Writing – original draft (Proof, Simulation Study), Writing – review and editing; **E.D.**: Formulation of mathematical model, Writing – review and editing; **A.S.**: Evaluation of application examples, Writing – review and editing; **B.N.**: Formal analysis of existence and uniqueness, Writing – review and editing; **J.H.**: Conceptualization, Formulation of mathematical model, Formal analysis of existence and uniqueness, Writing – review and editing, Funding acquisition

Declaration of competing interest

The authors declared no competing interests.

Funding Acknowledgements

This research was funded by the German Federal Ministry of Education and Research (INSIDe, 031L0297A; GENImmune, 031L0292F); Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (project IDs 390685813–EXC 2047 and 390873048–EXC 2151) and the University of Bonn via the Schlegel professorship to J.H.

Declaration of generative AI and AI-assisted technologies in the writing process

While preparing this work, the authors used ChatGPT 4o and Grammarly to improve the readability and language of the already drafted manuscript. After using this tool, the authors

reviewed and edited the content as needed and take full responsibility for the content of the published article.

Supplementary Figures

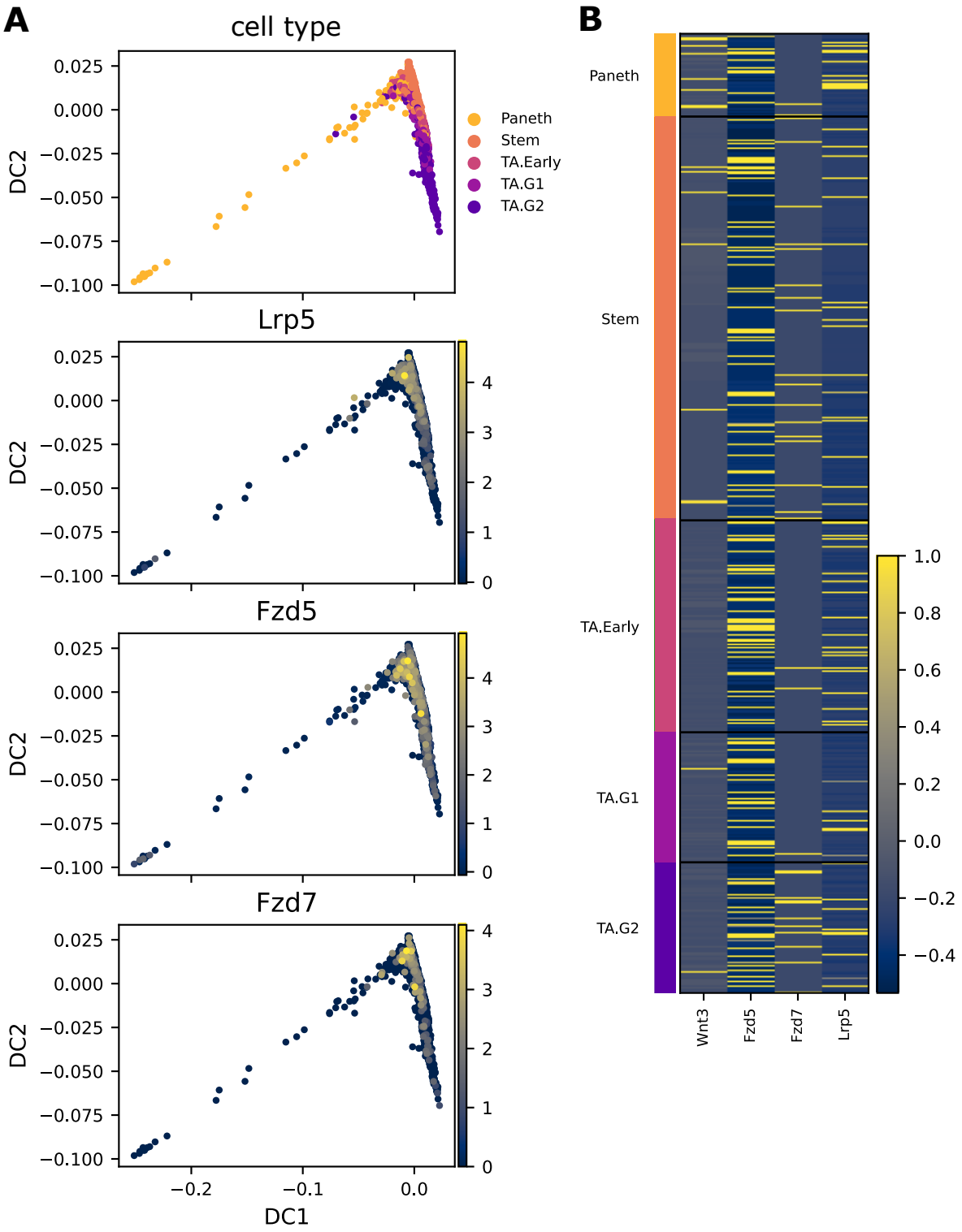


Figure S1: **Distributions of receptor and Wnt3 expressions across cell types. Data from [21]** (A) Diffusion pseudotime ordering with cell cluster annotation and z-scores of Fzd5-, Fzd7- and Lrp5-receptor expressions. (B) Heatmap showing z-scores of Wnt3, Fzd5, Fzd7, and Lrp5 expression across the cell state lineage.

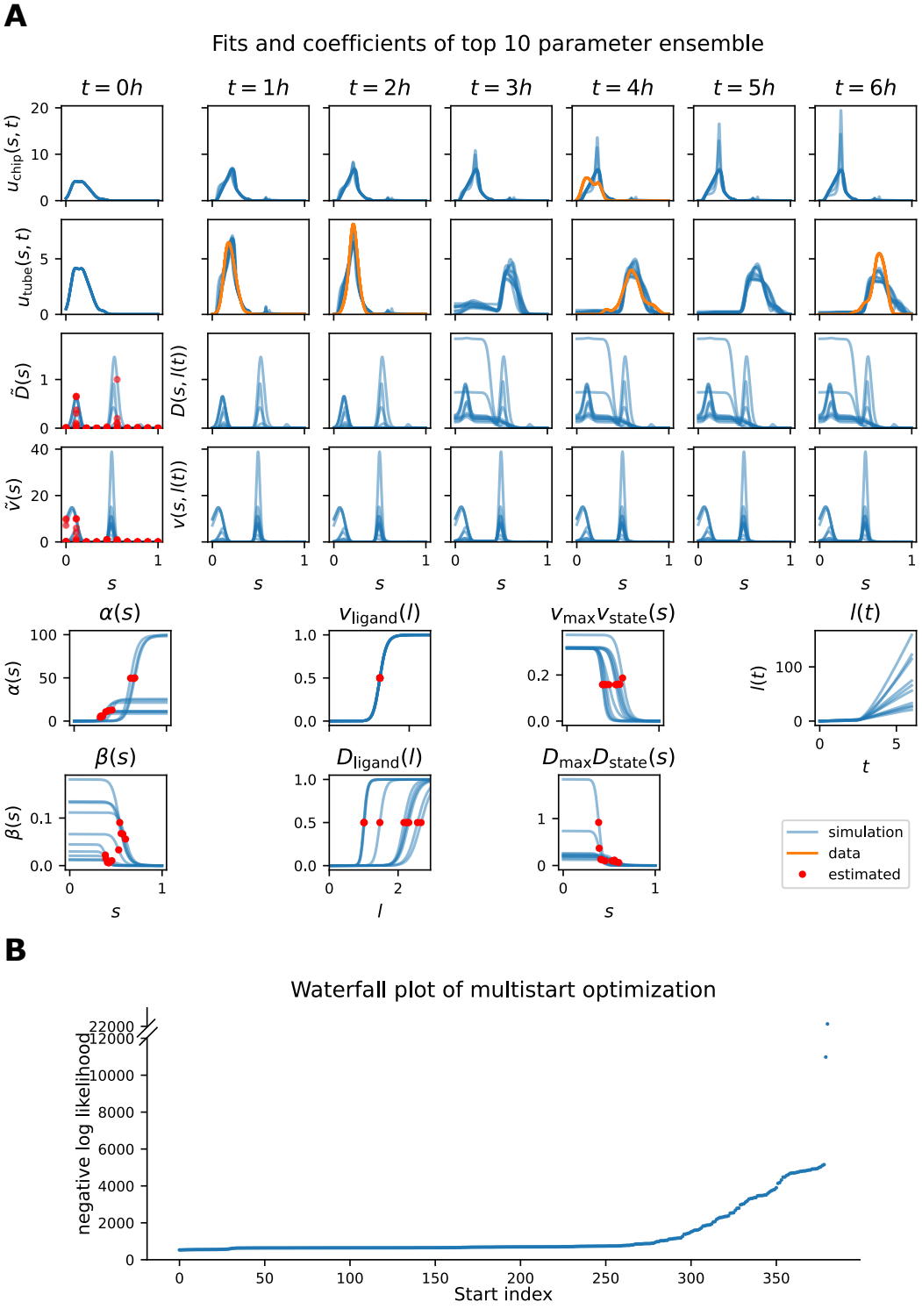


Figure S2: **Results of multistart optimization.** (A) Model and data comparison for the cell activation model. The 10 best parameter fits (out of 384 starts) are depicted in semi-transparent blue. The data are depicted in orange. (B) Waterfall plot for the cell activation model. Sorted objective function values of all 384 parameter estimation starts. Each start represented by blue point.

Supplementary Tables

Table S1: **Parameters used for stem cell recovery model described in Section 5.1.**

Cell Dynamics					
Main Branch				Side Branch	
$\hat{D}_{1,0} = -3$	$\hat{v}_{1,0} = 1.68$	$\hat{g}_{1,0} = 1.03$	$\hat{D}_{2,0} = -2.5$	$\hat{v}_{2,0} = -2$	$\hat{g}_{2,0} = 0.005$
$\hat{D}_{1,1} = -3$	$\hat{v}_{1,1} = 1.68$	$\hat{g}_{1,1} = 1.03$	$\hat{D}_{2,1} = -2.5$	$\hat{v}_{2,1} = -3$	$\hat{g}_{2,1} = 0.06$
$\hat{D}_{1,2} = -3$	$\hat{v}_{1,2} = 1.4$	$\hat{g}_{1,2} = 1.2$	$\hat{D}_{2,2} = -3$	$\hat{v}_{2,2} = -3$	$\hat{g}_{2,2} = 0.03$
$\hat{D}_{1,3} = -3$	$\hat{v}_{1,3} = 0.85$	$\hat{g}_{1,3} = 2.3$	$\hat{D}_{2,3} = -3.25$	$\hat{v}_{2,3} = -3$	$\hat{g}_{2,3} = -0.052$
$\hat{D}_{1,4} = -3$	$\hat{v}_{1,4} = 0.85$	$\hat{g}_{1,4} = 2.3$	$\hat{D}_{2,4} = -0.025$	$v_{2,4} = -8$	$\hat{g}_{2,4} = -0.06$
$\hat{D}_{1,5} = -3$	$\hat{v}_{1,5} = 0.85$	$\hat{g}_{1,5} = 2.3$	$\hat{D}_{2,5} = -0.025$	$\hat{v}_{2,5} = -8$	$\hat{g}_{2,5} = -0.07$
$\hat{D}_{1,6} = -3$	$\hat{v}_{1,6} = 0.45$	$\hat{g}_{1,6} = 2.3$	$\hat{D}_{2,6} = -0.025$	$v_{2,6} = -8$	$\hat{g}_{2,6} = -0.02$
$\hat{D}_{1,7} = -3$	$\hat{v}_{1,7} = 0$	$\hat{g}_{1,7} = -4$			$g_{1,2} = 0.44$
$\hat{D}_{1,8} = -3$	$\hat{v}_{1,8} = -2$	$\hat{g}_{1,8} = -4$			
$\hat{D}_{1,9} = -3$	$\hat{v}_{1,9} = -8$	$\hat{g}_{1,9} = -3$			
Space and Ligand Dependencies of v					
$K_{v,l} = 1.3$	$n_{v,l} = 14$	$k = 1.3$	$\kappa = \frac{8}{13}$	$K_{v,s} = 0.25$	$n_{v,s} = 10$
Ligand Dynamics					
$K_\alpha = 0.66$	$\alpha_{\max} = 188$	$n_\alpha = 13$	$K_\beta = 0.25$	$\beta_{\max} = 440$	$n_\beta = 10$
$\gamma = 0.05$					

Table S2: **Parameter estimation results of cell activation model described in Section 5.2.** Ensemble results of 10 best runs out of 384 starts provided for each parameter as median, minimum and maximum.

Parameter	Best Estimation	Ensemble (Top 10 of 484) Median (Min; Max)	Bounds
$\hat{D}_0$	-2.04	-3.21 $(-4.50; -1.90)$	$[-9.0, 0]$
$\hat{D}_1$	-0.19	-0.23 $(-0.76; -0.16)$	$[-9.0, 0]$
$\hat{D}_2$	-2.89	-2.94 $(-3.11; -2.85)$	$[-9.0, 0]$
$\hat{D}_3$	-6.99	-6.99 $(-7.39; -6.01)$	$[-11.0, 0]$
$\hat{D}_4$	-2.38	-4.00 $(-6.51; -1.53)$	$[-9.0, 0]$
$\hat{D}_5$	-0.96	-0.83 $(-4.92; 0.00)$	$[-9.0, 0]$
$\hat{D}_6$	-8.16	-3.59 $(-8.16; -2.02)$	$[-9.0, 0]$
$\hat{D}_7$	-1.69	-2.10 $(-3.01; -1.69)$	$[-9.0, 0]$
$\hat{D}_8$	-3.18	-2.89 $(-3.18; -2.61)$	$[-9.0, 0]$
$\hat{D}_9$	-5.14	-6.04 $(-7.69; -4.99)$	$[-9.0, 0]$
$\hat{v}_0$	0.85	-0.19 $(-1.13; 1.00)$	$[-5.0, 1]$
$\hat{v}_1$	1.00	0.96 $(0.44; 1.00)$	$[-5.0, 1]$
$\hat{v}_2$	-2.44	-2.43 $(-2.56; -2.34)$	$[-10.0, 0]$
$\hat{v}_3$	-8.00	-7.62 $(-8.00; -5.73)$	$[-8.0, 0]$
$\hat{v}_4$	-0.00	-0.02 $(-0.07; 0.00)$	$[-8.0, 0]$
$\hat{v}_5$	-0.00	-0.56 $(-2.03; -0.00)$	$[-8.0, 0]$
$\hat{v}_6$	-3.26	-3.46 $(-6.13; -2.34)$	$[-8.0, 0]$
$\hat{v}_7$	-6.48	-4.43 $(-7.94; -1.06)$	$[-8.0, 0]$
$\hat{v}_8$	-3.31	-3.67 $(-5.12; -2.84)$	$[-8.0, 0]$
$\hat{v}_9$	-3.84	-4.10 $(-4.95; -3.84)$	$[-8.0, 0]$
$K_{D,l}$	0.00	0.23 $(0.00; 0.36)$	$[0.001, 1]$
$\log_{10} D_{\max}$	-0.70	-0.72 $(-0.82; -0.22)$	$[-2.0, 2]$
$n_{D,l}$	19.94	17.94 $(16.69; 20.00)$	$[10.0, 20]$
$K_{v,l}$	0.50	0.50 $(0.50; 0.50)$	$[0.001, 1]$
$\log_{10} v_{\max}$	-0.50	-0.50 $(-0.52; -0.50)$	$[-2.0, 1]$
$n_{v,l}$	15.08	14.99 $(14.05; 16.30)$	$[10.0, 20]$
K_α	0.61	0.41 $(0.32; 0.66)$	$[0.001, 1]$
$\log_{10} \alpha_{\max}$	2.00	1.65 $(1.11; 2.00)$	$[-2.0, 2]$
n_α	15.20	15.04 $(14.41; 15.73)$	$[10.0, 20]$
K_β	0.46	0.47 $(0.41; 0.54)$	$[0.001, 1]$
$\log_{10} \beta_{\max}$	-1.68	-1.43 $(-1.94; -1.12)$	$[-2.0, 2]$
n_β	10.06	15.20 $(10.06; 18.17)$	$[10.0, 20]$
$\log_{10} \gamma$	-2.01	-1.77 $(-2.25; -1.36)$	$[-3.0, 2]$

Appendix A. Additional Proofs

For ease of notation, we simplify the ligand ODE by defining $f(u(t)) := \int_\Omega \alpha(s,t)u(s,t)ds$ and $h(u(t)) := \int_\Omega \beta(s,t)u(s,t)ds$ and dropping the time dependency of γ. Hence, the ligand

ODE can be rewritten as

$$\frac{dl(t)}{dt} = f(u(t)) - (h(u(t)) + \gamma)l(t). \tag{A.1}$$

Appendix A.1. Proof of existence of a unique solution l^u for fixed u

Lemma 1. *Assume that Condition 1 holds. Then, for a fixed $u \in X$, there is a unique solution l^u for the ODE system* (A.1)*. This solution is continuous in t and bounded on any closed interval.*

Proof: For a fixed $u \in X$, we have that the mappings $f \circ u$ and $h \circ u$ are continuous in t. Therefore, the ligand ODE (A.1) is a linear non-homogenous differential equation of first order and can be solved directly:

$$l^u(t) \equiv l(t) = e^{-\int_0^t h(s)+\gamma ds} \left(l_0 + \int_0^t f(s) e^{\int_s^t h(r)+\gamma dr} ds \right). \tag{A.2}$$

Now, we compute an upper bound for the solution l^u. Assuming $u \geq 0$, $l^u(t)$ can be bounded from above by the initial conditions and a constant. The non-negativity of u implies $h \geq 0$. Since additionally $\gamma \in \mathbb{R}_{>0}$ it follows that $\int_0^t h(\sigma) + \gamma d\sigma \geq 0$. Hence, $\int_r^t h(\sigma) + \gamma d\sigma \leq \int_0^t h(\sigma) + \gamma d\sigma$, for any $r \in [0, t]$. Therewith, we obtain

$$\begin{aligned}
l^u(t) &= e^{-\int_0^t h(\sigma)+\gamma d\sigma} \left(l_0 + \int_0^t f(r) e^{\int_r^t h(\sigma)+\gamma d\sigma} dr \right) \\
&\leq e^{-\int_0^t h(\sigma)+\gamma d\sigma} \left(l_0 + \int_0^t f(r) e^{\int_0^t h(\sigma)+\gamma d\sigma} dr \right) \\
&= e^{-\int_0^t h(\sigma)+\gamma d\sigma} l_0 + e^{-\int_0^t h(\sigma)+\gamma d\sigma} \int_0^t f(r) e^{\int_0^t h(\sigma)+\gamma d\sigma} dr \\
&\leq l_0 + \int_0^t f(r) dr,
\end{aligned}$$

where in the last step we used the fact that $\int_0^t h(r) + \gamma dr \geq 0$ implies $e^{-\int_0^t h(r)+\gamma dr} \leq 1$. Taking the absolute value, we obtain:

$$\begin{aligned}
|l^u(t)| &\leq |l_0| + \int_0^t |f(r)| dr \\
&\leq |l_0| + \int_0^t ||\alpha||_{L^\infty(\Omega)} |\Omega|^{\frac{1}{2}} ||u(r)||_{L^2(\Omega)} dr \\
&\leq |l_0| + ||\alpha||_{L^\infty(\Omega)} |\Omega|^{\frac{1}{2}} \int_0^T ||u(r)||_{L^2(\Omega)} dr.
\end{aligned}$$

Since $u \in X$, we obtain for all $t \in [0, T]$:

$$|l^u(t)| \leq |l_0| + ||\alpha||_{L^\infty(\Omega)} T |\Omega|^{\frac{1}{2}} C_X ||u_0||_{L^2(\Omega)}. \tag{A.3}$$

Additionally, it is clear from the analytic form of the solution that if $l_0 \in \mathbb{R}_{\geq 0}$ and $u \geq 0$, then $l^u(t) \geq 0$ for all $t \in [0,T]$.

$\square_{\text{Lemma 1}}$

Appendix A.2. Proof of existence of a weak solution $\hat{u}$ for fixed l^u

Lemma 2. *Assume that Condition 2 holds. For fixed $l^u \in C([0,T])$, the PDE (1)-(3) possesses a unique weak solution $\hat{u} \in L^2([0,T], H^1(\Omega)))$ with $\hat{u}' \in L^2([0,T], H^{-1}(\Omega))$. Moreover, $\hat{u} \in C([0,T], L^2(\Omega))$ and the following estimate holds for $\hat{u}$:*

$$\max_{0\leq t\leq T} ||\hat{u}(t)||_{L^2(\Omega)} + ||\hat{u}||_{L^2([0,T],H^1(\Omega))} + ||\hat{u}'||_{L^2([0,T],H^{-1}(\Omega))} \leq C||u_0||_{L^2(\Omega)},$$

with $C = C(T, \Omega, D, v, g)$. Additionally, if $\hat{u}(s,0) = u_0(s) \geq 0$, then $\hat{u}(s,t) \geq 0$ for all $t \in [0,T]$, i.e. non-negativity is preserved.

Proof: In the following, we are applying the results from Chapter 7 in [29]. Note that Ω is an open, bounded domain with piecewise C^1-boundary.

From Condition 2 (i) and since $l^u(t)$ is continuous and bounded in $t \in [0,T]$ (Lemma 1), it follows that $D(s, l^u(t), t),\ v(s, l^u(t), t),\ g(s, l^u(t), t) \in L^\infty(\Omega_T)$.

Suppose that $u \in H^2(\Omega)$ is a solution. Then, by multiplying with a test function $\phi \in H^1(\Omega)$, integrating over Ω, integration by parts, and the application of the boundary conditions, yields that for all $\phi \in H^1(\Omega)$ and for a.e. $t \in (0,T]$:

$$\begin{aligned}\int_\Omega \partial_t u(s,t)\phi(s,t)ds = &-\int_\Omega \left(D(s,l^u(t),t)\partial_s u(s,t) - v(s,l^u(t),t)u(s,t)\right)\partial_s\phi(s,t)ds \\ &+\int_\Omega g(s,l^u(t),t)u(s,t)\phi(s,t)ds,\end{aligned} \tag{A.4}$$

which is the weak formulation of the problem.

Chapter 7 of [29] gives an existence proof for a weak solution. Below, we briefly sketch their argument.

Since $H^1(\Omega)$ is compactly embedded in $L^2(\Omega)$, a common orthogonal basis exists. With this orthogonal basis, approximate solutions u_m are constructed that lie in the finite-dimensional subspaces generated by the first m basis functions, and solve the weak formulation with respect to test functions from those m-dimensional subspaces. Using energy estimates, one can show that the sequence $\{u_m\}$ is bounded in $L^2([0,T], H^1(\Omega))$ and that $\{u_m'\}$ is bounded in $L^2([0,T], H^{-1}(\Omega))$. By the Banach-Alaoglu Theorem, there exist subsequences converging weakly to some $\hat{u}$ in $L^2([0,T], H^1(\Omega))$ and $L^2([0,T], H^{-1}(\Omega))$, respectively. Making use of the L^2-weak convergence, it follows that $\hat{u}$ solves the weak formulation (A.4). As a consequence of Theorem 3 (Chapter 5 in [29]), $\hat{u} \in C([0,T], L^2(\Omega))$. The uniqueness follows directly as in Theorem 4 from Gronwall's inequality (Chapter 7 in [29]). To sum up:

If Condition 2 is satisfied, the PDE (1)-(3) possesses a unique weak solution $\hat{u} \in L^2([0,T], H^1(\Omega)))$ with $\hat{u}' \in L^2([0,T], H^{-1}(\Omega))$. Moreover, $\hat{u} \in C([0,T], L^2(\Omega))$.

By the energy estimate, $\hat{u}$ satisfies the following estimate:

$$\max_{0\leq t\leq T} ||\hat{u}(t)||_{L^2(\Omega)} + ||\hat{u}||_{L^2([0,T],H^1(\Omega))} + ||\hat{u}'||_{L^2([0,T],H^{-1}(\Omega))} \leq C||u_0||_{L^2(\Omega)}, \tag{A.5}$$

with $C = C(\Omega, D, v, g)\exp(C(\Omega, D, v, g)T)$.

In the definition of X, we notated this constant by C_X. It remains only to show that the solution $\hat{u}$ preserves non-negativity for a non-negative initial distribution to obtain that $\hat{u} \in X$.

To this end, we follow the idea of the proof of Theorem 1 in [60] and Chapter 3 in [28]. Consider $w := e^{\lambda t}\hat{u}$, where λ is chosen later. If $\hat{u}$ satisfies equation (A.4), then it must hold for w that:

$$\begin{aligned}\int_\Omega \partial_t w(s,t)\phi(s,t)ds = &-\int_\Omega \left(D(s,t)\partial_s w(s,t) - v(s,t)w(s,t)\right)\partial_s\phi(s,t)ds \\ &+ \int_\Omega \left(g(s,t)+\lambda\right)w(s,t)\phi(s,t)ds,\end{aligned} \tag{A.6}$$

for all $\phi \in H^1(\Omega)$ and a.e. $t \in [0,T]$. Since $u(s,0) \geq 0$ a.e. in Ω, we also have $w(s,0) \geq 0$ a.e. in Ω. We will omit the arguments s and t to facilitate notation in the following. We denote the positive and negative parts of w by w^+ and w^- respectively, satisfying $w = w^+ + w^-$, $w^+ \geq 0$ and $w^- \leq 0$.

To apply the theory of Sobolev functions, we need Stampacchia's lemma, which is proven in [61]:

Lemma 3 (Stampacchia's lemma)**.** *Let Ω bounded, $w \in W^{1,p}(\Omega)$, $1 \leq p \leq \infty$. Then $w^+, w^- \in W^{1,p}(\Omega)$ and*

$$\partial_s w^+(s) = \begin{cases}\partial_s w(s), & \text{if } w(s) > 0\\ 0, & \text{else}\end{cases}$$
$$\partial_s w^-(s) = \begin{cases}\partial_s w(s), & \text{if } w(s) < 0\\ 0, & \text{else.}\end{cases}$$

The Lemma also implies that $w^+w^- = 0$, $\partial_s w \partial_s w^- = (\partial_s w^-)^2$ and $ww^- = (w^-)^2$ a.e. in Ω. Setting $\phi = w^-$ in (A.6), we obtain with Lemma 3 for a.e. $t \in [0,T]$:

$$\begin{aligned}&\int_\Omega (\partial_t w)\, w^- ds + \int_\Omega D\partial_s w\partial_s w^- ds - \int_\Omega vw\partial_s w^- ds - \int_\Omega (g+\lambda)\, ww^- ds = 0\\ \Rightarrow &\int_\Omega (\partial_t w)\, w^- ds + \int_\Omega D(\partial_s w^-)^2 ds - \int_\Omega vw\partial_s w^- ds - \int_\Omega (g+\lambda)\,(w^-)^2 ds = 0.\end{aligned}$$

For the first summand on the left side of the equation, we can now, for a.e. $t \in [0,T]$, compute:

$$\int_\Omega (\partial_t w)\, w^- ds = \int_{w>0} (\partial_t w) \underbrace{w^-}_{=0} ds + \int_{w\leq 0} \partial_t (\underbrace{w^+}_{=0} + w^-) w^- ds = \frac{1}{2}\int_\Omega \partial_t (w^-)^2 ds,$$

and we obtain

$$\frac{1}{2}\int_\Omega \partial_t (w^-)^2 ds + \int_\Omega D(\partial_s w^-)^2 ds - \int_\Omega v w \partial_s w^- ds - \int_\Omega (g+\lambda)\,(w^-)^2 ds = 0. \tag{A.7}$$

By Young's inequality, for every $\epsilon > 0$ and a.e. $t \in [0,T]$:

$$\begin{aligned} v(s,t)|w\partial_s w^-| &\leq |\,||v||_{L^\infty(\Omega_T)} w \partial_s w^-| = |\,||v||_{L^\infty(\Omega_T)} w^- \partial_s w^-| \\ &\leq ||v||_{L^\infty(\Omega_T)} \left(\epsilon |\partial_s w^-|^2 + \frac{1}{4\epsilon}|w^-|^2\right). \end{aligned}$$

Using that for the diffusion term, we have $D \geq \theta > 0$, we get from equation (A.7) that

$$\frac{1}{2}\partial_t \int_\Omega (w^-)^2 ds + \int_\Omega (\theta - ||v||_{L^\infty(\Omega_T)}\epsilon)(\partial_s w^-)^2 ds + \int_\Omega -\left(g + \lambda + \frac{||v||_{L^\infty(\Omega_T)}}{4\epsilon}\right)(w^-)^2 ds \leq 0.$$

Choose $\epsilon = \frac{\theta}{2||v||_{L^\infty(\Omega_T)}}$ and $\lambda \in \mathbb{R}$ such that $-\left(g + \lambda + \frac{||v||_{L^\infty(\Omega_T)}}{4\epsilon}\right) \geq 0$. Note that there is such a λ since g is bounded by assumption. Then, for a.e. $t \in [0,T]$:

$$\begin{aligned} \frac{1}{2}\partial_t \int_\Omega (w^-)^2 ds + \frac{\theta}{2}\int_\Omega (\partial_s w^-)^2 ds &\leq 0 \\ \Rightarrow \frac{1}{2}\partial_t \int_\Omega (w^-)^2 ds &\leq 0. \end{aligned}$$

Integrating over $[0,t]$ and using the fact that $w(s,0) \geq 0$ a.e. in Ω implies $w^-(s,0) \equiv 0$ a.e. in Ω, we obtain for all $t \in [0,T]$:

$$\int_\Omega (w^-(s,t))^2 ds \leq \int_\Omega (w^-(s,0))^2 ds = 0.$$

This implies that a.e. in Ω:

$$w^-(t) \equiv 0 \ \forall t \geq 0 \Leftrightarrow\ w(t) \geq 0.$$

Since $\hat{u}$ and w have the same sign, it follows $\hat{u} \geq 0$ a.e. in Ω and for all $t \in [0,T]$.

□Lemma 2

Appendix A.3. Proof of Estimate (6)

Lemma 4. *For all $t \in [0,T]$, in particular for $t = T$, there exists a constant $K(T) > 0$ such that*

$$||\hat{u}_1 - \hat{u}_2||_X \leq K(T)||u_1 - u_2||_X^2,$$

with $K(T) = CT(1+T)\exp(TC\exp(CT))$ and $C = C(\Omega, D, v, g)$.

Proof: The proof can be divided into two parts, where we need to find the following two estimations

(a) $|l^{u_1} - l^{u_2}| \leq C(1+T)||u_1 - u_2||_X$

(b) $||\hat{u}_1 - \hat{u}_2||_X \leq \alpha||u_1 - u_2||_X$ for $0 < \alpha < 1$.

(a) Define $\delta l = l^{u_1} - l^{u_2}$ where

$$\begin{aligned}\partial_t l^{u_1} &= f(u_1) - (h(u_1) + \gamma)l^{u_1} \text{ with } l^{u_1}(0) = l_0\\ \partial_t l^{u_2} &= f(u_2) - (h(u_2) + \gamma)l^{u_2} \text{ with } l^{u_2}(0) = l_0.\end{aligned}$$

Note that f and h are Lipschitz continuous in u with constants L_f, L_h. Computing $\partial_t \delta l$, yields

$$\partial_t \delta l = f(u_1) - f(u_2) - (h(u_1) + \gamma)\delta l + (h(u_2) - h(u_1))l^{u_2}.$$

Multiplying with δl and using that f and h are Lipschitz and that $h(u_1) + \gamma \geq 0$:

$$\begin{aligned}&(\partial_t \delta l)\delta l \leq (L_f + L_h\|l^{u_2}|)||\delta u||_{L^2(\Omega}|\delta l|\\ &\Rightarrow \partial_t|\delta l| \leq (L_f + L_h|l^{u_2}|)||\delta u||_{L^2}.\end{aligned}$$

Define $K_1 := ||\alpha||_{L^\infty}||1||_{L^2(\Omega)}C_X$. Taking the square on both sides, integrating over t, and applying Jensen's inequality yields the estimate:

$$\forall t \in [0,T]: \ |\delta l(t)|^2 \leq (L_f + L_h(|l_0| + K_1 T||u_0||_{L^2(\Omega)}))^2||\delta u||_X^2. \tag{A.8}$$

(b) Define $\delta\hat{u} := \hat{u}_1 - \hat{u}_2 = \mathcal{B}(u_1) - \mathcal{B}(u_2)$ where $\hat{u}_1, \hat{u}_2$ are solutions of

$$\begin{aligned}\partial_t \hat{u}_1 &= \partial_s\left(D(s, l^{u_1}, t)\partial_s \hat{u}_1\right) - \partial_s\left(v(s, l^{u_1}, t)\hat{u}_1\right) + g(s, l^{u_1}, t)\hat{u}_1 \text{ with } \hat{u}_1(s,0) = u_0(s),\\ \partial_t \hat{u}_2 &= \partial_s\left(D(s, l^{u_2}, t)\partial_s \hat{u}_2\right) - \partial_s\left(v(s, l^{u_2}, t)\hat{u}_2\right) + g(s, l^{u_2}, t)\hat{u}_2 \text{ with } \hat{u}_2(s,0) = u_0(s).\end{aligned}$$

Computing $\partial_t \delta\hat{u}$ yields:

$$\begin{aligned}\partial_t \delta\hat{u} =&\partial_s\left(D\left(s, l^{u_1}, t\right)\partial_s \delta\hat{u} + \left(D\left(s, l^{u_1}, t\right) - D\left(s, l^{u_2}, t\right)\right)\partial_s \hat{u}_2\right) - \partial_s\left(v\left(s, l^{u_1}, t\right)\delta\hat{u}\right.\\ &\left.+\left(v\left(s, l^{u_1}, t\right) - v\left(s, l^{u_2}, t\right)\right)\hat{u}_2\right) + g\left(s, l^{u_1}, t\right)\delta\hat{u} + \left(g\left(s, l^{u_1}, t\right) - g\left(s, l^{u_2}, t\right)\right)\hat{u}_2.\end{aligned}$$

Multiplying with $\delta\hat{u}$, and integrating over Ω, we obtain

$$\begin{aligned}\int_\Omega \partial_t\left(\delta\hat{u}\right)\delta\hat{u}\ ds = \int_\Omega &\partial_s\left[D\left(s, l^{u_1}, t\right)\partial_s\delta\hat{u} + \left(D\left(s, l^{u_1}, t\right) - D\left(s, l^{u_2}, t\right)\right)\partial_s\hat{u}_2\right.\\ &\left.-v\left(s, l^{u_1}, t\right)\delta\hat{u} - \left(v\left(s, l^{u_1}, t\right) - v\left(s, l^{u_2}, t\right)\right)\hat{u}_2\right]\delta\hat{u}\ ds\\ &+\int_\Omega g\left(s, l^{u_1}, t\right)\left(\delta\hat{u}\right)^2 ds + \int_\Omega\left(g\left(s, l^{u_1}, t\right) - g\left(s, l^{u_2}, t\right)\right)\hat{u}_2\delta\hat{u}\ ds.\end{aligned}$$

Integration by parts, using Conditions 2 (i)+(v), and applying Estimate (A.5), yields:

$$\begin{aligned}&\frac{1}{2}\partial_t||\delta\hat{u}||^2_{L^2(\Omega)} + \int_\Omega D\left(s, l^{u_1}, t\right)\left(\partial_s\delta\hat{u}\right)^2 ds \\ \leq& K_2||u_0||_{L^2(\Omega)}|\delta l|||\partial_s\delta\hat{u}||_{L^2(\Omega)} + ||v||_{L^\infty(\Omega)}||\delta\hat{u}||_{L^2(\Omega)}||\partial_s\delta\hat{u}||_{L^2} \\ &+ ||g||_{L^\infty}||\delta\hat{u}||^2_{L^2} + K_3||u_0||_{L^2(\Omega)}|\delta l|||\delta\hat{u}||_{L^2},\end{aligned} \tag{A.9}$$

with $K_2 := (L_D + L_v)\, C\exp(CT)$ and $K_3 := L_g C\exp(CT)$, where L_D, L_v, L_g are the respective Lipschitz constant of D, v, g with respect to l and $C = C(\Omega, D, v, g)$. Integrating in time and using $D \geq \eta$ as well as Young's inequality, we obtain:

$$||\delta\hat{u}(.,t)||^2_{L^2} + \frac{\eta}{2}\int_0^t\int_\Omega |\partial_s\delta\hat{u}|^2 dsdt \leq C\left(||u_0||^2_{L^2(\Omega)}\int_0^t |\delta l|^2 d\tau + \int_0^t ||\delta\hat{u}||^2_{L^2(\Omega)} d\tau\right),$$

with $C = C(\Omega, D, v, g)\exp(C(\Omega, D, v, g)T)$. Using Equation (A.8), we find

$$\begin{aligned}||\delta\hat{u}(.,t)||^2_{L^2(\Omega)} + \frac{\eta}{2}\int_0^t\int_\Omega |\partial_s\delta\hat{u}|^2 dsdt &\leq C\int_0^T ||\delta\hat{u}||^2_{L^2(\Omega)} d\tau + C||u_0||^2_{L^2(\Omega)}T(1+T)||\delta u||^2_X \\ \Rightarrow ||\delta\hat{u}(.,t)||^2_{L^2(\Omega)} &\leq \tilde{C}||\delta u||^2_X + C\int_0^T ||\delta\hat{u}||^2_{L^2(\Omega)} d\tau.\end{aligned}$$

Then Gronwall's inequality implies

$$||\delta\hat{u}||^2_X \leq CT(1+T)\exp(CT\exp(CT))||\delta u||^2_X,$$

with $C = C(\Omega, D, v, g)$.

□Lemma 4

Appendix A.4. Global solution

Lemma 5. *Given $T > 0$, there exists $C = C(D, v, g)$ such that $||u(.,t)||^2_{L^2(\Omega)} \leq e^{CT}||u_0||^2_{L^2(\Omega)}$.*

Proof: Using the weak formulation Equation (A.4) with u as test function, we apply Condition 2 (iii) and use Young's inequality to obtain

$$\begin{aligned}\frac{d}{dt}\int_\Omega \frac{u^2}{2} ds = \int_\Omega (\partial_t u) u ds &= -\int_\Omega \underbrace{D}_{\geq\eta>0} |\partial_s u|^2 ds + \int_\Omega vu\partial_s u ds + \int_\Omega gu^2 ds \\ &\leq -\eta\int_\Omega |\partial_s u|^2 ds + \int_\Omega vu\partial_s u ds + \int_\Omega gu^2 ds \\ &\leq -\eta\int_\Omega |\partial_s u|^2 ds + \frac{1}{2}\eta\int_\Omega |\partial_s u|^2 ds + \left(C_\eta||v||^2_{L^\infty(\Omega)} + ||g||^2_{L^\infty}\right)\int_\Omega u^2 ds(\Omega) \\ &\leq -\frac{1}{2}\eta\int_\Omega |\partial_s u|^2 ds + \left(C_\eta||v||^2_{L^\infty(\Omega)} + ||g||^2_{L^\infty}\right)\int_\Omega u^2 ds(\Omega) \\ &\leq C(\eta, v, g)\int_\Omega u^2 ds.\end{aligned}$$

Applying the differential form of Gronwall's inequality, we can conclude that $||u(.,t)||^2_{L^2(\Omega)} \leq e^{CT}||u_0||^2_{L^2(\Omega)}$.

□Lemma 5

References

[1] L. Zhu, A. I. Skoultchi, Coordinating cell proliferation and differentiation, Current Opinion in Genetics & Development 11 (1) (2001) 91–97. doi:10.1016/S0959-437X(00)00162-3.

[2] S. Tay, J. J. Hughey, T. K. Lee, T. Lipniacki, S. R. Quake, M. W. Covert, Single-cell nf-$\kappa\beta$ dynamics reveal digital activation and analogue information processing, Nature 466 (7303) (2010) 267–271. doi:10.1038/nature09145.

[3] G. Camussi, M. C. Deregibus, S. Bruno, V. Cantaluppi, L. Biancone, Exosomes/microvesicles as a mechanism of cell-to-cell communication, Kidney International 78 (9) (2010) 838–848. doi:10.1038/ki.2010.278.

[4] M. A. Basson, Signaling in cell differentiation and morphogenesis, Cold Spring Harb Perspect Biol 4 (6) (Jun. 2012). doi:10.1101/cshperspect.a008151.

[5] S. Toda, N. W. Frankel, W. A. Lim, Engineering cell–cell communication networks: programming multicellular behaviors, Current Opinion in Chemical Biology 52 (2019) 31–38. doi:https://doi.org/10.1016/j.cbpa.2019.04.020.

[6] E. Armingol, A. Officer, O. Harismendy, N. E. Lewis, Deciphering cell-cell interactions and communication from gene expression, Nature Reviews Genetics 22 (2) (2021) 71–88. doi:10.1038/s41576-020-00292-x.

[7] C. H. Waddington, The Strategy of the Genes, Vol. George Allen and Unwin, George Allen and Unwin, London, UK, 1957.

[8] J. Ferrell, Bistability, bifurcations, and waddington's epigenetic landscape, Current Biology 22 (11) (2012) R458–R466. doi:10.1016/j.cub.2012.03.045.

[9] M. J. Casey, P. S. Stumpf, B. D. MacArthur, Theory of cell fate, WIREs Systems Biology and Medicine 12 (2) (2020) e1471. doi:10.1002/wsbm.1471.

[10] M. Sáez, J. Briscoe, D. A. Rand, Dynamical landscapes of cell fate decisions, Interface Focus 12 (4) (2022) 20220002. doi:10.1098/rsfs.2022.0002.

[11] M. Sáez, R. Blassberg, E. Camacho-Aguilar, E. D. Siggia, D. A. Rand, J. Briscoe, Statistically derived geometrical landscapes capture principles of decision-making dynamics during cell fate transitions, Cell Systems 13 (1) (2022) 12–28.e3. doi:10.1016/j.cels.2021.08.013.

[12] D. S. Fischer, A. K. Fiedler, E. M. Kernfeld, R. M. J. Genga, A. Bastidas-Ponce, M. Bakhti, H. Lickert, J. Hasenauer, R. Maehr, F. J. Theis, Inferring population dynamics from single-cell rna-sequencing time series data, Nature Biotechnology 37 (4) (2019) 461–468. doi:10.1038/s41587-019-0088-0.

[13] H. Cho, Y.-H. Kuo, R. C. Rockne, Comparison of cell state models derived from single-cell RNA sequencing data: graph versus multi-dimensional space, Math Biosci Eng 19 (8) (2022) 8505–8536. doi:10.3934/mbe.2022395.

[14] J. Shi, K. Aihara, T. Li, L. Chen, Energy landscape decomposition for cell differentiation with proliferation effect, National Science Review 9 (8) (2022-06) nwac116. doi:10.1093/nsr/nwac116.

[15] G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, L. Lee, J. Chen, J. Brumbaugh, P. Rigollet, K. Hochedlinger, R. Jaenisch, A. Regev, E. S. Lander, Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming, Cell 176 (4) (2019) 928–943.e22. doi:10.1016/j.cell.2019.01.006.

[16] Y. Sha, Y. Qiu, P. Zhou, Q. Nie, Reconstructing growth and dynamic trajectories from single-cell transcriptomics data, Nature Machine Intelligence 6 (1) (2024) 25–39. doi:10.1038/s42256-023-00763-w.

[17] D. Dimitrov, D. Türei, M. Garrido-Rodriguez, P. L. Burmedi, J. S. Nagai, C. Boys, R. O. Ramirez Flores, H. Kim, B. Szalai, I. G. Costa, A. Valdeolivas, A. Dugourd, J. Saez-Rodriguez, Comparison of methods and resources for cell-cell communication inference from single-cell rna-seq data, Nature Communications 13 (1) (2022) 3224. doi:10.1038/s41467-022-30755-0.

[18] S. Jin, C. F. Guerrero-Juarez, L. Zhang, I. Chang, R. Ramos, C.-H. Kuan, P. Myung, M. V. Plikus, Q. Nie, Inference and analysis of cell-cell communication using cellchat, Nature Communications 12 (1) (2021) 1088. doi:10.1038/s41467-021-21246-9.

[19] M. Efremova, M. Vento-Tormo, S. A. Teichmann, R. Vento-Tormo, Cellphonedb: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes, Nature Protocols 15 (4) (2020) 1484–1506. doi:10.1038/s41596-020-0292-x.

[20] R. Browaeys, W. Saelens, Y. Saeys, Nichenet: modeling intercellular communication by linking ligands to target genes, Nature Methods 17 (2) (2020) 159–162. doi:10.1038/s41592-019-0667-5.

[21] A. L. Haber, M. Biton, N. Rogel, R. H. Herbst, K. Shekhar, C. Smillie, G. Burgin, T. M. Delorey, M. R. Howitt, Y. Katz, et al., A single-cell survey of the small intestinal epithelium, Nature 551 (7680) (2017) 333–339. doi:10.1038/nature24489.

[22] A. K. Shalek, R. Satija, J. Shuga, J. J. Trombetta, D. Gennert, D. Lu, P. Chen, R. S. Gertner, J. T. Gaublomme, N. Yosef, et al., Single-cell rna-seq reveals dynamic paracrine control of cellular variation, Nature 510 (7505) (2014) 363–369. doi:10.1038/nature13437.

[23] D. Jovic, X. Liang, H. Zeng, L. Lin, F. Xu, Y. Luo, Single-cell rna sequencing technologies and applications: A brief overview, Clinical and Translational Medicine 12 (3) (2022) e694. doi:10.1002/ctm2.694.

[24] D. Deshpande, K. Chhugani, Y. Chang, A. Karlsberg, C. Loeffler, J. Zhang, A. Muszyńska, V. Munteanu, H. Yang, J. Rotman, L. Tao, B. Balliu, E. Tseng, E. Eskin, F. Zhao, P. Mohammadi, P. P. Łabaj, S. Mangul, Rna-seq data science: From raw data to effective interpretation, Frontiers in Genetics 14 (2023). doi:10.3389/fgene.2023.997383.

[25] P. S. Sonal M Manohar, A. Nair, Flow cytometry: Principles, applications and recent advances, Bioanalysis 13 (3) (2021) 181–198. doi:10.4155/bio-2020-0267.

[26] J. P. Robinson, R. Ostafe, S. N. Iyengar, B. Rajwa, R. Fischer, Flow cytometry: The next revolution, Cells 12 (14) (2023). doi:10.3390/cells12141875.
URL https://www.mdpi.com/2073-4409/12/14/1875

[27] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, S. W. Zucker, Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps, Proceedings of the National Academy of Sciences 102 (21) (2005) 7426–7431. doi:10.1073/pnas.0500334102.

[28] S. Eisenhofer, A coupled system of ordinary and partial differential equations modeling the swelling of mitochondria, Ph.D. thesis, Technische Universität München (2013).

[29] L. Evans, Partial differential equations (1998) graduate studies in mathematics, 19, Amer. Math. Soc., Providence, Rhode (1998).

[30] S. Merkt, L. Fuhrmann, E. Dudkin, A. Schlitzer, B. Niethammer, J. Hasenauer, Supplementary files for "a dynamic model for waddington's landscape accounting for cell-to-cell communication" (2024). doi:10.5281/zenodo.14295650.

[31] R. D. Lazarov, I. D. Mishev, P. S. Vassilevski, Finite Volume Methods for Convection-Diffusion Problems, Vol. 33, Society for Industrial and Applied Mathematics, 1996.

[32] R. Eymard, T. Gallouët, R. Herbin, Finite volume methods, in: Solution of Equation in $\mathbb{R}^n$ (Part 3), Techniques of Scientific Computing (Part 3), Vol. 7 of Handbook of Numerical Analysis, Elsevier, 2000, pp. 713–1018. doi:10.1016/S1570-8659(00)07005-8.

[33] W. Hundsdorfer, J. Verwer, et al., Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations, Springer Verlag, Berlin, Germany, 2003.

[34] A. C. Hindmarsh, P. N. Brown, K. E. Grant, S. L. Lee, R. Serban, D. E. Shumaker, C. S. Woodward, SUNDIALS: Suite of nonlinear and differential/algebraic equation solvers, ACM Transactions on Mathematical Software (TOMS) 31 (3) (2005) 363–396. doi:10.1145/1089014.1089020.

[35] D. J. Gardner, D. R. Reynolds, C. S. Woodward, C. J. Balos, Enabling new flexibility in the SUNDIALS suite of nonlinear and differential/algebraic equation solvers, ACM Transactions on Mathematical Software (TOMS) (2022). doi:10.1145/3539801.

[36] F. Fröhlich, D. Weindl, Y. Schälte, D. Pathirana, Ł. Paszkowski, G. T. Lines, P. Stapor, J. Hasenauer, Amici: High-performance sensitivity analysis for large ordinary differential equation models, BioinformaticsBtab227 (04 2021). doi:10.1093/bioinformatics/btab227.

[37] J. E. Guyer, D. Wheeler, J. A. Warren, FiPy: Partial differential equations with Python, Computing in Science & Engineering 11 (3) (2009) 6–15. doi:10.1109/MCSE.2009.52.

[38] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, SciPy 1.0 Contributors, SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python, Nature Methods 17 (2020) 261–272. doi:10.1038/s41592-019-0686-2.

[39] Y. Schälte, F. Fröhlich, P. J. Jost, J. Vanhoefer, D. Pathirana, P. Stapor, P. Lakrisenko, D. Wang, E. Raimúndez, S. Merkt, L. Schmiester, P. Städter, S. Grein, E. Dudkin, D. Doresic, D. Weindl, J. Hasenauer, pyPESTO: a modular and scalable tool for parameter estimation for dynamic models, Bioinformatics 39 (11) (2023) btad711. doi:10.1093/bioinformatics/btad711.

[40] A. Wächter, L. T. Biegler, On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming, Mathematical Programming 106 (1) (2006) 25–57. doi:10.1007/s10107-004-0559-y.

[41] A. F. Villaverde, E. Raimúndez, J. Hasenauer, J. R. Banga, Assessment of prediction uncertainty quantification methods in systems biology, IEEE/ACM Transactions on Computational Biology and Bioinformatics 20 (3) (2023) 1725–1736. doi:10.1109/TCBB.2022.3213914.

[42] L. G. Van Der Flier, H. Clevers, Stem cells, self-renewal, and differentiation in the intestinal epithelium, Annual review of physiology 71 (2009) 241–260. doi:10.1146/annurev.physiol.010908.163145.

[43] A. Gregorieff, D. Pinto, H. Begthel, O. Destrée, M. Kielman, H. Clevers, Expression pattern of wnt signaling components in the adult intestine, Gastroenterology 129 (2) (2005) 626–638. doi:10.1016/j.gastro.2005.06.007.

[44] N. Barker, Adult intestinal stem cells: critical drivers of epithelial homeostasis and regeneration, Nature Reviews Molecular Cell Biology 15 (1) (2014) 19–33. doi:10.1038/nrm3721.

[45] N. Barker, J. H. van Es, J. Kuipers, P. Kujala, M. van den Born, M. Cozijnsen, A. Haegebarth, J. Korving, H. Begthel, P. J. Peters, H. Clevers, Identification of stem cells in small intestine and colon by marker gene lgr5, Nature 449 (7165) (2007) 1003–1007. doi:10.1038/nature06196.

[46] J. H. Van Es, T. Sato, M. Van De Wetering, A. Lyubimova, A. N. Y. Nee, A. Gregorieff, N. Sasaki, L. Zeinstra, M. Van Den Born, J. Korving, et al., Dll1+ secretory progenitor cells revert to stem cells upon crypt damage, Nature cell biology 14 (10) (2012) 1099–1104. doi:10.1038/ncb2581.

[47] T. Sato, J. H. Van Es, H. J. Snippert, D. E. Stange, R. G. Vries, M. Van Den Born, N. Barker, N. F. Shroyer, M. Van De Wetering, H. Clevers, Paneth cells constitute the niche for lgr5 stem cells in intestinal crypts, Nature 469 (7330) (2011) 415–418. doi:10.1038/nature09637.

[48] S. Schwitalla, A. A. Fingerle, P. Cammareri, T. Nebelsiek, S. I. Göktuna, P. K. Ziegler, O. Canli, J. Heijmans, D. J. Huels, G. Moreaux, et al., Intestinal tumorigenesis initiated by dedifferentiation and acquisition of stem-cell-like properties, Cell 152 (1-2) (2013) 25–38. doi:10.1016/j.cell.2012.12.012.

[49] A. J. Carulli, L. C. Samuelson, S. Schnell, Unraveling intestinal stem cell behavior with models of crypt dynamics, Integrative Biology 6 (3) (2014) 243–257. doi:10.1039/c3ib40163d.

[50] C. Crosnier, D. Stamataki, J. Lewis, Organizing cell renewal in the intestine: stem cells, signals and combinatorial control, Nature Reviews Genetics 7 (5) (2006) 349–359. doi:10.1038/nrg1840.

[51] F. A. Wolf, P. Angerer, F. J. Theis, Scanpy: large-scale single-cell gene expression data analysis, Genome Biology 19 (1) (2018) 15. doi:10.1186/s13059-017-1382-0.

[52] F. A. Wolf, F. K. Hamey, M. Plass, J. Solana, J. S. Dahlin, B. Göttgens, N. Rajewsky, L. Simon, F. J. Theis, Paga: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells, Genome biology 20 (1) (2019) 1–9. doi:10.1186/s13059-019-1663-x.

[53] L. Haghverdi, M. Büttner, F. A. Wolf, F. Buettner, F. J. Theis, Diffusion pseudotime robustly reconstructs lineage branching, Nature Methods 13 (10) (2016) 845–848. doi:10.1038/nmeth.3971.

[54] N. Barker, A. van Oudenaarden, H. Clevers, Identifying the stem cell of the intestinal crypt: strategies and pitfalls, Cell stem cell 11 (4) (2012) 452–460. doi:10.1016/j.stem.2012.09.009.

[55] C. S. Potten, Stem cells in gastrointestinal epithelium: numbers, characteristics and death, Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 353 (1370) (1998) 821–830. doi:10.1098/rstb.1998.0246.

[56] H. J. Snippert, L. G. Van Der Flier, T. Sato, J. H. Van Es, M. Van Den Born, C. Kroon-Veenboer, N. Barker, A. M. Klein, J. Van Rheenen, B. D. Simons, et al., Intestinal crypt homeostasis results from neutral competition between symmetrically dividing lgr5 stem cells, Cell 143 (1) (2010) 134–144. doi:10.1016/j.cell.2010.09.016.

[57] P. Buske, J. Galle, N. Barker, G. Aust, H. Clevers, M. Loeffler, A comprehensive model of the spatio-temporal stem cell and tissue organisation in the intestinal crypt, PLoS Comput Biol 7 (1) (2011) e1001045. doi:10.1371/journal.pcbi.1001045.

[58] B. Hie, B. Bryson, B. Berger, Efficient integration of heterogeneous single-cell transcriptomes using scanorama, Nature Biotechnology 37 (6) (2019) 685–691. doi:10.1038/s41587-019-0113-3.

[59] A. T. Kamath, S. Henri, F. Battye, D. F. Tough, K. Shortman, Developmental kinetics and lifespan of dendritic cells in mouse lymphoid organs, Blood 100 (5) (2002) 1734–1741. doi:10.1182/blood.V100.5.1734.h81702001734_1734_1741.

[60] D. Le, H. Smith, Strong positivity of solutions to parabolic and elliptic equations on nonsmooth domains, Journal of Mathematical Analysis and Applications 275 (1) (2002) 208–221. doi:https://doi.org/10.1016/S0022-247X(02)00314-1.

[61] D. Kinderlehrer, G. Stampacchia, An Introduction to Variational Inequalities and Their Applications, Society for Industrial and Applied Mathematics, 2000. doi:10.1137/1.9780898719451.