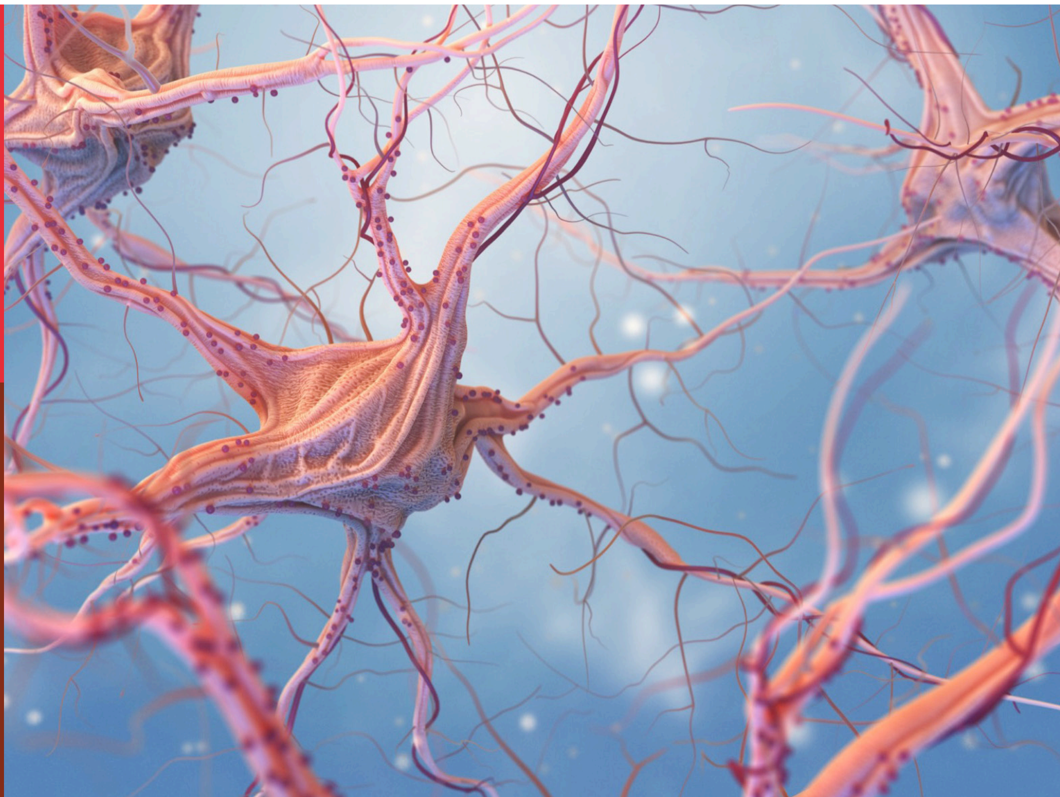
frontiers | Research Topics

Insights in systems biology research

Edited by
Gary An, Ioannis P. Androulakis, Eric H. Chang,
Rongling Wu, Shayn Peirce-Cottler and
Edoardo Saccenti

Published in
Frontiers in Systems Biology



ISSN 1664-8714
ISBN 978-2-8325-6432-5
DOI 10.3389/978-2-8325-6432-5

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Insights in systems biology research

Topic editors
Gary An — University of Vermont, United States
Ioannis P. Androulakis — Rutgers, The State University of New Jersey, United States
Eric H. Chang — Feinstein Institute for Medical Research, United States
Rongling Wu — The Pennsylvania State University (PSU), United States
Shayn Peirce-Cottler — University of Virginia, United States
Edoardo Saccenti — Wageningen University and Research, Netherlands

Citation
An, G., Androulakis, I. P., Chang, E. H., Wu, R., Peirce-Cottler, S., Saccenti, E., eds. (2025). *Insights in systems biology research*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-6432-5

Table of contents

TYPE Original Research
PUBLISHED 22 January 2024
DOI 10.3389/fsysb.2023.1283341

OPEN ACCESS

EDITED BY
Gary An,
University of Vermont, United States

REVIEWED BY
R. Chase Cockrell,
University of Vermont, United States
Giulia Russo,
University of Catania, Italy

*CORRESPONDENCE
Denise E. Kirschner,
✉ kirschne@umich.edu
Jennifer J. Linderman,
✉ linderma@umich.edu

RECEIVED 15 September 2023
ACCEPTED 14 December 2023
PUBLISHED 22 January 2024

CITATION
Michael CT, Almohri SA, Linderman JJ and Kirschner DE (2024), A framework for multi-scale intervention modeling: virtual cohorts, virtual clinical trials, and model-to-model comparisons.
Front. Syst. Biol. 3:1283341.
doi: 10.3389/fsysb.2023.1283341

A framework for multi-scale intervention modeling: virtual cohorts, virtual clinical trials, and model-to-model comparisons

Christian T. Michael[1,2], Sayed Ahmad Almohri[2], Jennifer J. Linderman[2]* and Denise E. Kirschner[1]*

[1]Department of Microbiology and Immunology, University of Michigan–Michigan Medicine, Ann Arbor, MI, United States, [2]Department of Chemical Engineering, University of Michigan, Ann Arbor, MI, United States

Computational models of disease progression have been constructed for a myriad of pathologies. Typically, the conceptual implementation for pathology-related *in silico* intervention studies has been ad hoc and similar in design to experimental studies. We introduce a multi-scale interventional design (MID) framework toward two key goals: tracking of disease dynamics from within-body to patient to population scale; and tracking impact(s) of interventions across these same spatial scales. Our MID framework prioritizes investigation of impact on individual patients within virtual pre-clinical trials, instead of replicating the design of experimental studies. We apply a MID framework to develop, organize, and analyze a cohort of virtual patients for the study of tuberculosis (TB) as an example disease. For this study, we use *HostSim*: our next-generation whole patient-scale computational model of individuals infected with *Mycobacterium tuberculosis*. *HostSim* captures infection within lungs by tracking multiple granulomas, together with dynamics occurring with blood and lymph node compartments, the compartments involved during pulmonary TB. We extend *HostSim* to include a simple drug intervention as an example of our approach and use our MID framework to quantify the impact of treatment at cellular and tissue (granuloma), patient (lungs, lymph nodes and blood), and population scales. Sensitivity analyses allow us to determine which features of virtual patients are the strongest predictors of intervention efficacy across scales. These insights allow us to identify patient-heterogeneous mechanisms that drive outcomes across scales.

KEYWORDS

model study design, digital partners, disease modeling, tuberculosis, computational biology, pharmacokinetic-pharmacodynamic model, sensitivity analysis, agent-based model

1 Introduction

Understanding the effectiveness of intervention measures in the context of patient-to-patient variability is a challenge in both drug and vaccine studies. Diseases such as cancer and infections such as COVID-19 and tuberculosis (TB) show patient variation in both infection outcomes and intervention efficacies. Actionable data–data that may help us determine efficacious interventions as well as understand patient variability–is limited by the frequency of patient visits, the quantity and quality of patient data, monitoring procedures, and resources.

Computational models are an additional approach toward gaining valuable insights into disease and accompanying interventions. Models applied in biomedicine have been used to disentangle the multitude of interconnected components of large complex systems such as cancer, HIV-1/AIDS, influenza and TB. Many modeling studies seek to: i) replicate experimental *in vivo, in vitro,* or *in situ* studies by using *in silico* experiments while maintaining experimental design, such as experimental interventional studies (Aggarwal and Ranganathan, 2019); ii) determine mechanistic impacts of model components and perturbations/treatments/interventions on output, e.g., by using sensitivity analyses; and/or iii) develop model extensions or reductions to determine the relative importance of detailed components (Kirschner et al., 2014).

In order for a model to credibly perform credible *in silico* experiments requires rigorous validation against available data (Tatka et al., 2023). The precision and rigor required are system-specific and adapted to the expected use of the model's output (Fogarty et al., 2022), and consequences of incorrect model predictions (Aldieri et al., 2023). Various standards exist to codify model validation (Fogarty et al., 2022; Tatka et al., 2023); including the ten rules for model credibility developed by the Multi-scale Modeling Consortium (Erdemir et al., 2020; Fogarty et al., 2022; Nanda et al., 2023; Tatka et al., 2023) for systems biology approaches, as well as the ASME VandV40 standards (ASME, 2018; Aldieri et al., 2023; Tatka et al., 2023), and NASA standards for models and simulation (NASA, 2016; Tatka et al., 2023). Each of these standards establishes a series of assessments by which we can establish the appropriateness of a model to address a given question of interest relative to the model's context of use. Here we describe a framework for using a validated computational model, for example, in a virtual clinical trial.

When we design virtual clinical trials from computational models, we find one luxury in that the definition of a "virtual patient" is flexible. For example, if a pharmacokinetic-pharmacodynamic (PK-PD) model is being implemented, then a patient's pharmacokinetic identity is entirely defined by a set of PK-PD parameters. In many individual-scale computational approaches, every population generated by a model is independent, which reflects the design that motivates experimental interventional studies. However, that same virtual patient can serve in multiple "what-if" scenarios, such as determining effects of model stochasticity or perturbed biological influences or as a negative control (no drug treatment). The experimental analogue to this approach would be tantamount to running different experimental interventions on the same patient under the same conditions and scenarios.

With our ability to select amongst many types of models that can credibly represent the same system, we need a methodology to compare models in an implementation-agnostic way. We have seen a recent push to standardize modeling approaches with modeling ecosystems such as CompuCell3D (Poplawski et al., 2008; Shirinifard et al., 2009), VCell (Blinov et al., 2008; Schaff et al., 2016), PhysiCell (Ghaffarizadeh et al., 2018), as well as standardized language for ODE model implementation such as SBML (Keating et al., 2020), SED-ML (Bergmann et al., 2017; Smith et al., 2021), COMBINE, OMEX (Bergmann et al., 2014; Neal et al., 2020), and others (Tatka et al., 2023). With this variety of platforms, software, computational frameworks, and databases available (computational models, medical digital twins, etc.), it is likely impossible to develop a single computational package to automate analysis or comparison methodologies that account for the myriad of modeling approaches possible without overly constraining their use context. One component common to all models is the representation a real patient by a virtual one (with varying degrees of accuracy and refinement), hence we can create a broadly-applicable methodological framework to perform model-to-model comparisons.

In this work we propose a generally applicable methodological framework, which we refer to as a *multi-scale interventional design* (MID) framework: a method of developing a cohort of virtual patients that we use to examine impacts of interventions on each virtual patient within a virtual cohort by tracking dynamics across physiological scales, from within-patient, through whole-patient, and up to the population scale (Figure 1A). Using a MID framework requires three key components: i) a cohort of *virtual patients*, along with a biological justification as to why the *same virtual patient* is able to be represented in multiple models; ii) a set of two related and validated *model versions*, such as a control model and an experimental model if representing, for example, a treatment intervention; and iii) an *impact quantification method* by which the outcomes of both model versions can be meaningfully compared.

Consider TB, a disease caused by an infectious bacterium *Mycobacterium tuberculosis* (Mtb) that has infected one-fourth of the current world's population (WHO, 2022). In 2020, TB had a comparable annual death-toll to COVID-19 (WHO, 2020), and concurrent infection with COVID-19 or HIV has increased mortality for TB patients (WHO, 2022). Patients infected with Mtb may eliminate infection, control infection (resulting in latent TB disease) or fail to control infection (resulting in active TB disease), yet the factors determining those outcomes are not fully understood. Note, it is important to distinguish that Mtb are the bacteria that cause infection, whereas tuberculosis (TB) is the disease that results from infection. Data for analysis of Mtb infection progression typically comes from low-resolution measurements in patients (e.g., sputum analysis (Portevin et al., 2014; Guzzetta et al., 2015; Esmail et al., 2016)) or at necropsy when studying non-human primates (NHPs) or other animal models (Barry et al., 2009; Martin et al., 2017; Lin and Flynn, 2018; Wong et al., 2020; Grant et al., 2022). As a result, deriving mechanistic insights to time-evolution of Mtb infections and its interplay with patient heterogeneity across populations is a crucial step in improving our ability to study TB as well as other diseases.

Pulmonary TB, the most common form of the disease, is a highly complex disease with multiple interacting systems determining patient fate (note that we will also refer to patients as hosts as this is common terminology for an infectious disease). There is heterogeneity in lung granulomas, the focal structures of Mtb-host interaction, within individual TB hosts that is critical to prediction of host outcomes (Cadena et al., 2017; Lyadova, 2017; Cicchese et al., 2020). Host-scale dynamics are also heterogeneous and fall into at least three groups that exist on a spectrum: hosts that will clear the infection, control the infection, or fail to control infection and thus suffer active disease (Lin and Flynn, 2018). The dynamics of Mtb infected cohorts are

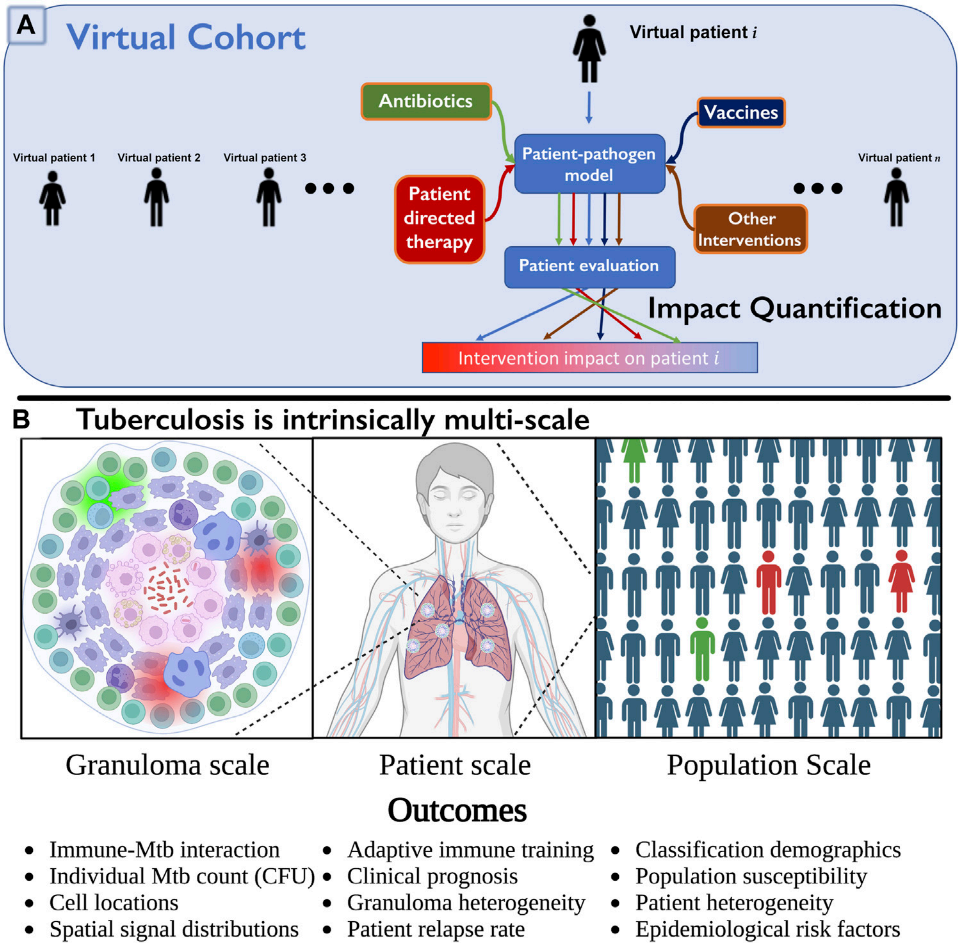


FIGURE 1
(A) Multi-scale intervention design to study over three physiological scales. We include a collection of virtual patients, a virtual cohort, that can each be represented by a control model or represented under various interventions applied (e.g., *HostSim* and a perturbed version, such as with drugs or vaccines). The virtual patient can be evaluated in each scenario, and impact level quantified by observing differences in specific patient outcomes. This can be quickly repeated for many patients in parallel to determine an overall population-scale impact (cohort effect), or to examine which subpopulations respond to interventions. **(B)** We illustrate three of the operative scales critical to understand TB. Lung granulomas encompass the complex dynamics of *Mycobacterium tuberculosis* (Mtb) populations and their interactions with various lymphocyte populations. Clinical classification of the patient (active or latent disease) is determined by multiple granulomas interacting with the patient's lymphatic system. At the population scale, patients within a cohort vary in their susceptibility to infection and response to treatment, complicating our understanding and prediction of the demographic of clinical classifications. Note: we created Panel **(B)** using BioRender.com.

also heterogeneous, e.g., some hosts improve with drug treatment rapidly while others do not. Thus to understand host infection progression and treatment, it is imperative to study TB at multiple scales and decipher how small-scale interactions influence large-scale findings (Figure 1B), making it an ideal candidate to test the MID framework.

To demonstrate our ability to study virtual cohorts using a MID framework, we implemented and tested our framework on multiple versions of *HostSim*, our next-generation, within-host to whole-host scale computational model of Mtb infection. These versions include a negative control version of *HostSim*, wherein infection of virtual TB hosts is left untreated, as well as three simple drug intervention versions for comparison. We implemented and tested these drug interventions in our virtual cohort and demonstrated that MID is an effective framework type to yield multi-scale virtual patient insights on complex biological problems that both include and explain patient heterogeneity at each scale.

2 Methods

Creating a MID framework requires three interconnected components: 1) a virtual cohort $\{VH\}$, 2) a pair of related model versions: a control model M_0 and an intervention model M_P to represent these hosts, and 3) an impact quantification: a method of evaluating and comparing the projected trajectories and final states of the virtual hosts between model versions. We present these components in the context of TB as an example. We also describe an updated version of *HostSim*, our previously published model of a whole-host, which captures the immune response to infection with Mtb within 3 physiological compartments: lungs, lymph nodes, and blood that represent pulmonary TB.

2.1 Creating the virtual cohort - a collection of 500 virtual hosts, $\{VH\}$

In our virtual cohort, each virtual host represents a typical host infected with Mtb with no comorbidities, and our virtual cohort will be generated to well-represent the demographic range of untreated patient outcomes observed in the biological context. We give our virtual hosts as an infection inoculum, 13 founding Mtb and one to five resting macrophages on day 0. Our virtual hosts represent Mtb infection progression in individuals up to 400 days post-infection, tracking granuloma cellular and bacterial composition once per day. In practice, each virtual host (VH_i) in the cohort ($\{VH\}$) is recorded as a granuloma and whole-host scale parameter set P_i that is preserved between all versions of that virtual host (whether disease, treatment, etc.), which we refer to as the *virtual patient (host) identity.* We choose our virtual cohort of 500 virtual hosts ($\{VH\}$) such that we capture the demographic of clinical outcomes observed in reality (Cadena et al., 2017). We select these parameter values by using the Latin Hypercube Sampling (LHS) method to generate values within a biologically viable range that we calibrate to multiple datasets (Section 2.3), ensuring that we accurately capture the heterogeneous spectrum of host outcomes. Note that the LHS method of parameter selection promotes stochastic and stratified coverage of the parameter space under the assumption of uniform distribution of each parameter within the experimental ranges (Helton and Davis, 2003; Cacuci and Ionescu-Bujor, 2004; Marino et al., 2008).

2.2 TB virtual host model*: HostSim* as M_0

Briefly, the *HostSim* model is based on known biology of pulmonary TB. When inhaled, Mtb is phagocytosed by macrophages. These inactive macrophages are unable to fully digest Mtb, which slowly replicates inside of them. Eventually, the macrophage bursts after reaching a carrying capacity of internal Mtb, and the cycle continues. In part due to the slow Mtb replication rate, inflammatory signals and antigen presentation occurs more slowly - and in NHPs, the lymph nodes (LNs) show no metabolic activity until 2–4 weeks post-infection (Coleman et al., 2014; Ganchua et al., 2018; Ganchua et al., 2020). Multiple granulomas form, typically one for each Mtb colony forming unit (CFU) (i.e., an individual Mtb bacterium) that lands within the lung (Martin et al., 2017). Mtb-specific T-cells arrive from LNs to activate macrophages and allow them to destroy intracellular Mtb and induce apoptosis of infected macrophages. These dynamics result in the development of a complex structure called a granuloma that comprises Mtb, live immune cells, and dead tissue (caseum).

HostSim, our untreated virtual host model, is a multi-scale computational model of an individual host that represents both the tissue-scale and whole-host scale response to pulmonary Mtb infection (Joslyn et al., 2022a; Joslyn et al., 2022b). We created a next-generation version of *HostSim* herein to include additional biological features and better capture Mtb infection immunobiology (see Supplementary Material S1 Section 2 for model updates, and Supplementary Material S2 for a complete model description and list of equations). We represent three physiological compartments in our hybrid computational model *HostSim*: lungs, LNs, and blood. The lung compartment captures a collection of lung granulomas represented as agents in an agent-based model. Each agent is itself comprised of a system of 22 nonlinear ordinary differential equations (ODEs) describing interactions between macrophages, three subpopulations of Mtb - intracellular, extracellular, and non-replicating; cytokine signals (e.g., IL-4, IL-10, IL-12, and TNF-α), and different T-cells in various states of differentiation (Figure 2). Granulomas allow antigen-presenting cells to travel to LNs proportional to the Mtb burden within a granuloma, and the LN clonally expands Mtb-specific T-cells. T-cells are released from the LN compartment (described by ODEs) into blood (also represented by ODEs) where they may be recruited into lung granulomas. Since each granuloma has its own instantiation and parameterization within our ODE system, and formation of new granulomas makes the number of granuloma ODE trajectories variable, we consider *HostSim* to be a hybrid agent-based model. *HostSim* is simulated in MATLAB using the ode15s variable order ODE solver for time-stepping the ODE portions of *HostSim.*

When running simulations, cytokine signals and antigen presenting cells circulate to a virtual host's LN compartment, which selectively clones Mtb-specific CD4$^+$ and CD8$^+$ T-cells. We have newly-calibrated parameter ranges to a variety of data from both NHPs (Gideon et al., 2015; Marino et al., 2016; Cadena et al., 2018; Darrah et al., 2019) and our fine-grained model of a single granuloma, *GranSim*, to capture the heterogeneity both between hosts and between granulomas within a single host (see Section 2.3).

2.3 Calibrating the virtual cohort $\{VH\}$ to be represented in M_0

We first need to calibrate *HostSim* in order for it to be a credible M_0 in our MID framework. With 201 varied parameters and 3 compartments, *HostSim* requires careful calibration that leverages known constraints and biological ranges. As in (Joslyn et al., 2022a; Joslyn et al., 2022b), we calibrate our model by comparing its outputs to data taken from 646 NHP granulomas assembled over the last 15 years (Gideon et al., 2015; Marino et al., 2016; Cadena et al., 2018; Darrah et al., 2019). We use our previously published calibration method, CaliPro (Joslyn et al., 2023), to refine both granuloma and LN parameter ranges from our previously calibrated values (Joslyn et al., 2022b). Our calibration criteria are implemented at the granuloma scale, and each criterion tests the proximity of simulated granulomas to granuloma data collected from NHPs (Gideon et al., 2015; Marino et al., 2016; Cadena et al., 2018; Darrah et al., 2019), as well as synthetic data from our fine-grain model of a single granuloma, *GranSim* (Segovia-Juarez et al., 2004; Pienaar et al., 2017; Warsinske et al., 2017; Sarathy et al., 2019; Cicchese et al., 2020; Budak et al., 2023).

Briefly, CaliPro is a calibration method that incorporates a broad range of model parameters and multiple and varied datasets. Using LHS, we choose a stratified collection of parameter values out of a broad parameter (Marino et al., 2008). CaliPro evaluates the model at each of these parameter values and determines whether the outputs are sufficiently close to the given dataset(s) to be admitted to a "pass set", i.e., meeting heuristic criteria that suggest that model output is biologically relevant. CaliPro then

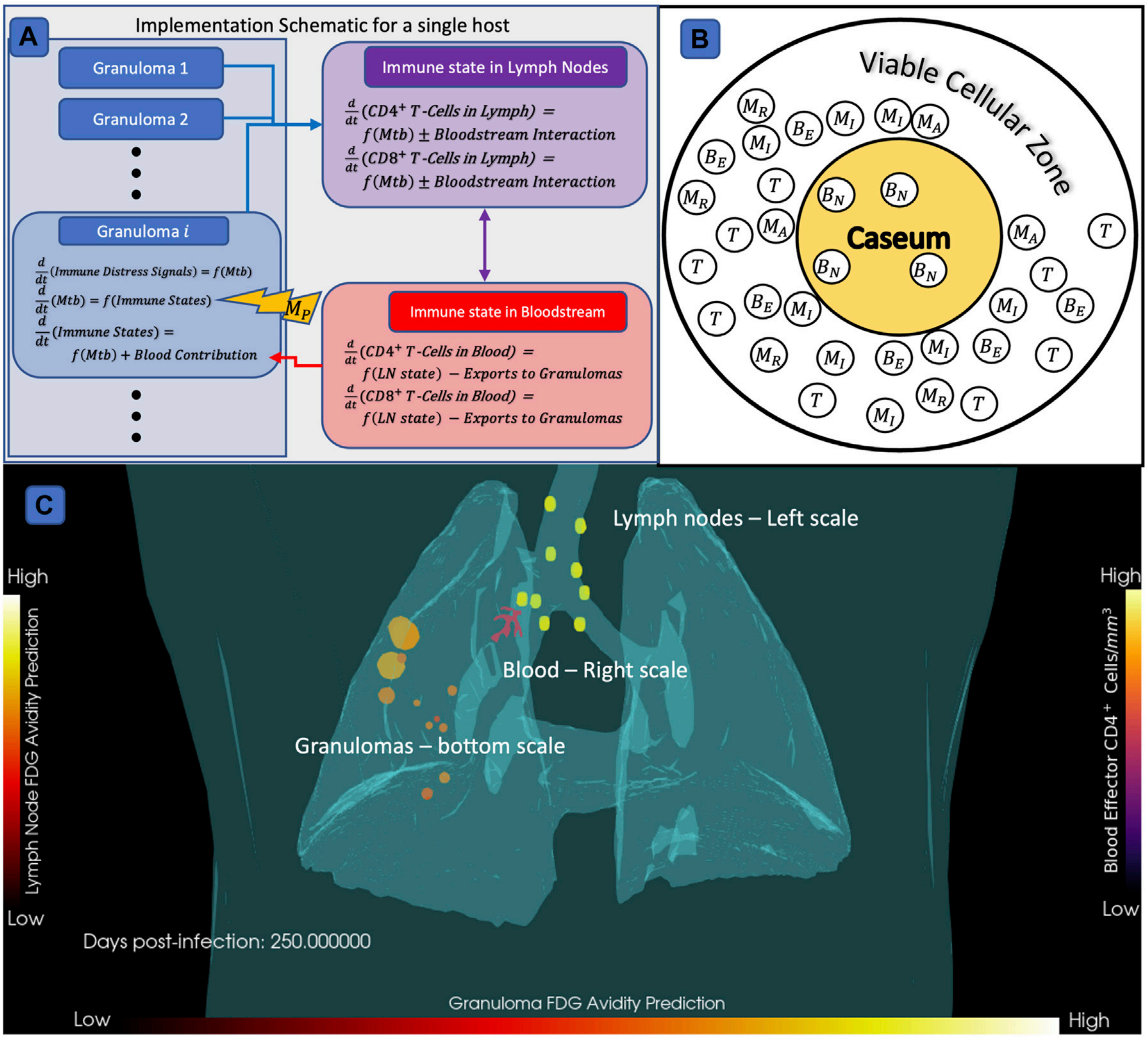

A
Implementation Schematic for a single host
Granuloma 1
Granuloma 2
Granuloma i
$\frac{d}{dt}$(Immune Distress Signals) = f(Mtb)
$\frac{d}{dt}$(Mtb) = f(Immune States)
$\frac{d}{dt}$(Immune States) = f(Mtb) + Blood Contribution
M_P
Immune state in Lymph Nodes
$\frac{d}{dt}$(CD4$^+$ T-Cells in Lymph) = f(Mtb) ± Bloodstream Interaction
$\frac{d}{dt}$(CD8$^+$ T-Cells in Lymph) = f(Mtb) ± Bloodstream Interaction
Immune state in Bloodstream
$\frac{d}{dt}$(CD4$^+$ T-Cells in Blood) = f(LN state) − Exports to Granulomas
$\frac{d}{dt}$(CD8$^+$ T-Cells in Blood) = f(LN state) − Exports to Granulomas
B
Viable Cellular Zone
Caseum
M_R M_I M_A B_E B_N T
C
Lymph nodes – Left scale
Blood – Right scale
Granulomas – bottom scale
High
Lymph Node FDG Avidity Prediction
Low
High
Blood Effector CD4$^+$ Cells/mm^3
Low
Days post-infection: 250.000000
Granuloma FDG Avidity Prediction
Low
High

FIGURE 2
Diagram of *HostSim* model construction, M_0. **(A)** Granulomas within lungs (blue compartments) are linked to lymph node (purple) and blood (red) compartments (details in Supplementary Material S2). We represent interventions as being applied to the equations governing Mtb development. **(B)** Diagram of a simplified granuloma as represented in *HostSim*. In the central caseum sub-compartment, nonreplicating bacteria are trapped within a hypoxic/necrotic core. All other species, including macrophages, T-cells, and extracellular bacteria, are in the viable cellular zone. Note that in *HostSim*, the viable cellular zone is treated as well-mixed for the sake of cell-cell interactions. **(C)** Lung granulomas and lymph nodes of virtual TB host at t = 250 days post-infection shown within the context of a lung and body triangulation of a nonhuman primate [courtesy of Henry J. Borish in JoAnne Flynn Lab, University of Pittsburgh]. Cylinders on the trachea represent the lymph node compartments, and spheres (colored by their CFU count and sized based on their cellular composition) represent granulomas. The branching blood vascular surface is colored based on blood effector CD4$^+$ T-cell concentration. (Details of visualization are in Supplementary Material S1 Section 3).

shrinks the parameter ranges to exclusively capture the pass set while still covering the broadest possible set of parameters. This process is iterated multiple times. After calibration, 90% of granulomas "passed" all tests against calibration criteria, which are described in Supplementary Material S1 Section 2.2. Our calibrated parameter ranges are listed in Supplementary Material S2 Section 5.

2.4 Validation that our virtual cohort hosts capture population demographics for TB

Our goal is to use *HostSim* simulations to determine, on 3 physiological scales: population-scale, host-scale, and granuloma-scale, which features drive both granuloma and whole host infection outcomes. As such, our virtual cohort should reflect epidemiological contexts for TB (Joslyn et al., 2022b; WHO, 2022), and our use case of *HostSim* is to generate a collection of virtual hosts whose trajectories agree with distributions of available global data on humans for TB. To do this, we define virtual host classifications in a clinically interpretable way. For studying TB, our classifications are clinically latent, bacteria sterilizing, and active disease. We classify virtual hosts as having active TB if either 1) they have higher total lung CFU than an active-host cutoff of $3.2 \cdot 10^5$; or 2) they have at least one granuloma which increases by more than 10% CFU between days 100 and 150 post-infection. We chose these times post-infection because primary infection sites have a transient peak of increasing bacterial numbers around day 30 before the immune system responds and forms a proper granuloma (Cadena et al., 2017). We define active disease based on data taken from 4 NHPs that were necropsied early due to severe Mtb infection (see Supplementary Material S1 Section 1, courtesy of the JoAnne Flynn lab). We classify virtual hosts as sterilizing hosts if their total lung CFU count has dropped to zero at or before 400 days post-infection. We consider all other hosts as having clinically latent TB, and there is a spectrum of outcomes within this group as is observed in humans and NHPs (Lin and Flynn, 2018). After performing calibration on the host and granuloma scales, our virtual cohort had a distribution of outcomes: approximately 90% of virtual hosts classified with latent TB, 6% with active TB, and 4% of virtual hosts sterilizing their infection entirely. This indicates that our virtual cohort reflects observed trends in patient outcomes at the population scale (Cadena et al., 2017; Lin and Flynn, 2018). Note our classifications are flexible as new data become available.

2.5 Developing intervention models M_P for our virtual cohort

Our goal is to create a cohort that we can test different model perturbations such as antibiotic treatment, vaccines, or other interventions. To do this, our goal is to build versions of our model that represent a control version M_0 and an intervention version, M_P. The intervention version should i) observe both the host and granuloma scale mechanisms from M_0, and thereby maintains credibility, and ii) sufficiently represent intervention dynamics to identify key drivers of host-response. Here, in the interest of demonstrating the MID framework and its use, we use a highly-simplified model of antibiotic treatment of TB as an example of an intervention model M_P. Our objective is to capture heterogeneity in the host-response to treatment over multiple physiological scales. To establish this approach, we use coarse-grained representation of 3 TB antibiotics, where we qualitatively represent the impact on bacterial burden in time by capturing the known modes of action of different antibiotics. These simplified TB drugs represent 3 different classes of drugs that are currently used to treat TB: isoniazid (INH), bedaquiline (BDQ), and pyrazinamide (PZA). While these drugs are typically used in combination therapy, here, for example, purposes, we implement each one individually. We model these drugs based only on their known killing (bactericidal) or bacteriostatic behaviors (Zhang and Mitchison, 2003; Jayaram et al., 2004; Sarathy et al., 2018; Budak et al., 2023), omitting for this simple model any consideration of pharmacokinetics or transport limitations in accessing portions of the granuloma as we have done in other work (Budak et al., 2023). We define an INH-like intervention version M_{INH}, a PZA-like intervention M_{PZA}, and a BDQ-like intervention M_{BDQ} (each version representing an M_P). Here, we note some differences in these drugs' mechanisms that we will phenomenologically capture: i) INH is able to penetrate into caseum and kill bacteria but is not taken up by infected macrophages (Jayaram et al., 2004; Prideaux et al., 2015; Nahid et al., 2016; Sarathy et al., 2016; Sarathy et al., 2018); ii) BDQ kills bacteria that it can reach more effectively but takes much longer to penetrate into caseum (Dhillon et al., 2010; Chahine et al., 2014; Prideaux et al., 2015; Sarathy et al., 2016; Sarathy et al., 2018); and iii) PZA is a bacteriostatic drug that slows bacterial replication but takes a long time to penetrate into caseum (Zhang and Mitchison, 2003; Prideaux et al., 2015; Nahid et al., 2016; Sarathy et al., 2016; Sarathy et al., 2018).

We represent dosing our virtual hosts by modifying the equations governing bacterial growth with the following unitless treatment values A_i after intervention time $t = 200$ days.

$$\frac{d}{dt}B_E = A_1(\text{Replication}) \pm (\text{conversion to } B_I) - A_2(\text{Death})$$

$$\frac{d}{dt}B_I = A_3(\text{Replication}) \pm (\text{conversion to } B_E,\ B_N) - A_4(\text{Death})$$

$$\frac{d}{dt}B_N = (\text{conversion from } B_I) - A_5(\text{Death})$$

where our control model M_0 *HostSim* is recovered if each $A_i = 1$. For M_{BDQ}, we set $A_2 = 5 \cdot 10^7$, $A_4 = 5000$ and $A_5 = 10$; intervention parameters $A_1 = 1$ and $A_3 = 1$ since we do not treat BDQ as bacteriostatic. In M_{INH}, we define these action coefficients relative to M_{BDQ} - in M_{INH}, we set $A_2 = 2500$ since INH is less effective at killing extracellular bacteria; we set $A_5 = 5$ since more INH ends up in caseum though it is less effective at a given concentration than BDQ, and $A_4 = 1$ because our simplified INH does not get taken into macrophages; for INH we also set $A_1 = 1$ and $A_3 = 1$ as it is not bacteriostatic. For M_{PZA}, the bacteriostatic effect is captured by setting $A_1 = A_3 = 0.5$, halving bacterial replication rates for all hosts. We set $A_2 = 1$, $A_4 = 1$, and $A_5 = 1$ since PZA is not bactericidal. It is important to remember that the *virtual patient (host) identity* parameter values (P_i) used to define the virtual cohort $\{VH\}$ are independent of M_0 and M_P. By running simulations using either M_0 or M_P with the same parameters P_i and initial conditions - and thus each virtual host VH_i - the entire virtual cohort can be represented in every model version, while the treatment values A_i are preserved across the cohort.

2.6 Impact quantification method for our MID framework

The final component of our MID framework is an *impact quantification* method that directly quantifies and compares the impact of the intervention model versions M_{INH}, M_{PZA}, and M_{BDQ} against the negative-treatment M_0 at multiple physiological scales. In principle, comparisons between virtual hosts and model versions may use any outcomes and measurements that may be relevant to the system under study. Importantly, the selection of impact quantification is implicitly related to the model's question of interest and context of use, since models may have different levels of credibility depending on which outcome is being observed. The multi-scale component of a MID framework comes from comparing the outcome of VH_i represented in M_P (i.e., $M_P(VH_i)$) to VH_i represented in M_0 (i.e., $M_0(VH_i)$) for each virtual patient in the virtual cohort. Here, we perform this quantification by directly comparing CFU counts between model versions over time. Since M_0 and M_P have identically formatted and nonnegative outputs - time-series data of all *HostSim* variables computed once per day - the ratio of the outputs may be considered at all scales. On the host scale, we examine the ratio of total lung CFUs as

$$\text{Host Impact score} = H_S(t;\, VH_i) = \log\left(\frac{M_0\, CFU(VH_i) + 1}{M_P\, CFU(VH_i) + 1}\right)(t) \tag{1}$$

In this way, hosts with $H_S \approx 0$ have very little treatment effectiveness, $H_S > 0$ have a positive influence on the system outcome, and $H_S < 0$ have a deleterious effect. Capturing impact score over time informs many aspects of the score, including projected time until expected results of intervention. We can also compute an impact score at other physiological scales. For example, at the granuloma scale, we compute an impact score for each granuloma in the lung to obtain the granuloma impact,

$$\begin{aligned}\text{Granuloma Impact Score} &= G_S(t; VH_i)\\ &= \log\left(\frac{M_0\, CFU(VH_i) + 1}{M_P\, CFU(VH_i) + 1}\right)(t).\end{aligned} \tag{2}$$

2.7 Sensitivity analyses

As an additional form of impact quantification in a MID framework, we can also evaluate the impact of M_P via *sensitivity analysis,* which allows us to identify parameters and initial conditions that drive specific features of model output. We use the partial rank correlation coefficient (PRCC) method, which is a computationally efficient and accurate method for performing sensitivity analysis on high-dimensional models (Marino et al., 2008; Renardy et al., 2019; Renardy et al., 2021). When given a set of model runs and a numerical output of that model, the PRCC method determines for each input parameter: i) a coefficient that measures the correlation between that parameter and the model output and ii) a p-value determining the statistical significance of that measurement. We typically use this method to understand the impact of parameter impacts on M_0 outputs. However, since the same virtual cohort $\{VH\}$ is being represented in both models, ($M_0(\{VH\})$ and $M_P(\{VH\})$), sensitivity analysis methods apply to composite models $f[M_0, M_P](\{VH\})$. Since impact quantification methods such as expressions [1] and [2] satisfy the requirements of a composite model, we can perform sensitivity analysis on these scores as well to determine what patient characteristics correlate with intervention scores.

3 Results

3.1 Constructing a MID framework

If we run thousands of simulations, allowing for patient-to-patient variability and representation of each virtual host with and without interventions, we refer to our collection as a *virtual cohort.* We introduce our MID framework, our goal for which is to create an easily implementable layer for most computational modeling systems that represent individual patient dynamics. MID is a framework for making meaningful comparisons between the outcomes of individual virtual patients' outcomes in between a negative control model M_0 and a perturbed *intervention version* of the model M_P (see Figure 1A for a schematic).

To be specific, we require three interconnected components to create our MID framework, and they are: 1) a virtual cohort $\{VH\}$, 2) a pair of related and validated models to represent those patients: a control model M_0 and an intervention model M_P, and 3) a method of evaluating and comparing the projected trajectories and final states of the virtual hosts in either model version (see Section 2.6). One of the only model prerequisites is that there be a natural representation, or a biological justification, for how the same virtual patient VH_i is represented by M_0 and M_P. For example, if M_0 contains a simplified representation of pathogen replication time and M_P contains a detailed pathogen life cycle interacting with a drug intervention, we must ensure that M_P matches the "control limit" as the drug level approaches 0. We must also use caution if two model versions have notably different representations of the same biological process. There must be some biologically-rooted justification as to why we can reasonably assume that the same host is being represented in both model versions.

Lastly, an impact quantification method should be specified that compares trajectories of individual virtual hosts represented in both the M_0 and M_P versions in a biologically-interpretable manner. These should be specific to the particular model system and made to ensure that comparisons between the models are relevant to the intended goal of the intervention. For example, a drug intervention may have an outcome evaluation that weighs time to sterilization, pathogen burden, and drug toxicity. The MID framework components are simple enough that they can be applied to many models from multiple biomedical applications. We list a few examples of potential MID framework implementations in Table 1. Note that if we want to perform a MID framework study using highly stochastic models, we must take care in defining virtual host outcomes. For example, we might work with $\text{Mean}(M)$ and $\text{Var}(M)$. Measurable features for impact quantification should be able to capture differences in dynamics between M_0 and M_P at the scale at which the intervention is applied. As *HostSim* is deterministic (except for rare dissemination events) once the initial agent properties are defined, we omit such considerations from our TB application.

3.2 *HostSim* provides M_0 for MID to study TB over multiple scales

A key step of developing our MID framework study is to declare a control model, M_0. This model represents the unperturbed system that we are interested in studying, which in our case is Mtb infection. We want this model to be well calibrated and validated, and to mechanistically represent our system at the scale that our intervention is going to perturb. For M_0, we use an updated version of *HostSim*, our whole-host model of Mtb infection.

We update our TB simulation *HostSim* (Joslyn et al., 2022b) and recalibrate it to additional published datasets from NHPs across granuloma, host, and population scales (Gideon et al., 2015; Marino et al., 2016; Cadena et al., 2018; Darrah et al., 2019). We calibrated using our *CaliPro* procedure (Joslyn et al., 2023), integrating these data by using a population of 500 virtual hosts $\{VH\}$ sampled from within experimentally viable parameter ranges (see Section 2). Supplementary Figure S2 shows several state variable trajectories over time for a single representative virtual host with latent Mtb infection.

TABLE 1 Examples of potential application of the MID framework to biomedical systems. The virtual patient definition can be flexibly adapted and generalized to a broad set of virtual subjects and intervention types. Note that in all cases, M_0 and M_P should be validated such that they may make credible claims about outcomes used in impact quantification. In some cases, finding a small impact may provide useful results (e.g., that a proposed treatment will not impact patient outcomes, or that a model simplification is sufficient to capture outcomes).

Model	Virtual cohort members	Model versions	Example impact quantification
HostSim (Joslyn et al., 2022a; Joslyn et al., 2022b) (Tuberculosis)	Virtual host: a vector of parameters describing host PK/PD in each granuloma; Initial conditions of each granuloma and lymph node	*HostSim*, which encompasses all equations, dynamics, component interactions M_0 - No treatment M_P - With antibiotic treatment	Ratio of bacteria load between untreated host and host with antibiotic intervention for each granuloma and host; demographic of clinically Latent, Active, or Sterilizing patients
Drug Interventions in GranSim (Pienaar et al., 2017; Sarathy et al., 2019; Cicchese et al., 2020; Budak et al., 2023) (Tuberculosis)	Virtual granuloma: vector of parameters for individual granuloma's immune response; initial conditions and grid configuration	*GranSim*, which encompasses all agent probabilities, dynamics and cell behaviors, agent interactions M_0 - No treatment M_P - With antibiotic treatment	Function designating granulomas as controlling, non-controlling, or sterilizing as a function of their end state; expected bacterial counts by subpopulation
Tuneable Resolution with GranSim (Segovia-Juarez et al., 2004; Fallahi-Sichani et al., 2012a; Fallahi-Sichani et al., 2012b; Kirschner et al., 2014; Pienaar et al., 2016) (Tuberculosis)	Virtual granuloma: vector of parameters for individual granuloma's immune response; initial conditions and grid configuration	*GranSim*, which encompasses all agent probabilities, dynamics and cell behaviors, agent interactions M_0 - Coarse grained representation of TNF-α, NF-κB, or metabolism M_P - Fine grained representation of TNF-α, NF-κB, or metabolism	Function designating granulomas as controlling, non-controlling, or sterilizing as a function of their end state; the predicted growth phenotypes of bacteria and activation levels of immune cells
Antibody-drug conjugate simulation (Menezes et al., 2020; Menezes et al., 2022) (Solid tumor)	Virtual tumor: vector of parameters for individual tumor's vascularization state, immune response, and initial grid conditions	Model that encompasses all agent probabilities, dynamics and cell behaviors, agent interactions M_0 - Untreated control M_P - Added antibody-drug conjugate treatment	Function designating tumors as growing or shrinking as a function of structure and cancerous cell count
Quantitative systems pharmacology simulation (Norton and Popel, 2014) (Hepatocellular carcinoma)	Virtual patient: vector of parameters for virtual patient's pharmacological parameters in the untreated case	Quantitative systems pharmacology simulation which describes immune-cancer interactions M_0 - Untreated control M_P - Added immune checkpoint inhibitors	Function quantifying the relative shrinkage of carcinoma with immune checkpoint inhibitors
Immune Response Agent-based Model (Cockrell and An, 2017; Larie et al., 2021) (Sepsis)	Virtual patient, wound, and environment: parameters determining of distributions of i) sustained endothelial tissue damage, ii) patient response thereto, iii) initial microstate, iv) external variables known to affect patient prognosis	Stochastic and mechanistic model of inflammatory response M_0 - Untreated control model M_P - Model of proposed treatment or medical intervention for clinical sepsis	Functions comparing the expectations, variances, and other descriptive distribution parameters of predicted oxygen deficit or cytokine levels in time with vs without treatment
Fibrin contraction simulation (Britton et al., 2019; Michael et al., 2023) (*in vitro* Blood clot contraction)	Virtual clot: collection of spatially arranged platelets embedded within a fibrin mesh	Subcellular element model that represents multiple platelets pulling on fibrin fibers to cause the emergent contraction of a blood clot M_0 - Untreated control M_P - Blebbistatin treatment to weaken contractile forces	Function quantifying the relative amount of contraction of the blood clot and distribution of multi-platelet clusters

We validated our virtual hosts at multiple scales according to the ten simple rules credibility standard (Erdemir et al., 2020; Fogarty et al., 2022; Nanda et al., 2023; Tatka et al., 2023). Figures 3A–E shows trajectories of 6,500 primary granulomas and whole-host CFU counts taken from 500 virtual hosts generated after we calibrated to multiple datasets from different NHP studies. At the population scale, clinically latent hosts had a range of 1–12 primary granulomas that eliminated all bacteria while

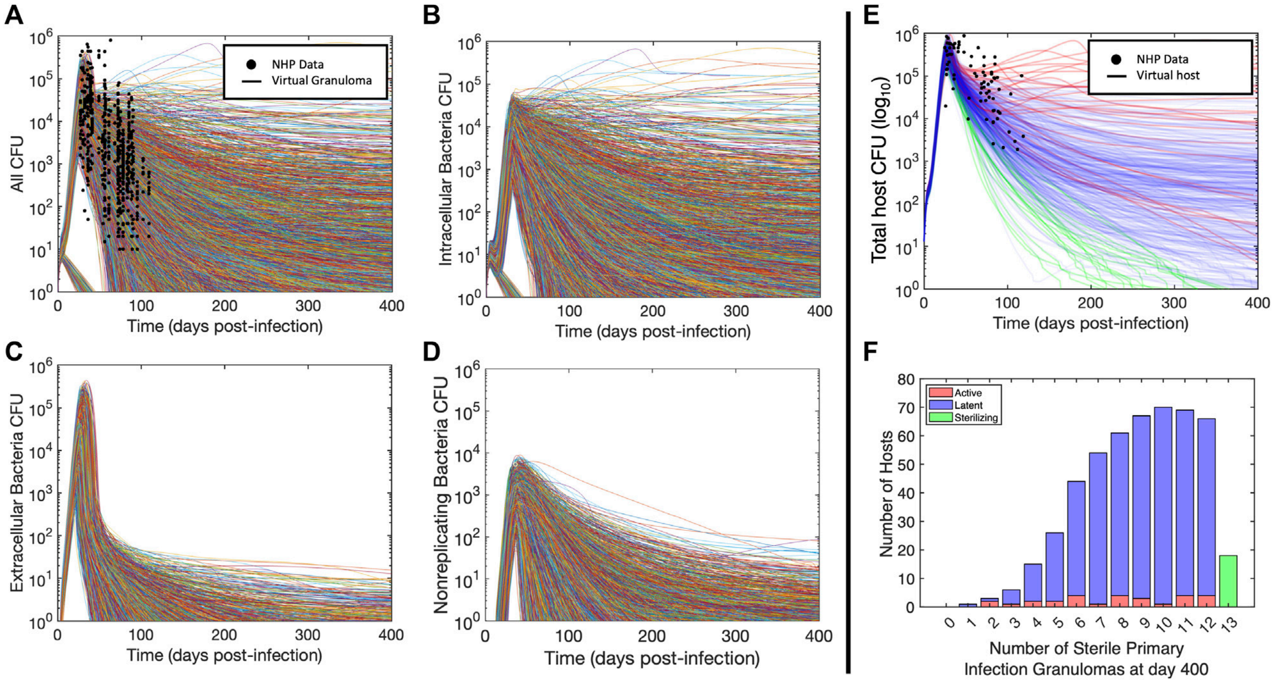


FIGURE 3
Virtual hosts and cohort for Mtb infection using *HostSim*. **(A–D)** Bacteria loads (CFU) for the total bacterial population and subpopulation trajectories for each granuloma. Curves showing granuloma CFU over time for each of the 13 primary granulomas in 500 hosts for 400 days post infection. Panel **(A)** shows total CFU per granuloma as well as the analogous measurements from NHPs at specific points (Gideon et al., 2015; Marino et al., 2016; Cadena et al., 2018; Darrah et al., 2019), **(B)** shows intracellular bacteria, **(C)** shows extracellular bacteria, and **(D)** shows nonreplicating bacteria. **(E)** Curves showing total lung CFU for each of 500 virtual hosts. Trajectories are colored by the virtual host classification as sterilizing, latent, or active. We have also show whole-lung CFU counts from published NHP studies (Gideon et al., 2015; Marino et al., 2016; Cadena et al., 2018; Darrah et al., 2019) by summing CFU across all lung granulomas. **(F)** Population scale histogram of the number of sterilizing granulomas per virtual host out of 500 virtual hosts.

hosts with active infection had 2–12 sterilized granulomas (Figure 3F). This recapitulates the common thinking that a single high-burden granuloma may determine the state of the Mtb host (i.e., active infection) (Lin et al., 2016). *HostSim* predicts that a small portion of granulomas are able to clear infection with innate response during early infection, which is presently not feasible to test *in vivo*. This is a feature of all of our computational and mathematical models of TB and it is believed to be a phenomenon that occurs in humans. On both the granuloma and host scale, we witness the presence of a transient high-CFU peak at approximately day 20, consistent with experimental observations (Gideon et al., 2015; Marino et al., 2016; Cadena et al., 2018; Darrah et al., 2019) (see Supplementary Material S1 Section 2 for details). Our updated *HostSim* model also is able to examine predictions that would match a PET-CT scan of a primate (human or NHP). We refer to this as FDG avidity, one of the only sources of time-series granuloma-scale data from live hosts and obtained via PET-CT scans (see Supplementary Material S1 Section 3 for details). FDG avidity is a measure of immune cell activity at the infection site within granulomas (Lin et al., 2013; Esmail et al., 2016). Supplementary Videos S1, S2 show the same representative latent virtual host developing granulomas over 400 days post-infection, with coloration based on their predicted FDG avidity values (comparable to NHP PET-CT images in Figure 1A of Ganchua et al. (Ganchua et al., 2018)).

3.3 Generating a virtual host *VH* at the whole-host scale and a virtual cohort $\{VH\}$ at the population scale

Our goal is to create a cohort of virtual hosts that mechanistically produce the trajectories of bacterial burden in time. We will use this virtual cohort to test interventions - either a treatment intervention, e.g., drugs, vaccines, etc. (or in some cases, a model modification). For an experimental treatment study, a cohort can be defined as both an infection population and an uninfected (negative) control population. Our "healthy" state is represented by steady-state levels of T-cells in the blood and LNs and resting macrophages within lungs (as we currently do not track host toxicity or tolerability in *HostSim*, we only use the infection model for drug studies). In our MID framework, we use a unique virtual population, our virtual cohort, on which we test our interventions to compare against the same virtual cohort against the untreated treatment control scenario M_0. To that end, we want to have a virtual cohort whose members, (i.e., the virtual hosts) have a meaningful identity that can be interpreted independently from any specific model version.

We represent our virtual hosts, members of our virtual cohort $\{VH\}$, by a collection of model parameters, P_i chosen from a set of biologically valid ranges. We created a virtual cohort of 500 virtual hosts in this way by sampling from calibrated experimental ranges, as described in Section 2. Since our model versions share all parameter fields (except intervention parameters A_i that do not vary between to virtual hosts), the same virtual host can be easily represented in either model version, known as the *virtual patient (host) identity*.

3.4 Drug interventions using *HostSim* (M_P)

A key aspect of creating a MID framework is to test interventions. For example, given the large drug regimen design space for diseases like TB, where multiple drugs are given for long periods of time, the possible combinations are on the order of 10^{17} (Cicchese et al., 2017)! The ability to explore the effects of drugs at the tissue, host, and population scales simultaneously in a virtual cohort is necessary to help screen this large space with the goal of identifying candidates that will be the best to test within a human cohort. A key goal of creating a MID framework is to use the impact of an intervention over multiple scales and to examine the wealth of synthetic data by comparing the outcomes of our virtual cohort with and without interventions.

To create an example intervention companion model M_P to *HostSim,* we will define a single-drug-like intervention. We will assume that these drugs solely affect either bacterial replication and/or death rates depending on their known drug actions (Figure 2). INH, BDQ and PZA are three antibiotics that are commonly used to treat TB (Chahine et al., 2014; Prideaux et al., 2015; Nahid et al., 2016; Sarathy et al., 2016). INH and BDQ are known to have bactericidal activity, although BDQ is more efficient at killing Mtb within the necrotic caseum region of granulomas and can also be taken inside of infected macrophages. PZA is a bacteriostatic drug whose action we represent by halving the bacterial replication rate (see Section 2.5). Our simple representations of drug interventions here do not include consideration of pharmacokinetics, or the ability of drugs to penetrate well into granulomas as we have done previously (Pienaar et al., 2017; Budak et al., 2023). We define our impact quantifications in this setting to be a host impact score H_S and a granuloma impact score G_S. These are derived from CFU ratios between M_P and M_0, where zero-score is zero-efficacy, and positive scores indicate a beneficial intervention for virtual hosts (see Section 2.6). Importantly, the outcomes being measured are credible from both M_0 and M_P respectively as i) mechanistically predicting CFU trajectories falls within their context of use, and ii) our goal for using our example M_P to calculate H_s and G_s is to examine qualitative behavior of CFU reduction by Mtb subpopulation.

We described above how we generate our virtual cohort $\{VH\}$. We then represent and simulate this virtual cohort in both the control version M_0 and drug intervention versions of *HostSim*: M_{INH}, M_{BDQ}, and M_{PZA}, and we calculate the granuloma and host impact scores (see Section 2.6, expressions [1] and [2]). Together, these components give us a way to study the impact of interventions on our virtual cohort, allowing us to analyze intervention efficacy across the cell/tissue, whole-host, and population scales.

3.5 Granuloma and host scale analyses of drug intervention capture mechanistic insights

As the final component of our MID framework, we want to understand how the perturbation or treatment M_P affects our virtual cohort over multiple physiological scales. With our drug interventions defined above, we use the impact quantification method described in Section 2 to compare outcomes of granulomas and hosts in the non-treatment control scenario against the three drug treatment scenarios. Figure 4 shows the impact quantification of the 3 different drug interventions at all three physiological scales. We begin treatment at day 200 and treat for 200 days post-infection. At all three scales, the impact scores suggest that M_{BDQ} is the most effective drug, which is consistent with how we defined it as compared to M_{INH}. Interestingly, there is a wide range of impact scores on both the granuloma and host scales, even if statistics on CFU counts at the population scale would not directly reveal this (Figure 4; Table 2). In many granulomas, treatment did not help much - indicated by an impact score near 0. Many granulomas with low impact scores either cleared in both model versions or cleared in the M_P version only (as a result of granulomas starting treatment with low CFU burden in the control case). However, we observed many granulomas with low impact scores (<0.5) that remained infected in both M_0 and M_P, indicating that some granulomas resist treatment more than others. This may depend on the action of a drug, on the host immune response or on the bacterial levels at the start of treatment. The population scale comparison between the control and intervention cases suggests that bactericidal interventions (as in the case of M_{BDQ} and M_{INH}) are a more effective action for a drug intervention (middle and bottom row panels). We observe, however, that the pooled cohort data (top row panels) cannot be used as accurately to predict whether or not a drug will help an individual host. This demonstrates the importance of developing a MID framework that captures both granuloma-scale and host-scale intervention responses that cannot be detected purely at a population level.

Another way to explore intervention impact scores is to understand variance of intervention efficacy. We analyzed host and granuloma impact scores as model outputs using a sensitivity analysis that considers non-linear correlations, called partial rank correlation (see Section 2.7). This method correlates non-linear model parameters to outputs of interest, and in this case, we can use both scale impact scores as a readout. The results shown in Table 3 suggest that many host model parameters impact the BDQ-like drug intervention. As BDQ is shown to have the largest possible intervention impact score of the three drugs that we studied (Figure 4; Table 2) as well as the widest variance of impact scores, we found it surprising that BDQ also interacts with the highest numbers of host parameters. It may be that interventions that interact with many model components may both reach higher efficacy but also have a more complex range of host responses. Moreover, we find that parameters that correlate with the impact of drug interventions also overlap with the parameters that impact FDG avidity outputs (i.e., a measure of host immune activity) (Supplementary Material S1 Section 3.2). What this tells us is that expressions FDG avidity, as predicted by expressions [S1-S2], is driven by the same parameters that drive our impact score. This may suggest that FDG avidity is a good predictor of projected intervention efficacy, or that both quantities are affected by the same mechanisms.

4 Discussion

We introduce a model analysis framework that can be used to track a virtual cohort and the impacts of interventions or other model perturbations across multiple physiological scales patient, that we refer to as a MID framework. The three components of a

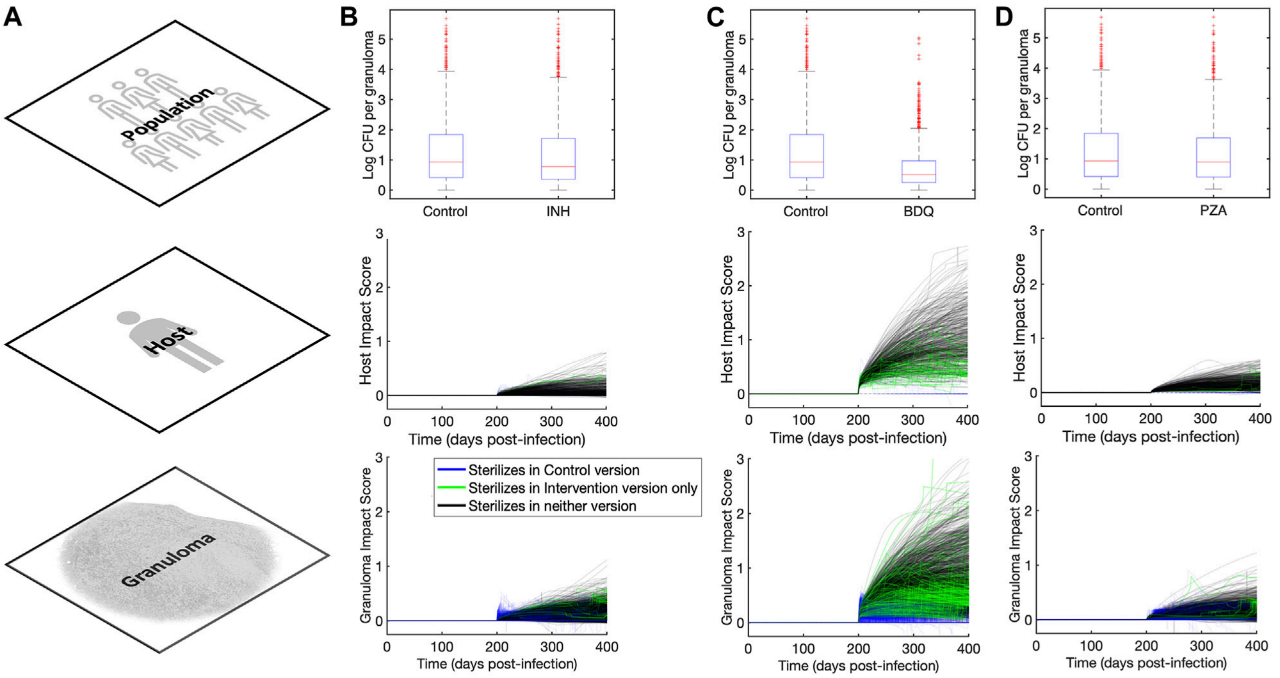


FIGURE 4
Impact quantification of three single-drug-like interventions across granuloma, host, and population scales. **(A)** Column showing the three scales (across rows) at which we analyze outcomes in our MID framework study. **(B–D)** Columns showing population, host, and granuloma scale impact quantification scores for **(B)** M_{INH}, **(C)** M_{BDQ}, and **(D)** M_{PZA} versions of *HostSim*. Granuloma and host scale plots show the granuloma and host impact scores (Eqs 1, 2) across time for each granuloma and host, respectively. An impact score of 0 indicates equal CFU in M_0 and M_P and higher impact scores indicate more favorable host outcomes. Blue lines show granuloma and host trajectories that are sterilized in the control group by day 400, green lines show granulomas and hosts that sterilized only in the intervention version, and black lines indicate trajectories that sterilized in neither control nor intervention cases. The population scale bar plots show a direct comparison between the control version and the intervention version at day 400, highlighting that the variation of the impact efficacy is obfuscated if individual host trajectories are not tracked.

TABLE 2 Impact of interventions of three different drugs on a virtual cohort with 500 hosts across multiple scales.

Feature	M_0	M_{INH}	M_{PZA}	M_{BDQ}
Percentage of sterilizing hosts in population	3.6%	4.2%	5.0%	12.0%
Percentage of hosts with active infection in population	5.6%	4.4%	4.0%	3.8%
Hosts that reduced CFU by 200 days post-intervention	-	53%	91%	96%
Granulomas that reduced CFU by 200 days post-intervention	-	16%	26%	32%
Granulomas that sterilized	67%	68%	69%	77%

TABLE 3 Descriptions of parameters significantly driving variance in granuloma impact scores for three different treatments. PRCC values remained unchanged qualitatively between days 200 and 400 so, for simplicity, only the trends are shown. We use + to indicate a positive correlation after intervention, and - to indicate a negative correlation, and "n/a" indicates no significant correlation. Trends indicated correspond to PRCC values that were filtered by PCC z-test (Marino et al., 2008) to control for the absolute magnitude of the intervention impact.

Parameter description	Efficacy correlation with $\boldsymbol{M_{INH}}$	Efficacy correlation with $\boldsymbol{M_{BDQ}}$	Efficacy correlation with M_{PZA}
In-macrophage carrying capacity of Mtb	++	++	++
Resting macrophage infection rate constant	n/a	+	++
Half-saturation of Mtb in infected macrophages	n/a	-	-
Decay rate constant of Interleukin-10	n/a	-	n/a

MID framework are i) a cohort of a virtual patients (or virtual hosts) consistent across model versions; ii) validated control and intervention model versions; and iii) an explicitly defined method of impact quantification. A MID framework leverages the ability of models to perform "what-if" experiments on the same virtual patient under different interventions and is able to decompose the spectrum of patient responses to predict system parameters - and thereby also individual model components - as being principally responsible for patient placement within a spectrum.

As part of creating a MID framework, we developed an updated version of our whole-host model of TB, *HostSim*, which ranges from the cell/tissue scale to the population scale. We calibrated this model to both experimental data from the Flynn lab (Gideon et al., 2015; Marino et al., 2016; Cadena et al., 2018; Darrah et al., 2019), and to synthetic data from our fine-grained model *GranSim*, which is an agent-based model that represents formation and function of individual granulomas. TB is an ideal candidate for implementation of a MID framework as it is complex and intrinsically multi-scale, which necessarily requires many parameters. Moreover, model outcomes from *HostSim* (e.g., CFU count and FDG avidity) are directly comparable to existing data and can be used to create and interrogate intuitive impact quantification measures.

We presented an example MID framework implementation to generate examples of quantitative, mechanism-based outcome predictions for interventions that are challenging to obtain experimentally and may be used to forecast outcome heterogeneity for future experiments. We used our TB-focused MID framework to analyze the impact of three different drug interventions–each of which phenomenologically represents a drug commonly used to treat TB–on a virtual cohort of 500 virtual hosts. In doing so, we captured and quantified the impact of different interventions at multiple scales, which is typically inaccessible to an experimental-like research design that usually occurs over a single scale. Our method shows that the parameters - and thereby mechanisms - most correlated with host responsiveness to drugs overlap with the parameters most that correlate with our prediction a non-invasive, spatial measurement of TB infection progression, FDG avidity.

Though we use a MID framework to study virtual human patients in the context of virtual clinical trials, the method is not tied to this application. Given a model system, one may develop intervention model versions for other forms of interventions after you have a suitable control version–e.g., host-directed therapies, vaccines, or booster efficacies. Indeed, there are existing model studies that employ virtual-cohort-like methods of analyses. However, without specific attention paid to each of the three components of a MID framework, *ad hoc* approaches may face i) an ill-defined notion of a virtual patient (or subject), such that it is difficult to determine whether the "same subject" is being represented in both model versions; ii) non-rectifiable or non-credible model versions, where the control version M_0 and the intervention version M_P are intractably different as in the case of a singular perturbation, and iii) an improperly constructed intervention quantification method which may bias or overly-abstract model outputs and thus preclude meaningful interpretation. Improper impact quantification selection may cause us to use model output outside of its context of use, and lead to subtly non-credible comparisons.

Another use of the MID framework may be to examine impacts of model updates, allowing us to demonstrate model consistency. If a model is updated extensively, we could use the original model as M_0 and the updated model as a new version $M_0^{'}$ instead of an M_P. In this case, minimal deviations would suggest that very little changed by way of introducing the new components–perhaps ideal for surrogate modeling, or more informative about the impact of fine-graining a model (Kirschner et al., 2014). Any simplification or re-representation of a model subcomponent could be examined in this way if model outputs and classifications are able to be meaningfully compared.

There are other advantages to having a MID framework. First, a calibrated virtual cohort annotated with MID-framework outcomes may be used to store virtual reference cases. That is, for computationally intensive models, it may be useful to store virtual hosts across a heterogeneous virtual cohort along with their control (M_0) and intervention (M_P) outcomes for comparison to quantitatively-similar real hosts. If a clinical patient or an experimental subject can be measured in such a way that we can find their nearest *digital partners*, then pre-simulated fine-grained virtual patients may be used to approximate both their untreated and treated outcomes. In this way, we may quantitatively rank the most effective treatment for a given real host, scaled with some confidence measure representing the "closeness" of the clinical host to their nearest digital partner. If the model is not entirely identifiable given live patient data, this will yield a twofold benefit: 1) a *family of nearest digital partners* identified by what data is available together with a forecast cone, which quantifies how those partners diverge over time; and 2) a clear and immediate use for new, multi-modal data. Including new data will whittle down the family of digital partners and narrow the forecast cone. We may also use the digital partner framework with existing models to best identify what modes of new data will best improve patient forecasting and illuminate what types of data will best improve parameter identification. This is particularly important when using mechanistic models for generating synthetic data for other applications (An and Cockrell, 2023). Lastly, we can continue to add more and more virtual patients to virtual cohorts as needed: generating virtual patients around a given human subject whose nearest digital partner lies in a sparsely-sampled region of parameter space will allow us to dynamically populate the virtual patient cohort to the needs of the real patient population.

Related efforts have been made to create tools that leverage computational models for medical decision supplementation and research (Vodovotz and An, 2019; Foy et al., 2020; Joshi et al., 2020; Wright and Davidson, 2020; Laubenbacher et al., 2022; Venkatapurapu et al., 2022), or for autonomous medical decision-making (Hoffmann et al., 2020; Singh et al., 2022; Yang, 2022) as a form of personalized medicine. A digital twin is a tool that predicts future states within a specific, real, complex biomedical system using a flexible, calibrated multi-scale computational model that integrates available real-time host-specific data. Medical digital twins (MDTs) have been developed to replicate and predict the trajectory of specific patients' diseases (Wright and Davidson, 2020; Masison et al., 2021; Laubenbacher et al., 2022). With recent demand for standardization of and development of MDT validation, uncertainty analysis, model linkage, and interpretable outcomes (Wright and Davidson, 2020; Laubenbacher et al., 2022), the ability to find digital partners within virtual cohorts created from digital twins and the associated response to treatment would be a powerful decision supplement tool.

It is worth noting the distinction between a MID framework and several related sensitivity analysis tools. Existing sensitivity analyses, including both local and global methods, uncover dependencies between a model's input variables and outcomes. While these tools (such as PRCC, used in this paper) are extremely valuable, they often include comparison of many parameter values distributed through a range. Often in experiments, intervention methods are defined regimens - a procedure applied to all subjects of the study (e.g., having multiple patients test the same FDA-approved drug dosage). In these cases, it is preferable to have an in-depth look at the impact of a single intervention regimen on an individual, as opposed to sampling a "gradient of intervention magnitude" - e.g., testing with/without drug, as opposed to various dosages. This is also true in the case of MDTs: having more detailed information on the projected impact of two mechanistically distinct interventions on a single patient may be invaluable. Moreover, M_P and M_0 may differ by more than a single parameter perturbation (e.g., a new cell type being considered in M_P). In these cases, comparison between M_0 and M_P is substantially distinct from a local sensitivity analysis.

Importantly, using a MID framework is not a substitute for rigorous and validated model construction, nor do we wish anybody to consider our MID framework as such. Instead, it is a method to analyze differences between two highly-related, credible, multi-scale models by separating out those components that are patient-specific and those components that are intervention-specific. Each individual model version should be considered as a trial procedure - such as experimental or a placebo group protocol-that is being applied to the *same* virtual host. Each model version should be able to make credible claims about host outcomes in each intervention scenario; and the MID framework is a systematic method for examining drivers of heterogeneity of the response to those interventions.

In the future, our *HostSim*-derived virtual cohort may be improved by the use of experimental distributions for each parameter in the model, instead of uniformly sampling from each range. This would ensure that the virtual cohorts in our MID framework capture the demographic of host heterogeneity in more detail. This may grant us more insights both into subtle differences between common presentations of TB at each scale, or it may allow us to predict outlier or unusual host presentations or responses. It is also worth noting that the three steps of creating a MID framework, while conceptually simple, must be considered carefully. Creation of intervention models may be straightforward in some cases, but there should be a limiting case where the control case can be recovered by reducing the intervention's amplitude. Representing a real-world entity (e.g., in the case of MDTs) in each model version may embed assumptions about that host that could be inconsistent between the model versions if each version's assumptions are not stated explicitly. Finally, the intervention impact quantification method should be free of biases that might favor one phenotype as more easily impacted than another and should not overreach the context of use of either model version.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

CM: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing–original draft, Writing–review and editing. SA: Software, Writing–review and editing. JL: Conceptualization, Formal Analysis, Funding acquisition, Methodology, Project administration, Supervision, Validation, Writing–original draft, Writing–review and editing. DK: Conceptualization, Formal Analysis, Funding acquisition, Methodology, Project administration, Resources, Supervision, Validation, Writing–original draft, Writing–review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was supported by NIH grants R01 AI50684 (DK) and U01 HL131072 (DK, JL) and is supported in part by funding from the Wellcome Leap Delta Tissue Program (DK, JL).

Acknowledgments

We would like to thank the lab of JoAnne L. Flynn for the use of her data and provision of the data given in Supplementary Material S1 Section 1; as well as Henry J. Borish for providing us with the updated spatial data for the NHP lung triangulation presented in Figure 2 and our Supplementary Videos. We also thank Paul Wolberg for his support in the construction and maintenance of *HostSim*.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fsysb.2023.1283341/full#supplementary-material

References

Aggarwal, R., and Ranganathan, P. (2019). Study designs: Part 4 - interventional studies. *Perspect. Clin. Res.* 10 (3), 137–139. doi:10.4103/picr.PICR_91_19

Aldieri, A., Curreli, C., Szyszko, J. A., La Mattina, A. A., and Viceconti, M. (2023). Credibility assessment of computational models according to ASME V&V40: application to the bologna biomechanical computed tomography solution. *Comput. Methods Programs Biomed.* 240, 107727. doi:10.1016/j.cmpb.2023.107727

An, G., and Cockrell, C. (2023). Generating synthetic multidimensional molecular time series data for machine learning: considerations. *Front. Syst. Biol.* 3. doi:10.3389/fsysb.2023.1188009

ASME (2018). *Assessing credibility of computational modeling through verification and validation: application to medical devices*. New York, NY, USA: ASME.

Barry, C. E., 3rd, Boshoff, H. I., Dartois, V., Dick, T., Ehrt, S., Flynn, J., et al. (2009). The spectrum of latent tuberculosis: rethinking the biology and intervention strategies. *Nat. Rev. Microbiol.* 7 (12), 845–855. doi:10.1038/nrmicro2236

Bergmann, F. T., Adams, R., Moodie, S., Cooper, J., Glont, M., Golebiewski, M., et al. (2014). COMBINE archive and OMEX format: one file to share all information to reproduce a modeling project. *BMC Bioinforma.* 15 (1), 369. doi:10.1186/s12859-014-0369-z

Bergmann, F. T., Nickerson, D., Waltemath, D., and Scharm, M. (2017). SED-ML web tools: generate, modify and export standard-compliant simulation studies. *Bioinformatics* 33 (8), 1253–1254. doi:10.1093/bioinformatics/btw812

Blinov, M. L., Ruebenacker, O., and Moraru, , II (2008). Complexity and modularity of intracellular networks: a systematic approach for modelling and simulation. *IET Syst. Biol.* 2 (5), 363–368. doi:10.1049/iet-syb:20080092

Britton, S., Kim, O., Pancaldi, F., Xu, Z., Litvinov, R. I., Weisel, J. W., et al. (2019). Contribution of nascent cohesive fiber-fiber interactions to the non-linear elasticity of fibrin networks under tensile load. *Acta Biomater.* 94, 514–523. doi:10.1016/j.actbio.2019.05.068

Budak, M., Cicchese, J. M., Maiello, P., Borish, H. J., White, A. G., Chishti, H. B., et al. (2023). Optimizing tuberculosis treatment efficacy: comparing the standard regimen with Moxifloxacin-containing regimens. *PLOS Comput. Biol.* 19 (6), e1010823. doi:10.1371/journal.pcbi.1010823

Cacuci, D. G., and Ionescu-Bujor, M. (2004). A comparative review of sensitivity and uncertainty analysis of large-scale systems - II: statistical methods. *Nucl. Sci. Eng.* 147 (3), 204–217. doi:10.13182/04-54cr

Cadena, A. M., Fortune, S. M., and Flynn, J. L. (2017). Heterogeneity in tuberculosis. *Nat. Rev. Immunol.* 17 (11), 691–702. doi:10.1038/nri.2017.69

Cadena, A. M., Hopkins, F. F., Maiello, P., Carey, A. F., Wong, E. A., Martin, C. J., et al. (2018). Concurrent infection with *Mycobacterium tuberculosis* confers robust protection against secondary infection in macaques. *PLOS Pathog.* 14 (10), e1007305. doi:10.1371/journal.ppat.1007305

Chahine, E. B., Karaoui, L. R., and Mansour, H. (2014). Bedaquiline: a novel diarylquinoline for multidrug-resistant tuberculosis. *Ann. Pharmacother.* 48 (1), 107–115. doi:10.1177/1060028013504087

Cicchese, J. M., Dartois, V., Kirschner, D. E., and Linderman, J. J. (2020). Both pharmacokinetic variability and granuloma heterogeneity impact the ability of the first-line antibiotics to sterilize tuberculosis granulomas. *PLoS Comput. Biol.* 11, 333. doi:10.3389/fphar.2020.00333

Cicchese, J. M., Pienaar, E., Kirschner, D. E., and Linderman, J. J. (2017). Applying optimization algorithms to tuberculosis antibiotic treatment regimens. *Cell Mol. Bioeng.* 10 (6), 523–535. doi:10.1007/s12195-017-0507-6

Cockrell, C., and An, G. (2017). Sepsis reconsidered: identifying novel metrics for behavioral landscape characterization with a high-performance computing implementation of an agent-based model. *J. Theor. Biol.* 430, 157–168. doi:10.1016/j.jtbi.2017.07.016

Coleman, M. T., Maiello, P., Tomko, J., Frye, L. J., Fillmore, D., Janssen, C., et al. (2014). Early Changes by (18)Fluorodeoxyglucose positron emission tomography coregistered with computed tomography predict outcome after *Mycobacterium tuberculosis* infection in cynomolgus macaques. *Infect. Immun.* 82 (6), 2400–2404. doi:10.1128/IAI.01599-13

Darrah, P. A., DiFazio, R. M., Maiello, P., Gideon, H. P., Myers, A. J., Rodgers, M. A., et al. (2019). Boosting BCG with proteins or rAd5 does not enhance protection against tuberculosis in rhesus macaques. *NPJ Vaccines* 4, 21. doi:10.1038/s41541-019-0113-9

Dhillon, J., Andries, K., Phillips, P. P. J., and Mitchison, D. A. (2010). Bactericidal activity of the diarylquinoline TMC207 against *Mycobacterium tuberculosis* outside and within cells. *Tuberculosis* 90 (5), 301–305. doi:10.1016/j.tube.2010.07.004

Erdemir, A., Mulugeta, L., Ku, J. P., Drach, A., Horner, M., Morrison, T. M., et al. (2020). Credible practice of modeling and simulation in healthcare: ten rules from a multidisciplinary perspective. *J. Transl. Med.* 18 (1), 369. doi:10.1186/s12967-020-02540-4

Esmail, H., Lai, R. P., Lesosky, M., Wilkinson, K. A., Graham, C. M., Coussens, A. K., et al. (2016). Characterization of progressive HIV-associated tuberculosis using 2-deoxy-2-[18F]fluoro-D-glucose positron emission and computed tomography. *Nat. Med.* 22 (10), 1090–1093. doi:10.1038/nm.4161

Fallahi-Sichani, M., Flynn, J. L., Linderman, J. J., and Kirschner, D. E. (2012a). Differential risk of tuberculosis reactivation among anti-TNF therapies is due to drug binding kinetics and permeability. *J. Immunol.* 188 (7), 3169–3178. doi:10.4049/jimmunol.1103298

Fallahi-Sichani, M., Kirschner, D. E., and Linderman, J. J. (2012b). NF-κB signaling dynamics play a key role in infection control in tuberculosis. *Front. Physiol.* 3, 170. doi:10.3389/fphys.2012.00170

Fogarty, L., Ammar, M., Holding, T., Powell, A., and Kandler, A. (2022). Ten simple rules for principled simulation modelling. *PLOS Comput. Biol.* 18 (3), e1009917. doi:10.1371/journal.pcbi.1009917

Foy, B. H., Gonçalves, B. P., and Higgins, J. M. (2020). Unraveling disease pathophysiology with mathematical modeling. *Annu. Rev. Pathology Mech. Dis.* 15 (1), 371–394. doi:10.1146/annurev-pathmechdis-012419-032557

Ganchua, S. K. C., Cadena, A. M., Maiello, P., Gideon, H. P., Myers, A. J., Junecko, B. F., et al. (2018). Lymph nodes are sites of prolonged bacterial persistence during *Mycobacterium tuberculosis* infection in macaques. *PLOS Pathog.* 14 (11), e1007337. doi:10.1371/journal.ppat.1007337

Ganchua, S. K. C., White, A. G., Klein, E. C., and Flynn, J. L. (2020). Lymph nodes-The neglected battlefield in tuberculosis. *PLoS Pathog.* 16 (8), e1008632. doi:10.1371/journal.ppat.1008632

Ghaffarizadeh, A., Heiland, R., Friedman, S. H., Mumenthaler, S. M., and PhysiCell, M. P. (2018). PhysiCell: an open source physics-based cell simulator for 3-D multicellular systems. *PLOS Comput. Biol.* 14 (2), e1005991. doi:10.1371/journal.pcbi.1005991

Gideon, H. P., Phuah, J., Myers, A. J., Bryson, B. D., Rodgers, M. A., Coleman, M. T., et al. (2015). Variability in tuberculosis granuloma T cell responses exists, but a balance of pro- and anti-inflammatory cytokines is associated with sterilization. *PLoS Pathog.* 11 (1), e1004603. doi:10.1371/journal.ppat.1004603

Grant, N. L., Maiello, P., Klein, E., Lin, P. L., Borish, H. J., Tomko, J., et al. (2022). T cell transcription factor expression evolves over time in granulomas from Mycobacterium tuberculosis-infected cynomolgus macaques. *Cell Rep.* 39 (7), 110826. doi:10.1016/j.celrep.2022.110826

Guzzetta, G., Ajelli, M., Yang, Z., Mukasa, L. N., Patil, N., Bates, J. H., et al. (2015). Effectiveness of contact investigations for tuberculosis control in Arkansas. *J. Theor. Biol.* 380, 238–246. doi:10.1016/j.jtbi.2015.05.031

Helton, J. C., and Davis, F. J. (2003). Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. *Reliab Eng. Syst. Safe* 81 (1), 23–69. doi:10.1016/s0951-8320(03)00058-9

Hoffmann, K., Cazemier, K., Baldow, C., Schuster, S., Kheifetz, Y., Schirm, S., et al. (2020). Integration of mathematical model predictions into routine workflows to support clinical decision making in haematology. *BMC Med. Inf. Decis. Mak.* 20 (1), 28. doi:10.1186/s12911-020-1039-x

Jayaram, R., Shandil, R. K., Gaonkar, S., Kaur, P., Suresh, B. L., Mahesh, B. N., et al. (2004). Isoniazid pharmacokinetics-pharmacodynamics in an aerosol infection model of tuberculosis. *Antimicrob. Agents Chemother.* 48 (8), 2951–2957. doi:10.1128/AAC.48.8.2951-2957.2004

Joshi, A., Wang, D. H., Watterson, S., McClean, P. L., Behera, C. K., Sharp, T., et al. (2020). Opportunities for multiscale computational modelling of serotonergic drug effects in Alzheimer's disease. *Neuropharmacology* 174, 108118. doi:10.1016/j.neuropharm.2020.108118

Joslyn, L., Kirschner, D., and Linderman, J. (2023). CaliPro: a calibration protocol that utilizes parameter density estimation to explore parameter space and calibrate complex biological models. *Cel. Mol. Bioeng.* 14, 31–47. doi:10.1007/s12195-020-00650-z

Joslyn, L. R., Flynn, J. L., Kirschner, D. E., and Linderman, J. J. (2022a). Concomitant immunity to *M. tuberculosis* infection. *Sci. Rep.* 12 (1), 20731. doi:10.1038/s41598-022-24516-8

Joslyn, L. R., Linderman, J. J., and Kirschner, D. E. (2022b). A virtual host model of *Mycobacterium tuberculosis* infection identifies early immune events as predictive of infection outcomes. *J. Theor. Biol.* 539, 111042. doi:10.1016/j.jtbi.2022.111042

Keating, S. M., Waltemath, D., König, M., Zhang, F., Dräger, A., Chaouiya, C., et al. (2020). SBML Level 3: an extensible format for the exchange and reuse of biological models. *Mol. Syst. Biol.* 16 (8), e9110. doi:10.15252/msb.20199110

Kirschner, D. E., Hunt, C. A., Marino, S., Fallahi-Sichani, M., and Linderman, J. J. (2014). Tuneable resolution as a systems biology approach for multi-scale, multi-compartment computational models. *Wiley Interdiscip. Rev. Syst. Biol. Med.* 6 (4), 289–309. doi:10.1002/wsbm.1270

Larie, D., An, G., and Cockrell, R. C. (2021). The use of artificial neural networks to forecast the behavior of agent-based models of pathophysiology: an example utilizing an agent-based model of sepsis. *Front. Physiology* 12, 716434. doi:10.3389/fphys.2021.716434

Laubenbacher, R., Niarakis, A., Helikar, T., An, G., Shapiro, B., Malik-Sheriff, R. S., et al. (2022). Building digital twins of the human immune system: toward a roadmap. *npj Digit. Med.* 5 (1), 64. doi:10.1038/s41746-022-00610-z

Lin, P. L., Coleman, T., Carney, J. P., Lopresti, B. J., Tomko, J., Fillmore, D., et al. (2013). Radiologic responses in cynomolgus macaques for assessing tuberculosis chemotherapy regimens. *Antimicrob. agents Chemother.* 57, 4237–4244. doi:10.1128/AAC.00277-13

Lin, P. L., and Flynn, J. L. (2018). The end of the binary era: revisiting the spectrum of tuberculosis. *J. Immunol.* 201 (9), 2541–2548. doi:10.4049/jimmunol.1800993

Lin, P. L., Maiello, P., Gideon, H. P., Coleman, M. T., Cadena, A. M., Rodgers, M. A., et al. (2016). PET CT identifies reactivation risk in cynomolgus macaques with latent *M. tuberculosis*. M. tuberculosis. *PLoS Pathog.* 12 (7), e1005739. doi:10.1371/journal.ppat.1005739

Lyadova, I. V. (2017). Neutrophils in tuberculosis: heterogeneity Shapes the way? *Mediat. Inflamm.* 2017, 8619307. doi:10.1155/2017/8619307

Marino, S., Gideon, H. P., Gong, C., Mankad, S., McCrone, J. T., Lin, P. L., et al. (2016). Computational and empirical studies predict Mycobacterium tuberculosis-specific T cells as a biomarker for infection outcome. *PLoS Comput. Biol.* 12 (4), e1004804. doi:10.1371/journal.pcbi.1004804

Marino, S., Hogue, I. B., Ray, C. J., and Kirschner, D. E. (2008). A methodology for performing global uncertainty and sensitivity analysis in systems biology. *J. Theor. Biol.* 254 (1), 178–196. doi:10.1016/j.jtbi.2008.04.011

Martin, C. J., Cadena, A. M., Leung, V. W., Lin, P. L., Maiello, P., Hicks, N., et al. (2017). Digitally barcoding *Mycobacterium tuberculosis* reveals *in vivo* infection dynamics in the macaque model of tuberculosis. *MBio* 8 (3), e00312-17. doi:10.1128/mBio.00312-17

Masison, J., Beezley, J., Mei, Y., Ribeiro, H., Knapp, A. C., Sordo Vieira, L., et al. (2021). A modular computational framework for medical digital twins. *Proc. Natl. Acad. Sci. U. S. A.* 118 (20), e2024287118. doi:10.1073/pnas.2024287118

Menezes, B., Cilliers, C., Wessler, T., Thurber, G. M., and Linderman, J. J. (2020). An agent-based systems pharmacology model of the antibody-drug conjugate kadcyla to predict efficacy of different dosing regimens. *AAPS J.* 22 (2), 29. doi:10.1208/s12248-019-0391-1

Menezes, B., Linderman, J. J., and Thurber, G. M. (2022). Simulating the selection of resistant cells with bystander killing and antibody coadministration in heterogeneous human epidermal growth factor receptor 2-positive tumors. *Drug Metab. Dispos.* 50 (1), 8–16. doi:10.1124/dmd.121.000503

Michael, C., Pancaldi, F., Britton, S., Kim, O. V., Peshkova, A. D., Vo, K., et al. (2023). Combined computational modeling and experimental study of the biomechanical mechanisms of platelet-driven contraction of fibrin clots. *Commun. Biol.* 6 (1), 869. doi:10.1038/s42003-023-05240-z

Nahid, P., Dorman, S. E., Alipanah, N., Barry, P. M., Brozek, J. L., Cattamanchi, A., et al. (2016). Executive summary: official American thoracic society/centers for disease control and prevention/infectious diseases society of America clinical practice guidelines: treatment of drug-susceptible tuberculosis. *Clin. Infect. Dis.* 63 (7), 853–867. doi:10.1093/cid/ciw566

Nanda, P., Budak, M., Michael, C. T., Krupinsky, K., and Kirschner, D. E. (2023). Development and analysis of multiscale models for tuberculosis: from molecules to populations. *bioRxiv* 11, 2023.11.13.566861. doi:10.1101/2023.11.13.566861

NASA (2016). *Standard for models and simulation*. Washington, D.C., United States: NASA.

Neal, M. L., Gennari, J. H., Waltemath, D., Nickerson, D. P., and König, M. (2020). Open modeling and exchange (OMEX) metadata specification version 1.0. *J. Integr. Bioinforma.* 17 (2-3), 20200020. doi:10.1515/jib-2020-0020

Norton, K. A., and Popel, A. S. (2014). An agent-based model of cancer stem cell initiated avascular tumour growth and metastasis: the effect of seeding frequency and location. *J. R. Soc. Interface* 11 (100), 20140640. doi:10.1098/rsif.2014.0640

Pienaar, E., Matern, W. M., Linderman, J. J., Bader, J. S., and Kirschner, D. E. (2016). Multiscale model of *Mycobacterium tuberculosis* infection maps metabolite and gene perturbations to granuloma sterilization predictions. *Infect. Immun.* 84 (5), 1650–1669. doi:10.1128/IAI.01438-15

Pienaar, E., Sarathy, J., Prideaux, B., Dietzold, J., Dartois, V., Kirschner, D. E., et al. (2017). Comparing efficacies of moxifloxacin, levofloxacin and gatifloxacin in tuberculosis granulomas using a multi-scale systems pharmacology approach. *PLoS Comput. Biol.* 13 (8), e1005650. doi:10.1371/journal.pcbi.1005650

Poplawski, N. J., Shirinifard, A., Swat, M., and Glazier, J. A. (2008). Simulation of single-species bacterial-biofilm growth using the Glazier-Graner-Hogeweg model and the CompuCell3D modeling environment. *Math. Biosci. Eng.* 5 (2), 355–388. doi:10.3934/mbe.2008.5.355

Portevin, D., Moukambi, F., Clowes, P., Bauer, A., Chachage, M., Ntinginya, N. E., et al. (2014). Assessment of the novel T-cell activation marker-tuberculosis assay for diagnosis of active tuberculosis in children: a prospective proof-of-concept study. *Lancet Infect. Dis.* 14 (10), 931–938. doi:10.1016/S1473-3099(14)70884-9

Prideaux, B., Via, L. E., Zimmerman, M. D., Eum, S., Sarathy, J., O'Brien, P., et al. (2015). The association between sterilizing activity and drug distribution into tuberculosis lesions. *Nat. Med.* 21 (10), 1223–1227. doi:10.1038/nm.3937

Renardy, M., Joslyn, L. R., Millar, J. A., and Kirschner, D. E. (2021). To Sobol or not to Sobol? The effects of sampling schemes in systems biology applications. *Math. Biosci.* 337, 108593. doi:10.1016/j.mbs.2021.108593

Renardy, M., Wessler, T., Blemker, S., Linderman, J., Peirce, S., and Kirschner, D. (2019). Data-driven model validation across dimensions. *Bull. Math. Biol.* 81 (6), 1853–1866. doi:10.1007/s11538-019-00590-4

Sarathy, J., Blanc, L., Alvarez-Cabrera, N., O'Brien, P., Dias-Freedman, I., Mina, M., et al. (2019). Fluoroquinolone efficacy against tuberculosis is driven by penetration into lesions and activity against resident bacterial populations. *Antimicrob. Agents Chemother.* 63 (5), e02516-18. doi:10.1128/AAC.02516-18

Sarathy, J. P., Via, L. E., Weiner, D., Blanc, L., Boshoff, H., Eugenin, E. A., et al. (2018). Extreme drug tolerance of *Mycobacterium tuberculosis* in caseum. *Antimicrob. Agents Chemother.* 62 (2), e02266-17. doi:10.1128/AAC.02266-17

Sarathy, J. P., Zuccotto, F., Hsinpin, H., Sandberg, L., Via, L. E., Marriner, G. A., et al. (2016). Prediction of drug penetration in tuberculosis lesions. *ACS Infect. Dis.* 2 (8), 552–563. doi:10.1021/acsinfecdis.6b00051

Schaff, J. C., Vasilescu, D., Moraru, II, Loew, L. M., and Blinov, M. L. (2016). Rule-based modeling with virtual cell. *Bioinformatics* 32 (18), 2880–2882. doi:10.1093/bioinformatics/btw353

Segovia-Juarez, J. L., Ganguli, S., and Kirschner, D. (2004). Identifying control mechanisms of granuloma formation during *M. tuberculosis* infection using an agent-based model. *J. Theor. Biol.* 231 (3), 357–376. doi:10.1016/j.jtbi.2004.06.031

Shirinifard, A., Gens, J. S., Zaitlen, B. L., Poplawski, N. J., Swat, M., and Glazier, J. A. (2009). 3D multi-cell simulation of tumor growth and angiogenesis. *PLoS One* 4 (10), e7190. doi:10.1371/journal.pone.0007190

Singh, D., Nagaraj, S., Mashouri, P., Drysdale, E., Fischer, J., Goldenberg, A., et al. (2022). Assessment of machine learning–based medical directives to expedite care in pediatric emergency medicine. *JAMA Netw. Open* 5 (3), e222599. doi:10.1001/jamanetworkopen.2022.2599

Smith, L. P., Bergmann, F. T., Garny, A., Helikar, T., Karr, J., Nickerson, D., et al. (2021). The simulation experiment description markup language (SED-ML): language specification for level 1 version 4. *J. Integr. Bioinforma.* 18 (3), . doi:10.1515/jib-2021-0021

Tatka, L. T., Smith, L. P., Hellerstein, J. L., and Sauro, H. M. (2023). Adapting modeling and simulation credibility standards to computational systems biology. *J. Transl. Med.* 21 (1), 501. doi:10.1186/s12967-023-04290-5

Venkatapurapu, S. P., Iwakiri, R., Udagawa, E., Patidar, N., Qi, Z., Takayama, R., et al. (2022). A computational platform integrating a mechanistic model of crohn's disease for predicting temporal progression of mucosal damage and healing. *Adv. Ther.* 39 (7), 3225–3247. doi:10.1007/s12325-022-02144-y

Vodovotz, Y., and An, G. (2019). Agent-based models of inflammation in translational systems biology: a decade later. *WIREs Syst. Biol. Med.* 11 (6), e1460. doi:10.1002/wsbm.1460

Warsinske, H. C., Pienaar, E., Linderman, J. J., Mattila, J. T., and Kirschner, D. E. (2017). Deletion of TGF-β1 increases bacterial clearance by cytotoxic T cells in a tuberculosis granuloma model. *Front. Immunol.* 8, 1843. doi:10.3389/fimmu.2017.01843

WHO (2020). *Global tuberculosis report 2020*. Geneva, Switzerland: World Health Organization.

WHO (2022). *Global tuberculosis report 2022*. Geneva, Switzerland: World Health Organization.

Wong, E. A., Evans, S., Kraus, C. R., Engelman, K. D., Maiello, P., Flores, W. J., et al. (2020). IL-10 impairs local immune response in lung granulomas and lymph nodes during early *Mycobacterium tuberculosis* infection. *Mycobacterium Tuberc. Infect.* 204 (3), 644–659. doi:10.4049/jimmunol.1901211

Wright, L., and Davidson, S. (2020). How to tell the difference between a model and a digital twin. *Adv. Model. Simul. Eng. Sci.* 7 (1), 13. doi:10.1186/s40323-020-00147-4

Yang, C. C. (2022). Explainable artificial intelligence for predictive modeling in healthcare. *J. Healthc. Inf. Res.* 6 (2), 228–239. doi:10.1007/s41666-022-00114-1

Zhang, Y., and Mitchison, D. (2003). The curious characteristics of pyrazinamide: a review. *Int. J. Tuberc. Lung Dis.* 7 (1), 6–21.

frontiers | Frontiers in Systems Biology

TYPE Perspective
PUBLISHED 27 February 2024
DOI 10.3389/fsysb.2024.1363884

OPEN ACCESS

EDITED BY
Edoardo Saccenti,
Wageningen University and Research,
Netherlands

REVIEWED BY
Daniela Besozzi,
University of Milano-Bicocca, Italy
Guillermo Ramirez-Santiago,
National Autonomous University of Mexico,
Mexico

*CORRESPONDENCE
Rahuman S. Malik Sheriff,
sheriff@ebi.ac.uk

RECEIVED 31 December 2023
ACCEPTED 09 February 2024
PUBLISHED 27 February 2024

CITATION
Malik Sheriff RS, Asari H, Hermjakob H, Huber W, Quail T, Santos SDM, Smith AM and Uhlmann V (2024), BioModels' Model of the Year 2023. *Front. Syst. Bisol.* 4:1363884.
doi: 10.3389/fsysb.2024.1363884

BioModels' Model of the Year 2023

Rahuman S. Malik Sheriff[1]*, Hiroki Asari[2], Henning Hermjakob[1], Wolfgang Huber[3], Thomas Quail[4], Silvia D. M. Santos[5], Amber M. Smith[6] and Virginie Uhlmann[1]

[1]European Bioinformatics Institute, European Molecular Biology Laboratory (EMBL-EBI), Cambridge, United Kingdom, [2]Epigenetics and Neurobiology Unit, EMBL Rome, European Molecular Biology Laboratory, Monterotondo, Italy, [3]Genome Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany, [4]Cell Biology and Biophysics Unit, European Molecular Biology Laboratory, Heidelberg, Germany, [5]Quantitative Stem Cell Biology Laboratory, The Francis Crick Institute, London, United Kingdom, [6]Department of Pediatrics, University of Tennessee Health Science Center, Memphis, TN, United States

Mathematical modeling is a pivotal tool for deciphering the complexities of biological systems and their control mechanisms, providing substantial benefits for industrial applications and answering relevant biological questions. BioModels' Model of the Year 2023 competition was established to recognize and highlight exciting modeling-based research in the life sciences, particularly by non-independent early-career researchers. It further aims to endorse reproducibility and FAIR principles of model sharing among these researchers. We here delineate the competition's criteria for participation and selection, introduce the award recipients, and provide an overview of their contributions. Their models provide crucial insights into cell division regulation, protein stability, and cell fate determination, illustrating the role of mathematical modeling in advancing biological research.

KEYWORDS

mathematical modeling, BioModels, competition, cell cycle, protein turnover, cell-to-cell variability

1 Introduction

The mathematical modeling of biological systems plays a crucial role in understanding complex processes, their regulation, and answering relevant biological questions. It offers a broad spectrum of industrial applications. To recognize and encourage advancements in this field, BioModels (Malik-Sheriff et al., 2020) launched the "Model of the Year 2023" competition, with supporters such as the EMBL Theory Transversal Theme, SBML, COMBINE, and ICSB. The competition aims to highlight the work of non-independent early-career researchers who have made significant contributions within the last two years (2021–2022) to the field through exciting mathematical modeling-based research. Applications were accepted from researchers including, but not limited to, PhD students, postdocs, staff scientists, and research assistants from both academia and industry. Models developed by PIs before becoming independent were also considered.

To facilitate the sharing of the model with the broader community and their potential reuse, competition participants were required to submit their models to public model-sharing repositories such as BioModels, CellCollective (Helikar et al., 2012), and Physiome (Yu et al., 2011). The Physiome repository provides a curated collection of physiological models primarily in CellML format, whereas CellCollective is a collaborative modeling building and sharing platform. BioModels is a leading repository of curated biological models which allows the submission of models in diverse modeling formats that are further manually reproduced and semantically annotated. The participants had the flexibility to submit their models in any format, including SBML, CellML, COMBINE archive, MATLAB, Mathematica, R, Python, or C++ (Malik-Sheriff et al., 2020), to enter the competition. However, participants were encouraged to submit models with well-commented or annotated code that adhered to MIRIAM guidelines (Le Novère et al., 2005).

The selection process had a strong emphasis on scientific excellence, along with the ability of the models to yield insights into complex biological phenomena or practical applications. Factors such as model reproducibility, adherence to community standards, and good code-sharing practices were also considered. The top models, regardless of their submission format—whether in a community standard like SBML or as documented code—underwent a manual verification process to ensure that their results could be reproduced. Any non-reproducible models were disqualified. Among the 25 submissions, the winning models listed below were selected on the basis of the above criteria.

- **Dr Jan Rombouts** Advisor: Gelens, L., KU Leuven, Belgium "Modular approach to modeling the cell cycle" (De Boeck et al., 2021) BioModels submission: BIOMD0000001079 BIOMD0000001080.
- **Dr Eva-Maria Geissen** Advisor: Hammarén, H.M., EMBL, Germany "Protein turnover and post-translational modification" (Hammarén et al., 2022) BioModels submission BIOMD0000001078.
- **Dr Lorenz Adlung** Advisor: Schilling M, DKFZ, Germany "Cell-to-cell variability in JAK2/STAT5 pathway" (Adlung et al., 2021) BioModels submission: BIOMD0000001077.

2 Overview of winning models

2.1 Modular approach to modeling the cell cycle

These models address the fundamental biological question of how cells regulate their division cycle. The mathematical crux of the model lies in its simulation of bistable switches, which are critical for understanding the robust and rapid transitions between different phases of the cell cycle. The models capture the essence of these bistable switches by applying a modified Hill-type ultrasensitive response. This approach allows the exploration of how cellular mechanisms, like the accumulation and degradation of cyclin B, govern the cell cycle. This approach is illustrated in two models: (1) the early embryonic cell cycle of *Xenopus laevis* (BIOMD0000001079) and (2) the somatic cell cycle with different cell cycle phases (BIOMD0000001080). The models effectively decipher the intricate balance and feedback loops involved in cell cycle regulation and have the potential to offer a profound understanding of cell proliferation and its dysregulation in diseases.

2.2 Protein turnover and post-translational modification

This model helps us understand whether protein turnover data from metabolic labeling experiments can reveal the impact of post-translational modifications (PTMs) on protein stability. Through its reaction rate equations framework, the model dissects the influence of the dynamics of interconvertible proteo-forms—different forms of the same protein differentiated only by their PTMs—on the measured protein turnover dynamics. The model revealed that these dynamics mask the actual stability-related dynamics of proteins. However, the model highlighted the order of PTM addition and/or removal relative to protein synthesis. This insight is vital to the accurate interpretation of PTM-resolved turnover data and an understanding of protein modification in the context of its lifecycle.

2.3 Cell-to-cell variability in JAK2/STAT5 pathway

This model addresses the crucial question of how erythroid progenitor cells decide between proliferation, differentiation, and apoptosis. The model employs a sophisticated series of coupled ordinary differential equations to unravel the JAK2/STAT5 signaling pathway's role in this decision-making process. The mathematical modeling here is pivotal in identifying the specific thresholds of STAT5 activation that determine cell fate and it addresses a significant gap in our understanding of erythropoiesis. The model's strength lies in its ability to handle the inherent cell-to-cell variability within a population, providing insights critical for developing targeted therapies for blood disorders.

3 Discussion

Through the "Model of the Year 2023" competition and its next edition, "Model of the Year 2024", BioModels aims to recognize outstanding contributions from early-career researchers to systems biology modeling and highlight the crucial role that these models play in answering fundamental biological questions or industrial applications. Each winning model exemplifies how mathematical modeling can be harnessed to dissect complex biological processes, providing insights that are pivotal for both basic biological understanding and potential therapeutic applications. These models are testament to the potential of integrating mathematical modeling with biological research.

Data availability statement

The codes of models presented in the article are publicly available in the BioModels repository; further inquiries can be directed to the corresponding author.

Author contributions

RMS: writing—original draft and writing-review and editing. HA: writing—original draft and writing-review and editing. HH: writing—original draft and writing-review and editing. WH: writing—original draft and writing-review and editing. TQ: writing—original draft and writing-review and editing. SS: writing—original draft and writing-review and editing. AS: writing—original draft and writing-review and editing. VU: writing—original draft and writing-review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. The competition was supported by EMBL Core Funding. SS at the Francis Crick Institute is supported from core funding from Cancer Research UK (FC0010596), the UK Medical Research Council (FC0010596) and the Wellcome Trust (FC0010596).

Acknowledgments

We thank Krishna Tiwari and Miroslav (Mirek) Kratochvil for their support in assessing the model code(s); organisers of ICSB 2022, ICSB 2023 as well as COMBINE, and SBML communities. RMS co-ordinated the MOY2023 competition; HA, WH, TQ, SS, AS, and VU served on the selection panel for the competition; HH supported the implementation of the competition.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers at the time of submission. This had no impact on the peer review process and the final decision.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article or claim that may be made by its manufacturer is not guaranteed or endorsed by the publisher.

References

Adlung, L., Stapor, P., Tönsing, C., Schmiester, L., Schwarzmüller, L. E., Postawa, L., et al. (2021). Cell-to-cell variability in JAK2/STAT5 pathway components and cytoplasmic volumes defines survival threshold in erythroid progenitor cells. *Cell. Rep.* 36 (6), 109507. doi:10.1016/j.celrep.2021.109507

De Boeck, J., Rombouts, J., and Gelens, L. (2021). A modular approach for modeling the cell cycle based on functional response curves. *PLoS Comput. Biol.* 17 (8), e1009008. doi:10.1371/journal.pcbi.1009008

Hammarén, H. M., Geissen, E. M., Potel, C. M., Beck, M., and Savitski, M. M. (2022). Protein-Peptide Turnover Profiling reveals the order of PTM addition and removal during protein maturation. *Nat. Commun.* 13 (1), 7431. doi:10.1038/s41467-022-35054-2

Helikar, T., Kowal, B., McClenathan, S., Bruckner, M., Rowley, T., Madrahimov, A., et al. (2012). The Cell Collective: toward an open and collaborative approach to systems biology. *BMC Syst. Biol.* 6, 96. doi:10.1186/1752-0509-6-96

Le Novère, N., Finney, A., Hucka, M., Bhalla, U. S., Campagne, F., Collado-Vides, J., et al. (2005). Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat. Biotechnol.* 23 (12), 1509–1515. doi:10.1038/nbt1156

Malik-Sheriff, R. S., Glont, M., Nguyen, T. V. N., Tiwari, K., Roberts, M. G., Xavier, A., et al. (2020). BioModels-15 years of sharing computational models in life science. *Nucleic acids Res.* 48 (D1), D407–D415. doi:10.1093/nar/gkz1055

Yu, T., Lloyd, C. M., Nickerson, D. P., Cooling, M. T., Miller, A. K., Garny, A., et al. (2011). The Physiome model repository 2. *Bioinforma. Oxf. Engl.* 27 (5), 743–744. doi:10.1093/bioinformatics/btq723

TYPE Original Research
PUBLISHED 01 March 2024
DOI 10.3389/fsysb.2024.1291293

Check for updates

OPEN ACCESS

EDITED BY
Ioannis P. Androulakis,
The State University of New Jersey,
United States

REVIEWED BY
Pasquale Palumbo,
University of Milano-Bicocca, Italy
Steven M. Abel,
The University of Tennessee, Knoxville,
United States

*CORRESPONDENCE
William E. Bentley,
✉ bentley@umd.edu

RECEIVED 19 September 2023
ACCEPTED 23 January 2024
PUBLISHED 01 March 2024

CITATION
Chun K, VanArsdale E, May E, Payne GF and Bentley WE (2024), Assessing electrogenetic activation via a network model of biological signal propagation.
Front. Syst. Biol. 4:1291293.
doi: 10.3389/fsysb.2024.1291293

Assessing electrogenetic activation via a network model of biological signal propagation

Kayla Chun[1,2,3], Eric VanArsdale[1,2,3], Elebeoba May[4], Gregory F. Payne[2] and William E. Bentley[1,2,3]*

[1]Fischell Department of Bioengineering, University of Maryland, College Parko, MD, United States, [2]Institute for Bioscience and Biotechnology Research, University of Maryland, College Park, MD, United States, [3]Robert E. Fischell Institute for Biomedical Devices, University of Maryland, College Park, MD, United States, [4]Medical Microbiology and Immunology Department, University of Wisconsin-Madison, Madison, WI, United States

Introduction: Molecular communication is the transfer of information encoded by molecular structure and activity. We examine molecular communication within bacterial consortia as cells with diverse biosynthetic capabilities can be assembled for enhanced function. Their coordination, both in terms of engineered genetic circuits within individual cells as well as their population-scale functions, is needed to ensure robust performance. We have suggested that "electrogenetics," the use of electronics to activate specific genetic circuits, is a means by which electronic devices can mediate molecular communication, ultimately enabling programmable control.

Methods: Here, we have developed a graphical network model for dynamically assessing electronic and molecular signal propagation schemes wherein nodes represent individual cells, and their edges represent communication channels by which signaling molecules are transferred. We utilize graph properties such as edge dynamics and graph topology to interrogate the signaling dynamics of specific engineered bacterial consortia.

Results: We were able to recapitulate previous experimental systems with our model. In addition, we found that networks with more distinct subpopulations (high network modularity) propagated signals more slowly than randomized networks, while strategic arrangement of subpopulations with respect to the inducer source (an electrode) can increase signal output and outperform otherwise homogeneous networks.

Discussion: We developed this model to better understand our previous experimental results, but also to enable future designs wherein subpopulation composition, genetic circuits, and spatial configurations can be varied to tune performance. We suggest that this work may provide insight into the signaling which occurs in synthetically assembled systems as well as native microbial communities.

KEYWORDS

network science, electrogenetics, synthetic biology, molecular information, microbial dynamics

1 Introduction

Synthetic biology has enabled the production and sensing of biomolecules through the design, testing and implementation of genetic circuits. In addition to guiding complex biosynthesis processes for therapeutic and industrial applications (Mimee et al., 2015; Jiang and Zhang, 2016; Cao et al., 2020), these engineered systems hold potential to communicate with and guide synthetic consortia and even native biomes (Hwang et al., 2017). Recently, synthetic consortia have been developed for leveraging the diversity of multi population systems in ways that expand biosynthetic potential and increase metabolic efficiency (Dinh et al., 2020; VanArsdale et al., 2022; Zhao et al., 2022; Gwon et al., 2023). The interactions within these engineered communities rely on robust cascades of molecular communication that convey information between cells (Quan et al., 2016; Servinsky et al., 2016). As such, system designs need to consider not only the genetic circuits within "designer" cells, but

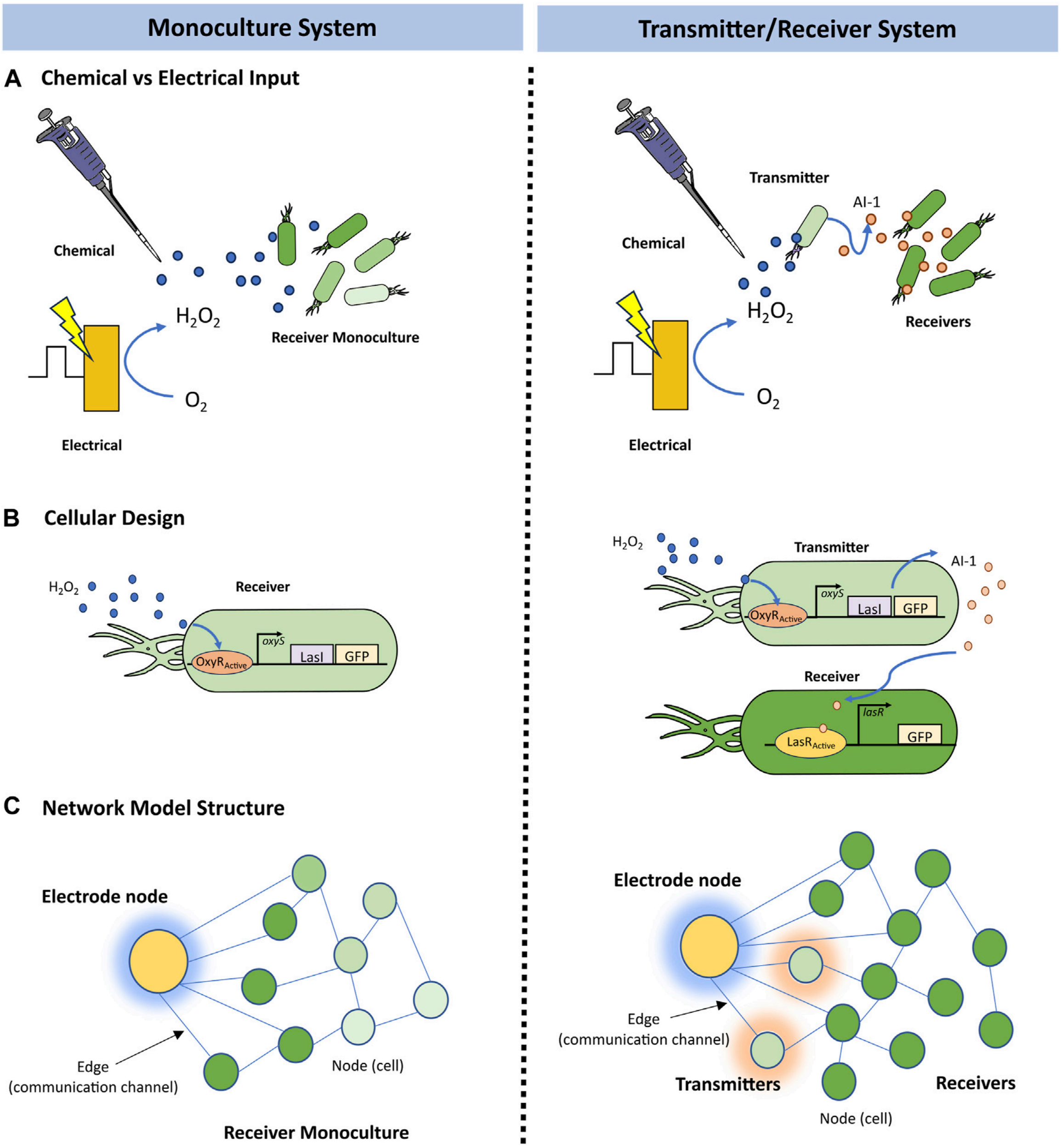


FIGURE 1
Systems overview. Schematic of the **(A)** cellular design of a Monoculture System (left) in which "Receiver" cells express LasI and GFP in response to hydrogen peroxide induction via the *oxyRS* regulon, and of a Transmitter/Receiver System (right) in which the same "Receiver" cells of the Monoculture system are repurposed as "Transmitter" cells that convey molecular information (AI-1 by the expression of *lasI*) to a second population also denoted "Receiver," but that only express GFP in response to AI-1. **(B)** electrogenetic experimental setup where a biased gold electrode creates hydrogen peroxide as an initial input signal to the cellular systems, and **(C)** an example network model structure for the Monoculture and Transmitter/Receiver systems, saturation of green representing GFP levels and shading around nodes representing inducer production at those nodes.

the communication networks that tie them together (Terrell et al., 2021).

In this study we wanted to mathematically characterize molecular signaling that guided previously published experimental results (VanArsdale et al., 2022) in which two cell-based systems synthesize a model product (green fluorescent protein, GFP) via chemical and electrical induction schemes exploiting different signaling pathways. These are depicted in Figure 1A. They are both based on induction by hydrogen peroxide. The base case (chemical induction) is actuated by the simple addition of hydrogen peroxide. Then, we had previously developed a means for electronically inducing cells; using simple electrodes, we altered the redox state of inducers (Tschirhart et al., 2017; Virgile et al., 2018; Kim et al., 2019; VanArsdale et al., 2022) and these activate genetic circuits. We refer to the genetic expression induced by electronic input as electrogenetics (Tschirhart et al., 2017) and have shown how one can electronically control gene expression, cell attributes (Virgile et al., 2018; VanArsdale et al., 2022), and even cell consortia (Tschirhart et al., 2017; Stephens et al., 2019; Bhokisham et al., 2020; VanArsdale et al., 2022; VanArsdale et al., 2023). In our experimental work (Figure 1A), we either added H_2O_2 (chemical induction) or we biased gold electrodes (2 mm diameter disk) immersed in the cultures with a—0.55 V vs. Ag/AgCl reductive potential (VanArsdale et al., 2022). This voltage is sufficient to electronically induce cells, it works by reducing oxygen dissolved in the growth media, creating hydrogen peroxide. Cells in the vicinity of the electrode genetically respond to the hydrogen peroxide through an engineered *oxyRS* regulon that activates a genetic circuit via the hydrogen peroxide sensitive transcriptional promoter, OxyR (Figure 1B). OxyR endogenously regulates oxidative stress management genes by repressing transcription until its cysteine groups are oxidized into disulfide bonds. The resulting conformation change stabilizes the transcription complex, inducing downstream gene expression (Pomposiello and Demple, 2001).

In Figure 1B, we illustrate the design of the two systems: (i) a receiver Monoculture and (ii) a Transmitter/Receiver co-culture. In the former case, hydrogen peroxide stimulates LasI and GFP production (Figure 1B). GFP is the model product in both cases and is easily measured by its fluorescence. LasI synthesizes the quorum sensing molecule N-3-oxo-dodecanoyl-L-homoserine lactone, which we refer to as autoinducer-1 (i.e., AI-1). Quorum sensing signaling molecules enable a collection of cells to take on a population-wide phenotype. In the Transmitter/Receiver system, the same cells used in the Monoculture system are repurposed as “transmitters,” where the hydrogen peroxide-induced quorum sensing signal is secreted and then encountered by the “receiver” cells and these respond by producing GFP (VanArsdale et al., 2022). Hence, in this two-strain culture one subpopulation turns the electronic signal into a biological signal for subsequent genetic activation and product synthesis in the second subpopulation. Autoinducer-1 is a very strong signaling molecule in that it activates gene expression at nanomolar amounts (Stephens et al., 2019). This amplifies the original signal to increase gene expression of the desired molecular product.

In this work, we employed a graphical modeling approach which enables a coarse grain interpretation of multicellular systems (Barabasi, 2013), thus, allowing us to capture agent-based intercellular interactions that fit population dynamics (Gosak et al., 2018). In Figure 1C, we depict our model in which each node represents a cell that possesses several weighted attributes: (i) local substrate concentration, (ii) the local inducer molecule concentration, and (iii) GFP expression level. The edges connecting nodes represent a communication channel where signaling molecules may transfer information between nodes. To characterize the movement of these signaling molecules, we implemented a previously developed overlay that approximates a formal diffusion model onto the network architecture (Sayama, 2015). This dramatically reduces computational demand while retaining dynamics of molecular communication and cellular connectivity.

With this model, we then characterized system performance in response to chemical and electrical induction by evaluating GFP production in both schemes. We further explored the effects of spatially fixed cultures (biofilms) in comparison to continuously stirred cultures by varying the edge dynamics in our model. Edges that are fixed reflect static cells, like would exist in a biofilm. Edges that are continually reconnecting between nodes reflect stirred cultures. Then, by utilizing modularity, a graph measure of a network’s subcommunity structure (Newman and Girvan, 2004), we related the network’s spatial organization to its signal output. Overall, our model enables a kinetic understanding of signal propagation and GFP production among spatially varied bacterial populations that, in turn, exploit different signaling processes. This provides new hypotheses regarding modes of information transmission and their effectiveness, ultimately leading to new designs.

2 Materials and methods

2.1 Model formalism

Network initialization was performed by generating a random undirected G (n, m) graph (Barabasi, 2013) in which there are n total nodes and m total edges that are randomly distributed amongst the nodes. In this network, each node represents an individual cell and edges represent communication channels by which signaling molecules can be transferred between nodes. Each node N_i possesses the following dynamic node weights: $s_i(t)$, $H_2O_2\ _i(t)$, $AI\text{-}1_i(t)$, and $GFP_i(t)$ corresponding to the cell’s substrate, hydrogen peroxide, autoinducer-1, and green fluorescent protein concentrations at time t, respectively. In this graph, edges are unweighted and undirected, meaning they do not possess quantitative attributes, nor do they follow any directionality in their connections, i.e., signaling molecules can flow to in either direction between two connected nodes. In our model, time is discrete and represented by natural numbers, evolving forward with each iteration of the simulation as depicted in Figure 2A. At each timestep a transition is applied in which each attribute of the network is sequentially updated via the following modules: (i) Gene activation, (ii) Molecular production, (iii) Signal diffusion, (iv) Growth, and (v) Edge randomization. That is, a gene activation module is applied, and then activated nodes carry out their respective molecular production models, resulting in increased molecular concentrations at these nodes. Next a signal diffusion

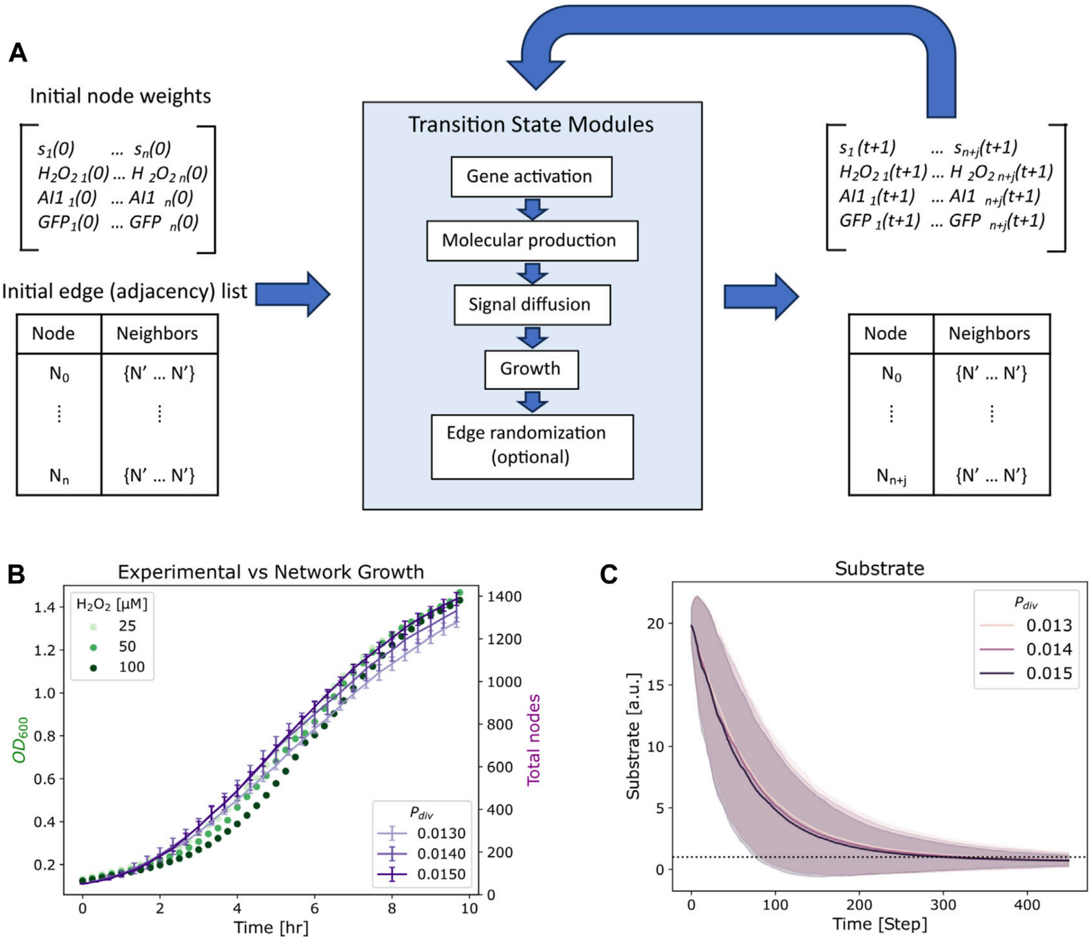


FIGURE 2
Simulation process and growth fit. **(A)** Overview of the simulation iterations, where initial state variables and edge structure are updated via transition state modules at each timestep. The output length of the state variables matrix and edge list increase by *j* new nodes. **(B)** Growth measurements for *E. coli* strain OxyR-LasI-GFP (transmitters) with various hydrogen peroxide induction concentrations (chemical addition) are plotted in green, alongside average total nodes of 10 simulation repeats at various division probabilities (P_{div}) over time in purple. Bars represent standard deviation. **(C)** Average substrate per node for P_{div} in **(B)**, the horizontal dashed line indicates a user-specified substrate threshold, $k = 1$, below which a node will no longer divide. The shaded zones indicate standard deviation of the substrate concentration across the network for each probability.

TABLE 1 Equations for gene activation and subsequent protein production for the inducers: hydrogen peroxide and AI-1.

Description	Equation
Hydrogen peroxide induced gene activation probability	(1) $Prob_{H_2O_2}(H_2O_2 > 0) = \frac{1}{1+e^{-\frac{[H_2O_2]-8}{2}}}$
	(2) $Prob_{H_2O_2}(0) = 0$
AI-1 induced gene activation probability	(3) $Prob_{AI1}(AI1 > 0) = \frac{1}{1+e^{-50([AI1]-0.25)}}$
	(4) $Prob_{AI1}(0) = 0$
Hydrogen peroxide induced molecular production rate	(5) $Rate_{H2O2} = \frac{2[S]}{1+e^{-[H_2O_2]}}$
AI-1 induced molecular production rate	(6) $Rate_{AI1} = 0.2[AI1][S]$

module is applied, and molecular concentrations are updated based on the calculated exchange of molecules. Lastly, a growth module is applied to nodes with available substrate; the concentrations of divided nodes are amended. After this, edge randomization may be applied to stirred culture simulations, and the time is forwarded to the next timestep. Thus, the state of the system can be described at any point by the nodes, each with their own set of state variables described by their weights and edges as depicted in Figure 2A.

2.1.1 Gene activation and molecular production

To capture genetic induction and subsequent molecular production we implemented a two step mechanism at each node

TABLE 2 State variable dynamics equations.

Variable	Equation
H_2O_2	$H_2O_{2\ i}(t+1) = H_2O_{2i}(t) + \alpha[\sum_{j\epsilon N_i} H_2O_{2j}(t) - H_2O_{2i}(t)\deg(i)]$
$AI\text{-}1$	$AI\text{-}1_i(t+1) = AI1_i(t) + Prob_{H_2O_2} {*} Rate_{H_2O_2} + \alpha[\sum_{j\epsilon N_i} AI1_j(t) - AI1_i(t)\deg(i)]$
s	$s_i(t+1) = \frac{s_i(t)}{2}$ (if node i divides)
GFP	$GFP_i(t+1) = Prob_{AI1} {*} Rate_{AI1} + GFP_i(t)$ (AI-1 induced) or $GFP_i(t+1) = Prob_{H_2O_2} {*} Rate_{H_2O_2} + GFP_i(t)$ ($\mathrm{H_2O_2}$ induced)

at every timestep. First, the probability of gene activation is a function of inducer concentration (see Table 1; Equations 1–6). H_2O_2 induced gene expression is described by a logistic curve (Table 1; Equations 1, 2; Supplementary Figure S1A) with a threshold of 12.5 μM. AI-1 dependent gene expression is implemented using a steeper and linear step function (Table 1: Equations 3, 4; Supplementary Figure S1A) reflecting the nanomolar requirements for induction (Chun et al., 2021; VanArsdale et al., 2022).

If a gene is activated at a node for a timestep (via probability function based on inducer concentration), it will produce the specified molecular product (GFP or AI-1) at a set expression rate based on the prevailing inducer concentration and substrate availability (Table 1: Equations 5, 6; Supplementary Figures S1B, C). For production based on AI-1, the rate is linear while for hydrogen peroxide it is a saturation function so that at low concentrations there is a steep peroxide dependence and at high concentrations the rate is saturated (Stephens et al., 2019; Terrell et al., 2021; VanArsdale et al., 2022). These transitions occur at each timestep prior to the diffusion and growth modules, such that molecular production occurs with the concentrations from the previous timestep. The discrete equations are described in Table 2.

2.1.2 Signal diffusion

Signaling between nodes occurs across edges, such that only nodes connected by an edge may transfer H_2O_2 and AI-1. Signal molecule movement across edges are defined by a discrete approximation of diffusion derived by the following equations as previously described by Sayama (Sayama, 2015):

$$\frac{dc_i}{dt} = \alpha \sum_{j\epsilon N_i} \left(c_j - c_i\right) \tag{1}$$

$$c_i(t + \Delta t) - c_i(t) = \left[\alpha \sum_{j\epsilon N_i} \left(c_j - c_i\right)\right] \Delta t \tag{2}$$

$$c_i(t + \Delta t) = c_i(t) + \alpha \left[\sum_{j\epsilon N_i} c_j(t) - c_i(t)\deg(i)\right] \Delta t \tag{3}$$

where c_i is the concentration of signaling molecule at a given node i, c_j is the concentration at that node's neighbor j, $deg(i)$ is the number of edges at node i, and α is a diffusion coefficient (See Supplementary Table S1 for all coefficient values). In Eq. 1, diffusion is generalized to the change in concentration at a node with respect to the difference between its own concentration and its neighbors. This can be discretized (Eq. 2) and solved to find that the change in concentration at a node is determined by the difference between the sum of its neighbors' concentrations and the product of its own concentration and number of edges (Eq. 3). At every timestep, the concentration is calculated from Eq. 3 for each node and updated prior to growth module implementation. This process applies to the following state variables and occurs prior to the calculation of network growth: H_2O_2 (t) and $AI\text{-}1(t)$. The equations for these variables prior to network growth can be found in Table 2.

2.1.3 Network growth

The network grows with time, depending on substrate availability and growth probability, P_{div}. Initially, each node is assigned the same initial substrate weight, s_0. At each time step, if a node has a substrate level above a minimum threshold, k, the node has the probability P_{div}, that it may divide into two. Following a division event, the substrate (Table 2), H_2O_2 and AI-1 node weights are divided equally between daughter nodes at each timestep. As noted above this occurs after the diffusion module, such that the newly calculated H_2O_2 and AI-1 concentrations may be divided in two upon a division event. As depicted in Figure 2A, with each iteration the network will increase by j nodes, determined by substrate availability at each node and P_{div}. We note that daughter nodes maintain fluorescence (GFP) of their parent's. This assumption is in agreement with previous experiments (Servinsky et al., 2016). We additionally neglect protein degradation, again in agreement with experimental results (Servinsky et al., 2016).

After a division event, the resulting daughter nodes share an edge and maintain their parent's edges, limited to a maximum of 10 neighbors. Note, as commonly defined within the field of network science, we refer to neighbors as nodes which share an edge (Newman et al., 2006). These 10 neighbors are randomly sampled from the parent's neighbors including those that have divided at that timestep. In a case where a dividing node has 10 neighbors that also all divide at that time step, out of the 20 surrounding nodes only 10 will be randomly selected to share an edge with each daughter.

In Figure 2B, we depict growth curves for the *Escherichia coli* strain OxyR-LasI-GFP grown with various hydrogen peroxide concentrations (VanArsdale et al., 2022). These cells are the receivers in the Monoculture case and transmitters in the Transmitter/Receiver case (Terrell et al., 2021) (Figure 1). Alongside we show the total number of nodes over time for a simulated network of with 50 initial nodes, an s_0 of 20, and a k of 1 for various P_{div}. With these s_0 and k values, each node can divide five times during the growth phase, allowing us to fit the initial node count to 50 and total possible number of nodes to 1,600 which approximates 1 node to 0.001 OD_{600}. The P_{div} values assigned helped to ensure that the growth phase of the network translated well to experimental results, such that 45 timesteps represented ~1 h of cell culture. Our simulation thus mimicked the log phase growth of the cell cultures. We note that flexibility for fitting experiments is enabled by altering the division probability, P_{div}. Additionally, we note that as the network grows the average substrate per node decreases over time (Figure 2C), until reaching below the threshold value of $k = 1$. As described above, below this threshold, nodes may no longer divide. The shadowed area in Figure 2C represents the full ranges of substrate levels across the network and for each division probability. While the substrate defined in our model represents general nutrient availability, the trend shown in Figure 2C emulates the decrease in glucose over time in *E. coli* cultures demonstrated experimentally (Shiloach et al., 1996). That is, while the network model formalism does not include a typical deterministic Monod model for growth with a maximum specific growth rate and saturation constant, the configuration here well represents the overall culture dynamics.

2.1.4 Edge randomization

To describe the spatiotemporal effects of various modes of cell culture such as stirred, immobilized biofilms (static), and combinations thereof, we implemented edge randomization. In the absence of stirring, edges which are assigned during network initialization and at each node division, remain fixed for the duration of a simulation. To simulate a stirred batch culture, we randomized the edges amongst all nodes at every timestep. We simulated two base cases with either static or randomized edges and with or without network growth to demonstrate the effects signaling dynamics: one case where inducers may come from a highly concentrated source node and another case where an electrode may generate inducers at its surface over a specific time period (Supplementary Figure S2). As anticipated, cases that include network growth and edge randomization resulted in faster homogeneity of signaling molecule concentration across the network than non-growing networks or those growing with static edges. From these tests, we found a set of parameters that when used, enabled reasonable agreement between our previously published data ($s_0 = 20$, $k = 1$, $P_{div} = 0.015$, $\alpha = 1$ and an initial average of 4 edges per node).

2.1.5 Electrical hydrogen peroxide generation

To mathematically characterize the production of hydrogen peroxide at the surface of a biased electrode as a mode of information transmission into bacterial cells, we model the input as a signal generated from an individual source node, then link this source to the various nodes. In our network architecture, the electrode is represented by a single node which produces hydrogen peroxide at each time step that it is turned "on." To simulate the actual experimental conditions in which electrical stimulus resulted in negligible growth during the time of induction (VanArsdale et al., 2022), we set the growth probability parameter, P_{div}, to zero when the electrode is "on" until that time when the growth was observed to increase. We fit the hydrogen peroxide production for an initial network size of 100 to produce 46 µM hydrogen peroxide per timestep to approximate experimental results (Supplementary Figure S3A).

Previously reported experimental results demonstrated that electrical induction yielded lower GFP output compared to a chemical addition, suggesting that the spatiotemporal heterogeneity resulting from the localized inducer production at the electrode's surface effects output. To recapitulate these findings in our model we limited the number of nodes connected to the electrode to 5% of the total network at every timepoint. In Supplementary Figure S3B, we plotted the Monoculture response for chemically and electrically induced simulations to demonstrate that the limitation of electrode connectivity to the network reproduces experimental trends, via reduced GFP production compared to chemical induction.

2.2 Code and data availability

Graph simulations were performed in Python using NetworkX (Hagberg et al., 2008), and modifying and implementing the Simulation class from A First Course in Network Science (Menczer et al., 2020). Graph generation and initialization and graph transition states were defined and are contained in supplemental notebooks. Visualizations were performed using Python's matplotlib and seaborn libraries (Hunter, 2007; Waskom, 2021). Experimental data used for parameter fitting are from (VanArsdale et al., 2022).

Python notebooks and simulation data are available online at github.com/kaychun29/bio-network-simulations.

3 Results

3.1 Chemical and electrical induction of monoculture and transmitter/receiver systems

We first simulate the two cellular systems in response to the chemical addition of hydrogen peroxide. We aimed to capture the experimental results depicted in Figures 3A, B (reproduced with permission), where identical levels of hydrogen peroxide were added to the Monoculture system and to the Transmitter/Receiver System. We later measured GFP expression in all cells via flow cytometry after 3 h (VanArsdale et al., 2022). Flow cytometry provides for the distribution of GFP among all cells in a population. Especially at high concentrations, a chemical addition of hydrogen peroxide should result in a homogeneous input (VanArsdale et al., 2022) wherein there is little "noise" accompanying induction. In the Monoculture system, increases in GFP became obvious at initial concentrations of 12.5 uM H_2O_2. Further increases in H_2O_2 had relatively little effect on GFP. Interestingly, for the Transmitter/Receiver system, lower initial concentrations of H_2O_2 resulted in significant GFP expression owing to the AI-1 signal propagation. In the end, the yield of GFP for this Transmitter/Receiver system was nearly 10-fold higher than the case with just H_2O_2 added to the monoculture, even at the highest concentrations (VanArsdale et al., 2022).

To simulate these results, we assigned each node the same initial hydrogen peroxide weight based on the initial experimental concentration. We set initial GFP weights randomly using a Gaussian distribution with a mean of 500 and standard deviation of 250. This allows for all nodes to have fluorescence background, which fit our previously published experimental distribution for uninduced cells, Figures 3A, B (VanArsdale et al., 2022). For the following simulations the initial network size was 100 nodes, with an average of four edges per node. These initial conditions enabled

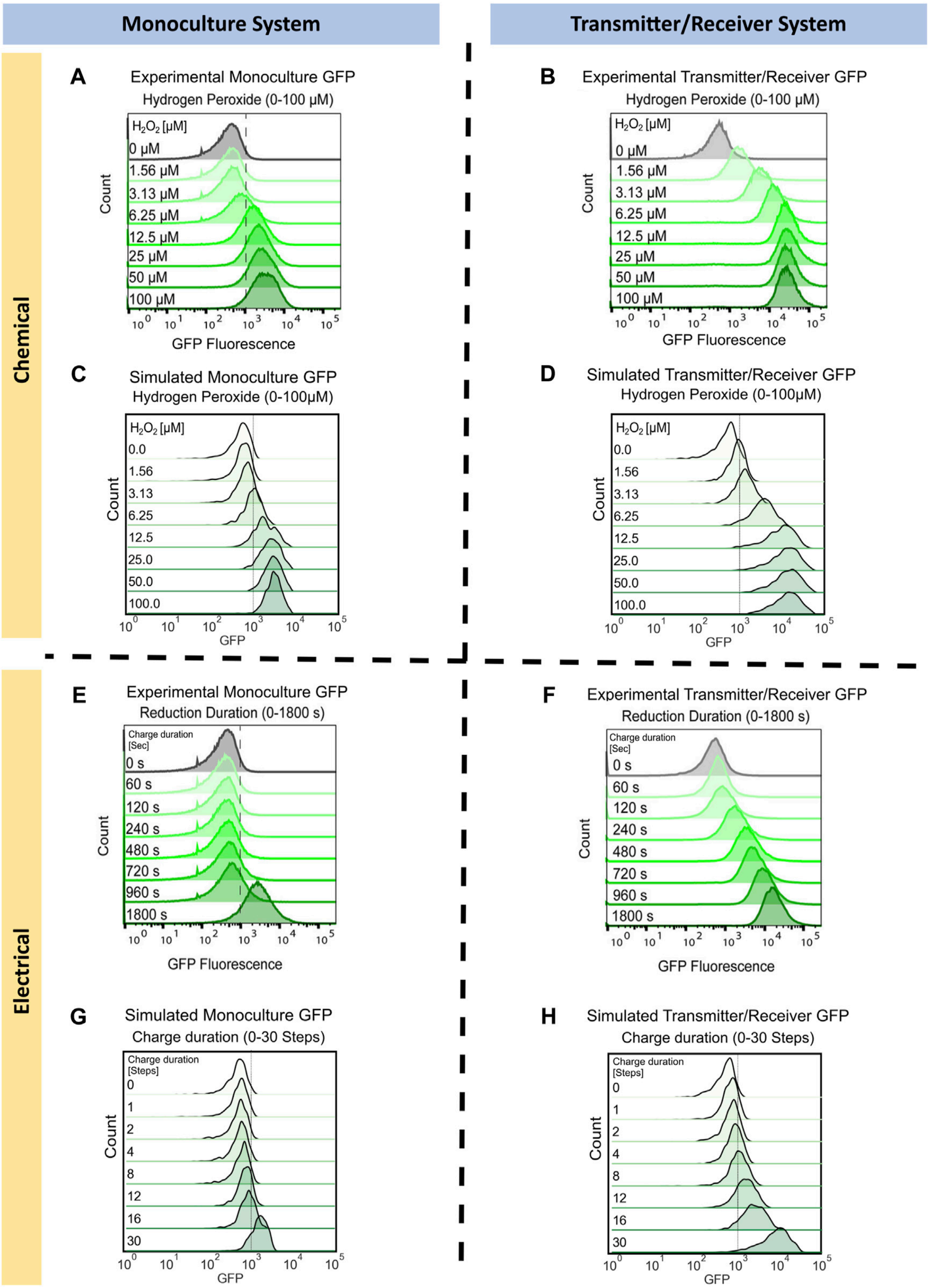


FIGURE 3
Monoculture and Transmitter/Receiver GFP distributions for chemically and electrically induced edge randomized networks. Chemically induced **(A)** Monoculture and **(B)** Transmitter/Receiver system at 3 h hydrogen peroxide addition, reproduced with permission from VanArsdale et al. (2022). **(C)** The simulated monoculture system GFP distribution at 180 timesteps is shown for an aggregate of 10 replicates, with initial hydrogen peroxide concentration ranging from 0 to 100 μM. **(D)** The simulated Transmitter/Receiver system's GFP distribution across all nodes at 180 timesteps is shown for an aggregate of 10 replicates, with initial hydrogen peroxide concentration ranging from 0 to 100 μM. Experimental flow cytometry data of the **(E)** Monoculture and **(F)** Transmitter/Receiver system at 3 h post charge application, reproduced with permission from VanArsdale et al. (2022). GFP distributions of simulated electrical induction for the **(G)** Monoculture system and **(H)** Transmitter/Receiver's receiver GFP distributions across all nodes at 180 timesteps post charge application. Distributions shown are an aggregate of 10 simulated replicates, with charge durations ranging from 0-30 timesteps.

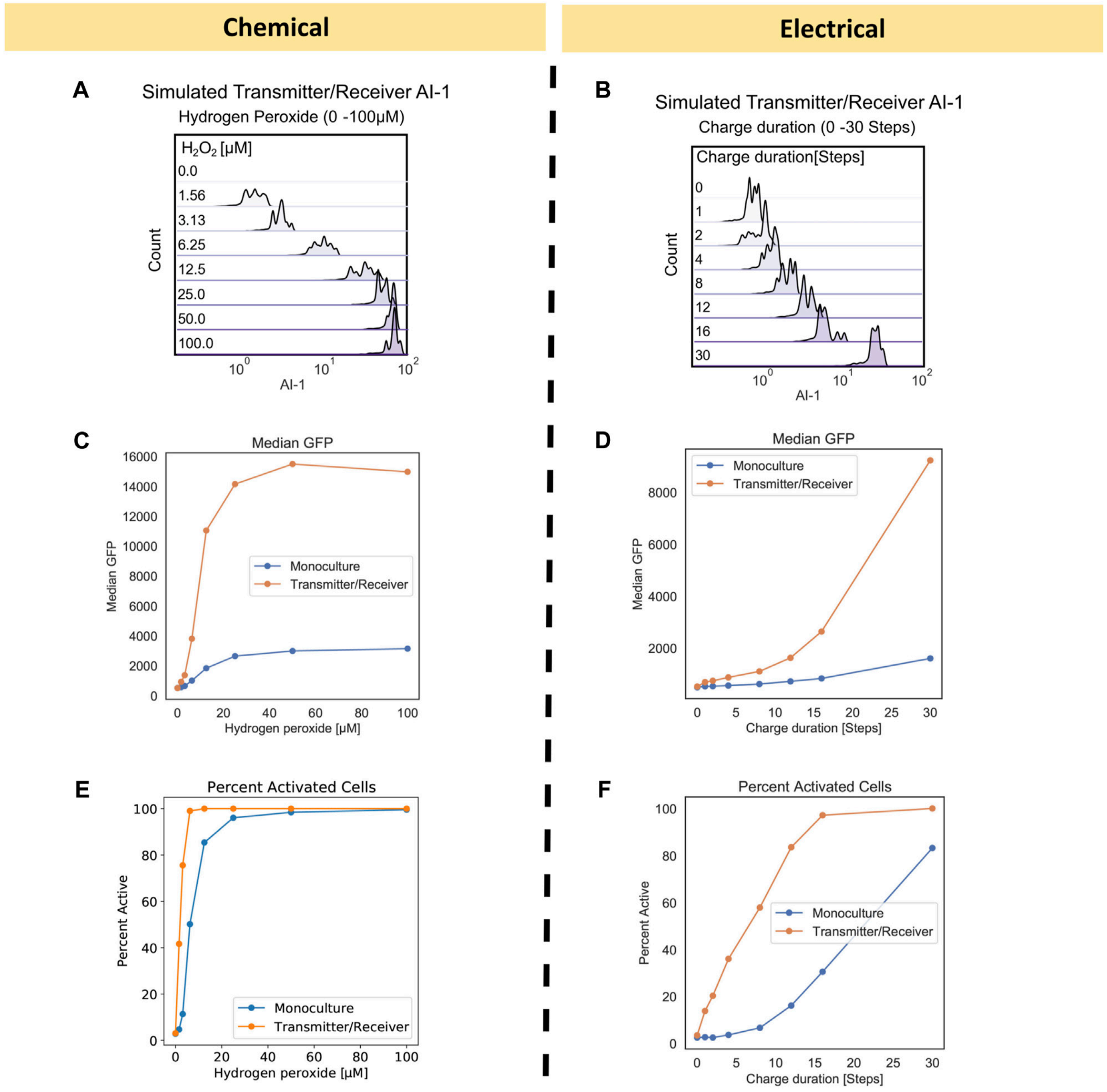


FIGURE 4
Transmitter/Receiver AI-1 distributions and signal metrics of chemically and electrically induced edge randomized networks. **(A)** The AI-1 distribution amongst all nodes in the Transmitter/Receiver network at 180 timesteps is shown for an aggregate of 10 replicates. **(B)** The AI-1 distribution amongst all of nodes in the Transmitter/Receiver network at 180 timesteps post charge application is shown for an aggregate of 10 replicates. **(C, D)** Calculated median GFP from the distributions data shown in Figure 3 plotted over their initial inducer concentration **(C)** and charge duration **(D)**. **(E, F)** Calculated percent active nodes from the distributions data shown in Figure 3 plotted over their initial inducer concentration **(E)** and charge duration **(F)**, threshold for activation was defined at 1000 GFP.

reproducible network propagation, while conserving computational time. We implemented network growth and edge randomization at each timestep to recapitulate the well-mixed growing culture, according to the methods previously described. For the Monoculture system, chemical induction was simulated using the gene activation probability ($Prob_{H_2O_2}$, Table 1: Equations 1, 2) and the molecular production rate ($Rate_{H2O2}$, Table 1: Equation 5). To model the Transmitter/Receiver system in which a two-strain co-culture is used to amplify the initial hydrogen peroxide signal, we partitioned the initial network into 10 percent transmitter nodes, which function the same as the Monoculture's receivers, and 90 percent receiver nodes which activate GFP production by AI-1 induction. In both systems, AI-1 freely diffuses between nodes at each timestep (Stephens et al., 2019), while in neither case does the GFP diffuse out of the cell[47]. In this Transmitter/Receiver system, GFP production is probabilistically activated ($Prob_{AI1}$, Table 1; Equations 3, 4) and produced at a rate ($Rate_{AI1}$, Table 1: Equation 6) dependent on AI-1 and substrate concentration. In

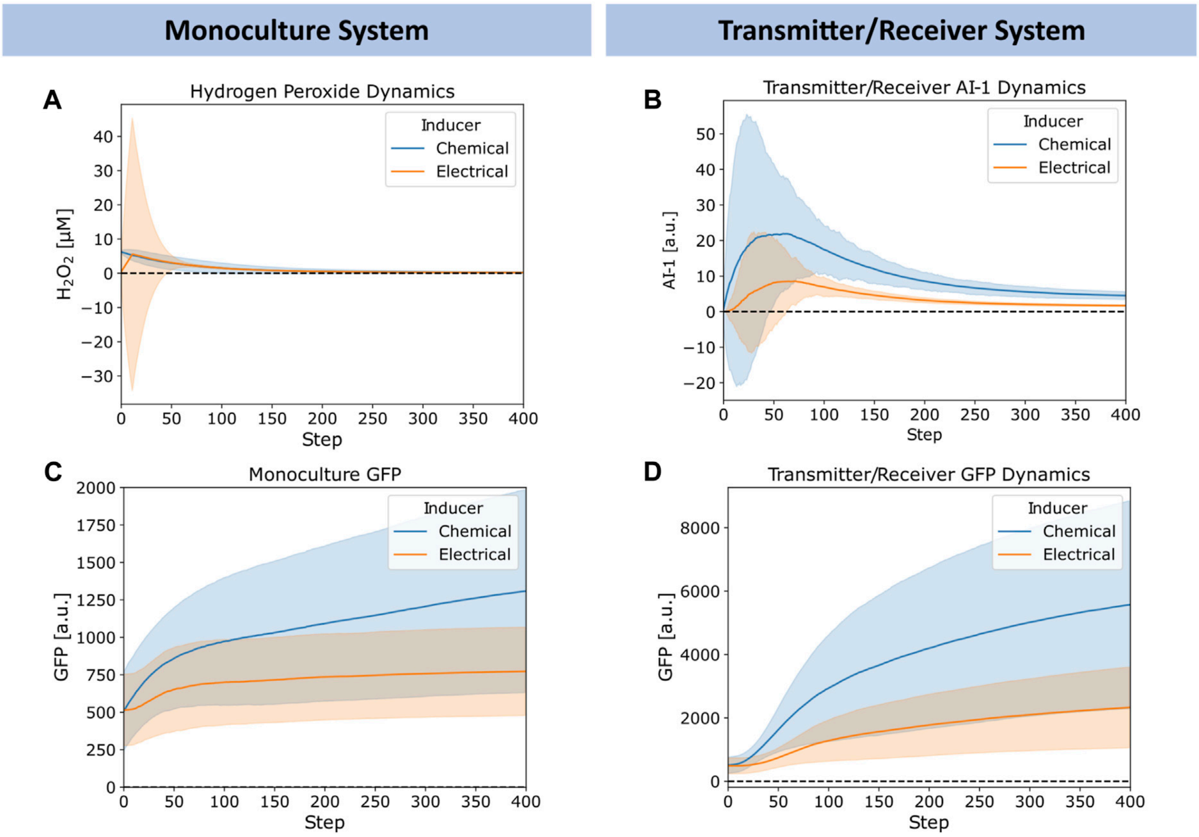


FIGURE 5
Chemical *versus* electrical signaling dynamics for edge randomized networks. **(A)** Average hydrogen peroxide, **(B)** Average Monoculture GFP, **(C)** Average Transmitter/Receiver GFP, and **(D)** Average Transmitter/Receiver AI-1 concentrations over time for a 6.25 μM hydrogen peroxide induced chemical addition (blue) and 12 step charge duration (orange) across the entire network. Error bars appear as shaded regions, representing standard deviation of aggregated network data from 10 simulation replicates.

Figures 3C, D we plotted the simulated GFP distributions across the entire network for both the Monoculture and Transmitter/Receiver networks at mid-log growth (180 timesteps) as a function of initial H_2O_2 level. Consistent with the experimental results, the range of expression in the Transmitter/Receiver network reached 10^5, while the Monoculture network's maximum values were ten-fold lower.

We next simulated the electrogenetic approach wherein an applied reducing potential on the electrode generates hydrogen peroxide and this, in turn, stimulates the cells. Naturally, a major difference between this mode of induction is that the hydrogen peroxide is generated at the electrode and while the system is mixed, the peroxide level increases with the extent of its generation rate. The experimental results from earlier work are shown in Figures 3E, F (reproduced with permission) (VanArsdale et al., 2022). In the Monoculture system, small increases in GFP were observed until the cells were exposed to −0.55 V for 1,800 s. In the previous work, a solution exposed to this reduction duration produced approximately 15 μM of H_2O_2 (VanArsdale et al., 2022). Thus, the experimental results for the electrogenetic case were roughly equivalent to the chemical addition of H_2O_2. It was interesting to see that in the case of the Transmitter/Receiver system, a continuous increase in GFP was observed with increased charge. This was previously described as a result of cells near the electrode experiencing sufficient peroxide to induce AI-1, which, in turn, is stable and can be mixed throughout (VanArsdale et al., 2022).

To simulate electrical induction, we utilized the same model structure as described prior for chemical induction with the exception of initial hydrogen peroxide concentrations. For electrical induction, initial hydrogen peroxide weights were set to zero across the whole network and hydrogen peroxide was produced over a designated charge duration as described in **Methods**. In Figure 3G, we found the simulated GFP distribution of the electrically induced Monoculture system did not increase significantly until greater than 30 timesteps of applied charge (equivalent of 30 min), aligning with experimental results in Figure 3E. For the Transmitter/Receiver system (Figure 3H), activation increased nearly immediately, and full activation was attained with 30 steps of electrode charge. Our network model, in all cases, corresponded well with the actual data in Figures 3E, F, wherein the Monoculture distribution was essentially unchanged until over 960 s and the Transmitter/Receiver distribution increased across the span of 960 s to reach full activation.

An advantage of the network approach is that one can examine state variables that are otherwise difficult to obtain experimentally. Also, one can more easily align results with underlying mechanisms. In Figures 4A, B, we plotted the estimated AI-1 distributions for the

Transmitter/Receiver networks. While not measured experimentally (VanArsdale et al., 2022), these simulated values are consistent with expectations. The AI-1 distributions suggest significant heterogeneity within the network. We found this heterogeneity was a result of the variance in activation and spatial distribution of the transmitter nodes and we note this heterogeneity has been reported in chemically induced bacterial cell cultures (Servinsky et al., 2016). We also note that such heterogeneity is not characterized with commonly implemented population scale ODE models, but it can be manipulated experimentally via quorum sensing and genetic circuit design (Zargar et al., 2015). Our initial network model suggests that there is a level of heterogeneity that is innate to the system and is introduced when amplifying an initial homogenous input through a subpopulation of cells.

Interestingly, we found that the range of GFP for both Transmitter/Receiver systems was reflected in the AI-1 distributions in Figures 4A, B**.** In the chemically induced system, the AI-1 concentrations were between 10^1–10 (Jiang and Zhang, 2016) for initial H_2O_2 concentrations above 6.25 µM. Comparatively, for the electrically induced system the AI-1 distribution across the entire network increased incrementally with only the highest charge duration of 30 timesteps producing above 10^1 of AI-1. We further evaluated signal transmission by assessing the median GFP and fraction of activated cells for chemical and electrical induced systems. These serve as metrics for final signal output. The median GFP shows that with electronic induction, expression was generally lower than with chemical induction (Figures 4C vs. 4D), suggesting the signal was attenuated when the inducer was produced at a point source (the electrode node) and needed to diffuse outward among the cells to provide induction.

When comparing the Monoculture to Transmitter/Receiver systems, we observed the amplified response enabled by the Transmitter/Receiver system was readily apparent; the median GFP was above 1.4×10^4 *versus* 2.5×10^3 for the Monoculture (Figure 4C), an approximate 5-fold increase, when chemically induced with 100 µM. With electrical induction the median GFP of the Transmitter/Receiver system reached about 8.5×10^3 at the longest charge duration (30 steps), whereas the Monoculture system did not increase above 2.0×10^3, an approximate 4-fold difference. In addition to median GFP we also calculated the percent activated nodes in the network for each initial inducer concentration (by measuring the number of nodes with GFP above a 10^3 threshold). In Figure 4E, we plotted chemically induced systems and observed that although both systems ultimately reached 100% activity, the Transmitter/Receiver system reached this peak at lower H_2O_2. For the electrically induced systems, the portion of active nodes increased incrementally and monotonically with charge (Figure 4F). We note that the Monoculture system had a consistently lower percentage of active nodes than the Transmitter/Receiver system, as expected, and never reached 100% by with 30 timesteps of induction. Overall, our model simulations corresponded well with the previous data (Figures 3A, B, E, F). Our simulations also suggest that despite the heterogeneity or "noise" that is introduced by amplifying the initial signal through a subset of cells (electrode induction), the molecular amplification that was enabled by transforming the H_2O_2 into a stronger secondary signaling molecule, in particular one that evokes a quorum sensing response, overcame that disruption, and produced high levels of signal and activation.

In Figure 5, we explored further the dynamics of H_2O_2, AI-1, and GFP for the chemically and electrically actuated cases by plotting their average (lines) and standard deviation (shaded) across the network over time. We chose representative cases with similar average H_2O_2 concentrations. In Figure 5A, we depict the simulated H_2O_2 dynamics for the chemical addition of 6.25 µM H_2O_2 and for the electrical induction at 12 timesteps of applied charge (~6 µM of hydrogen peroxide generated). The widely distributed H_2O_2 level in the case of electrical induction was expected, but the average concentration simulated was quite similar. We note, Figure 5A depicts Transmitter/Receiver H_2O_2 dynamics, however Monoculture dynamics were nearly identical suggesting the type of cellular system does not affect hydrogen peroxide diffusion and generation. Despite the comparable average H_2O_2 levels in the systems over time, the AI-1 concentration of the Transmitter/Receiver system was nearly 2-fold higher that of the chemical induction (Figure 5B). In general, the GFP levels produced by both the Monoculture and Transmitter/Receiver systems (Figures 5C, D) were higher for the chemical addition relative to the electronically induced systems. This was understandable because the electrode produced H_2O_2 levels were found to be widely dispersed, indicating that many cells likely encountered minimal levels of inducer (Figure 5A). When comparing the Monoculture *versus* Transmitter/Receiver GFP dynamics (Figure 5C vs. Figure 5D), GFP expression in the Monoculture increased consistently over time whereas the Transmitter/Receiver network expression was slightly delayed initially during which time AI-1 was produced (~50 steps corresponding to peak AI-1) and subsequently accumulated. For both modes of induction, the Transmitter/Receiver GFP yields were higher irrespective of a delay in production.

Overall, we note that the large standard deviations depicted in Figure 5 reflect substantial heterogeneity within the network. We suggest this heterogeneity is rooted in the wide signaling molecule distribution that can occur when cell numbers are low (early on) and when electrodes are used to generate hydrogen peroxide. In the latter case, this signal molecule interacts with cells in a random and distributed manner. In the experimental system, an uninduced cell needs to be transported near an electrode to receive H_2O_2. At further distances the peroxide could be depleted so that cells far away never experience high levels. Interestingly, our network model seems to well characterize the extent of signal propagation and the effects of its design structure in determining system outcome. The tradeoffs between the delay in responses and expression levels provide insight on system design. They also suggest spatial heterogeneity, we explore this as a potential design feature as follows.

3.2 Spatial design via network topology: Graph modularity and edge dynamics' effect on signaling

Based on our successes in characterizing experimental data from both the chemical addition of hydrogen peroxide and its electrode-based generation for both the Monoculture and the Transmitter/Receiver systems, we decided to interrogate the design space for altered induction methodologies. Specifically, we next explored how

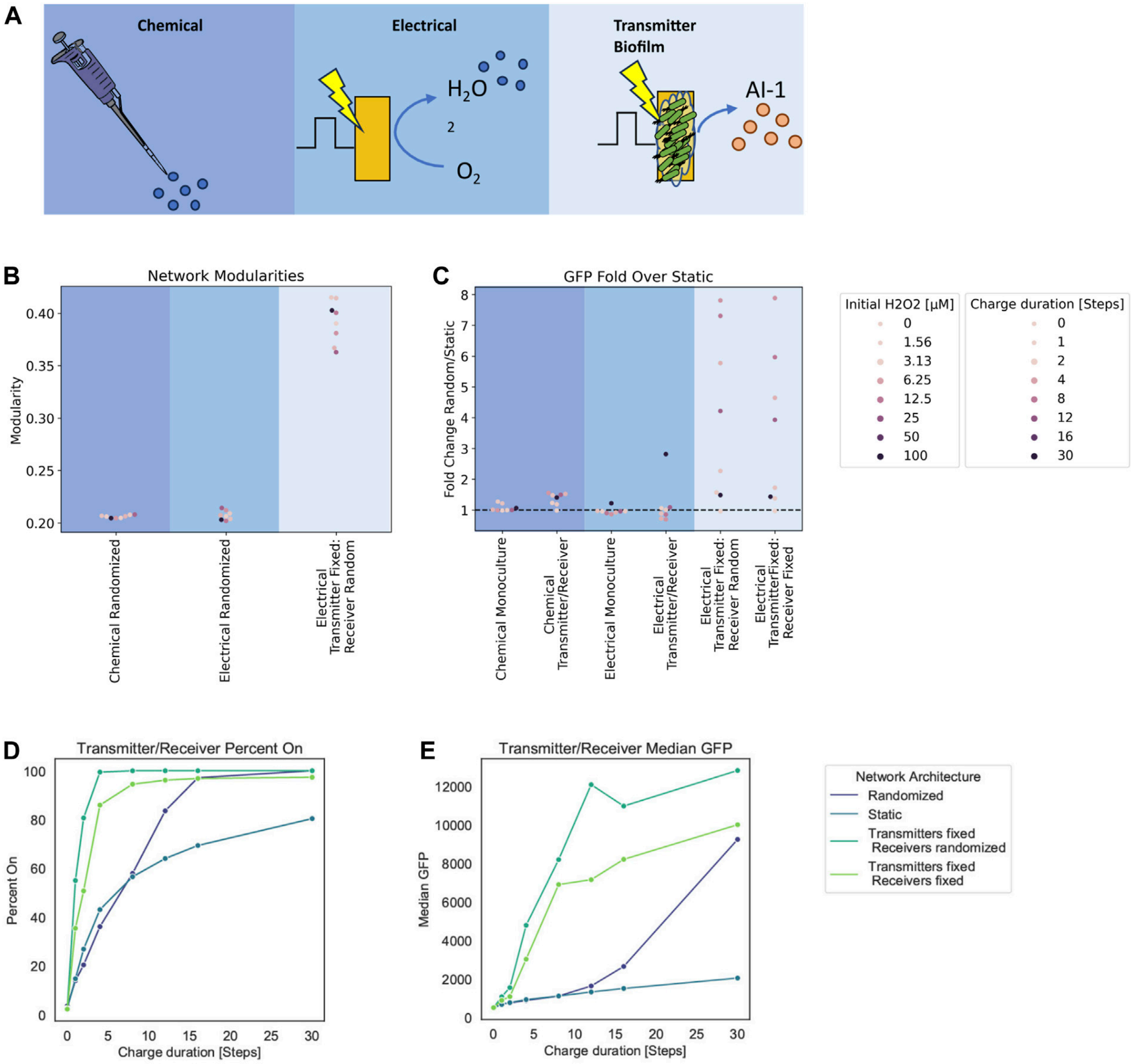


FIGURE 6
Modularity and fold change dependent on network structure. **(A)** Graph schematic of the three spatial configurations tested. **(B)** Network modularity of differing node arrangements and edge dynamics at timestep 180. **(C)** Fold change in GFP of randomized edge networks over static edge networks for the Monoculture and Transmitter/Receiver systems with either chemical or electrical induction at 180 timesteps post induction. **(D, E)** Signal transmission metrics for Transmitter/Receiver network architectures at 180 steps post 30 steps of electrical induction, calculated from an aggregate distribution of 10 simulation replicates. **(D)** Percent active nodes for varying charge duration times. **(E)** Median GFP across network for varying charge duration times.

the relative spatial distribution of cells (nodes) could affect the signaling. We decided to test a case where we retain transmitter cells directly onto the electrode. Thus, in Figure 6A, in addition to the (i) a chemical addition and (ii) electrical induction cases previously described, we added (iii) electrical induction of transmitter cells that are fixed to its surface. This last network structure captures experimental designs in which cells are either engineered to bind to gold electrodes (Terrell et al., 2021) or that are retained in an assembled hydrogel film (Li et al., 2020). Cells localized in this manner could receive electronic signals (hydrogen peroxide) and then transmit their "message" to cells outside of the film through signal synthesis, secretion, and transport to cells occupying the liquid proximal to the electrode and beyond (Li et al., 2020). For affixed cells, instead of randomizing edges at time steps, we fixed edges and maintained them throughout. This mimics a static system, representative of a biofilm (Li et al., 2007; Cornell et al., 2020) or a set of cells localized on an electrode (Terrell et al., 2021).

To quantify structural variation that emerges due to growth and edge dynamics, we used modularity (Newman and Girvan, 2004; Newman et al., 2006; Blondel et al., 2008) as a measure of network structure (Supplementary Figure S4A). In general, modularity describes how well a network is partitioned into various sub-communities (Newman et al., 2006). A single community wherein the connections are near random is represented by a modularity value of zero, while a network where all edges fall within the same community would have a modularity of 1 due to its strong community structure (Newman and Girvan, 2004). For our calculations, we use the Louvain method to calculate the modularity as it is computationally efficient in finding high modularity partitions of large networks (Blondel et al., 2008).

In Figure 6B, we depict the calculated Louvain modularity at 180 timesteps for the cases above (chemical and electrode induction for mixed cultures), as well as the new case where transmitter cells are fixed to the electrode (initial 10 nodes) and the receiver cells are not fixed. For the Transmitters fixed onto the electrode, dividing nodes inherit the edges from their parent nodes without further edge randomization. As expected, our results show that there was increased modularity calculated in the case where some cells are fixed (Transmitters) and some are free to move (Receivers). In general, we found that the modularity of randomized networks was lower than static networks (see Supplementary Figure S4B for simulations of completely fixed systems, not shown here). This is understandable because randomized distribution of edges among the nodes yields an unorganized network structure. In comparison, as static networks grow, they maintain structure.

We further ran simulations with fixed edges for different charge durations and hydrogen peroxide concentrations as in the earlier simulations, to examine static biofilm cultures relative to well stirred systems. We found differences in charge duration and initial hydrogen peroxide concentrations did not affect the modularity as molecular concentrations that are represented by node weights do not affect the spatial structure of the network. We then analyzed the output (i.e., GFP level) for these simulations. To compare the output of these static cultures we calculated the ratio of average GFP at 180 timesteps from randomized networks to the static networks. We use this as a way to measure the benefit of cells in the traditional well-mixed system to those in a fixed or partially fixed system (i.e., cells fixed to an electrode propagating signals to those in fluid nearby). In Figure 6C, we plotted these ratios for each inducer and system type. For the new case of transmitters fixed to an electrode, we also tested a case in which the receivers are also fixed to emulate multilayer deposition of cells onto an electrode as a potential design. This is more representative to a complete biofilm. The fold change calculated from these transmitter fixed cases were done relative to static networks of electrically induced Transmitter/Receiver simulations.

Here we see that for chemical addition, there was little difference between the network structures. This results from the fact that all nodes experience the same initial inducer concentrations. For electrically induced systems, there was minimal effect on the Monoculture at all charge durations. In the Transmitter/Receiver system, we found that for 30 steps of charge there was an approximately 3-fold increase in signal when randomizing the network. In fixed transmitter simulations, we found a substantially larger range for the overall system output. These fold increases are indicative of how edge randomization generally increases output while strategic spatial arrangement of the co-culture with respect to inducer sources can largely amplify signal throughput.

Finally, we assessed how topological effects leading to increased modularity affect signaling within the network. We calculated the percentage of cells that are active (GFP above a 10^3 threshold) and the median GFP for these Transmitter/Receiver simulations with various edge dynamics (Figures 6D, E). We observed that the static networks had both the lowest median GFP and the fraction of active cells (Figures 6D, E). Interestingly, our simulations suggest that introducing transmitters that are fixed to the electrode increases the overall activation and median GFP over completely randomized networks, and this is irrespective of receiver conformation (fixed or not). We suggest this is due to the faster and increased AI-1 production that is enabled by transmitter proximity to the electrode (Supplementary Figure S4C). Randomizing the receivers further increased estimated output. This is a consequence of allowing the whole receiver population's increased contact with the transmitter population, as the AI-1 source. This was evident as the network's GFP distributions where increasing static network components correlate to a wider range in GFP values (Supplementary Figure S4D). Overall, these results reveal that while high modularity yields increased signal heterogeneity, it also lowered signal output compared to low modularity networks. That said, the strategic or intentional organization of subpopulations can drastically increase output, despite increased modularity.

4 Discussion

In this work, we developed a graphical network approach for modeling multi-population bacterial cultures. By coarse graining the cell-to-cell signaling interactions that are known to occur in complex bacterial systems (Waters and Bassler, 2005) and leveraging intrinsic network properties that attempt to simulate spatial distributions, we have elucidated signal dynamics that would be very difficult to ascertain using traditional deterministic population scale multicellular modeling. The implementation of a graph-based model allowed us to vary network structures that we had previously implemented experimentally. We were able to determine network parameters (probabilities of growth, molecule production, gene activation) that when employed in the model, accurately recapitulated the experimental observations. Then node weights (other state variables such as inducer levels, substrate levels, etc.) were examined to better understand the experimental results. Perhaps more importantly, with this agreement we then tested hypotheses regarding the spatial composition of microbial systems. Further, by implementing various edge architectures, we attempted to mimic various engineered and endogenous culture structures. We mimicked stirred batch conditions common to biomanufacturing settings via edge randomization. Static edge conformations imitate biofilms found in nature and other immobilized or hydrogel-assembled cell systems. Additionally, we could easily accommodate varied edge profiles in our model so that we could test how relative spatial structures affect communication between different populations.

Owing to the natural tendency to think in terms of subpopulations and quorum sensing (Servinsky et al., 2016), we introduced the notion that network modularity would be a valuable tool in analyzing bacterial networks when organized in the various experimental configurations. In testing fixed spatial conformations we found that for increased modularity, meaning more subcommunities in the network, maximum signal throughput is reduced and delayed for simulations with an electrode as an input source. We suggest this is attributed to the need for the input signal to diffuse into each subcommunity for the secondary signal to then be produced and diffused back out for further signaling. We suggest this introduces an increase in noise at each step of signal transmission due to structural constraints. That said, these decreases in signal can be overcome by spatially orienting transmitter nodes close to the electrode as the input signal source. We further tested fixing all transmitter nodes to the input signal source (the electrode) and found that regardless of whether the receivers were fixed or randomized this restored signal in fixed networks and resulted in higher expression than in randomized simulations. Correspondingly, in Terrell et al. (2021), they demonstrated that by fixing microbes to a gold electrode they could produce AI-1 with electrochemical stimulation, and this was shown to be quite successful in signal propagation (more so than in VanArsdale et al. (2022), where the transmitter and receiver populations were fully mixed in a stirred vessel). Unfortunately, in neither case was it experimentally feasible to monitor the AI-1 diffusion and activation across the system boundaries (Terrell et al., 2021). Here, our work may provide theoretical insight into the signaling occurring in these types of experimental configurations and those found in natural biofilm systems, where measurements in real time and at small distances is difficult.

Additionally, we suggest that models such as this can be further extended to simulate other spatial conformations of cell populations to provide insight into how much input and signal transmission is necessary for successful outcomes (Chun et al., 2021). These include cases where synthetic assembled consortia of higher complexity may be cultured together in batch or spatially fixed within gels (Luo and Shoichet, 2004), between membranes or 3D printed niches (Duraj-Thatte et al., 2021), or within varying ecological niches (Li et al., 2007; Schiessl et al., 2019; Cornell et al., 2020; Ciccarese et al., 2022; Evans et al., 2023). For example, the field of biomaterials has implemented the spatial confinement of cells within hydrogel structures and microcapsules for the use in generating functional living materials and to recreate micro communities found in nature (Dsouza et al., 2022; Molinari et al., 2022; Wang et al., 2022; Yanamandra et al., 2022).

Data availability statement

The simulation datasets for this study can be found here - http://github.com/kaychun29/bio-network-simulations. Further inquiries can be made to the corresponding author - bentley@umd.edu.

Author contributions

KC: Conceptualization, Investigation, Methodology, Writing–original draft, Writing–review and editing. EV: Conceptualization, Writing–review and editing. EM: Writing–review and editing. GP: Writing–review and editing. WB: Conceptualization, Funding acquisition, Supervision, Writing–original draft, Writing–review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The authors would like to acknowledge partial support of this work by the National Science Foundation (MCB# 2227598, CBET# 1932963), the Department of Energy (BER#SCW1710), the Defense Threat Reduction Agency (HDTRA1-19-1-0021), and the Gordon and Betty Moore Foundation (#11395). As well as the COMBINE Fellowship (NSF Graduate Training Grant, DGE-1632976, to KC).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fsysb.2024.1291293/full#supplementary-material

References

Barabasi, A. L. (2013). Network science. *Philos. Trans. A Math. Phys. Eng. Sci.* 371 (1987), 20120375. doi:10.1098/rsta.2012.0375

Bhokisham, N., VanArsdale, E., Stephens, K. T., Hauk, P., Payne, G. F., and Bentley, W. E. (2020). A redox-based electrogenetic CRISPR system to connect with and control biological information networks. *Nat. Commun.* 11 (1), 2427. doi:10.1038/s41467-020-16249-x

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* 2008 (10), P10008. doi:10.1088/1742-5468/2008/10/p10008

Cao, M., Gao, M., Suastegui, M., Mei, Y., and Shao, Z. (2020). Building microbial factories for the production of aromatic amino acid pathway derivatives: from commodity chemicals to plant-sourced natural products. *Metab. Eng.* 58, 94–132. doi:10.1016/j.ymben.2019.08.008

Chun, K., Stephens, K., Wang, S., Tsao, C. Y., Payne, G. F., and Bentley, W. E. (2021). Parsed synthesis of pyocyanin via co-culture enables context-dependent intercellular redox communication. *Microb. Cell Fact.* 20 (1), 215. doi:10.1186/s12934-021-01703-2

Ciccarese, D., Micali, G., Borer, B., Ruan, C., Or, D., and Johnson, D. R. (2022). Rare and localized events stabilize microbial community composition and patterns of spatial self-organization in a fluctuating environment. *ISME J.* 16 (5), 1453–1463. doi:10.1038/s41396-022-01189-9

Cornell, W. C., Zhang, Y., Bendebury, A., Hartel, A. J. W., Shepard, K. L., and Dietrich, L. E. P. (2020). Phenazine oxidation by a distal electrode modulates biofilm morphogenesis. *Biofilm* 2, 100025. doi:10.1016/j.bioflm.2020.100025

Dinh, C. V., Chen, X., and Prather, K. L. J. (2020). Development of a *quorum*-Sensing based circuit for control of coculture population composition in a naringenin production system. *ACS Synth. Biol.* 9 (3), 590–597. doi:10.1021/acssynbio.9b00451

Dsouza, A., Constantinidou, C., Arvanitis, T. N., Haddleton, D. M., Charmet, J., and Hand, R. A. (2022). Multifunctional composite hydrogels for bacterial capture, growth/elimination, and sensing applications. *ACS Appl. Mater Interfaces* 14 (42), 47323–47344. doi:10.1021/acsami.2c08582

Duraj-Thatte, A. M., Manjula-Basavanna, A., Rutledge, J., Xia, J., Hassan, S., Sourlis, A., et al. (2021). Programmable microbial ink for 3D printing of living materials produced from genetically engineered protein nanofibers. *Nat. Commun.* 12 (1), 6600. doi:10.1038/s41467-021-26791-x

Evans, C. R., Smiley, M. K., Thio, S. A., Wei, M., Price-Whelan, A., Min, W., et al. (2023). Spatial heterogeneity in biofilm metabolism elicited by local control of phenazine methylation. *bioRxiv* 120, e2313208120. doi:10.1073/pnas.2313208120

Gosak, M., Markovic, R., Dolensek, J., Slak Rupnik, M., Marhl, M., Stozer, A., et al. (2018). Network science of biological systems at different scales: a review. *Phys. Life Rev.* 24, 118–135. doi:10.1016/j.plrev.2017.11.003

Gwon, D.-a., Seo, E., and Lee, J. W. (2023). Construction of synthetic microbial consortium for violacein production. *Biotechnol. Bioprocess Eng.* 28, 1005–1014. doi:10.1007/s12257-022-0284-5

Hagberg, A. A., Schult, D. A., and Swart, P. J. (2008). "Exploring network structure, dynamics, and function using NetworkX," in Proceedings of the 7th Python in Science Conference 2008.

Hunter, J. D. (2007). *MATPLOTLIB: a 2D graphics environment*. Scientific Programming.

Hwang, I. Y., Koh, E., Wong, A., March, J. C., Bentley, W. E., Lee, Y. S., et al. (2017). Engineered probiotic *Escherichia coli* can eliminate and prevent *Pseudomonas aeruginosa* gut infection in animal models. *Nat. Commun.* 8, 15028. doi:10.1038/ncomms15028

Jiang, M., and Zhang, H. (2016). Engineering the shikimate pathway for biosynthesis of molecules with pharmaceutical activities in *E. coli*. *Curr. Opin. Biotechnol.* 42, 1–6. doi:10.1016/j.copbio.2016.01.016

Kim, E., Li, J., Kang, M., Kelly, D. L., Chen, S., Napolitano, A., et al. (2019). Redox is a global biodevice information processing modality. *Proc. IEEE Inst. Electr. Electron Eng.* 107 (7), 1402–1424. doi:10.1109/JPROC.2019.2908582

Li, J., Attila, C., Wang, L., Wood, T. K., Valdes, J. J., and Bentley, W. E. (2007). *Quorum* sensing in *Escherichia coli* is signaled by AI-2/LsrR: effects on small RNA and biofilm architecture. *J. Bacteriol.* 189 (16), 6011–6020. doi:10.1128/JB.00014-07

Li, J., Kim, E., Gray, K. M., Conrad, C., Tsao, C. Y., Wang, S. P., et al. (2020). Multifunctional artificial artery from direct 3D printing with built-in ferroelectricity and tissue-matching modulus for real-time sensing and occlusion monitoring. *Adv. Funct. Mater.* 30 (30), 2002868. doi:10.1002/adfm.202002868

Luo, Y., and Shoichet, M. S. (2004). A photolabile hydrogel for guided three-dimensional cell growth and migration. *Nat. Mater* 3 (4), 249–253. doi:10.1038/nmat1092

Menczer, F., Fortunato, S., and Davis, C. A. (2020). *A first Course in network science*.

Mimee, M., Tucker, A. C., Voigt, C. A., and Lu, T. K. (2015). Programming a human commensal bacterium, Bacteroides thetaiotaomicron, to sense and respond to stimuli in the murine gut microbiota. *Cell Syst.* 1 (1), 62–71. doi:10.1016/j.cels.2015.06.001

Molinari, S., Tesoriero, R. F., Jr., Li, D., Sridhar, S., Cai, R., Soman, J., et al. (2022). A *de novo* matrix for macroscopic living materials from bacteria. *Nat. Commun.* 13 (1), 5544. doi:10.1038/s41467-022-33191-2

Newman, M., Barabási, A.-L., and Watts, D. J. (2006). *The Structure and dynamics of networks*. Princeton: Princeton University Press.

Newman, M. E. J., and Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E* 69, 026113. doi:10.1103/PhysRevE.69.026113

Pomposiello, P. J., and Demple, B. (2001). Redox-operated genetic switches: the SoxR and OxyR transcription factors. *Trends Biotechnol.* 19, 109–114. doi:10.1016/s0167-7799(00)01542-0

Quan, D. N., Tsao, C. Y., Wu, H. C., and Bentley, W. E. (2016). *Quorum* sensing desynchronization leads to bimodality and patterned behaviors. *PLoS Comput. Biol.* 12 (4), e1004781. doi:10.1371/journal.pcbi.1004781

Sayama, H. (2015). "Simulating dynamics on networks," in *Introduction to the modeling and analysis of complex systems*. Open SUNY Textbooks, Milne Library

Schiessl, K. T., Hu, F., Jo, J., Nazia, S. Z., Wang, B., Price-Whelan, A., et al. (2019). Phenazine production promotes antibiotic tolerance and metabolic heterogeneity in *Pseudomonas aeruginosa* biofilms. *Nat. Commun.* 10 (1), 762. doi:10.1038/s41467-019-08733-w

Servinsky, M. D., Terrell, J. L., Tsao, C. Y., Wu, H. C., Quan, D. N., Zargar, A., et al. (2016). Directed assembly of a bacterial *quorum*. *ISME J.* 10 (1), 158–169. doi:10.1038/ismej.2015.89

Shiloach, J., Kaufman, J., Guillard, A. S., and Fass, R. (1996). Effect of glucose supply strategy on acetate accumulation, growth, and recombinant protein production by *Escherichia coli* BL21 (lambdaDE3) and *Escherichia coli* JM109. *Biotechnol. Bioeng.* 49, 421–428. doi:10.1002/(SICI)1097-0290(19960220)49:4<421::AID-BIT9>3.0.CO;2-R

Stephens, K., Pozo, M., Tsao, C. Y., Hauk, P., and Bentley, W. E. (2019). Bacterial co-culture with cell signaling translator and growth controller modules for autonomously regulated culture composition. *Nat. Commun.* 10 (1), 4129. doi:10.1038/s41467-019-12027-6

Terrell, J. L., Tschirhart, T., Jahnke, J. P., Stephens, K., Liu, Y., Dong, H., et al. (2021). Bioelectronic control of a microbial community using surface-assembled electrogenetic cells to route signals. *Nat. Nanotechnol.* 16 (6), 688–697. doi:10.1038/s41565-021-00878-4

Tschirhart, T., Kim, E., McKay, R., Ueda, H., Wu, H. C., Pottash, A. E., et al. (2017). Electronic control of gene expression and cell behaviour in *Escherichia coli* through redox signalling. *Nat. Commun.* 8, 14030. doi:10.1038/ncomms14030

VanArsdale, E., Navid, A., Chu, M. J., Halvorsen, T. M., Payne, G. F., Jiao, Y., et al. (2023). Electrogenetic signaling and information propagation for controlling microbial consortia via programmed lysis. *Biotechnol. Bioeng.* 120 (5), 1366–1381. doi:10.1002/bit.28337

VanArsdale, E., Pitzer, J., Wang, S., Stephens, K., Chen, C. Y., Payne, G. F., et al. (2022). Electrogenetic signal transmission and propagation in coculture to guide production of a small molecule, tyrosine. *ACS Synth. Biol.* 11 (2), 877–887. doi:10.1021/acssynbio.1c00522

Virgile, C., Hauk, P., Wu, H. C., Shang, W., Tsao, C. Y., Payne, G. F., et al. (2018). Engineering bacterial motility towards hydrogen-peroxide. *PLoS One* 13 (5), e0196999. doi:10.1371/journal.pone.0196999

Wang, L., Zhang, X., Tang, C., Li, P., Zhu, R., Sun, J., et al. (2022). Engineering consortia by polymeric microbial swarmbots. *Nat. Commun.* 13 (1), 3879. doi:10.1038/s41467-022-31467-1

Waskom, M. (2021). seaborn: statistical data visualization. *J. Open Source Softw.* 6 (60), 3021. doi:10.21105/joss.03021

Waters, C. M., and Bassler, B. L. (2005). *Quorum* sensing: cell-to-cell communication in bacteria. *Annu. Rev. Cell Dev. Biol.* 21, 319–346. doi:10.1146/annurev.cellbio.21.012704.131001

Yanamandra, A. K., Bhusari, S., del Campo, A., Sankaran, S., and Qu, B. (2023). *In vitro* evaluation of immune responses to bacterial hydrogels for the development of living therapeutic materials. *Biomaterials Advances*, 153, 213554. doi:10.1016/j.bioadv.2023.213554

Zargar, A., Quan, D. N., Emamian, M., Tsao, C. Y., Wu, H. C., Virgile, C. R., et al. (2015). Rational design of 'controller cells' to manipulate protein and phenotype expression. *Metab. Eng.* 30, 61–68. doi:10.1016/j.ymben.2015.04.001

Zhao, S., Li, F., Yang, F., Ma, Q., Liu, L., Huang, Z., et al. (2022). Microbial production of valuable chemicals by modular co-culture strategy. *World J. Microbiol. Biotechnol.* 39 (1), 6. doi:10.1007/s11274-022-03447-6

TYPE Methods
PUBLISHED 21 March 2024
DOI 10.3389/fsysb.2024.1355595

OPEN ACCESS

EDITED BY
Rongling Wu,
The Pennsylvania State University (PSU),
United States

REVIEWED BY
Xiaoxi Shen,
Texas State University, United States
Luis Diambra,
National University of La Plata, Argentina

*CORRESPONDENCE
Li Zhang,
✉ li.zhang@ucsf.edu
Tao He,
✉ hetao@sfsu.edu

†These authors share senior authorship

RECEIVED 14 December 2023
ACCEPTED 05 March 2024
PUBLISHED 21 March 2024

CITATION
Le P, Gong X, Ung L, Yang H, Keenan BP, Zhang L and He T (2024), A robust ensemble feature selection approach to prioritize genes associated with survival outcome in high-dimensional gene expression data.
Front. Syst. Biol. 4:1355595.
doi: 10.3389/fsysb.2024.1355595

A robust ensemble feature selection approach to prioritize genes associated with survival outcome in high-dimensional gene expression data

Phi Le[1], Xingyue Gong[2], Leah Ung[1], Hai Yang[1], Bridget P. Keenan[1,3], Li Zhang[1,3,4*†] and Tao He[5*†]

[1]Division of Hematology/Oncology, Department of Medicine, University of California, San Francisco, San Francisco, CA, United States, [2]Department of Physiological Nursing, School of Nursing, University of California, San Francisco, San Francisco, CA, United States, [3]Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, San Francisco, CA, United States, [4]Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA, United States, [5]Department of Mathematics, San Francisco State University, San Francisco, CA, United States

Exploring features associated with the clinical outcome of interest is a rapidly advancing area of research. However, with contemporary sequencing technologies capable of identifying over thousands of genes per sample, there is a challenge in constructing efficient prediction models that balance accuracy and resource utilization. To address this challenge, researchers have developed feature selection methods to enhance performance, reduce overfitting, and ensure resource efficiency. However, applying feature selection models to survival analysis, particularly in clinical datasets characterized by substantial censoring and limited sample sizes, introduces unique challenges. We propose a robust ensemble feature selection approach integrated with group Lasso to identify compelling features and evaluate its performance in predicting survival outcomes. Our approach consistently outperforms established models across various criteria through extensive simulations, demonstrating low false discovery rates, high sensitivity, and high stability. Furthermore, we applied the approach to a colorectal cancer dataset from The Cancer Genome Atlas, showcasing its effectiveness by generating a composite score based on the selected genes to correctly distinguish different subtypes of the patients. In summary, our proposed approach excels in selecting impactful features from high-dimensional data, yielding better outcomes compared to contemporary state-of-the-art models.

KEYWORDS

colorectal cancer, ensemble feature selection, high-dimensional data, time-to-event outcome, pseudo variables, group lasso

Introduction

Next-generation sequencing (NGS) techniques (Hu et al., 2021) can provide us with information on the expression of more than 30,000 genes, which helps researchers understand gene regulations and interactions to find treatments for diseases. However, the number of genes associated with a particular disease is small (Yang et al., 2005).

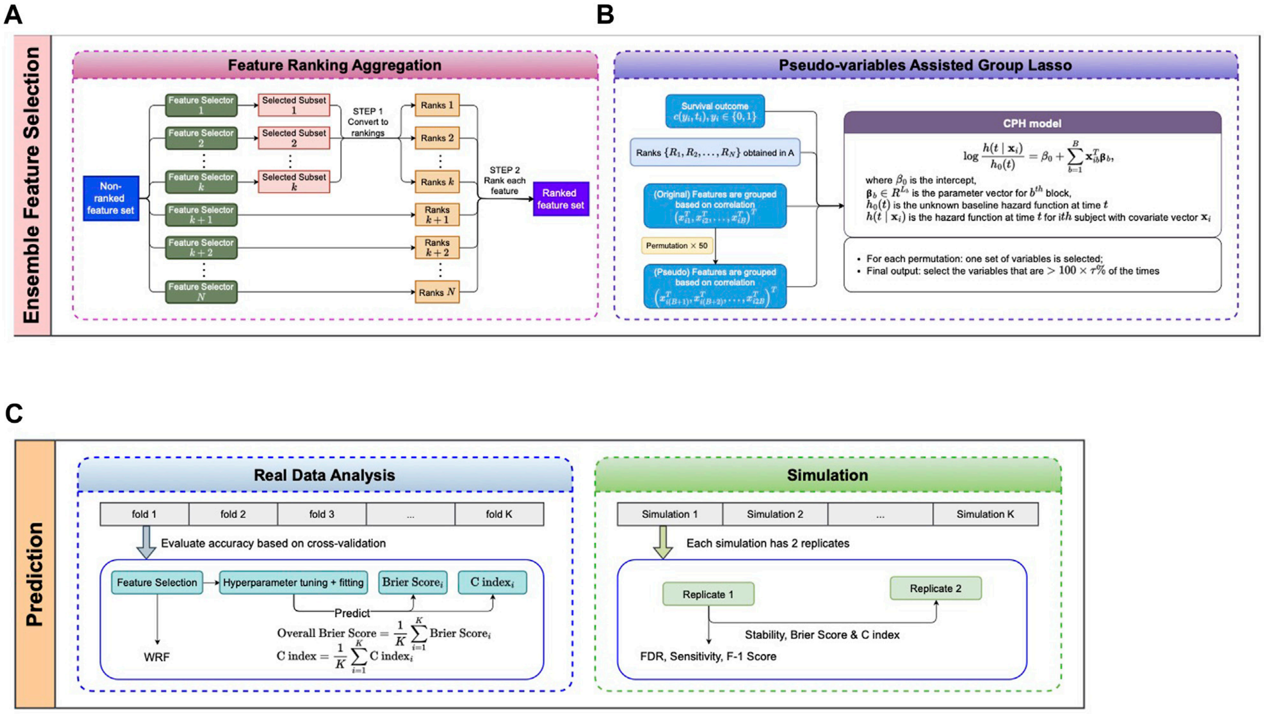

A
B
Ensemble Feature Selection
Feature Ranking Aggregation
Non-ranked feature set
Feature Selector 1
Feature Selector 2
Feature Selector k
Feature Selector $k+1$
Feature Selector $k+2$
Feature Selector N
Selected Subset 1
Selected Subset 2
Selected Subset k
STEP 1 Convert to rankings
Ranks 1
Ranks 2
Ranks k
Ranks $k+1$
Ranks $k+2$
Ranks N
STEP 2 Rank each feature
Ranked feature set
Pseudo-variables Assisted Group Lasso
Survival outcome $c(y_i, t_i), y_i \in \{0, 1\}$
Ranks $\{R_1, R_2, \dots, R_N\}$ obtained in A
(Original) Features are grouped based on correlation $(x_{i1}^T, x_{i2}^T, \dots, x_{iB}^T)^T$
Permutation × 50
(Pseudo) Features are grouped based on correlation $\left(x_{i(B+1)}^T, x_{i(B+2)}^T, \dots, x_{i2B}^T\right)^T$
CPH model
$\log \frac{h(t \mid \mathbf{x}_i)}{h_0(t)} = \beta_0 + \sum_{b=1}^{B} \mathbf{x}_{ib}^T \boldsymbol{\beta}_b,$
where β_0 is the intercept,
$\boldsymbol{\beta}_b \in R^{L_b}$ is the parameter vector for b^{th} block,
$h_0(t)$ is the unknown baseline hazard function at time t
$h(t \mid \mathbf{x}_i)$ is the hazard function at time t for ith subject with covariate vector $\mathbf{x}_i$
• For each permutation: one set of variables is selected;
• Final output: select the variables that are $> 100 \times \tau\%$ of the times
C
Prediction
Real Data Analysis
fold 1
fold 2
fold 3
...
fold K
Evaluate accuracy based on cross-validation
Feature Selection
Hyperparameter tuning + fitting
Brier Score$_i$
C index$_i$
Predict
WRF
Overall Brier Score $= \frac{1}{K}\sum_{i=1}^{K}$ Brier Score$_i$
C index $= \frac{1}{K}\sum_{i=1}^{K}$ C index$_i$
Simulation
Simulation 1
Simulation 2
...
Simulation K
Each simulation has 2 replicates
Replicate 1
Replicate 2
Stability, Brier Score & C index
FDR, Sensitivity, F-1 Score

FIGURE 1
Proposed pipeline. **(A, B)** Ensemble feature selection. **(A)** Feature selection based on different methods and aggregation of selected features. **(B)** Pseudo-variables assisted group lasso. **(C)** Prediction for real data and simulation datasets.

Therefore, we need to develop powerful tools to select genes that work as a group and are associated with clinical outcomes. Feature selection approaches were developed to choose the most relevant and informative features for research questions from the original raw set of features; therefore, they can help avoid overfitting, reduce training time, handle the challenge of dimensionality, and simplify data representations.

Survival analysis (Klein et al., 1992) is a statistical model studying time-to-event data in which the event may not be observed (censored) during the study because of loss to follow-up or early end of the study. Due to the presence of censoring, which is a unique characteristic in survival analysis, there is a need to develop novel techniques to work with feature selections for survival data, especially for high-throughput gene expression data in which most of the potential predictors are unimportant, with nearly no effect on the outcome (Friedman et al., 2010). The Cox proportional hazards model is the most commonly used technique for analyzing survival data. However, it was not designed for high-dimensional datasets with a large number of predictors. Lasso (Least Absolute Shrinkage and Selection Operator) introduces a penalty term to the Cox model's likelihood function, which penalizes the absolute values of the regression coefficients. By forcing some coefficients to be exactly zero, Lasso effectively performing variable selection. In addition, there are models tailored to effectively handle situations where the number of features outweighs the number of observations (Li et al., 2018). Machine learning techniques that inherently handle high-dimensional data have been adapted to handle censored data, offering more flexible alternatives for analyzing high-dimensional, right-censored, heterogeneous data. However, unlike statistical models based on a mathematical framework, machine learning approaches do not impose a specified relationship on the predictors and outcomes and rely mainly on–data-driven algorithms, which makes it hard to interpret results. Furthermore, a lot of feature selection methods for survival analysis use a scoring model (Neums et al., 2019) to measure variations of features to select important features. Since the scoring algorithm was developed specifically to take care of the data censors and tie events of survival data, the results are biased (Munson et al., 2009) which may lead to selecting nonimportant features and provide a less accurate prediction.

We introduce a robust and effective "Pseudo-variables Assisted Group Lasso" method built on the ensemble idea, i.e., "more heads are better than one", where features obtained from different selectors are aggregated to enhance the final selection. Moreover, we incorporated pseudo-variables which we know are irrelevant to the outcome and the permutation technique to assist the selection. The ensemble and pseudo-variables are nicely embedded into the Group Lasso model to yield the final output. Among aggregated features, only the features that consistently show stronger signals than the pseudo-variables (known noises) across permutations will be selected. We used colorectal cancer data from The Cancer Genome Atlas (TCGA) for illustration of our proposed approach. In addition, we performed simulation studies based on two different settings, where the first one mimicked the colorectal cancer data, and the second considered more complicated situations under various scenarios. For each simulation, we first simulated gene expression data for hundreds of genes and then generated survival outcomes based on some causal genes. The proposed feature

TABLE 1 Parameters used for feature selection methods.

Approach	R package	Parameter	Description	Value
MIM (select top k)	praznik	k	Select top k features	25
MRMR (select top k)				
RF Min Depth (select top k)	randomForestSRC	ntree	Number of trees	1,000
RF Var Imp (select top k)		mtry	Number of variables to possibly split at each node	default
		nodesize	Minimum size of terminal node	15
RF Var Hunt (select top k)		k	Select top k feature	25
		nsplit	Number of random splits for splitting a variable	10
Cox (select up to top k which have *p*-value less than α)	survival	k	Select top k feature	25
		alpha	*p*-value threshold	0.05
LASSO	glmnet	lamba	Tuning parameter grid values	$10^{(-10,-9.9,...,0,...,9.9,10)}$
Ensemble1		ρ_T	Minimum pairwise correlation within block	0.75
Ensemble2		K	Total number of permutations	50
		τ	Threshold of selection percentage	0.5

selection ensemble method was applied to “uncover” the causal genes and compared to the existing methods.

Materials and methods

Colorectal cancer data set from TCGA database

Raw gene expression counts were downloaded from colon cancer (The Cancer Genome Atlas Network, 2012) datasets using The Cancer Genomics Cloud (Lau et al., 2017); additional clinical metadata was downloaded from cBioportal (Cerami et al., 2012). The mRNA-Seq data from TCGA was produced using the Illumina HiSeq 2000 platform and processed by the RNAseqV2 pipeline, which used MapSplice for alignment and RSEM for quantification.

A robust feature selection ensemble

The proposed pseudo-variable-assisted feature ensemble procedure has two major steps: 1) aggregating the feature selection results from multiple feature selectors (Figure 1A) and 2) fitting a group Lasso model on the identified feature set with a new permutation-assisted tuning strategy (Figure 1B). In the second step, the group is defined based on the correlation structure, ensuring that features are highly correlated within each group.

Aggregating the results from different feature selection approaches is a critical step in ensemble learning. The outputs of the different approaches can be various, either the subsets of selected features, the rankings of all features, or both. We applied the same scheme as in (He et al., 2022) to obtain the ranked feature set depending on the types of outputs (Figure 1A), where the final rank is an aggregation from each ranking. We assume that the observations are $(\boldsymbol{x}_i, y_i), i = 1, \ldots, n$, where $\boldsymbol{x}_i$ is a G-dimensional vector in which each feature has its aggregated rank, and y_i is a survival outcome. Without loss of generality, we assume the G features are quantitative variables (e.g., gene expressions). However, the proposed method can be applied to categorical or mixed-type variables. Similar to (He et al., 2022), we can rewrite the G-dimensional vector $\boldsymbol{x}_i$ as $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}^T, \boldsymbol{x}_{i2}^T, \ldots, \boldsymbol{x}_{iB}^T)^T$ with $\boldsymbol{x}_{ib}$ of dimension $L_b, b = 1, \ldots, B$, $\sum_{b=1}^{B} L_b = G$, based on their correlation structure such that within each block, the absolute value of pairwise correlation is all greater than a correlation threshold ρ_T.

Next we consider a Group Lasso model (Utazirubanda et al., 2021) on the ranked feature set (Figure 1B) for survival outcomes. For commonly seen right censored survival data, $y_i = (\mathrm{T_i}, \Delta_i)$ is a survival outcome, where $T_i = \min(\mathrm{U_i}, V_i)$, $\Delta_i = I(U_i \leq V_i) \in \{0, 1\}$, with U_i and V_i denote the event time of interest and the censoring time for the i th subject, respectively. We model the relationship between the survival outcomes y_i and features $\boldsymbol{x}_i$ using the Cox proportional hazards model (Deo et al., 2021)

$$log \frac{h(t \mid \boldsymbol{x}_i)}{h_o(t)} = \beta_0 + \sum_{b=1}^{B} \boldsymbol{x}_{ib}^T \boldsymbol{\beta}_b \triangleq \gamma_{\boldsymbol{\beta}}(\boldsymbol{x}_i),$$

where β_0 is the intercept, and $\boldsymbol{\beta}_b \in R^{L_b}$ is the parameter vector for the bth block, $h_o(t)$ is the (unknown) baseline hazard function at time t, and $h(t \mid \boldsymbol{x}_i)$ is the hazard function at time t for the ith subject with covariate vector $\boldsymbol{x}_i$. We aim to identify which gene groups amongst the B groups associated with the survival outcomes.

Based on the partial likelihood function,

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left\{ \frac{\exp\left[\gamma_{\boldsymbol{\beta}}(\boldsymbol{x}_i)\right]}{\sum_{k \in Q_j} \exp\left[\gamma_{\boldsymbol{\beta}}(\boldsymbol{x}_k)\right]} \right\}^{\Delta_i},$$

TABLE 2 Simulation scenario.

Scenarios	Label	Sample size	# Of genes	Event rate	Sparsity (# of causal genes/# of genes)	β_1	β_2	β_3
1	n100_G1200_er0.3	100	1,200	0.3	30/1,200	−6	8	−10
2	n100_G1200_er0.5	100	1,200	0.5	30/1,200	−2	3	−4
3	n100_G1200_er0.7	100	1,200	0.7	30/1,200	−2	−3	4
4	n100_G600_er0.3	100	600	0.3	30/600	−6	8	−10
5	n100_G600_er0.5	100	600	0.5	30/600	−2	3	−4
6	n100_G600_er0.7	100	600	0.7	30/600	−2	−3	4
7	n200_G1200_er0.3	200	1,200	0.3	30/1,200	−6	8	−10
8	n200_G1200_er0.5	200	1,200	0.5	30/1,200	−2	3	−4
9	n200_G1200_er0.7	200	1,200	0.7	30/1,200	−2	−3	4
10	n200_G600_er0.3	200	600	0.3	30/600	−6	8	−10
11	n200_G600_er0.5	200	600	0.5	30/600	−2	3	−4
12	n200_G600_er0.7	200	600	0.7	30/600	−2	−3	4

$Q_j = \{k\colon T_k \geq T_j\}$, we can obtain the estimation of the complete parameter vector $\boldsymbol{\beta}$ by minimizing the following objective function.

$$Q_\lambda(\boldsymbol{\beta}) = -L(\boldsymbol{\beta}) + \lambda\sum_{b=1}^{B} s_b\|\boldsymbol{\beta}_b\|_2.$$

Recall that λ is the tuning parameter that controls the amount of shrinkage (larger λ shrinks more coefficients to zero), and s_b is used to rescale the penalty to each group. To ensure the top-ranked features are more likely to be selected, we put a small penalty on top-ranked feature sets by proposing using the product of the minimum rank among each feature set and $\sqrt{L_b}$.

The objective of this study is more about selecting the important features than improving the prediction accuracy. Therefore, we propose to use the pseudo-variables assisted tuning strategy (He et al., 2022) to facilitate the group-lasso tuning parameter selection. This strategy is built on the idea of combining the original and permutated input features (e.g., expressed genes), where the permutations work as a control to determine the significance of each group. Hence, we can select significantly important genes (not by chance).

It is known that the λ in group-lasso-type regularization controls the amount of shrinkage. As λ increases, fewer groups are selected. A group can be considered more important one if it is selected when λ is large. Based on these observations, we can define an importance measure $V_b = sup\{\lambda$: the coefficient for bth group is nonzero$\}$, for each of the 2B groups, including the B groups from original input features $(b = 1, \ldots, B)$ and their B groups of permutated copies $(b = B + 1, \ldots, 2B)$. For each permutation, groups from original input features are selected if their V_b is larger than $\max_{B+1 \leq b \leq 2B} V_b$, i.e., the strongest signal among permutated groups which we have known are irrelevant groups. After running K (e.g., K = 50) times of permutations, we selected the groups of features that have been selected more than a certain number of percentages τ (i.e., $\tau = 0.5$) among K permutations.

Feature selection and machine learning algorithms

In our study, we evaluated nine different feature selection methods, including seven existing feature selection methods and two robust ensemble feature selectors we constructed. The nine selectors can be divided into four major groups: (I) feature selection algorithms based on mutual information optimization: mutual information maximization (MIM) (Torkkola, 2003), minimum redundancy maximum relevance (MRMR) (Radovic et al., 2017); and (II) random forest-based approaches: a random forest minimal depth (RF Min Depth) (Ishwaran et al., 2008; Ishwaran et al., 2011), a random forest variable importance (RF Var Imp) (Archer and Kimes, 2008), a random forest variable hunting (RF Var Hunt) (Chen and Ishwaran, 2013); and (III) Cox-based approaches: Cox hazard proportional (Cox) (Deo et al., 2021) and ℓ_1 penalized Cox (Lasso) (Goeman, 2010); (IV) ensemble learners (Zhou, 2021). We created two feature ensembles, Ensemble 1 and Ensemble 2, where the first one is the ensemble of Lasso, Cox, and MIM, and the second is the ensemble of Lasso, Cox, MIM, and MRMR. Parameters used in the paper were included in Table 1.

To compare the results of our feature selection ensemble method with others, we tested the selected features on five well-known prediction models using machine learning and non-parametric techniques: (I) the Cox model with ℓ_1 regularization (Lasso) (Binder, 2015); (II) models based on boosted trees: xgboost (XGB) (Chen and Guestrin, 2016) (III) boosted gradient linear models: xgboost based on linear learner (XGB linear) (Chen and Guestrin, 2016) and (IV) random forest-based methods: random survival forest (RF) (Segal, 2004) and ranger (Wright and ranger, 2017). All feature selection methods and machine learning algorithms assessed here can handle the time-to-event outcome.

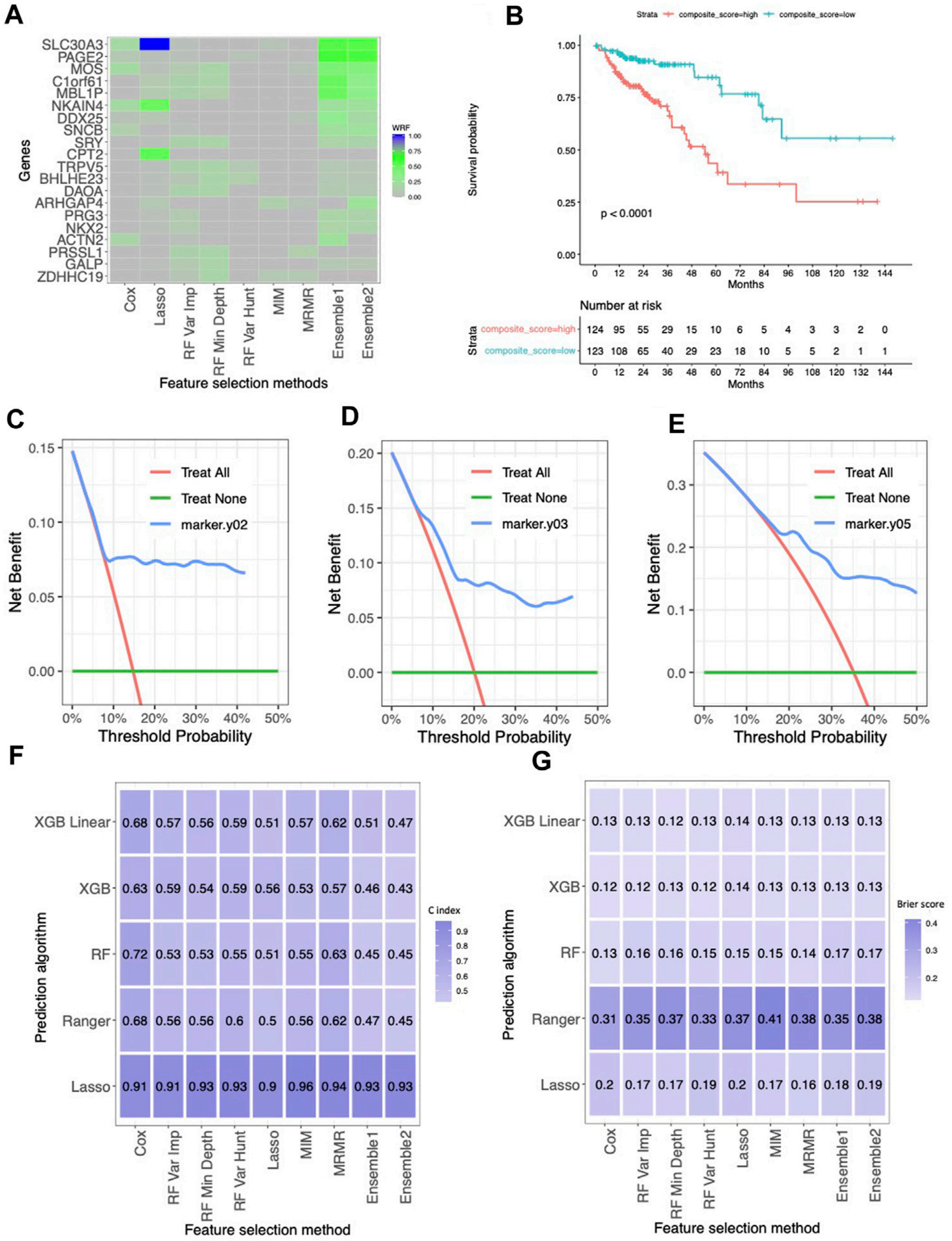


FIGURE 2
The results for the TCGA colorectal cancer dataset. **(A)** Normalized selection frequency of the top 20 selected genes by each feature selection approach. Each row represents an individual single gene, and each column represents the feature selection approaches. **(B)** Kaplan-Meier survival curves. The low-risk group and high-risk group were defined by median of the composite score. The composite score was calculated as the linear combination of those genes selected by ensemble approach and their coefficients in the cox proportional hazard model. **(C–E)** DCA of 2 year, 3 year and 5 year. **(F)** Heatmap of concordance index (C-index). The heatmap shows the mean value of the C-Index across 5 repeats of 5-fold cross-validation for each combination of machine learning algorithms (rows) and feature selection methods (columns). **(G)** Heatmap of Brier Score. The heatmap shows the mean value of the Brier Score across 5 repeats of 5-fold cross-validation for each combination of machine learning algorithms (rows) and feature selection methods (columns).

Simulation

To mimic the correlation structure in real data, we conducted a simulation based on the colorectal cancer data. Considering in the real world, we usually do not often observe the causal variables directly, but rather the variables that are highly correlated with the causal variables, if any. Here we use a modified version of the simulation strategy as in (Degenhardt et al., 2019; He et al., 2022) to mimic this real-world situation. We first picked six correlated gene expression blocks from the colorectal cancer data, where each block included 6,7,8,7,8 and 9 highly correlated genes (correlation coefficient greater than 0.5) respectively (Supplementary Table S1). For each of the first three blocks, we randomly selected one of the genes as the unobserved causal variables (z_1, z_2 , and z_3) which are in the boldface in Supplementary Table S1 and the rest of the genes in the first three blocks as observed causal variables

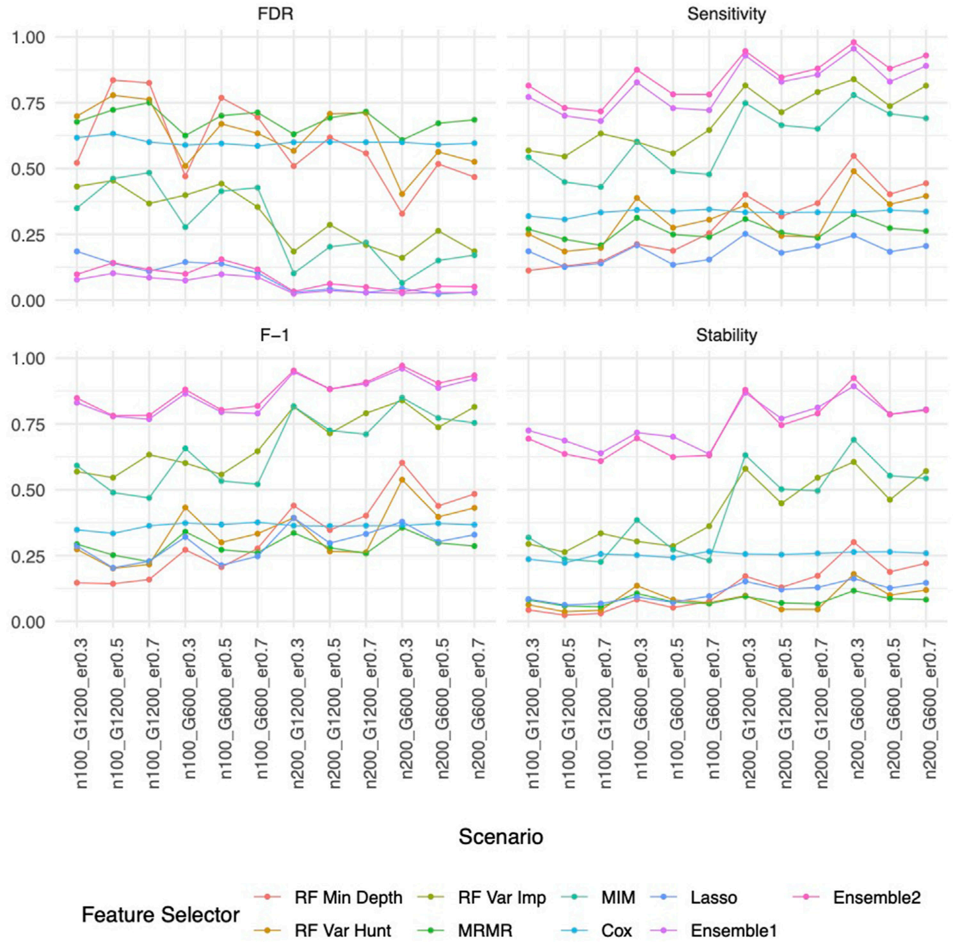


FIGURE 3
Feature selection performance based on Simulation B. In each panel, x-axis stands for different simulation listed in Table 1, y-axis stands for different evaluation metrics including FDR, Sensitivity, F-1 and Stability. For example, n100_G1200_eta0.3 stands for sample size is 100 with 1,200 candidate genes and the event rate is 0.5. Each colored curve stands for different feature selection approaches.

$\left\{v_i^{(j)}, i = 1, 2, 3; j = 1, \ldots, J_i - 1, J_1 = 6, J_2 = 7, J_3 = 8\right\}$, while considering the genes from the last three blocks $\left\{v_i^{(j)}, i = 4, 5, 6; j = 1, \ldots, J_i, J_4 = 7, J_5 = 8, J_6 = 9\right\}$ as observed noncausal variables. For i th block, the variables $\{v_i^{(j)}, J = 1, \ldots, J_i\}$ were generated using multivariate normal distribution with mean zero and the correlation matrix computed based on the real data. Then we generated survival outcomes using the three unobserved causal variables based on a Cox proportional hazards model using the *reda* R package (Fu et al., 2022) (*simEvent* function), with $h_o(t)$ set as 1,

$$log \frac{h(t \mid z)}{h_o(t)} = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3$$

In addition, we generated $G - 42$ independent predictor variables $w_k, k = 1, \ldots, G - 42$, which are uncorrelated with the base variables $\left\{v_i^{(j)}\right\}$, are simulated based on a uniform distribution of (0,1). The input $G = 1000$ features consisted of $\left\{v_i^{(j)},\ i = 1, 2, 3; j = 1, \ldots, J_i - 1, J_1 = 6, J_2 = 7, J_3 = 8\right\}$, $\left\{v_i^{(j)}, i = 4, 5, 6; j = 1, \ldots, J_i, J_4 = 7, J_5 = 8, J_6 = 9\right\}$ and $\{w_k,$ $k = 1, \ldots, G - 42\}$. We generated paired replicates (two $n \times G$ matrixes) with the first used for feature selection evaluation and the prediction models training, and the second used for assessing stability of feature selection and evaluating the prediction performance, and we repeated the processes for 100 times. The details of this real-data-based simulation, including the coefficients, full list of the gene blocks, and names of the unobserved causal genes, are provided in Supplementary Table S1. For ease of presentation, we will refer this real-data-based simulation as Simulation A below.

To further evaluate the performance of the proposed method under more diverse scenarios, we performed additional simulations (referred as Simulation B below). Similar to Simulation A, we first generate unobserved causal variables (z_1, z_2 , and z_3) and then the observed variables, where some are highly corrected with the causal variables (i.e., observed causal variables), and the rest are irrelevant (i.e., noise variables). The survival outcome is also simulated based on a Cox proportional hazards model using the *reda* R package (Fu et al., 2022) (*simEvent* function)

$$log \frac{h(t \mid z)}{h_o(t)} = \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3$$

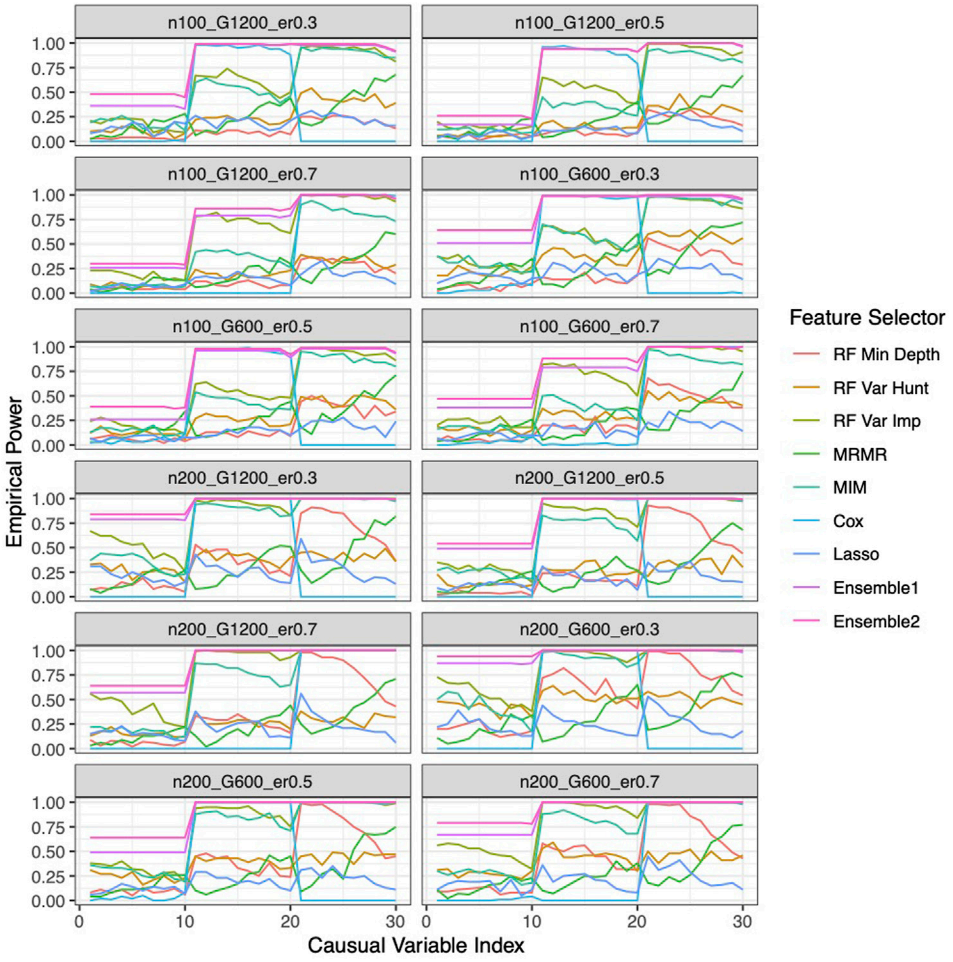


FIGURE 4
Empirical power of the feature selection approaches based on Simulation B. Each panel represents different simulation scenario listed in Table 1. For example, n100_G1200_eta0.3 stands for sample size is 100 with 1,200 candidate genes and the event rate is 0.5. In each panel, x-axis stands for the causal variable index.

where $h_o(t)$ is set as 1. The three base variables (z_1, z_2 , and z_3, the unobserved causal variables) and three additional independent base variables (z_4, z_5 and z_6, the unobserved non-causal variables) are independently sampled from a uniform distribution of (0,1). For each of the base variables z_i, we generate a set of 10 correlated predictor variables $v_i^{(j)}$, denoting the j th variable in group i, for $j = 1, \ldots, 10$ and $i = 1, \ldots, 6$, using the following formula:

$$v_i^{(j)} = z_i + \left(0.01 + \frac{0.5\,(j-1)}{9}\right) \times N\,(0, 0.3\Big),$$

The correlation between the base variable z_i and $v_i^{(j)}$ decreased as j increased. Please note that z_i, $i = 1, \ldots, 6$, are only used to simulate correlated variables $v_i^{(j)}$, and are not included for feature selection and classification. $G - 60$ independent predictor variables $w_k, k = 1, \ldots, G - 60$, which are uncorrelated with the base variables $\left\{v_i^{(j)}\right\}$, are also simulated based on a uniform distribution of (0,1). Here we assume that the base variables are not observed. Hence, the input features consist of 30 observed causal variables $\left\{v_i^{(j)}, i = 1, 2, 3; j = 1, \ldots, 10\right\}$ and 30 correlated, non-causal variables $\left\{v_i^{(j)}, i = 4, 5, 6; j = 1, \ldots, 10\right\}$ and $G - 60$ uncorrelated, non-causal variables $\{w_k, k = 1, \ldots, G - 60\}$, a total of G variables.

We consider twelve different simulation scenarios (Table 2) including 1) different event rates ($\eta = 0.3, 0.5, 0.7$) which are mainly determined by the coefficients in the Cox proportional hazards model; 2) sparsity of causal genes (2.5%, 5%) with a different number of genes (G = 600 and 1,200); and 3) different sample sizes (n = 100 and 200). Similar as in Simulation A, for each of the scenarios, we generated 100 paired replicates, where each pair is consisted of two $n \times G$ matrixes.

Model evaluation

In the real data studies, the causal variables are unknown. Moreover, due to different algorithm, we may have different lists of selected features across all methods. Therefore, to determine the important rank of features, we proposed using a weighted relative frequency (WRF) to measure the relative frequency that a feature is selected across five different folds as in (He et al., 2022). The weight

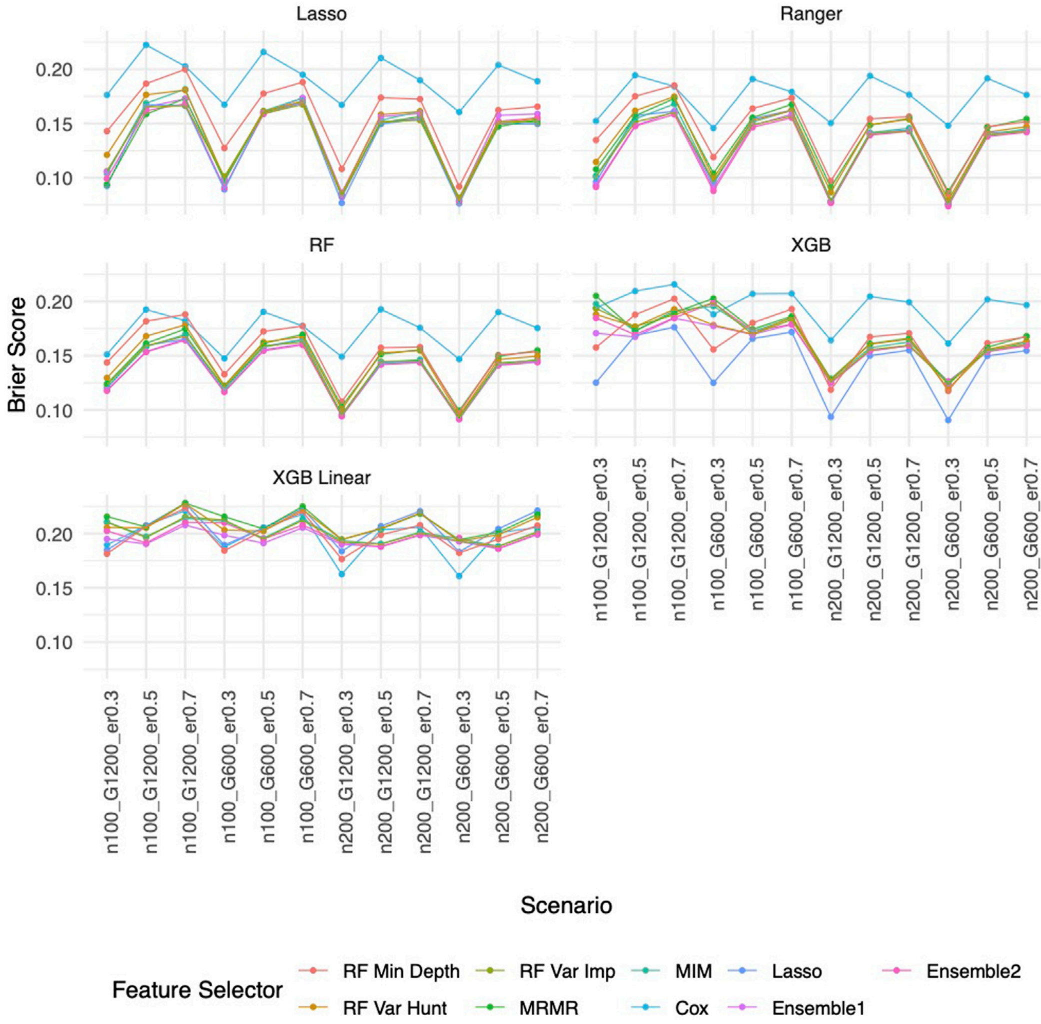


FIGURE 5
Brier score based on Simulation B. Panels present the Brier score for the corresponding prediction approach as indicated. In each panel, x-axis stands for different simulation scenario listed in Table 1. For example, n100_G1200_eta0.3 stands for sample size is 100 with 1,200 candidate genes and the event rate is 0.5. Each colored curve stands for different feature selection approaches.

of each selection is reciprocal to the number of features selected, i.e., larger set of selection adds less weight to each selected feature. A higher WRF indicates this feature is more consistently and sparsely selected across different folds.

Since the causal variables are known in simulation studies, we can evaluate the feature selection performance by comparing the selection to the truth (the known causal variables). Specifically, we used the following four commonly used metrics: false discovery rate (FDR), sensitivity, stability, F-1 score and empirical powers. FDR is the proportion of false-positive features in the selected feature set. Sensitivity is calculated as the proportion of selected causal variables among all the causal variables. Stability is calculated using Jaccard's index: the ratio of the length of the intersection and the length of the union of two sets, where the two sets are the selections from the paired replicates. F-1 score is calculated as $2\frac{\text{precision*sensitivity}}{\text{precision+sensitivity}}$, serving as a balanced metric (harmonic mean) between sensitivity and precision (1-FDR). Empirical power could be calculated for each of the causal variables. It is the ratio that this particular causal variable is selected among the simulation replicates. A power of 1 indicates this casual variable was identified in each replicate, and a power of 0 means it was never selected across replicates. For each feature selection method and each of the twelve scenarios, we reported the average FDR, sensitivity, stability, F-1 score, and empirical powers across the first replicate of each of the 100 simulations.

Furthermore, to check the effectiveness on the predictions of our selected features compared to other well-known models, we used the Integrated Brier score (Ishwaran et al., 2008; Moradian et al., 2017) to assess the accuracy of predicted survival probabilities over a specified time period of events. Lower values of the Integrated Brier Score indicate better predictive accuracy, with 0 being the optimal score (perfect prediction) and 1 representing a model with no predictive ability. Harrell's C-statistic, also known as the concordance index (C-index), was used to evaluate discrimination with a higher value indicating better discrimination, meaning the model is better

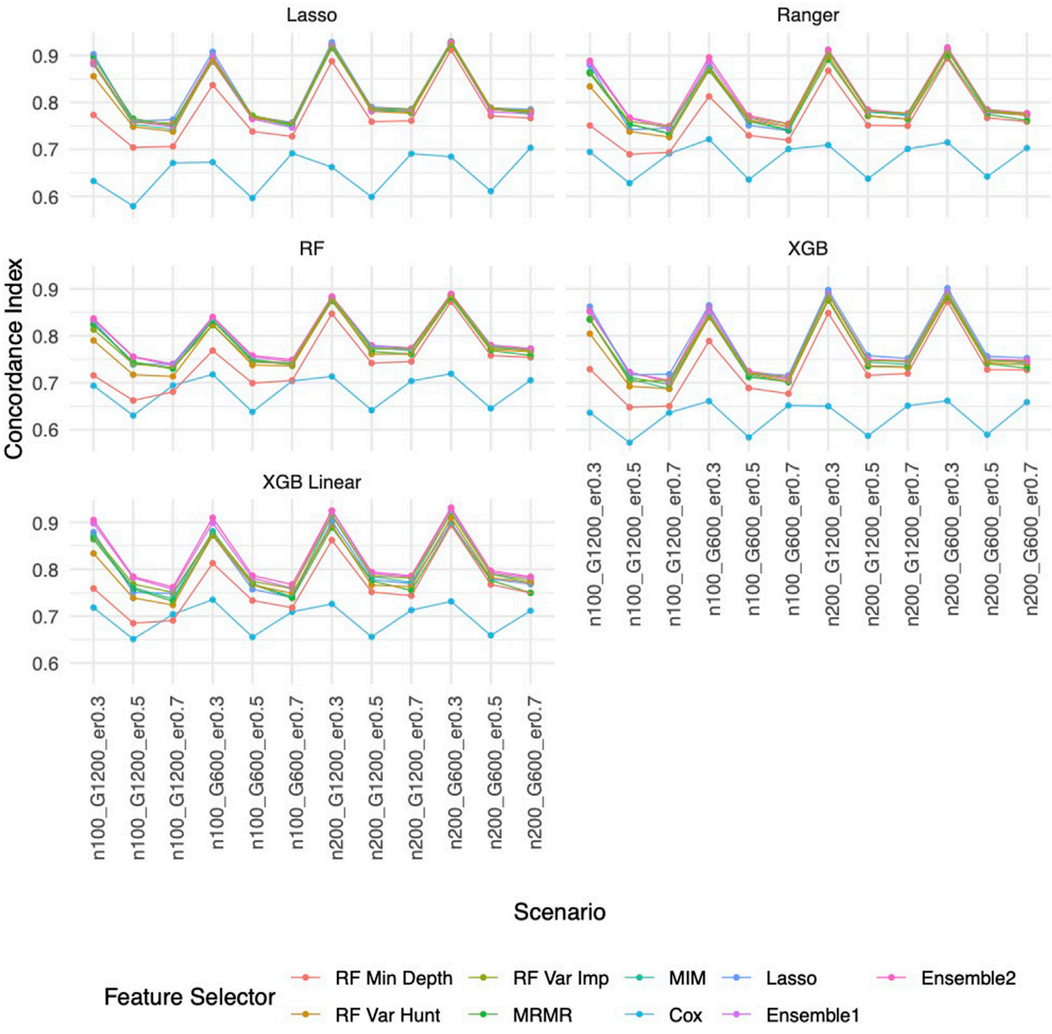


FIGURE 6
C-index based on Simulation B. Panels present the C-index for the corresponding prediction approach as indicated. In each panel, x-axis stands for different simulation scenario listed in Table 1. For example, n100_G1200_eta0.3 stands for sample size is 100 with 1,200 candidate genes and the event rate is 0.5. Each colored curve stands for different feature selection approaches.

at distinguishing between different outcomes. A C-index of 0.5 suggests that the model's predictions are no better than random chance, while a C-index of 1.0 indicates perfect discrimination.

Results

Key features selected by the ensemble feature selection approach on a colorectal cancer (CRC) dataset

In the cohort of $n = 253$ colorectal cancer subjects, encompassing 19,947 genes, the median overall survival (OS) was 83.2 months, with a median follow-up time of 22.5 months. We identified $G = 2{,}303$ genes with p-value less than 0.05 based on the univariable Cox proportional hazard model to further evaluate different feature selection and prediction approaches. We then applied our proposed ensemble approach, where the groups were defined based on the correlation structure of the $G = 2{,}303$ genes, such that within each block, the absolute value of pairwise correlation is all greater than 0.75. The proposed ensemble approaches (Ensemble 1 and Ensemble 2) show the consistency of selected genes and their important rankings compared to all genes, while other methods can only recognize some of them based on WRF (Figure 2A). Notably, the gene *SLC30A3*, although selected by Lasso with the highest WRF, was not identified by other methods. However, it attained the top rank in our proposed ensemble approach, showing the strength of the ensemble approach. Conversely, several genes (*MOS*, *C1ORF61*, and *MBL1P*) that did not rank highest in Lasso achieved top positions in random forest approaches, contributing to higher WRF in the ensemble approaches. Within the top five genes based on WRF (Ensemble 1 and Ensemble 2), *SLC30A3*, *MOS*, *C1ORF61*, and *MBL1P* genes were found to have an association with CRC (Lin et al., 2007; Zheng et al., 2015; Yin et al., 2020; Peng et al., 2021; Cui et al., 2022), gene *PAGE2*, a gene from cancer-germline genes, was found to be upregulated in Caco-2 colorectal cancer cell line (Yilmaz-Ozcan et al., 2014). On the other hand, other methods identified some of the above genes that has connections with colorectal cancers. Using these five genes, we created a composite score by calculating a linear combination of the gene expressions multiplied by their respective coefficients in a multivariable Cox proportional hazards model. Figure 2B presents the Kaplan Meier curves for the subjects with a composite score above and below the median composite score (which is −0.40), with a median OS of 54.6 months and not reachable (log-rank test *p*-value <0.001), respectively. The DCA (Decision Curve Analysis) curves based on 2 years, 3 years and 5 years (Figures 2C–E) consistently show that the net benefit curve outperforms reference lines across various threshold probabilities, indicating clinical utility. As shown in C-index (Figure 2F) and Brier scores (Figure 2G), in general, the prediction approaches have the most impact on the prediction performance rather than the feature selectors. Lasso has a higher C index, and random forest, XGB, and XBG linear yield the lowest Brier scores, while Ranger demonstrates relatively poorer performance.

Improved performance by the ensemble feature selection approach based on simulation studies

Our ensemble approaches consistently demonstrated superior feature selection performance compared to other methods (Supplementary Figure S1A; Figure 3) with Ensemble 1 and Ensemble 2 exhibiting similar performance based on both Simulation A and Simulation B. Although the Lasso method also had low FDRs, it had the lowest sensitivity, reduced F-1 and lower stability. The random forest approaches overall showed poor performance. As expected, in general, a larger sample size (200 vs. 100) resulted in improved performance for all feature selection approaches. However, the impact of the gene sparsity of (2.5% vs. 5%) and event rates (0.3, 0.5, 0.7) on prediction performance was minimal, with slightly better performance observed at lower sparsity. In Supplementary Figure S1B; Figure 4, the empirical power of our ensemble approaches is consistently higher than or at least equivalent to that of other feature selectors across all thirteen scenarios (1 scenario for simulation A, and 12 for simulation B) for all 30 causal variables.

Similar to the real data analysis, the overall impact on prediction performance is predominantly driven by the choice of prediction approaches rather than the feature selectors due to models' bias. This observation is expected, as feature selection does not guarantee an improvement in prediction performance. Nevertheless, feature selection proves valuable by reducing the dimensionality and complexity of predictive models, leading to quicker model training times and improved convergence. Predictably, across all prediction approaches, feature selection based on the univariate Cox proportional hazards model consistently exhibited the least favorable performance, while the various selector approaches appeared quite similar. Notably, a higher event rate corresponded to larger Brier scores (Figure 5) and smaller C-index (Figure 6), indicating poorer prediction performance. A larger sample size contributed to slightly improved prediction performance in terms of Brier score and C-index. Interestingly, gene sparsity did not exert a notable impact on prediction performance. While our feature selection models may not have surpassed others in terms of accuracy measurements, we observed that they provided a stable and consistent accuracy across all measurements (as shown in Figure 5, 6; Supplementary Figures S1C, S1D). This suggests that the features we selected are significant and exhibit less bias, contributing to the reliability of our selected features. We also performed Simulation C with smaller effect sizes (Supplementary Table S2) with the same setting as Simulation B. The results (Supplementary Figures S2–S4) were consistent with all the observations mentioned above.

Conclusion and discussion

This paper proposes a robust ensemble feature selection approach tailored explicitly for survival analysis. The ensemble feature selection approach is built on enhancing the feature selection process by combining different feature selection algorithms, ultimately improving the quality of feature selection and providing stabilized results. This is accomplished through a novel ranking algorithm integrated with a group lasso model, which is particularly advantageous when dealing with feature groups. Therefore, our proposed model is well-suited for applications in genetic data studies, where it is imperative to analyze genes as cohesive groups rather than individual entities. The proposed approach demonstrates a unique ability to select the most compelling features from top-tier models.

The key benefits of ensemble feature selections are 1) Robustness: by aggregating the results from diverse feature selection methods, the final ensemble is less likely to be influenced by the biases or limitations of a single feature selector; 2) Improved Generalization: the ensemble of multiple feature selectors, each built on a different algorithm, can lead to improved generalization and better performance on unseen data; 3) Model Agnosticism: feature selection ensembles are usually model-agnostic, meaning it is not tied to or dependent on a specific machine learning model. Instead, they can be applied across various feature selection models without favoring one over the other, making them widely applicable.

Though we only applied the proposed method to gene expression data, our method can be applied to a wide variety of data having very large number of features in genetics/genomics studies and medical research in general, such as genomic data, transcriptomic data, epigenomic data, proteomic data, clinical and phenotypical data and so on. Besides, the proposed method can smoothly take care of the correlated structure, and even utilize the natural set from certain biological knowledge such as pathway. Moreover, ensemble feature selection can be applied to different response variable, including quantitative, qualitative and time-to-event responses. Although the prediction gain is incremental, the benefits of feature selection are still significant. Firstly, it can enhance the interpretability particularly in the biomedical field and aims the discovery of meaning biological insights. Secondly, it can greatly improve the computational efficiency of downstream analysis, making it more feasible to handle large-scale data sets. Thirdly, it can help filter out irrelevant noise variable, avoid overfitting and enhance the reliability of the analyses.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://portal.gdc.cancer.gov/projects/TCGA-COAD.

Ethics statement

The studies involving humans were approved by the Institutional Review Boards or Ethics Committees of the participating centers. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

PL: Formal Analysis, Investigation, Software, Visualization, Writing–original draft, Writing–review and editing. XG: Formal Analysis, Investigation, Software, Visualization, Writing–review and editing. LU: Formal Analysis, Investigation, Software, Visualization, Writing–review and editing. HY: Data curation, Investigation, Writing–review and editing. BK: Data curation, Investigation, Writing–review and editing. LZ: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Visualization, Writing–original draft, Writing–review and editing. TH: Conceptualization, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing–original draft, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. PL, XG, LU, YH, BK, LZ, and TH are partially supported by National Cancer Institute, National Institutes of Health, through grant number R21CA264381. PL, LU, YH, BK, LZ, and TH are partially supported by the National Library of Medicine, National Institutes of Health, through grant number R01LM013763-01A1. TH and LU are partially supported by National Science Foundation through grant number DMS-2137983. BK is supported by Helen Diller Family Comprehensive Cancer Center Cancer Physician Scholar Program in Clinical Oncology (PSPCO) Career Development Award (1K12CA260225-01).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fsysb.2024.1355595/full#supplementary-material

References

Archer, K. J., and Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. *Comput. Statistics Data Analysis* 52 (4), 2249–2260. doi:10.1016/j.csda.2007.08.015

Binder, H. (2015). R package "CoxBoost". Available from: https://github.com/binderh/CoxBoost.

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* 2 (5), 401–404. doi:10.1158/2159-8290.CD-12-0095

Chen, T., and Guestrin, C. (2016). "XGBoost: a scalable tree boosting system," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016 (San Francisco California USA: ACM), 785–794. doi:10.1145/2939672.2939785

Chen, X., and Ishwaran, H. (2013). Pathway hunting by random survival forests. *Bioinformatics* 29 (1), 99–105. doi:10.1093/bioinformatics/bts643

Cui, J., Guo, F., Yu, Y., Ma, Z., Hong, Y., Su, J., et al. (2022). Development and validation of a prognostic 9-gene signature for colorectal cancer. *Front. Oncol.* 12, 1009698. doi:10.3389/fonc.2022.1009698

Degenhardt, F., Seifert, S., and Szymczak, S. (2019). Evaluation of variable selection methods for random forests and omics data sets. *Briefings Bioinforma.* 20 (2), 492–503. doi:10.1093/bib/bbx124

Deo, S. V., Deo, V., and Sundaram, V. (2021). Survival analysis—part 2: cox proportional hazards model. *Indian J. Thorac. Cardiovasc Surg.* 37 (2), 229–233. doi:10.1007/s12055-020-01108-7

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* 33 (1), 1–22. doi:10.18637/jss.v033.i01

Fu, H., Han, S., and Yan, J. (2022). Reda R package. Available from: https://github.com/wenjie2wang/reda.

Goeman, J. J. (2010). L_1 penalized estimation in the cox proportional hazards model. *Biom. J.* 52 (1), 70–84. doi:10.1002/bimj.200900028

He, T., Baik, J. M., Kato, C., Yang, H., Fan, Z., Cham, J., et al. (2022). Novel ensemble feature selection approach and application in repertoire sequencing data. *Front. Genet.* 13, 821832. 821832. doi:10.3389/fgene.2022.821832

Hu, T., Chitnis, N., Monos, D., and Dinh, A. (2021). Next-generation sequencing technologies: an overview. *Hum. Immunol.* 82 (11), 801–811. doi:10.1016/j.humimm.2021.02.012

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *Ann. Appl. Stat.* 2 (3). doi:10.1214/08-aoas169

Ishwaran, H., Kogalur, U. B., Chen, X., and Minn, A. J. (2011). Random survival forests for high-dimensional data. *Stat. Anal.* 4 (1), 115–132. doi:10.1002/sam.10103

Klein, J. P., and Goel, P. K. (1992). "Survival analysis: state of the art," in *NATO ASI series. Series E, Applied sciences* Editors J. P. Klein and P. K. Goel (Dordrecht ; Boston: Kluwer Academic Publishers), 451.

Lau, J. W., Lehnert, E., Sethi, A., Malhotra, R., Kaushik, G., Onder, Z., et al. (2017). The cancer genomics Cloud: collaborative, reproducible, and democratized—a new paradigm in large-scale computational research. *Cancer Res.* 77 (21), e3–e6. doi:10.1158/0008-5472.CAN-17-0387

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., et al. (2018). Feature selection: a data perspective. *ACM Comput. Surv.* 50 (6), 1–45. doi:10.1145/3136625

Lin, H. M., Chatterjee, A., Lin, Y. H., Anjomshoaa, A., Fukuzawa, R., McCall, J., et al. (2007). Genome wide expression profiling identifies genes associated with colorectal liver metastasis. *Oncol. Rep.* 17, 1541–1549. doi:10.3892/or.17.6.1541

Moradian, H., Larocque, D., and Bellavance, F. (2017). L_1 splitting rules in survival forests. *Lifetime Data Anal.* 23 (4), 671–691. doi:10.1007/s10985-016-9372-1

Munson, M. A., and Caruana, R. (2009). "On feature selection, bias-variance, and bagging," in *Machine learning and knowledge discovery in databases* Editors W. Buntine, M. Grobelnik, D. Mladenić, and J. Shawe-Taylor (Berlin, Heidelberg: Springer Berlin Heidelberg), 144–159. (Lecture Notes in Computer Science; vol. 5782). Available from: http://link.springer.com/10.1007/978-3-642-04174-7_10.

Neums, L., Meier, R., Koestler, D. C., and Thompson, J. A. (2019). "Improving survival prediction using a novel feature selection and feature reduction framework based on the integration of clinical and molecular data," in *Biocomputing 2020* (Kohala Coast, Hawaii, USA: World Scientific), 415–426. doi:10.1142/9789811215636_0037

Peng, J., Peng, J., Wang, R., Liu, C., and Wang, Z. (2021). Expression of MOS gene and its correlations with clinicopathological features and prognosis of patients with colorectal cancer. *Chin. General Pract.* 24 (24), 3077. doi:10.12114/j.issn.1007-9572.2021.00.434

Radovic, M., Ghalwash, M., Filipovic, N., and Obradovic, Z. (2017). Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. *BMC Bioinforma.* 18 (1), 9. doi:10.1186/s12859-016-1423-9

Segal, M. (2004). *Machine learning benchmarks and random forest regression*. UCSF: Center for Bioinformatics and Molecular Biostatistics. Available at: https://escholarship.org/uc/item/35x3v9t4

The Cancer Genome Atlas Network (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 487 (7407), 330–337. doi:10.1038/nature11252

Torkkola, K. (2003). Feature extraction by non-parametric mutual information maximization. *J. Mach. Learn. Res.* 3, 1415–1438. Mar.

Utazirubanda, J. C., León T, M., and Ngom, P. (2021). Variable selection with group LASSO approach: application to Cox regression with frailty model. *Commun. Statistics - Simul. Comput.* 50 (3), 881–901. doi:10.1080/03610918.2019.1571605

Wright, M. N., and ranger, Z. A. (2017). A fast implementation of random forests for high dimensional data in *C++* and *R. J. Stat. Soft* 77 (1). doi:10.18637/jss.v077.i01

Yang, Q., Khoury, M. J., Friedman, J., Little, J., and Flanders, W. D. (2005). How many genes underlie the occurrence of common complex diseases in the population? *Int. J. Epidemiol.* 34 (5), 1129–1137. doi:10.1093/ije/dyi130

Yilmaz-Ozcan, S., Sade, A., Kucukkaraduman, B., Kaygusuz, Y., Senses, K. M., Banerjee, S., et al. (2014). Epigenetic mechanisms underlying the dynamic expression of cancer-testis genes, PAGE2, -2B and SPANX-B, during mesenchymal-to-epithelial transition. *PLoS ONE* 9(9), e107905. doi:10.1371/journal.pone.0107905

Yin, Z., Yan, X., Wang, Q., Deng, Z., Tang, K., Cao, Z., et al. (2020). Detecting prognosis risk biomarkers for colon cancer through multi-omics-based prognostic analysis and target regulation simulation modeling. *Front. Genet.* 11, 524. doi:10.3389/fgene.2020.00524

Zheng, Y., Zhou, J., and Tong, Y. (2015). Gene signatures of drug resistance predict patient survival in colorectal cancer. *Pharmacogenomics J.* 15 (2), 135–143. doi:10.1038/tpj.2014.45

Zhou, Z. H. (2021). *Machine learning*. Singapore: Springer, 458.

frontiers | Frontiers in Systems Biology

TYPE Perspective
PUBLISHED 03 April 2024
DOI 10.3389/fsysb.2024.1385112

OPEN ACCESS

EDITED BY
Eric H. Chang,
Feinstein Institute for Medical Research,
United States

REVIEWED BY
Mohamad Zakaria Koubeissi,
George Washington University, United States

*CORRESPONDENCE
Glenn D. R. Watson,
✉ gdw13@duke.edu

RECEIVED 11 February 2024
ACCEPTED 15 March 2024
PUBLISHED 03 April 2024

CITATION
Watson GDR, Meletti S, Mahavadi AK, Besson P, Bandt SK and Smith JB (2024), Is there room in epilepsy for the claustrum?
Front. Syst. Biol. 4:1385112.
doi: 10.3389/fsysb.2024.1385112

Is there room in epilepsy for the claustrum?

Glenn D. R. Watson[1]*, Stefano Meletti[2,3], Anil K. Mahavadi[4], Pierre Besson[5], S. Kathleen Bandt[6] and Jared B. Smith[7]

[1]Department of Psychology and Neuroscience, Duke University, Durham, NC, United States, [2]Department of Biomedical, Metabolic and Neural Sciences, Center for Neuroscience and Neurotechnology, University of Modena and Reggio Emilia, Modena, Italy, [3]Department of Neurosciences, OCSAE Hospital, Modena, Italy, [4]Department of Neurosurgery, University of Alabama at Birmingham, Birmingham, AL, United States, [5]Department of Radiology, Northwestern University Feinberg School of Medicine, Chicago, IL, United States, [6]Department of Neurologic Surgery, Northwestern University Feinberg School of Medicine, Chicago, IL, United States, [7]Molecular Neurobiology Laboratory, Salk Institute for Biological Studies, La Jolla, CA, United States

The function of the claustrum and its role in neurological disorders remains a subject of interest in the field of neurology. Given the claustrum's susceptibility to seizure-induced damage, there is speculation that it could serve as a node in a dysfunctional epileptic network. This perspective article aims to address a pivotal question: Does the claustrum play a role in epilepsy? Building upon existing literature, we propose the following hypotheses for the involvement of the claustrum in epilepsy: (1) Bilateral T2/FLAIR magnetic resonance imaging (MRI) hyperintensity of the claustrum after status epilepticus represents a radiological phenomenon that signifies inflammation-related epileptogenesis; (2) The ventral claustrum is synonymous with a brain area known as 'area tempestas,' an established epileptogenic center; (3) The ventral subsector of the claustrum facilitates seizure generalization/propagation through its connections with limbic and motor-related brain structures; (4) Disruption of claustrum connections during seizures might contribute to the loss of consciousness observed in impaired awareness seizures; (5) Targeting the claustrum therapeutically could be advantageous in seizures that arise from limbic foci. Together, evidence from both clinical case reports and animal studies identify a significant role for the ventral claustrum in the generation, propagation, and intractable nature of seizures in a subset of epilepsy syndromes.

KEYWORDS

claustrum, claustrum sign, area tempestas, seizure generalization, impaired awareness seizures, epilepsy, status epilepticus

1 Introduction

In epilepsy, a disorder characterized by recurring, spontaneous seizures, the challenge of diverse etiologies complicates both seizure localization and treatment decisions (Watson et al., 2021a). Although significant progress has been made in understanding the neural underpinnings of epilepsy by viewing it as a brain network disorder, the ability to leverage this network for therapeutic purposes remains elusive. To overcome these shortfalls and identify novel targets for epilepsy therapeutics, recent studies have delved into the molecular and epigenetic changes that occur in key brain regions associated with seizure activity, namely, the hippocampus (Conboy et al., 2021; Pires et al., 2021). But another avenue has emerged based on neuroimaging results pointing to an obscure brain region, the claustrum,

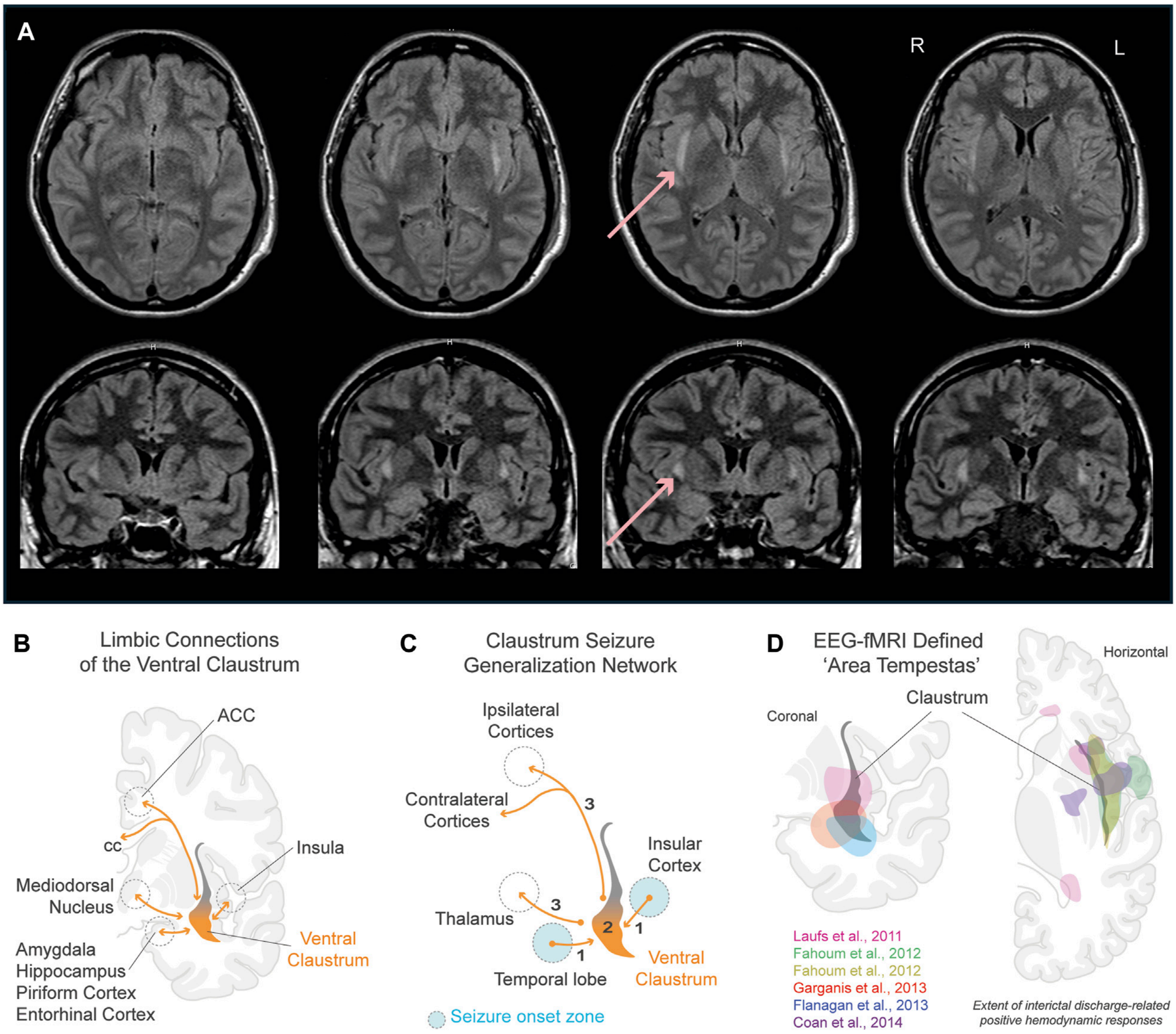


FIGURE 1
Role of the claustrum in seizures. **(A)** Claustrum sign in a 24 year old female 7 days from onset of status epilepticus after febrile illness refractory to antiseizure medications (Meletti et al., 2015). Horizontal (top row) and coronal (bottom row) sections show extent of bilateral claustrum FLAIR MRI hyperintensity. Red arrows point to claustrum hyperintensity in one section. **(B)** Limbic connections of the ventral claustrum. Reciprocal connections (orange lines) are shown with the hippocampal system, basolateral amygdala, piriform cortex, entorhinal cortex, insular cortex, medial PFC, and ACC. **(C)** Claustrum seizure generalization network. Schematic illustrates the claustrum as a secondary node propagating seizures arising from limbic brain structures to ipsi- and contralateral motor-related cortices. **(D)** Physiologically defined 'area tempestas' in epilepsy patients using EEG-fMRI. Overlayed colored areas correspond to approximated areas of interictal discharge-related positive hemodynamic responses in studies color coded within panel. Pink: Laufs et al., 2011, group EEG-fMRI analysis for a mixed cohort of focal epilepsy patients (n = 19). Green and yellow: Fahoum et al., 2012, group EEG-fMRI analysis results of temporal lobe epilepsy group (n = 32) for hemodynamic response functions peaking at 3s (green) and 5s (yellow) after interictal epileptic discharges. Red: Garganis et al., 2013, discharge-correlated BOLD change in a patient experiencing recurrent focal seizures following temporal lobectomy. Blue: Flanagan et al., 2013, group EEG-fMRI random effects analysis for a mixed epilepsy cohort (n = 27). Purple: Coan et al., 2014, group EEG-fMRI T-maps from mesial temporal lobe epilepsy patients with hippocampal sclerosis (n = 13). Refer to studies for original EEG-fMRI overlays t-score values, and *p*-values. Abbreviations: ACC, anterior cingulate cortex; BOLD, blood-oxygen-level-dependent; cc, corpus callosum; EEG, electroencephalogram; fMRI, functional magnetic resonance imaging; FLAIR, fluid-attenuated inversion recovery; PFC, prefrontal cortex.

which appears to have a role in aberrant networks that give rise to the hyperexcitability underlying epilepsy.

Situated beneath the insular cortex, between the external and extreme capsules, the claustrum is a distinctive subcortical structure whose geometry can be described as a thin sheet of glutamatergic projection neurons, extending across the anteroposterior extent of the forebrain. The claustrum's connectivity is extensive, innervating the entire cerebral cortex, including the contralateral cortex, as well as receiving inputs from both hemispheres (for review see Smith et al., 2019a). These connections are topographically organized based on modality across the claustrum's dorsoventral extent, with limbic connections concentrated in its ventral portion (Smith and Alloway, 2014; Watson et al., 2017; Marriott et al., 2021). In fact, renewed interest in the role of the claustrum in epilepsy stems from these limbic connections with brain regions frequently identified as seizure foci, such as the mediodorsal

TABLE 1 Appearance of Claustrum Sign in Cases of Generalized Tonic-Clonic Seizures and Status Epilepticus.

Case Study	Patient	Seizure Type	EEG	MRI Findings	Diagnosis
Ayatollahi et al. (2021) (North America)	18 y/o F	GTCS	Theta/delta slow-wave activity	Unremarkable day 7, bilateral claustrum T2/FLAIR hyperintensity day 21, near-complete resolution one month after	COVID-19 post-infectious encephalitis
Di Dier et al. (2023) (Europe)	39 y/o F	Focal evolving into SE	N/A	Bilateral claustrum T2/FLAIR hyperintensity	FIRES
Guo and Hong (2023) (Asia)	19 y/o F	SE	Bilateral multifocal discharges, left hemisphere predominance	Bilateral claustrum T2/FLAIR hyperintensity day 19	FIRES
Humayun et al. (2023) (Asia)	6 y/o F	GTCS	Moderate amplitude 4 Hz theta, intermixed delta	Bilateral claustrum T2/FLAIR hyperintensity	COVID-19 post-infectious encephalitis
Hwang et al. (2014) (Asia)	28 y/o F	SE	Generalized spike and waves at 1-1.5 Hz	Bilateral claustrum T2/FLAIR hyperintensity day 27†	SE with unknown etiology
Ishii et al. (2011) (Asia)	21 y/o M	GTCS evolving into SE	Slow basic rhythms with epileptic discharges	Unremarkable day 7, bilateral claustrum T2/FLAIR hyperintensity day 13, resolution day 26	Mumps encephalitis
Muccioli et al. (2022) (Europe)	40 y/o F	SE	Bilateral asymmetric lateralized periodic discharges, predominance in right fronto-temporal region	Bilateral claustrum T2/FLAIR hyperintensity	FIRES
Nixon et al. (2001)* (Europe)	35 y/o M	GTCS evolving into SE	Generalized slow wave	Unremarkable day 7, bilateral claustrum T2/FLAIR hyperintensity day 13 (4 days after SE onset)	SE with unknown etiology
Safan et al. (2023) (Asia)	30 y/o M	GTCS evolving into SE	Continuous left-sided epileptiform discharges, left middle temporal predominance	Bilateral external/extreme capsule hyperintensities with bilateral claustrum sparing day 9, resolution day 37	Seronegative limbic encephalitis
Silva et al. (2018) (Europe)	6 y/o F	SE	Occipital intermittent rhythmic delta activity	Bilateral external/extreme capsule hyperintensities day 22, reduction at month 3	NORSE
Silva and Sousa (2019) (South America)	16 y/o F	SE	N/A	Bilateral claustrum T2/FLAIR hyperintensity day 21†, resolution 4 months later	N/A
Sperner et al. (1996) (Europe)	12 y/o F	SE followed by focal impaired awareness	Severe generalized slowing, right-sided sharp slow waves	Bilateral claustrum T2/FLAIR hyperintensity and T1 hypointensity day 21, resolution day 25, normal MRI at week 7	SE with unknown etiology

Dates of MRI findings extrapolated from first instance of symptoms reported within case studies. See articles for list of negative laboratory findings. Diagnoses are listed as presented in case studies. Note that all cases show T2/FLAIR hyperintensity restricted to the claustrum with diffusion into the external and extreme capsules without involvement of other brain regions. See Meletti et al., 2015; Meletti et al., 2017 for cohort population reports in patients with FIRES and NORSE. See Atilgan et al., 2022 for additional case studies involving hyperintensities in other brain regions appearing with claustrum sign. Abbreviations: EEG, electroencephalogram; F, female; FIRES, febrile infection-related epilepsy syndrome; FLAIR, fluid-attenuated inversion recover; GTCS, generalized tonic-clonic seizures; M, male; MRI, magnetic resonance imaging; NORSE, new-onset refractory status epilepticus; SE, status epilepticus. *Case resulted in death. † Timeframe approximated based on article text.

thalamus, hippocampus, and amygdala (Jackson et al., 2020; Smith et al., 2020; Benarroch, 2021).

Clinically, claustrum lesions that cause seizures often encompass surrounding structures, complicating the identification of the claustrum's specific role. Recent structural magnetic resonance imaging (MRI) studies have identified a distinctive signature that prominently features the claustrum in patients with intractable seizures, providing compelling evidence of its involvement in epilepsy. To support this perspective, we begin by reviewing case study evidence of radiological signals in the claustrum of new-onset seizure patients with acute encephalopathies. Subsequently, we turn to rodent models of epilepsy to precisely define the claustrum's involvement, allowing us to distinguish subregions implicated in seizure generation and propagation, particularly its ventral portion known to have significant connections with the limbic system (Watson et al., 2017). The culmination of these separate approaches has led us to the novel hypothesis that the ventral claustrum is in fact synonymous with the so called 'area tempestas,' a non-circumscribed brain area traditionally implicated in seizure propagation and epileptogenesis. We also speculate that disruption of claustrum connections with the thalamus and cortex may impair consciousness during certain seizure subtypes. With this evidence, we conclude by exploring therapeutic development centered on the claustrum and identifying the indications most likely to benefit from claustrum remediation.

2 Misleading signs? The claustrum's role in *de novo* status epilepticus

Case reports of patients with status epilepticus (SE) have provided valuable insights into the involvement of the claustrum in seizures (Meletti et al., 2015; Meletti et al., 2017; Atilgan et al., 2022). In the acute phase of SE, a distinct hyperintensity localized to the claustrum emerges in T2-weighted-fluid-attenuated inversion recovery (T2/FLAIR) MRI images latent from seizure onset, as illustrated in Figure 1A. This radiological occurrence, termed 'claustrum sign,' is notable for its association with generalized tonic-clonic seizures and its reversibility following SE resolution (Table 1). Intriguingly, claustrum-related imaging abnormalities are rare in patients with SE but are strongly linked to a *de novo* SE that typically develops in young, healthy patients that are refractory to antiseizure medications. In some cases, autoimmune antibody positive encephalitic syndromes have been reported, including some cases of COVID-19 post-infection encephalitis (Ayatollahi et al., 2021; Humayun et al., 2023). However, the etiology in most cases remains undetermined with patients being described in the context of febrile infection-related epilepsy syndrome (FIRES) and new-onset refractory status epilepticus (NORSE) (see updated terminology in Hirsch et al., 2018). Thus, the claustrum sign serves as a distinctive radiological biomarker, suggesting a potential link to inflammation-related epileptogenesis and cytokine-mediated neuroinflammation. But does this hyperintensity indicate a *causative* role for the claustrum in inflammation-related SE, or does the claustrum sign simply signify inflammation?

Unlike claustrum damage resulting from a hemorrhagic stroke or penetrating head injury, which infrequently leads to seizures, viral and autoimmune encephalitic etiologies can manifest claustrum sign (Table 1; see Atilgan et al., 2022, for additional case summaries). Two hypotheses may explain the appearance of this radiological phenomenon. The first involves postinfection neuronal loss, encompassing gliosis, spongiform degeneration, and demyelination during the recovery phase (Kimura et al., 1994; Sperner et al., 1996; Nomoto et al., 2007; Ishii et al., 2011). Lending credence to this hypothesis, the presence of ischemic cell changes and acute astrocytic reaction (astrogliosis) were observed in the claustra during histopathological analysis of a patient's brain after a fatal SE case (refer to Table 1; Nixon et al., 2001). Conversely, a comprehensive neuropathological study found no abnormalities in the claustra of patients with chronic epilepsy and SE (Margerison and Corsellis, 1966). Another hypothesized mechanism is focal edema, gaining support from recent case studies of claustral edema in the context of refractory SE following consumption of Sugihiratkae mushrooms (Kuwabara et al., 2005; Nishizawa, 2005; Nomoto et al., 2007). Edema localized to the claustrum may therefore contribute to an aspect of the refractory nature of SE.

The claustrum sign may not solely be a structural abnormality but could instead signify network dysfunction, although it is seldom observed outside of encephalopathies (Steriade et al., 2017; Silva et al., 2018; Altigan et al., 2022). This prompts an exploration into whether viral-induced connectional changes are causative factors behind the appearance of this hyperintensity. A clue may reside in the claustrum's strikingly high density of inhibitory kappa-opioid receptors (KORs) compared to other subcortical brain regions (Peckys and Landwehrmeyer, 1999; Stiefel et al., 2014; Cahill et al., 2022). The potential link between viral-induced KOR dysfunction and the claustrum sign, potentially driven by runaway excitation due to reduced dynorphin expression, necessitates careful consideration (Solbrig and Koob, 2004; Solbrig et al., 2006; Silva et al., 2018). More intriguingly, most case studies reporting claustrum sign in the literature originate from Asia and Europe, raising questions about the veracity of this signal's physiological significance, or whether it represents an underreported, time-dependent radiological phenomenon appearing around one to three weeks from symptom onset (Table 1).

3 Gene expression changes in claustrum during seizures

The hypothesis that MRI hyperintensities may indicate aberrant hyperactivity within nodes of an epileptic network has been proposed (Silva et al., 2018; Ayatollahi et al., 2021). Examining the claustrum's role as a node within a limbic epileptic network is a potential avenue to clarify its relationship with seizure activity. As mentioned previously, the ventral most region of the claustrum connects to various limbic brain structures that are implicated in seizure generation and epileptic pathology (Smith et al., 2020). These regions include the piriform, medial prefrontal, orbitofrontal, and entorhinal cortices, the amygdala (basolateral, central, and medial nuclei), and the anterior and mediodorsal nuclei of the thalamus (Fernandez-Miranda et al., 2008; Watson et al., 2017; Smith et al., 2019b) (Figure 1B). Despite its anatomical significance, the involvement of this limbic subsector of the claustrum has largely been overlooked in seizure research, potentially due to its obscurity and the ability to selectively modulate it without affecting neighboring white matter (Watson and Kopell, 2022).

An alternative method to delve into the claustrum's potential involvement in seizures involves measuring its neuronal activity in validated animal models of epileptogenesis. Numerous studies investigating c-fos expression in temporal proximity to SE induced by various methods consistently show increased expression in limbic regions connected to the claustrum such as the hippocampus, piriform cortex, medial prefrontal cortex, entorhinal cortex, amygdala, and anterior nucleus of the thalamus (Morgan et al., 1987; Sitcoske O'Shea et al., 2000; Szyndler et al., 2009; Barros et al., 2015; Siow et al., 2020) (Figure 1B). Studies utilizing kainic acid, pentylenetetrazol, lithium-pilocarpine, and kindling exhibit increased c-fos expression and evidence of neuronal cell death in the claustrum itself (Willoughby et al., 1997; Zhang et al., 1997; Covolan and Mello, 2000; Sitcoske O'Shea et al., 2000; Zhang et al., 2001; Siow et al., 2020; Druga et al., 2024). In fact, a region we recently delineated as being a part of the ventral claustrum in rodents, the dorsal endopiriform nucleus, shows c-fos expression during SE only after the first convulsive seizure, corresponding to the appearance of claustrum sign after the onset of SE in humans (Table 1; Smith et al., 2020; see Majak and Moryś, 2007 for review).

In light of the emerging evidence supporting the ventral claustrum's role in epilepsy through gene expression studies, a compelling avenue of exploration lies in understanding its potential influence on specific aspects of seizures. Building upon these insights, we next delve into a distinct aspect of claustrum involvement - its potential role in impaired awareness seizures.

4 A case for involvement of the claustrum in impaired awareness seizures

Brain regions with changes in c-fos expression during seizures provide a biological anchor by which to interpret resting-state functional MRI (rs-fMRI) data. In one of our recent rs-fMRI studies, we observed functional connections between the claustrum and the thalamus, amygdala, and prefrontal cortex that are weakened under isoflurane anesthesia (Smith et al., 2017). Building on these results and findings in human rs-fMRI studies, we implicated the ventral claustrum as a critical node within both the salience and default-mode intrinsic connectivity networks (ICNs): interconnected brain regions that are functionally co-activated or co-deactivated during specific cognitive activities that are found to be impaired throughout epileptogenesis (Luo et al., 2011a; b; Smith et al., 2017; Smith et al., 2019b). Interestingly, a decrease in default-mode ICN activity is shown during generalized tonic-clonic seizures, and selective impairment to this ICN during seizures is associated with loss of consciousness (Danielson et al., 2011; Crone et al., 2015). This presents the possibility that ventral claustrum output could be impaired during seizures, and in turn alter ICN-mediated consciousness.

The claustrum's role in consciousness, speculated by Crick and Koch (2005) has sparked renewed interest in research on the subject. A study on a refractory epilepsy patient undergoing stimulation mapping demonstrated that electrical stimulation near the claustrum could reversibly disrupt consciousness (Koubeissi et al., 2014). However, a later study involving several epilepsy neurosurgical patients contradicted this finding, as electrical stimulation of the claustrum did not lead to a loss of consciousness (Bickel and Parvizi, 2019). Nevertheless, we do not entirely rule out the possibility of the claustrum's involvement in seizures that impair awareness via ICN alterations. Cases with claustrum sign often report impairment in consciousness (see Atilgan et al., 2022 for review). Furthermore, most focal impaired awareness seizures arise from the temporal lobe, where many of its structures directly project to the ventral claustrum, with more than half evolving into focal to bilateral generalized seizures (Kumar and Sharma, 2023).

An ongoing clinical trial (NCT04897776, 2024) stimulating the intralaminar thalamus to restore arousal in temporal lobe epilepsy patients with impaired conscious awareness may offer mechanistic insight. As shown in Figure 1C, seizures emanating from limbic brain structures could impair interactions between the claustrum, cortex, and thalamic nuclei. A robust and common target of both the intralaminar thalamus and the claustrum is the anterior cingulate cortex: where seizures often lead to impaired consciousness and motor manifestations, often involving the temporal lobe (Alkawadri et al., 2016; Benarroch, 2021). We previously hypothesized that the connectivity between the claustrum and cingulate cortex plays a major role in salience and default-mode ICNs (Smith et al., 2019a; Kou et al., 2023). Building upon this, we further hypothesize that disruption to this critical network connection may impair awareness during seizures originating from temporal lobe structures through its interaction with motor-related cortical areas (cingulate cortex) and the thalamus. We therefore support the viewpoint that while the claustrum can influence the consciousness "master switch" of a brainstem and diencephalic origin, it is not the master switch itself (Blumenfeld, 2014; Gummadavelli et al., 2015).

5 The ventral claustrum is synonymous with 'area tempestas': a brain region imlicated in seizure generation and propagation

It is plausible that the claustrum is a node by which seizures can generalize or propagate from limbic-connected structures to cortical regions considering the functional connectivity data discussed above. Supporting this possibility, from animal data, amygdaloid kindling studies reveal that claustrum lesions destabilize or entirely block seizure generalization (Wada and Kudo, 1997; Wada and Tsuchimochi, 1997; Mohapel et al., 2000). Interestingly, a non-circumscribed anatomical region termed 'area tempestas' traditionally described within the deep piriform cortex (primary olfactory) demonstrates strikingly similar kindling results (Löscher et al., 1995). Upon further research, the dorsal endopiriform (rodents) and pre-endopiriform (human) nuclei (i.e., ventral claustrum) correspond to this physiologically defined area (see Majak and Moryś, 2007; Vaughan and Jackson, 2014 for review).

Insight into the exact functional relationship amongst the ventral claustrum and limbic brain structures during seizures can be further gleaned from a formative electroencephalogram (EEG)-fMRI study involving focal epilepsy patients (Laufs et al., 2011). Regardless of the localization of interictal and ictal activity, the study identified a common, tightly localized brain region attributed to be the "human equivalent of area tempestas," exhibiting increased hemodynamic responses in relation to interictal epileptiform discharges. Based on the reported Talairach coordinates, we previously hypothesized that this area corresponds to the ventral claustrum (Meletti et al., 2015). To further support our hypothesis, we show EEG-fMRI results from this and subsequent studies that attribute interictal discharge-related hemodynamic responses to 'area tempestas' that correspond to the location of the ventral claustrum (Figure 1D).

Interestingly, Laufs et al. also found reduced benzodiazepine-GABA$_A$ receptor binding complex expression in 'area tempestas,' as measured by flumazenil positron emission tomography, in patients experiencing more frequent seizures. The claustrum notably harbors a significant population of GABAergic interneurons, which are influenced by anesthetic agents that interact with benzodiazepine GABA$_A$ receptor binding complexes. Consequently, reductions observed in the expression of these complexes may in fact occur within the claustrum. This reduction may also be linked to the decreased effectiveness of benzodiazepines observed in cases of refractory SE (Singh et al., 2014; Borroto-Escuela and Fuxe, 2020; Kim et al., 2020; Luo et al., 2023).

In line with these results, in Figure 1C we illustrate a putative seizure generalization network from limbic-associated brain structures to motor-related cortical areas via the ventral claustrum. We hypothesize that seizures arising from temporal lobe structures can generalize broadly across ipsi- and contralateral neocortices (e.g., generalized tonic-clonic seizures) through ventral claustrum projections. Considering this subcortical generalization network, we further speculate that anterior temporal lobectomy and other temporal lobe resection techniques utilized in epilepsy could, in some cases, resect the ventral portion of the claustrum or transect limbic fibers to this subregion (Feindel et al., 2009; Borger et al., 2021; Dalio et al., 2022). The incidence, benefits, and/or altered outcomes of these surgical possibilities are unknown. However, a recent case study provides strong evidence that not fully resecting 'area tempestas' may cause seizure recurrence following temporal lobectomy (Garganis et al., 2013; Figure 1D).

6 Is the claustrum a suitable therapeutic target in epilepsy?

Our perspective on available evidence implicates the ventral claustrum as a key node within a dysfunctional epileptic network. To review, four key pieces of evidence from clinical and animal studies point to the claustrum as a useful target for therapeutic development in epilepsy: (1) Neuroimaging data that show increased activation in a region that stereotactically corresponds to the ventral claustrum; (2) Histopathological data in both animals and humans that reveal neuronal cell death and cellular alterations in the ventral claustrum after uncontrolled seizures; (3) Electrical kindling data revealing that the ventral claustrum has a low threshold and high susceptibility to seizure induction, and lesions to this region can profoundly mitigate or block seizure generalization; (4) Seizure-induced disruptions to claustro-cortico-thalamic interactions that constitute brain wide ICNs could impair consciousness during certain seizure subtypes. Considering this evidence, we conclude that the ventral claustrum represents a viable target in correcting a dysfunctional epileptic network. Below we review several therapeutic possibilities.

Targeting endogenous opioids has recently gained attention as a promising therapy to treat temporal lobe epilepsy (Zangrandi and Schwarzer, 2022; Lankhuijzen and Ridler, 2024). As described earlier, the claustrum has a high density of KORs, presenting a unique opportunity to target this brain region with KOR agonists to reduce neuronal excitability, especially during SE (Kumar et al., 2023). Therefore, the use of KOR agonists as anticonvulsants, specifically for refractory SE, should be explored further. More work is also needed to explore how the use of benzodiazepine and non-benzodiazepine GABA$_A$ modulators can be used to selectively target claustrum interneurons during seizures.

We previously discussed data hinting at the possibility that resecting the ventral claustrum could, theoretically, provide benefit in patients with generalized seizures that arise from temporal lobe structures (Feindel et al., 2009; Borger et al., 2021; Dalio et al., 2022). However, this viewpoint is highly speculative and requires formal investigation to support or refute. Magnetic resonance-guided focused ultrasound to selectively ablate the ventral claustrum may provide a starting point to test this hypothesis (Ranjan et al., 2019). Long-standing neuromodulation techniques can also be used to target the claustrum, especially with closed-loop, state-dependent stimulation (Wong et al., 2021; Watson and Kopell, 2022). It is conceivable that claustrum electrical stimulation may help correct an aberrant epilepsy network or prevent seizure generalization, but confounding variables such as the alteration of consciousness seen in the N-of-1 study discussed (Koubeissi et al., 2014), and the possibility of off-target white matter stimulation effects may make this modality less favorable (Kurada et al., 2019). Owing to the claustrum's unique anatomy, more advanced cell- and pathway-specific neuromodulation techniques to effectively target this structure are warranted (Watson et al., 2021b).

An emerging tool that could selectively target the claustrum to treat epilepsy is gene therapy (Shaimardanova et al., 2022; Boileau et al., 2023; Miyakawa et al., 2023). Diffuse or more targeted use of viral promoters (e.g., adeno-associated viruses, AAVs) are being used to restrict vector expression to select populations of neuronal subtypes. Gene therapy would address the issue of non-specific neuromodulation and the systemic targeting of many antiseizure medications. For example, selectively attenuating glutamatergic projection neurons in the ventral claustrum through gene therapy may prevent seizure generalization or impaired awareness as previously discussed. Even developmental and epileptic encephalopathies such as Dravet syndrome may benefit from selective targeting of Nav1.1 parvalbumin neurons in the claustrum (Vormstein-Schneider et al., 2020; Niibori et al., 2023).

As we contemplate the therapeutic potential of targeting the claustrum, the prospect of correcting a dysfunctional epileptic network becomes both promising and challenging. This perspective article opens new avenues for understanding the intricate interplay between the claustrum and limbic brain structures, providing a foundation for future research and potential breakthroughs in epilepsy therapeutics.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

GW: Conceptualization, Visualization, Writing–original draft, Writing–review and editing. SM: Conceptualization, Visualization, Writing–original draft, Writing–review and editing. AM: Conceptualization, Writing–review and editing. PB: Conceptualization, Writing–review and editing. SB: Conceptualization, Writing–review and editing. JS: Conceptualization, Visualization, Writing–original draft, Writing–review and editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

GW was an employee of SK Life Science, Inc. when drafting this article. JS was an employee of REGENXBIO when drafting this article.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Alkawadri, R., So, N. K., Van Ness, P. C., and Alexopoulos, A. V. (2016). Cingulate epilepsy: report of 3 electroclinical subtypes with surgical outcomes. *JAMA Neurol.* 70 (8), 995–1002. doi:10.1001/jamaneurol.2013.2940

Atilgan, H., Doody, M., Oliver, D. K., McGrath, T. M., Shelton, A. M., Echeverria-Altuna, I., et al. (2022). Human lesions and animal studies link the claustrum to perception, salience, sleep and pain. *Brain* 145, 1610–1623. doi:10.1093/brain/awac114

Ayatollahi, P., Tarazi, A., and Wennberg, R. (2021). Possible autoimmune encephalitis with claustrum sign in case of acute SARS-CoV-2 infection. *Can. J. Neurol. Sci.* 48 (3), 430–432. doi:10.1017/cjn.2020.209

Barros, V. N., Mundim, M., Galindo, L. T., Bittencourt, S., Porcionatto, M., and Mello, L. E. (2015). The pattern of c-fos expression and its refractory period in the brain of rats and monkeys. *Front. Cell Neurosci.* 12, 72. doi:10.3389/fncel.2015.00072

Benarroch, E. E. (2021). What is the role of the claustrum in cortical function and neurologic disease?. *Neur.* 96 (3), 110–113. doi:10.1212/WNL.0000000000011280

Bickel, S., and Parvizi, J. (2019). Electrical stimulation of the human claustrum. *Epilepsy Behav.* 97, 296–303. doi:10.1016/j.yebeh.2019.03.051

Blumenfeld, H. (2014). A master switch for consciousness? *Epilepsy Behav.* 37, 234–235. doi:10.1016/j.yebeh.2014.07.008

Boileau, C., Deforges, S., Peret, A., Scavarda, D., Bartolomei, F., Giles, A., et al. (2023). GluK2 is a target for gene therapy in drug-resistant temporal lobe epilepsy. *Ann. Neurol.* 94 (4), 745–761. doi:10.1002/ana.26723

Borger, V., Schneider, M., Taube, J., Potthoff, A., Keil, V. C., Hamed, M., et al. (2021). Resection of piriform cortex predicts seizure freedom in temporal lobe epilepsy. *Ann. Clin. Transl. Neurol.* 8 (1), 177–189. doi:10.1002/acn3.51263

Borroto-Escuela, D. O., and Fuxe, J. (2020). On the G protein-coupled receptor neuromodulation of the claustrum. *Neurochem. Res.* 45 (1), 5–15. doi:10.1007/s11064-019-02822-4

Cahill, C., Tejeda, H. A., Spetea, M., Chen, C., and Liu-Chen, L. (2022). Fundamentals of the dynorphins/kappa opioid receptor system: from distribution to signaling and function. *Handb. Exp. Pharmacol.* 271, 3–21. doi:10.1007/164_2021_433

Coan, A. C., Campos, B. M., Beltramini, G. C., Yasuda, C., Covolan, R. J. M., and Cendes, F. (2014). Distinct functional and structural MRI abnormalities in mesial temporal lobe epilepsy with and without hippocampal sclerosis. *Epilepsia* 55 (8), 1187–1196. doi:10.1111/epi.12670

Conboy, K., Henshall, D. C., and Brennan, G. P. (2021). Epigenetic principles underlying epileptogenesis and epilepsy syndromes. *Neurobiol. Dis.* 148, 105179. doi:10.1016/j.nbd.2020.105179

Covolan, L., and Mello, L. E. (2000). Temporal profile of neuronal injury following pilocarpine or kainic acid-induced status epilepticus. *Epilepsy Res.* 39 (2), 133–152. doi:10.1016/s0920-1211(99)00119-9

Crick, F. C., and Koch, C. (2005). *Philos. Trans. R. Soc. Lond B Biol. Sci.* 360 (1458), 1271–1279. doi:10.1098/rstb.2005.1661

Crone, J. S., Schurz, M., Holler, Y., Bergmann, J., Monti, M., Schmid, E., et al. (2015). Impaired consciousness is linked to changes in effective connectivity of the posterior cingulate cortex within the default mode network. *Neuroimage* 110, 101–109. doi:10.1016/j.neuroimage.2015.01.037

Dalio, M. T. R. P., Velasco, T. R., Feitosa, I. D. F., Assirati Junior, J. A., Carlotti Junior, C. G., Leite, J. P., et al. (2022). Long-term outcome of temporal lobe epilepsy surgery in 621 patients with hippocampal sclerosis: clinical and surgical prognostic factors. *Font. Neurol.* 13, 833293. doi:10.3389/fneur.2022.833293

Danielson, N. B., Guo, J. N., and Blumenfeld, H. (2011). The default mode network and altered consciousness in epilepsy. *Behav. Neurol.* 24 (1), 55–65. doi:10.3233/BEN-2011-0310

Di Dier, K., Dekesel, L., and Dekeyzer, S. (2023). The claustrum sign in febrile infection-related epilepsy syndrome (FIRES). *J. Belg Soc. Radiol.* 107 (1), 45. doi:10.5334/jbsr.3142

Druga, R., Mares, P., Salaj, M., and Kubova, H. (2024). Degenerative changes in the claustrum and endopiriform nucleus after early-life status epilepticus in rats. *Int. J. Mol. Sci.* 25, 1296. doi:10.3390/ijms25021296

Fahoum, F., Lopes, R., Pittau, F., Dubeau, F., and Gotman, J. (2012). Widespread epileptic networks in focal epilepsies: EEG-fMRI study. *Epilepsia* 53 (9), 1618–1627. doi:10.1111/j.1528-1167.2012.03533.x

Feindel, W., Leblanc, R., and Almeida, A. N. (2009). Epilepsy surgery: historical highlights 1909-2009. *Epilepsia* 50 (Suppl. 3), 131–151. doi:10.1111/j.1528-1167.2009.02043.x

Fernandez-Miranda, J. C., Rhoton Jr, A. L., Kakizawa, Y., Choi, C., and Alvarez-Linera, J. (2008). The claustrum and its projection system in the human brain: a microsurgical and tractographic anatomical study. *J. Neurosurg.* 108 (4), 764–774. doi:10.3171/JNS/2008/108/4/0764

Flanagan, D., Badawy, R. A. B., and Jackson, G. D. (2014). EEG-fMRI in focal epilepsy: local activation and regional networks. *Clin. Neurophysiol.* 125, 21–31. doi:10.1016/j.clinph.2013.06.182

Garganis, K., Kokkinos, V., and Zountsas, B. (2013). EEG-fMRI findings in late seizure recurrence following temporal lobectomy: a possible contribution of area tempestas. *Epilepsy Behav. Case Rep.* 12 (1), 157–160. doi:10.1016/j.ebcr.2013.09.001

Gummadavelli, A., Kundishora, A. J., Willie, J. T., Andrews, J. P., Gerrard, J. L., Spencer, D. D., et al. (2015). Neurostimulation to improve level of consciousness in patients with epilepsy. *Neurosurg. Focus* 38 (6), E10. doi:10.3171/2015.3.FOCUS1535

Guo, K., and Hong, Z. (2023). Claustrum sign in febrile infection-related epilepsy syndrome (FIRES). *Neurol. Sci.* 44 (9), 3357–3359. doi:10.1007/s10072-023-06887-6

Hirsch, L. J., Gaspard, N., van Baalen, A., Nabbout, R., Demeret, S., Loddenkemper, T., et al. (2018). Proposed consensus definitions for new-onset refractory status epilepticus (NORSE), febrile infection-related epilepsy syndrome (FIRES), and related conditions. *Epilepsia* 59 (4), 739–744. doi:10.1111/epi.14016

Humayun, M. B., Khalid, S., Khalid, H., Zahoor, W., and Malik, W. T. (2023). Post-COVID-19 encephalitis with claustrum sign responsive to immunomodulation. *Cureus* 15 (2), e35363. doi:10.7759/cureus.35363

Hwang, K. J., Park, K., Yoon, S. S., and Ahn, T. (2014). Unusual lesion in the bilateral external capsule following status epilepticus: a case report. *J. Epilepsy Res.* 4 (2), 88–90. doi:10.14581/jer.14019

Ishii, K., Tsuji, H., and Tamaoka, A. (2011). Mumps virus encephalitis with symmetric claustrum lesions. *Am. J. Neuroradiol.* 32 (7), E139. doi:10.3174/ajnr.A2603

Jackson, J., Smith, J. B., and Lee, A. K. (2020). The anatomy and physiology of claustrum-cortex interactions. *Annu. Rev. Neurosci.* 43, 231–247. doi:10.1146/annurev-neuro-092519-101637

Kim, J. J., Gharpure, A., Teng, J., Zhuang, Y., Howard, R. J., Zhu, S., et al. (2020). Shared structural mechanisms of general anaesthetics and benzodiazepines. *Nature* 585 (7824), 303–308. doi:10.1038/s41586-020-2654-5

Kimura, S., Nezu, A., Osaka, H., and Saito, K. (1994). Symmetrical external capsule lesions in a patient with herpes simplex encephalitis. *Neuropediatrics* 25 (3), 162–164. doi:10.1055/s-2008-1073016

Kou, Z., Chen, C., Abdurahman, M., Weng, X., Hu, C., and Geng, H. (2023). The claustrum controls motor activity through anterior cingulate cortex input and local circuit synchronization in a preparatory manner. *Neurosci. Bull.* 39 (10), 1591–1594. doi:10.1007/s12264-023-01079-w

Koubeissi, M. Z., Bartolomei, F., Beltagy, A., and Picard, F. (2014). Electrical stimulation of a small brain area reversibly disrupts consciousness. *Epilepsy Behav.* 37, 32–35. doi:10.1016/j.yebeh.2014.05.027

Kumar, A., and Sharma, S. (2023). *Focal impaired awareness seizures. StatPearls* Treasure Island (FL): StatPearls Publishing. 2024 Jan 31. PMID: 30085572.

Kumar, H., Katyal, J., and Gupta, Y. K. (2023). Effect of U50488, a selective kappa opioid receptor agonist and levetiracetam against lithium-pilocarpine-induced status epilepticus, spontaneous convulsive seizures and related cognitive impairment. *Neurosci. Lett.* 815, 137477. doi:10.1016/j.neulet.2023.137477

Kurada, L., Bayat, A., Joshi, S., and Koubeissi, M. Z. (2019). The claustrum in relation to seizures and electrical stimulation. *Front. Neuroanat.* 12 (13), 8. doi:10.3389/fnana.2019.00008

Kuwabara, T., Arai, A., Honma, N., and Nishizawa, M. (2005). Acute encephalopathy among patients with renal dysfunction after ingestion of "sugihiratake", angel's wing mushroom--study on the incipient cases in the northern area of Niigata Prefecture. *Rinsho Shinkeigaku* 45 (3), 239–245.

Lankhuijzen, L. M., and Ridler, T. (2024). Opioids, microglia, and temporal lobe epilepsy. *Front. Neurol.* 14, 1298489. doi:10.3389/fneur.2023.1298489

Laufs, H., Richardson, M. P., Salek-Haddadi, A., Vollmar, C., Duncan, J. S., Gale, K., et al. (2011). Converging PET and fMRI evidence for a common area involved in human focal epilepsies. *Neur.* 77 (9), 904–910. doi:10.1212/WNL.0b013e31822c90f2

Löscher, W., Ebert, U., Wahnschaffe, U., and Rundfeldt, C. (1995). Susceptibility of different cell layers of the anterior and posterior part of the piriform cortex to electrical stimulation and kindling: comparison with the basolateral amygdala and "area tempestas.". *Neuroscience* 66 (2), 265–276. doi:10.1016/0306-4522(94)00614-b

Luo, C., Li, Q., Lai, Y., Xia, Y., Qin, Y., Liao, W., et al. (2011a). Altered functional connectivity in default mode network in absence epilepsy: a resting-state fMRI study. *Hum. Brain Mapp.* 32 (3), 438–449. doi:10.1002/hbm.21034

Luo, C., Qiu, C., Guo, Z., Fang, J., Li, Q., Lei, X., et al. (2011b). Disrupted functional brain connectivity in partial epilepsy: a resting-state fMRI study. *PLoS One* 7 (1), e28196. doi:10.1371/journal.pone.0028196

Luo, T., Li, L., Li, J., Cai, S., Wang, Y., Zhang, L., et al. (2023). Claustrum modulates behavioral sensitivity and EEG activity of propofol anesthesia. *CNS Neurosci. Ther.* 29 (1), 378–389. doi:10.1111/cns.14012

Majak, K., and Moryś, J. (2007). Endopiriform nucleus connectivities: the implications for epileptogenesis and epilepsy. *Folia Morphol. Warsz.* 66 (4), 267–271.

Margerison, J. H., and Corsellis, J. A. (1966). Epilepsy and the temporal lobes. A clinical, electroencephalographic and neuropathological study of the brain in epilepsy, with particular reference to the temporal lobes. *Brain* 89 (3), 499–530. doi:10.1093/brain/89.3.499

Marriott, B. A., Do, A. D., Zahacy, R., and Jackson, J. (2021). Topographic gradients define the projection patterns of the claustrum core and shell in mice. *J. Comp. Neurol.* 529 (7), 1607–1627. doi:10.1002/cne.25043

Meletti, S., Giovannini, G., d'Orsi, G., Toran, L., Monti, G., Guha, R., et al. (2017). New-onset refractory status epilepticus with claustrum damage: definition of the clinical and neuroimaging features. *Front. Neurol.* 27 (8), 111. doi:10.3389/fneur.2017.00111

Meletti, S., Slonkova, J., Mareckova, I., Monti, G., Specchio, N., Hon, P., et al. (2015). Claustrum damage and refractory status epilepticus following febrile illness. *Neurology* 85 (14), 1224–1232. doi:10.1212/WNL/.0000000000001996

Miyakawa, N., Nagai, Y., Hori, Y., Mimura, K., Orihara, A., Oyama, K., et al. (2023). Chemogenetic attenuation of cortical seizures in nonhuman primates. *Nat. Commun.* 14 (1), 971. doi:10.1038/s41467-023-36642-6

Mohapel, P., Hannesson, D. K., Armitage, L. L., Gillespie, G. W., and Corcoran, M. E. (2000). Claustral lesions delay amygdaloid kindling in the rat. *Epilepsia* 41 (9), 1095–1101. doi:10.1111/j.1528-1157.2000.tb00313.x

Morgan, J. I., Cohen, D. R., Hempstead, J. L., and Curran, T. (1987). Mapping patterns of c-fos expression in the central nervous system after seizure. *Science* 10 (4811), 192–197. doi:10.1126/science.3037702

Muccioli, L., Pensato, U., Di Vito, L., Messia, M., Nicodemo, M., and Tinuper, P. (2022). Teaching neuroimage: claustrum sign in febrile infection-related epilepsy syndrome. *Neurology* 98 (10), e1090–e1091. doi:10.1212/WNL.0000000000013261

NCT04897776 (2024). Stimulation of the thalamus for arousal restoral in temporal lobe epilepsy (START). Available at: https://clinicaltrials.gov/study/NCT04897776.

Niibori, Y., Duba-Kiss, R., Bruder, J. T., Smith, J. B., and Hampson, D. R. (2023). *In silico* prediction and *in vivo* testing of promoters targeting GABAergic inhibitory neurons. *Mol. Ther. Methods Clin. Dev.* 28, 330–343. doi:10.1016/j.omtm.2023.01.007

Nishizawa, M. (2005). Acute encephalopathy after ingestion of "sugihiratke" mushroom. *Rinsho Shinkeigaku* 45 (11), 818–820.

Nixon, J., Bateman, D., and Moss, T. (2001). An MRI and neuropathological study of a case of fatal status epilepticus. *Seizure* 10 (8), 588–591. doi:10.1053/seiz.2001.0553

Nomoto, T., Seta, T., Nomura, K., Shikama, Y., Katagiri, T., Katsura, K., et al. (2007). A case of reversible encephalopathy accompanied by demyelination occurring after ingestion of sugihiratake mushrooms. *J. Nippon. Med. Sch.* 74 (3), 261–264. doi:10.1272/jnms.74.261

Peckys, D., and Landwehrmeyer, G. B. (1999). Expression of mu, kappa, and delta opioid receptor messenger RNA in the human CNS: a 33P *in situ* hybridization study. *Neuroscience* 88 (4), 1093–1135. doi:10.1016/s0306-4522(98)00251-6

Pires, G., Leitner, D., Drummond, E., Kanshin, E., Nayak, S., Askenazi, M., et al. (2021). Proteomic differences in the hippocampus and cortex of epilepsy brain tissue. *Brain Commun.* 3 (2), fcab021. doi:10.1093/braincomms/fcab021

Ranjan, M., Boutet, A., Bhatia, S., Wilfong, A., Hader, W., Lee, M. R., et al. (2019). Neuromodulation beyond neurostimulation for epilepsy: scope for focused ultrasound. *Expert Rev. Neurother.* 19 (10), 937–943. doi:10.1080/14737175.2019.1635013

Safan, A. S., Al-Termanini, M., Abdelhady, M., Osman, Y., Elzouki, A. Y., and Abdussalam, A. L. (2023). Claustrum sparing sign in seronegative limbic encephalitis. *eNeurologicalSci* 16 (31), 100465. doi:10.1016/j.ensci.2023.100465

Shaimardanova, A. A., Chulpanova, D. S., Mullagulova, A. I., Afawi, Z., Gamirova, R. G., Solovyeva, V. V., et al. (2022). Gene and cell therapy for epilepsy: mini review. *Front. Mol. Neurosci.* 11: 15: 868531. doi:10.3389/fnmol.2022.868531

Silva, G., Jacob, S., Melo, C., Alves, D., and Costa, D. (2018). Claustrum sign in a child with refractory status epilepticus after febrile illness: why does it happen? *Acta Neurol. Belg* 118 (2), 303–305. doi:10.1007/s13760-017-0820-9

Silva, R. A. E., and Sousa, T. A. P. (2019). Isolated involvement of external capsules and claustrum in status epilepticus. *Arq. Neuropsiquiatr.* 77 (5), 369. doi:10.1590/0004-282X20190040

Singh, S. P., Agarwal, S., and Faulkner, M. (2014). Refractory status epilepticus. *Ann. Indian Acad. Neurol.* 17 (Suppl. 1), S32–S36. doi:10.4103/0972-2327.128647

Siow, P., Tsao, C., Chang, H., Chen, C., Yu, I., Lee, K., et al. (2020). Mice lacking connective tissue growth factor in the forebrain exhibit delayed seizure response, reduced c-fos expression and different microglial phenotype following acute PTZ injection. *Int. J. Mol. Sci.* 21 (14), 4921. doi:10.3390/ijms21144921

Sitcoske O'Shea, M., Rosen, J. B., Post, R. M., and Weiss, S. R. (2000). Specific amygdaloid nuclei are involved in suppression or propagation of epileptiform activity during transition stage between oral automatisms and generalized clonic seizures. *Brain Res.* 873 (1), 1–17. doi:10.1016/s0006-8993(00)02307-6

Smith, J. B., and Alloway, K. D. (2014). Interhemispheric claustral circuits coordinate sensory and motor cortical areas that regulate exploratory behavior. *Front. Syst. Neurosci.* 19: 8: 93. doi:10.3389/fnsys.2014.00093

Smith, J. B., Alloway, K. D., Hof, P. R., Orman, R., Reser, D. H., Watakabe, A., et al. (2019a). The relationship between the claustrum and endopiriform nucleus: a perspective towards consensus on cross-species homology. *J. Comp. Neurol.* 527 (2), 476–499. doi:10.1002/cne.24537

Smith, J. B., Lee, A. K., and Jackson, J. (2020). The claustrum. *Curr. Biol.* 30 (23), R1401–R1406. doi:10.1016/j.cub.2020.09.069

Smith, J. B., Liang, Z., Watson, G. D. R., Alloway, K. D., and Zhang, N. (2017). Interhemispheric resting-state functional connectivity of the claustrum in the awake and anesthetized states. *Brain Struct. Funct.* 222 (5), 2041–2058. doi:10.1007/s00429-016-1323-9

Smith, J. B., Watson, G. D. R., Liang, Z., Liu, Y., Zhang, N., and Alloway, K. D. (2019b). A role for the claustrum in salience processing? *Front. Neuroanat.* 19, 64. doi:10.3389/fnana.2019.00064

Solbrig, M. V., Adrian, R., Baratta, J., Lauterborn, J. C., and Koob, G. F. (2006). Kappa opioid control of seizures produced by a virus in an animal model. *Brain* 129 (Pt 3), 642–654. doi:10.1093/brain/awl008

Solbrig, M. V., and Koob, G. F. (2004). Epilepsy, CNS viral injury and dynorphin. *Trends Pharmacol. Sci.* 25 (2), 98–104. doi:10.1016/j.tips.2003.12.010

Sperner, J., Sander, B., Lau, S., Krude, H., and Scheffner, D. (1996). Severe transitory encephalopathy with reversible lesions of the claustrum. *Pediatr. Radiol.* 26 (11), 769–771. doi:10.1007/BF01396197

Steriade, C., Tang-Wai, D. F., Krings, T., and Wennberg, R. (2017). Claustrum hyperintensities: a potential clue to autoimmune epilepsy. *Epilepsia Open* 2 (4), 476–480. doi:10.1002/epi4.12077

Stiefel, K. M., Merrifield, A., and Holcombe, A. O. (2014). The claustrum's proposed role in consciousness is supported by the effect and target localization of Salvia divinorum. *Front. Integr. Neurosci.* 26 (8), 20. doi:10.3389/fnint.2014.00020

Szyndler, J., Maciejak, P., Turzynska, D., Sobolewska, A., Taracha, E., Skorzewska, A., et al. (2009). Mapping of c-fos expression in the rat brain during the evolution of pentylenetetrazol-kindled seizures. *Epilepsy Behav.* 16 (2), 216–224. doi:10.1016/j.yebeh.2009.07.030

Vaughan, D. N., and Jackson, G. D. (2014). The piriform cortex and human focal epilepsy. *Front. Neurol.* 5, 259. doi:10.3389/fneur.2014.00259

Vormstein-Schneider, D., Lin, J. D., Pelkey, K. A., Chittajallu, R., Guo, B., Arias-Garcia, M. A., et al. (2020). Viral manipulation of functionally distinct interneurons in mice, non-human primates and humans. *Nat. Neurosci.* 23 (12), 1629–1636. doi:10.1038/s41593-020-0692-9

Wada, J. A., and Kudo, T. (1997). Involvement of the claustrum in the convulsive evolution of temporal limbic seizure in feline amygdaloid kindling. *Electroencephalogr. Clin. Neurophysiol.* 103 (2), 249–256. doi:10.1016/s0013-4694(97)96160-5

Wada, J. A., and Tsuchimochi, H. (1997). Role of the claustrum in convulsive evolution of visual afferent and partial nonconvulsive seizure in primates. *Epilepsia* 38 (8), 897–906. doi:10.1111/j.1528-1157.1997.tb01255.x

Watson, G. D. R., Afra, P., Bartolini, L., Graf, D. A., Kothare, S. V., McGoldrick, P., et al. (2021a). A journey into the unknown: an ethnographic examination of drug-resistant epilepsy treatment and management in the United States. *Epilepsy Behav.* 124, 108319. doi:10.1016/j.yebeh.2021.108319

Watson, G. D. R., Hughes, R. N., Petter, E. A., Fallon, I. P., Kim, N., Severino, F. P. U., et al. (2021b). Thalamic projections to the subthalamic nucleus contribute to movement initiation and rescue of parkinsonian symptoms. *Sci. Adv.* 7 (6), eabe9192. doi:10.1126/sciadv.abe9192

Watson, G. D. R., and Kopell, B. H. (2022). Editorial: all roads lead to Rome: Harnessing thalamic neuromodulation for difficult-to-treat neurological disorders. *Front. Hum. Neurosci.* 14, 1155605. doi:10.3389/fnhum.2023.1155605

Watson, G. D. R., Smith, J. B., and Alloway, K. D. (2017). Interhemispheric connections between the infralimbic and entorhinal cortices: the endopiriform nucleus has limbic connections that parallel the sensory and motor connections of the claustrum. *J. Comp. Neurol.* 15 (6), 1363–1380. doi:10.1002/cne.23981

Willoughby, J. O., Mackenzie, L., Medvedev, A., and Hiscock, J. J. (1997). Fos induction following systemic kainic acid: early expression in hippocampus and later widespread expression correlated with seizure. *Neuroscience* 77 (2), 379–392. doi:10.1016/s0306-4522(96)00462-9

Wong, K. L. L., Nair, A., and Augustine, G. J. (2021). Changing the cortical conductor's tempo: neuromodulation of the claustrum. *Front. Neural Circuits* 13, 658228. doi:10.3389/fncir.2021.658228

Zangrandi, L., and Schwarzer, C. (2022). The kappa opioid receptor system in temporal lobe epilepsy. *Handb. Exp. Pharmacol.* 271, 379–400. doi:10.1007/164_2021_444

Zhang, X., Hannesson, D. K., Saucier, D. M., Wallace, A. E., Howland, J., and Corcoran, M. E. (2001). Susceptibility to kindling and neuronal connections of the anterior claustrum. *J. Neurosci.* 15 (10), 3674–3687. doi:10.1523/JNEUROSCI.21-10-03674.2001

Zhang, X., Le Gal La Salle, G., Ridoux, V., Yu, P. H., and Ju, G. (1997). Prevention of kainic acid-induced limbic seizures and fos expression by the GABA-A receptor agonist muscimol. *Eur. J. Neurosci.* 9 (1), 29–40. doi:10.1111/j.1460-9568.1997.tb01350.x

TYPE Mini Review
PUBLISHED 03 June 2024
DOI 10.3389/fsysb.2024.1402664

Check for updates

OPEN ACCESS

EDITED BY
Rongling Wu,
The Pennsylvania State University (PSU),
United States

REVIEWED BY
Shaoyu Li,
University of North Carolina at Charlotte,
United States

*CORRESPONDENCE
Jianrong Wang,
✉ wangj164@msu.edu

RECEIVED 18 March 2024
ACCEPTED 15 May 2024
PUBLISHED 03 June 2024

CITATION
Yang J, Das Adhikari S, Wang H, Huang B, Qi W, Cui Y and Wang J (2024), De novo prediction of functional effects of genetic variants from DNA sequences based on context-specific molecular information.
Front. Syst. Biol. 4:1402664.
doi: 10.3389/fsysb.2024.1402664

De novo prediction of functional effects of genetic variants from DNA sequences based on context-specific molecular information

Jiaxin Yang[1], Sikta Das Adhikari[1,2], Hao Wang[1], Binbin Huang[1], Wenjie Qi[1,3], Yuehua Cui[2] and Jianrong Wang[1]*

[1]Department of Computational Mathematics, Science and Engineering, Michigan State University, East Lansing, MI, United States, [2]Department of Statistics and Probability, Michigan State University, East Lansing, MI, United States, [3]Department of Biomedical Engineering, Michigan State University, East Lansing, MI, United States

Deciphering the functional effects of noncoding genetic variants stands as a fundamental challenge in human genetics. Traditional approaches, such as Genome-Wide Association Studies (GWAS), Transcriptome-Wide Association Studies (TWAS), and Quantitative Trait Loci (QTL) studies, are constrained by obscured the underlying molecular-level mechanisms, making it challenging to unravel the genetic basis of complex traits. The advent of Next-Generation Sequencing (NGS) technologies has enabled context-specific genome-wide measurements, encompassing gene expression, chromatin accessibility, epigenetic marks, and transcription factor binding sites, to be obtained across diverse cell types and tissues, paving the way for decoding genetic variation effects directly from DNA sequences only. The *de novo* predictions of functional effects are pivotal for enhancing our comprehension of transcriptional regulation and its disruptions caused by the plethora of noncoding genetic variants linked to human diseases and traits. This review provides a systematic overview of the state-of-the-art models and algorithms for genetic variant effect predictions, including traditional sequence-based models, Deep Learning models, and the cutting-edge Foundation Models. It delves into the ongoing challenges and prospective directions, presenting an in-depth perspective on contemporary developments in this domain.

KEYWORDS

genetic variants, deep learning, DNA sequence, disease genetics, systems genetics, cellular context specificity, foundation models

Introduction

Genetic variants have emerged as pivotal factors in the etiology of severe human diseases (Klein et al., 2005). Therefore, quantitative and systems-level understandings of the relationship between human diseases and genetic variants are critical in precision medicine and clinical care. Over the past decades, the Genome-wide Association Study (GWAS) (Hirschhorn and Daly, 2005; Visscher et al., 2012) has revolutionized the field of complex disease genetics, in which millions of single-nucleotide polymorphisms (SNPs) of individuals are tested to identify significant genotype-phenotype associations. However, GWAS grapples with two pronounced limitations that have spurred the quest for advanced

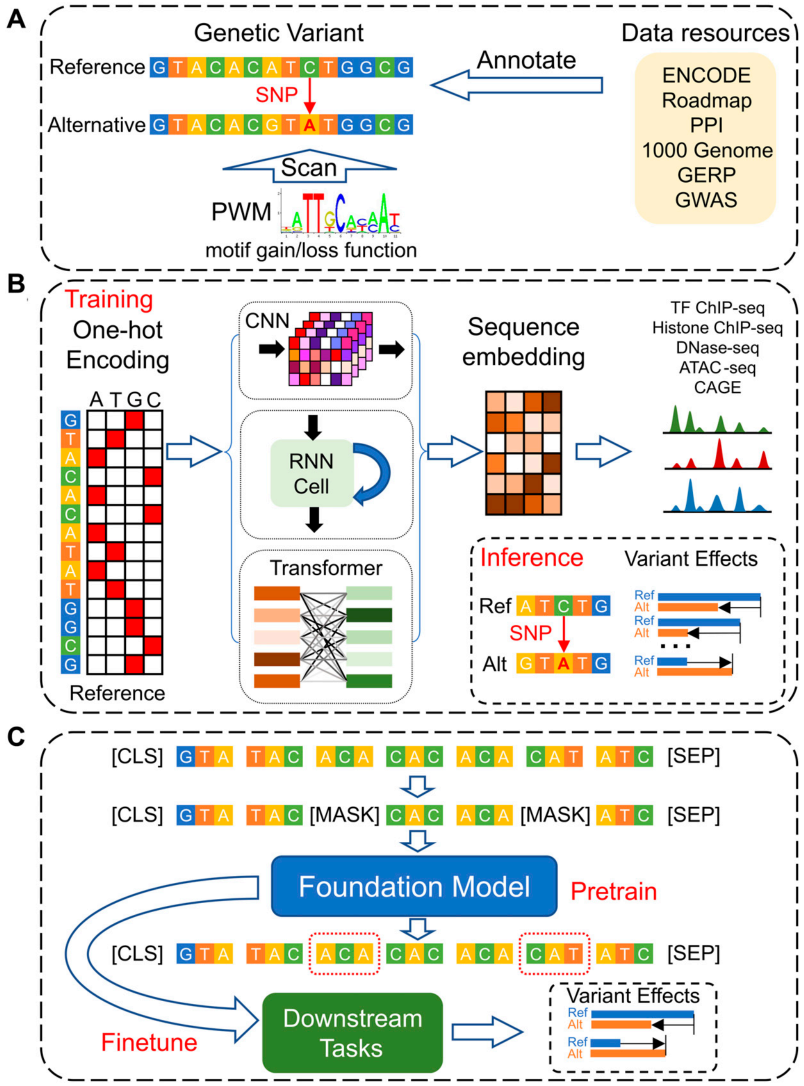


FIGURE 1
The development of models for genetic variants' effect predictions based on DNA sequences. **(A)** Traditional models leverage multi-omics data resources to annotate and prioritize genetic variants and use static motif PWMs to analyze the gain- and loss-function of TF bindings. **(B)** Deep Learning models, employing CNN, RNN, and Transformer architectures, are designed to predict functional genomics profiles across various cell types. They determine the effects of genetic variants by comparing the predicted genomic profiles for the reference *versus* alternative alleles. **(C)** Foundation Models utilize a self-supervised pre-training strategy based on DNA sequences only, enabling them to be efficiently fine-tuned for a range of downstream tasks, including the prediction of genetic variant effects across different cellular contexts.

methodologies (Tam et al., 2019). Firstly, it often limited by low statistical power, mainly stemming from the constraints imposed by limited sample sizes and the arduous multi-testing demands. Secondly, the causal relationships between specific genetic variants and diseases remain obscured, partly owing to the ambiguity induced by Linkage Disequilibrium (LD) (Bulik-Sullivan et al., 2015) and the paucity of insights into the underlying molecular mechanisms. Traditionally, human disease genetics research has centered around SNPs located in protein coding regions, a mere 1.2% of the human genome (Visscher et al., 2012). Next-generation Sequencing (NGS) (Buermans and den Dunnen, 2014) technologies like RNA-seq, DNase-seq, and ChIP-seq (Luo et al., 2020) have empowered researchers to measure gene expression, chromatin accessibility, and transcription factor (TF) binding genome-wide. This advance fuels an exploration of the vast non-coding genome and gives the potential to analyze the effect of genetic variants on nearby local regions.

Given the DNA sequence's fundamental role as the instruction manual for all aspects of life, understanding the function of regulatory genomic elements that control gene expression is paramount. Moving beyond population-based statistical analyses like GWAS and Transcriptome-Wide Association Studies (TWAS) (Wainberg et al., 2019), direct predictions of genetic variant effects from DNA sequences are pivotal for elucidating the underlying biological mechanisms. This review will explore the evolution of computational models for predicting genetic variant effects genome-wide. We first review the traditional annotation-based models that rely on simple sequence motifs to estimate variant impacts, then dive into the advancements achieved through *de novo* prediction models that leverage deep learning techniques (Figure 1). We conclude by discussing the current challenges in the field of systems genetics and proposing future research directions that hold promise for further breakthroughs.

Functional variant annotation and prioritization

The ENCODE (Luo et al., 2020) and the Roadmap Epigenomics Consortium (Bernstein et al., 2010) have significantly advanced our understanding of the human genome by profiling a wide array of functional noncoding elements through diverse assays. This wealth of data has enabled the functional annotation of genetic variants across the human genome (Figure 1A). GWAVA (Ritchie et al., 2014), by leveraging a comprehensive suite of genomic and epigenomic annotations, predicts the functional impact of noncoding variants. Its features encompass open chromatin regions, TF binding sites, histone modifications, RNA polymerase interactions, CpG islands, genomic segmentation, evolutionary conservation, genic context, and sequence context. These annotations are synthesized to mitigate the challenges posed by context dependency and the variability of evolutionary conservation signals within regulatory elements. Furthermore, pattern recognition algorithms help to identify DNA sequence motifs overrepresented in regulatory regions of co-expressed genes, enhancing our understanding of gene regulation (Stormo and Fields, 1998). The Position Weight Matrix (PWM) (Stormo and Fields, 1998) represents DNA binding sites of different TFs by scoring each potential base at a given genomic position, thereby quantifying the specificity of protein-DNA interactions and facilitating the prediction of new binding sites. An annotation-based approach, Funseq2 (Fu et al., 2014), integrates these methodologies to analyze loss-of-function and gain-of-function events in TF binding. It calculates motif-breaking scores for variants within TF binding motifs identified by ChIP-seq peaks, and motif-gaining scores for variants in promoters or regulatory elements significantly associated with genes, based on PWM *p*-values for the mutated allele. Funseq2 also incorporates annotation-based features such as conservation, enhancer-gene links, network centrality, and recurrence across samples. However, reliance solely on regulatory annotations and static PWMs has its drawbacks: many variants in non-coding regions do not overlap with regulatory annotations, and novel motifs cannot be discovered through static PWMs (Zhou and Troyanskaya, 2015; Kelley et al., 2016).

Addressing these limitations, kmer-SVM (Lee et al., 2011) emerged as a pioneering model for predicting regulatory elements directly from DNA sequences, bypassing the need for existing annotated motifs. It counts the frequencies of various k-mers within a piece of DNA sequence, employing a support vector machine (SVM) trained on these k-mer features to assess the likelihood of a sequence being a functional genomic regulatory element or a tissue-specific enhancer. Gapped k-mers, utilized as features in the gkm-SVM (Ghandi et al., 2014), have further enhanced model accuracy in enhancer identification and TF binding site prediction. Moreover, Delta-SVM (Lee et al., 2015) incorporates the gkm-SVM predictions to assess the disruptive impacts of genetic variants. Despite these advances, the complexity and non-linearity of the underlying regulatory grammar in DNA sequences require further improvements in model performance (Zhou and Troyanskaya, 2015; Kelley et al., 2016).

De novo prediction of genetic variants' effects based on deep learning

Deep learning excels in two key capabilities: 1) extracting and representing features, with enhanced flexibility and power, from semi-structured and unstructured data formats, such as texts and images, and 2) approximating various functions effectively through deep layering, with neural networks comprising stacks of linear transformations interspersed with non-linear activations. For the purpose of predicting the effects of genetic variants (Figure 1B), deep learning models typically represent reference DNA sequences using the one-hot encoding (where A = [1,0,0,0], C = [0,1,0,0], G = [0,0,1,0], T = [0,0,0,1], and N = [0,0,0,0]). The input DNA fragments are represented accordingly, $S \in \mathbb{R}^{4\times L}$, where L denotes the DNA sequence length. Feature extraction from these one-hot encoded sequences to produce sequence embeddings typically employs two foundational architectures: the 1D Convolutional Neural Network (CNN) (O'Shea and Nash, 2015) and the Recurrent Neural Network (RNN) (Sherstinsky, 2018), such as Long Short-Term Memory Network (LSTM) (Sherstinsky, 2018).

The CNN architecture focuses on local sequence information, with the initial layer acting as a position-weight matrix, so that the convolution operations are analogous to computing PWM scores across the DNA sequence within each sliding window. Subsequent deep CNN layers capture the non-linear and complex sequence signatures, by utilizing the pooling layers to reduce dimensions after each CNN layer. On the other hand, the LSTM architectures capture sequential dependencies in the genome, by incorporating an internal state that reflects the long-term sequential information. Following these feature representation layers, several fully connected layers are then utilized to generate the final predictions. CNNs, in particular, are adept at learning hierarchical layers of complex, nonlinear patterns without requiring strong prior biological assumptions, thus enabling the discovery of novel sequence motifs and their organizational sequence contexts (Zhou and Troyanskaya, 2015; Kelley et al., 2016; Quang and Xie, 2016).

Pioneering applications of deep neural networks in this field, such as DeepSEA (Zhou and Troyanskaya, 2015) and Basset (Kelley et al., 2016), have demonstrated the significant potential of CNNs for predicting genetic variants' effects based solely on DNA sequences. DeepSEA leverages a multi-task CNN model to predict TF ChIP-seq, DNase-seq, and histone mark ChIP-seq peaks across a variety of cell types, based on the data from the ENCODE and Roadmap Epigenomics projects. Basset focuses on chromatin accessibility, while DanQ (Quang and Xie, 2016) combines CNN and LSTM to enhance peak profile prediction performance. Trained on the large-scale multi-omics datasets across different cell types from the reference genome, these deep learning models are thus capable of predicting the peak profiles of distinct regulatory factors in a cell-type specific way. For a specific alternative allele of interest, the model's predictions based on the altered genome sequence are compared to those based on the reference genome. The differences in predictions are then used as indicators of the alternative allele's functional disruptions under specific cellular contexts, leading to mechanistic hypotheses of its downstream effects in complex human diseases.

Further advancements have seen models like Basenji (Kelley et al., 2018), which employs CNN architectures to predict a wider range of genomic signals, including DNase-seq, histone mark ChIP-seq, and CAGE signals across cell types. By using dilated convolution layers, Basenji is able to capture more contextual information around 32 kb DNA sequence windows, thereby identifying relevant regulatory sequences over a broader scope. Additionally, efforts to understand genetic variant effects have expanded from modeling the genomic and epigenomic levels to predicting target genes' expressions. For instance, ExPecto (Zhou et al., 2018) predicts the effects on nearby gene expression in a two-stage strategy. First, ExPecto forecasts histone marks, TF, and DNase profiles from DNA sequences, and second, it aggregates the forecasted signals to make predictions of tissue-specific expression. This approach allows for the interpretation of genetic variants' effects in the dysregulation of nearby genes. Moreover, BPNet (Avsec et al., 2021a) has pushed the boundaries further by predicting base-resolution genomic profiles, utilizing a CNN architecture without pooling layers to achieve the single-base pair resolution predictions.

Cross-species regulatory information and long-range variant effects

Expanding the training dataset is a well-regarded strategy to enhance the accuracy of deep learning models. While new genome-wide functional genomics profiles grow fast, these new datasets primarily provide information that has already been captured by the model from existing datasets in the human genome. The additional benefits of gathering more functional genomics datasets from additional human genomes may decrease, since the genotypes of different individuals are largely similar. In this context, the quest for significantly different training sequences becomes paramount, with a greater potential to develop and refine more sophisticated and precise models.

An intriguing solution lies in the exploration of non-human species as a reservoir of novel training data. The regulatory DNA sequences of species that are genetically related to humans possess sufficient similarities, enabling the application of machine learning models trained across these diverse genomes. Such cross-species training has the potential to enhance the models' understanding of regulatory sequence activities. An example of this approach is the expansion of the Basenji model to simultaneously process functional genomic signal tracks from both the mouse and human genomes (Kelley, 2020). This cross-species training strategy has been shown to yield more accurate predictions on the test set of sequences which has not been seen by the model previously, compared to those trained exclusively on data from a single species. This innovative approach underscores the utility of integrating diverse genomic data sources to significantly advance the precision of predictive models in functional genomics.

However, CNNs, the key architecture in previous models, often struggle with the problem of capturing semantic dependencies over long genomic distances due to their focus on localized feature extraction, which is limited by the filter size. Besides, RNNs can learn long-term dependencies but are hampered by issues like vanishing gradients and inefficiency in dealing with long genomic sequences. This limitation is particularly challenging in modeling complex cell-type specific gene regulation, where distal enhancers can influence gene expression over large distances (Lieberman-Aiden et al., 2009; Wang et al., 2021), underscoring the importance in predicting long-range effects of genetic variants. The Transformer model (Vaswani et al., 2017) has demonstrated remarkable success beyond its initial applications in natural language processing and computer vision, increasingly supplanting traditional CNN and RNN-based models across various domains. Its exceptional capability to capture long-range dependencies without relying on recurrent units renders it more scalable and adaptable for handling large datasets. At the heart of the Transformer architecture is the multi-head self-attention mechanism, which efficiently models dependencies between genomic locations, regardless of their distance (Vaswani et al., 2017). This ability allows deeper layers of the model to discern increasingly complex relationships, facilitating the prediction of distal genetic variant effects by capturing interactions between genomic locations separated by considerable distances.

Enformer (Avsec et al., 2021b), a state-of-the-art model leveraging both CNNs and the Transformer architecture, excels in predicting histone marks, TF binding sites, chromatin accessibility, and gene expression across diverse cell types, including those from the genomes of human and mouse. Its design significantly extends the model's receptive field, enabling the identification of distal regulatory elements up to 100 kb away. This expansive reach allows Enformer to integrate information from all pertinent regions, such as enhancers, thereby enhancing gene expression prediction. Moreover, the model's attention weights offer greater interpretability, shedding light on the underlying mechanisms of chromatin and gene regulation. With its superior performance of predictions across >5,000 functional genome profiles, including gene expressions, Enformer showcases an unparalleled capacity to forecast both local and distal genetic variant effects. This demonstrates the potential of Transformer-based models in advancing our understanding and prediction of genetic regulations underlying complex traits.

General sequence grammar of variants learned by foundation models

Traditional deep learning models have achieved impressive results in interpreting functional genomic profiles from DNA sequences through supervised learning, where the models are trained to accurately predict experimental genomic tracks based on the sequence representations. However, this approach necessitates a vast amount of labeled data, constraining the models' performance and utility in situations where labeled data is scarce. Obtaining high-quality, labeled datasets is often expensive and time-consuming. Moreover, the available data tends to be biased towards certain well-studied cell types with many tracks, neglecting a broad spectrum of cell types yet to be explored. This imbalance results in overrepresented genomic tracks overshadowing the DNA sequence representation, diminishing the efficacy of genomic variant effect prediction in less studied, underrepresented cell types.

In contrast, the development of Foundation Models originally in the fields such as text and image generation illustrates the potential

TABLE 1 Summary of computational models.

Tool	Model architecture	Required data	Link
GWAVA	Annotation-based	Experimental annotation	https://www.sanger.ac.uk/tool/gwava/
Funseq2	Annotation + PWM	Experimental annotation + DNA sequence	http://funseq2.gersteinlab.org/
Delta-SVM	SVM	DNA sequence	https://www.beerlab.org/deltasvm/
DeepSEA	CNN	DNA sequence + experiment peaks	https://hb.flatironinstitute.org/deepsea/
Basset	CNN	DNA sequence + experiment peaks	https://github.com/davek44/Basset
DanQ	CNN + LSTM	DNA sequence + experiment peaks	https://github.com/uci-cbcl/DanQ
Basenji	CNN	DNA sequence + experiment signals	https://github.com/calico/basenji
ExPecto	CNN + regression	DNA sequence + experiment signals	https://github.com/FunctionLab/ExPecto
BPNet	CNN	DNA sequence + experiment signals	https://github.com/kundajelab/bpnet/
Basenji2	CNN	DNA sequence + experiment signals	https://github.com/calico/basenji
Enformer	CNN + Transformer	DNA sequence + experiment signals across species	https://github.com/google-deepmind/deepmind-research/tree/master/enformer
DNABERT	Transformer	DNA sequence	https://github.com/jerryji1993/DNABERT
DNABERT2	Transformer	DNA sequence	https://github.com/MAGICS-LAB/DNABERT_2
DNABERTS	Transformer	DNA sequence	https://github.com/MAGICS-LAB/DNABERT_S
The Nucleotide Transformer	Transformer	DNA sequence	https://github.com/instadeepai/nucleotide-transformer
HyenaDNA	Hyena	DNA sequence	https://github.com/HazyResearch/hyena-dna

benefits of leveraging context information through a self-supervised pre-training strategy (Devlin et al., 2018; Brown et al., 2020). These models, trained on enormous datasets, have demonstrated capabilities surpassing human performance in certain tasks. The pre-training and fine-tuning framework of Foundation Models involves initial training on vast unlabeled datasets, followed by fine-tuning for specific downstream tasks (Devlin et al., 2018; Brown et al., 2020). Applied to disease genetics studies, this approach entails pre-training models on unlabeled genomic sequences, which are subsequently fine-tuned for specific genomic interpretation tasks (Figure 1C). This methodology not only mitigates the challenges associated with data scarcity and bias but also enhances the model's ability to understand and predict across a diverse range of cell types and genomic contexts (Ji et al., 2021).

DNABERT (Ji et al., 2021) is a pioneer encoder-based Foundation Model in genetics. It processes DNA sequences by breaking them down into k-mers. For input sequences with lengths up to 512 bp, 15% of k-mers are randomly replaced by a [MASK] token. The Transformer encoder then leverages context information to reconstruct these masked k-mers without additional information. By accurately reconstructing the masked k-mers, DNABERT captures the fundamental grammatical structures of DNA sequences, enabling it to generate meaningful representations for any given sequence. This model has demonstrated remarkable efficacy across numerous downstream applications (Ji et al., 2021), such as promoter identification, TF binding site prediction, and the detection of functional genetic variants. Building on DNABERT's foundation, subsequent iterations like DNABERT2 (Zhou et al., 2023) and DNABERTS (Zhou et al., 2024) have broadened the scope of Foundation Models to encompass a wider range of species beyond just humans.

The Nucleotide Transformer (Dalla-Torre et al., 2023), an advanced and larger encoder-based Foundation Model, is pre-trained on DNA sequences with over 2.5 billion parameters and can handle sequences up to 6 kb in length. This model has shown remarkable success in a variety of downstream tasks (Dalla-Torre et al., 2023) after fine-tuning, demonstrating the beneficial impacts of both increased model size and the ability to process longer sequences. Beyond the Transformer architecture, HyenaDNA (Nguyen et al., 2023) innovatively extends the contextual reach to up to 1 million tokens at the single nucleotide level through the use of global convolutional filters. This significant enhancement enables the model to effectively leverage long-range chromatin regulation at single base pair resolution. Additionally, HyenaDNA introduces novel downstream adaptation methods, such as a unique soft prompt technique. This approach allows for exceptional downstream results without the necessity of updates to the pre-trained model, thus facilitating the seamless application of the Foundation Model to various tasks, including the prediction of genetic variant effects. This revolution in model design and functionality marks a pivotal advancement in our capacity to understand and interpret complex genetic information.

Discussions

This review has explored the evolution of models dedicated to predicting the effects of genetic variants using only DNA sequences

(Table 1). Enabled by the widespread availability of multi-omics datasets and enhanced computational resources, researchers have transitioned from basic feature annotation and motif recognition to the development of sophisticated deep learning models. These models, trained through both supervised and self-supervised approaches, have progressively achieved more accurate predictions of the genetic variant effects across a variety of cell types.

Despite their advancements, deep learning models for predicting genetic variant effects face two significant challenges: Firstly, model training predominantly relies on labeled data at the cell type level, which limits their capability to discern the functional effects at the single-cell level. With the advent of single-cell sequencing technologies, such as scRNA-seq, scATAC-seq, and scHi-C, there is an influx of data providing detailed insights into gene expression, chromatin accessibility, and regulation at the single-cell level. This type of data, however, tends to be sparse and noisy. Foundation models, pre-trained on the fundamental sequence grammar, exhibit a strong potential for enhancing their performance through fine-tuning with minimal data, addressing the challenge of integrating single-cell level data. Secondly, the training of current models is anchored to the reference genome, neglecting the diversity and frequency of genetic variations across different genotypes. While these models may excel in predicting genetic profiles based on the reference genome, they primarily capture consensus information, which may not accurately represent the actual effects of genetic variants. The discrepancies between the reference and alternative alleles do not fully encapsulate the impact of genetic variants. CRISPR (Korkmaz et al., 2016; Fulco et al., 2019) technology, which elucidates the casual and real effects of genetic variants, offers valuable insights beyond the reference genomic context. The CRISPR-derived data is expected to help to fill the gap between model predictions and biological reality.

Author contributions

JY: Conceptualization, Writing–review and editing, Writing–original draft, Data curation, Formal Analysis, Visualization. SD: Conceptualization, Writing–review and editing. HW: Conceptualization, Writing–review and editing. BH: Conceptualization, Writing–review and editing. WQ: Conceptualization, Writing–review and editing. YC: Conceptualization, Writing–review and editing. JW: Conceptualization, Writing–review and editing, Funding acquisition, Project administration, Supervision, Writing–original draft.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work is funded, in part, by the awards R01GM131398 from the National Institutes of Health and NSF1942143 from the National Science Foundation

Acknowledgments

The authors thank Pronoy Kanti Mondal and Tairan Song for helpful inputs and discussions.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., et al. (2021b). Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* 18, 1196–1203. doi:10.1038/s41592-021-01252-x

Avsec, Ž., Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., Dalal, K., et al. (2021a). Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat. Genet.* 53, 354–366. doi:10.1038/s41588-021-00782-6

Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., et al. (2010). The NIH Roadmap epigenomics mapping Consortium. *Nat. Biotechnol.* 28, 1045–1048. doi:10.1038/nbt1010-1045

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Jared, K., Dario, A., et al. (2020). *Language models are few-shot learners*. arXiv:2005.14165. doi:10.48550/arXiv.2005.14165

Buermans, H. P. J., and den Dunnen, J. T. (2014). Next generation sequencing technology: advances and applications. *Biochimica Biophysica Acta (BBA) - Mol. Basis Dis.* 1842, 1932–1941. doi:10.1016/j.bbadis.2014.06.015

Bulik-Sullivan, B. K., Loh, P. R., Finucane, H. K., Ripke, S., Yang, J., et al. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* 47, 291–295. doi:10.1038/ng.3211

Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., Carranza, N. L., Grzywaczewski, A. H., Oteri, F., et al. (2023). The nucleotide transformer: building and evaluating robust foundation models for human genomics. *bioRxiv* 2023, 523679. doi:10.1101/2023.01.11.523679

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). *BERT: pre-training of deep bidirectional transformers for language understanding*. arXiv:1810.04805. doi:10.48550/arXiv.1810.04805

Fu, Y., Liu, Z., Lou, S., Bedford, J., Mu, X. J., Yip, K. Y., et al. (2014). FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol.* 15, 480. doi:10.1186/s13059-014-0480-5

Fulco, C. P., Nasser, J., Jones, T. R., Munson, G., Bergman, D. T., Subramanian, V., et al. (2019). Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* 51, 1664–1669. doi:10.1038/s41588-019-0538-0

Ghandi, M., Lee, D., Mohammad-Noori, M., and Beer, M. A. (2014). Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput. Biol.* 10, e1003711. doi:10.1371/journal.pcbi.1003711

Hirschhorn, J. N., and Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* 6, 95–108. doi:10.1038/nrg1521

Ji, Y., Zhou, Z., Liu, H., and Davuluri, R. V. (2021). DNABERT: pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. *Bioinformatics* 37, 2112–2120. doi:10.1093/bioinformatics/btab083

Kelley, D. R. (2020). Cross-species regulatory sequence activity prediction. *PLoS Comput. Biol.* 16, e1008050. doi:10.1371/journal.pcbi.1008050

Kelley, D. R., Reshef, Y. A., Bileschi, M., Belanger, D., McLean, C. Y., and Snoek, J. (2018). Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res.* 28, 739–750. doi:10.1101/gr.227819.117

Kelley, D. R., Snoek, J., and Rinn, J. L. (2016). Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* 26, 990–999. doi:10.1101/gr.200535.115

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., et al. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science* 308, 385–389. doi:10.1126/science.1109557

Korkmaz, G., Lopes, R., Ugalde, A. P., Nevedomskaya, E., Myacheva, K., et al. (2016). Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. *Nat. Biotechnol.* 34, 192–198. doi:10.1038/nbt.3450

Lee, D., Gorkin, D. U., Baker, M., Strober, B. J., Asoni, A. L., McCallion, A. S., et al. (2015). A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.* 47, 955–961. doi:10.1038/ng.3331

Lee, D., Karchin, R., and Beer, M. A. (2011). Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Res.* 21, 2167–2180. doi:10.1101/gr.121905.111

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326, 289–293. doi:10.1126/science.1181369

Luo, Y., Hitz, B. C., Gabdank, I., Hilton, J. A., Kagda, M. S., Lam, B., et al. (2020). New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res.* 48, D882–D889. doi:10.1093/nar/gkz1062

Nguyen, E., Poli, M., Faizi, M., Thomas, A., Aman, P., Re, C., et al. (2023) *HyenaDNA: long-range genomic sequence modeling at single nucleotide resolution*. arXiv:2306.15794. doi:10.48550/arXiv.2306.15794

O'Shea, K., and Nash, R. (2015). *An introduction to convolutional neural networks*.

Quang, D., and Xie, X. (2016). DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res.* 44, e107. doi:10.1093/nar/gkw226

Ritchie, G. R. S., Dunham, I., Zeggini, E., and Flicek, P. (2014). Functional annotation of noncoding sequence variants. *Nat. Methods* 11, 294–296. doi:10.1038/nmeth.2832

Sherstinsky, A. (2018). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Phys. D. Nonlinear Phenom.* 404, 132306. doi:10.1016/j.physd.2019.132306

Stormo, G. D., and Fields, D. S. (1998). Specificity, free energy and information content in protein–DNA interactions. *Trends Biochem. Sci.* 23, 109–113. doi:10.1016/s0968-0004(98)01187-6

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* 20, 467–484. doi:10.1038/s41576-019-0127-1

Vaswani, A., et al. (2017). "Attention is all you need," in *Advances in neural information processing systems*. Editor I. Guyon (Curran Associates, Inc.), 30.

Visscher, P. M., Brown, M. A., McCarthy, M. I., and Yang, J. (2012). Five years of GWAS discovery. *Am. J. Hum. Genet.* 90, 7–24. doi:10.1016/j.ajhg.2011.11.029

Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, A. N., Knowles, D. A., Golan, D., et al. (2019). Opportunities and challenges for transcriptome-wide association studies. *Nat. Genet.* 51, 592–599. doi:10.1038/s41588-019-0385-z

Wang, H., Yang, J., Zhang, Y., and Wang, J. (2021). Discover novel disease-associated genes based on regulatory networks of long-range chromatin interactions. *Methods* 189, 22–33. doi:10.1016/j.ymeth.2020.10.010

Zhou, J., Theesfeld, C. L., Yao, K., Chen, K. M., Wong, A. K., and Troyanskaya, O. G. (2018). Deep learning sequence-based *ab initio* prediction of variant effects on expression and disease risk. *Nat. Genet.* 50, 1171–1179. doi:10.1038/s41588-018-0160-6

Zhou, J., and Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. *Nat. Methods* 12, 931–934. doi:10.1038/nmeth.3547

Zhou, Z., Ji, Y., Li, W., Dutta, P., Ramana, D., Liu, H., et al. (2023) *DNABERT-2: efficient foundation model and benchmark for multi-species genome*. arXiv:2306.15006. doi:10.48550/arXiv.2306.1500

Zhou, Z., Wu, W., Ho, H., Wang, J., Shi, L., Liu, H., et al. (2024). *DNABERT-S: learning species-aware DNA embedding with genome foundation models*. arXiv:2402.08777. doi:10.48550/arXiv.2402.0877

frontiers | Frontiers in Systems Biology

TYPE Perspective
PUBLISHED 17 July 2024
DOI 10.3389/fsysb.2024.1417800

Life's building blocks: the modular path to multiscale complexity

Saúl Huitzil[1,2]* and Cristián Huepe[1,2,3]

[1]Engineering Sciences and Applied Mathematics, Northwestern University, Evanston, IL, United States, [2]Northwestern Institute on Complex Systems, Northwestern University, Evanston, IL, United States, [3]CHuepe Labs, Chicago, IL, United States

OPEN ACCESS

EDITED BY
Shayn Peirce-Cottler,
University of Virginia, United States

REVIEWED BY
Nathan Weinstein,
Copenhagen University Hospital,
Rigshospitalet, Denmark

*CORRESPONDENCE
Saúl Huitzil,
saulhuitzil@gmail.com

RECEIVED 15 April 2024
ACCEPTED 19 June 2024
PUBLISHED 17 July 2024

CITATION
Huitzil S and Huepe C (2024), Life's building blocks: the modular path to multiscale complexity.
Front. Syst. Biol. 4:1417800.
doi: 10.3389/fsysb.2024.1417800

Modularity, the structuring of systems into discrete, interconnected units, is a fundamental organizing principle in biology across multiple scales. Recent progress in understanding the role of modularity as an evolutionary mechanism and a key driver of biological complexity has highlighted its importance in shaping the structure and function of living systems. Here, we propose a unifying framework that identifies the potential evolutionary advantages of modularity in systems ranging from molecular networks to ecologies, such as facilitating evolvability, enhancing robustness, improving information flows, and enabling the emergence of higher-level functions. Our analysis reveals the pervasiveness of modularity in living systems and highlights its crucial role in the evolution of multiscale hierarchies of increasing complexity.

KEYWORDS

modularity, multiscale modeling, hierarchical organization, emergent complexity, evolutionary structures, modular evolution, evolutionary dynamics, biological complexity

1 Introduction

Modularity is a fundamental organizing principle in biological systems that manifests itself at multiple scales and levels of organization (Ravasz et al., 2002; Meunier et al., 2009; Lorenz et al., 2011). Although its precise definition can depend on the context, in a broad sense, modularity in biology has been connected to the capacity of living systems to be "near decomposable," (Simon, 1962), that is, to their ability to divide functions into different subunits known as modules, which perform specific tasks with a certain degree of autonomy (Wagner et al., 2007). These modules can be viewed as composed of parts that interact more closely with each other than with other modules, thus showing a degree of functional independence that allows them to perform specific functions efficiently (Klingenberg et al., 2003; Cheverud et al., 2004; Kadelka et al., 2023). Modularity is also closely related to the emergence of hierarchical organization, in which systems are organized into nested levels, where each level is composed of subsystems from lower levels and, in turn, forms part of supersystems at higher levels (Barabasi and Oltvai, 2004).

Modularity is a multifaceted concept that has been studied through diverse perspectives, including developmental, evolutionary, genetic, and morphological approaches, each with its own set of questions, methods, and insights (Zelditch and Goswami, 2021). For example, network theory provides a quantitative framework for analyzing modularity based on topological features, while other approaches focus on the physical structures found in living organisms, such as the organization of cells into tissues and organs or the arrangement of skeletal elements (Melo et al., 2016; Felice et al., 2018). Other studies have explored the modular functional interactions among components, such as gene regulatory networks and metabolic pathways (Raff, 1996; Wagner et al., 2007). While these approaches have different emphases and may not always fully address the origins, evolution, or implications of modularity, their collective findings highlight the ubiquity of modular organization in biological systems. This suggests the existence of a universal principle driving the emergence of complexity, whereby simpler subsystems agglomerate into stable combinations that become the building blocks of larger and more intricate structures and functions, potentially leading to the formation of hierarchical layers through successive combinations of components and subcomponents (Schaffer and Ideker, 2021). In this context, biological complexity is understood as the degree to which a system comprised of interrelated components can collectively exhibit emergent properties and behaviors that are more than the sum of its parts (Lobo, 2008). To fully understand the role of modularity in the organization of life, an integrative approach that synthesizes insights from different perspectives and considers the origins, evolution, and implications of modularity across multiple scales is necessary.

The fundamental role of modularity in the evolution of biological complexity is evidenced by its presence in a great diversity of living systems (at multiple scales). For example, the modular organization of cells is considered a crucial factor in the emergence of higher life forms. As highlighted by Lynn Margulis' groundbreaking work on endosymbiotic theory, the origin of eukaryotic cells is a prime example of how modularity has driven the emergence of more complex forms of life (Sagan, 1967; Gray, 2017). According to this theory, the modular integration of specialized organelles (such as mitochondria and chloroplasts), which evolved from symbiotic bacteria, allowed for greater efficiency in cellular processes and played a key role in the appearance of eukaryotic cells (Schliwa and van Blerkom, 1981).

The emergence of multicellularity is another notable example of how modularity has driven the evolution towards increasing complexity, as discussed by Smith and Szathmary (1997) in "The Major Transitions in Evolution." This seminal work explores the role of modularity in the evolution of life, from the integration of replicating molecules into chromosomes to the origin of societies. Organisms like *Volvox carteri*, which appear to be in a transitional stage towards multicellularity (Kirk, 2005), demonstrate how the organization of cells into modules can give rise to more complex life forms. In more advanced multicellular organisms, modular specialization extends to tissues and organs, thus enabling the emergence of highly complex adaptive systems (Bonner, 1988; Wagner and Altenberg, 1996).

The holobiont concept (increasingly relevant for systems biology) further illustrates how modularity and hierarchical organization enable the emergence of higher levels of complexity in biological systems. The holobiont refers to the collective biological entity formed by a host and its associated microbiome, functioning as an integrated and coherent unit of evolution (Bordenstein and Theis, 2015; Rosenberg and Zilber-Rosenberg, 2018). Just as the modular integration of organelles gave rise to eukaryotic cells, and the modular organization of cells led to multicellular organisms, the holobiont represents a higher level of modular organization, where the host and its microbiome form a collective organism that is more complex and adaptive (Huitzil et al., 2018; 2023). Its hierarchical organization allows for the emergence of novel properties and functions that are not present in the individual components (Huitzil et al., 2020; Huitzil et al., 2023), enabling holobionts to adapt to diverse environments and respond to challenges more effectively than either the host or the microbiome could alone.

At even larger scales, populations and ecosystems also exhibit modular organization, forming complex networks of interactions where groups of species interact more closely with each other than with other groups, (Pimm, 1991; Sole and Montoya, 2001). Moreover, superorganisms, such as bee and ant colonies, represent a further level of organization into modular structures and functions where groups of individuals specialized in different tasks contribute to the efficiency and adaptability of the colony as a whole (Holldobler and Wilson, 2009).

Multiscale modularity is not only a property observed in the structural organization of biological systems but must also have important implications for their evolution and adaptation. For example, modular organization allows for the evolution of new functions through the modification and recombination of existing modules, without disrupting the entire system, while a hierarchy of modules allows for evolution at multiple levels (Simon, 1962; Kashtan and Alon, 2005; Wagner et al., 2007). This flexibility may have been a key factor in generating the great diversity and complexity of life on Earth. Various models and conceptual foundations have been developed to better understand the evolutionary implications of multiscale modularity, which we briefly describe in the next section.

2 Models and theoretical foundations

The concept of modularity has been explored from various perspectives to understand its role in the organization and evolution of biological systems across multiple scales. One of the most influential contributions in this field is the work by Simon (1962), who introduced the idea of "nearly decomposable systems" described in the introduction. This seminal work laid the foundations for understanding how hierarchical modularity can facilitate the efficient evolution and adaptation of complex systems by reducing the interactions between subsystems. Building upon these ideas, the study of modularity has been approached from different angles, including network theory, evolutionary biology, and systems biology, to unveil the principles governing the emergence and maintenance of modular organization in living systems.

Further advances in the study of modularity have revealed its crucial role in shaping the structure and function of biological networks. For instance, Ravasz et al. (2002) demonstrated that metabolic networks exhibit a hierarchical modular organization, with highly connected modules composed of smaller, less connected modules. This hierarchical structure was shown to be related to the functional classification of metabolic reactions, suggesting that modularity and hierarchy are essential for the efficient functioning of metabolic systems.

The complexity of biological systems and their modular and hierarchical organization have inspired the development of mathematical and computational models that seek to capture fundamental principles underlying these phenomena. These minimal models have been crucial for understanding how modularity and hierarchy can emerge and evolve in complex adaptive systems (Hartwell et al., 1999; Alon, 2007; Solé and Valverde, 2008). Optimization-based models, in particular, have been instrumental in understanding the evolution of modularity (Kashtan and Alon, 2005; Clune et al., 2013; Mengistu et al., 2016). Kashtan and Alon (2005) demonstrated that modularity can evolve in networks when the environment changes in a modular fashion, suggesting that modularity is an adaptive response to certain features of the environment.

Another important line of theoretical research has focused on the evolutionary mechanisms that give rise to modularity in biological systems. Wagner et al. (2007) reviewed the concept of modularity from an evolutionary perspective, discussing how natural selection can favor the emergence of modular architectures. They argued that modularity enhances evolvability by allowing for the independent evolution of different functional modules, thus enabling the exploration of new adaptive solutions.

Network theory has provided a quantitative framework for analyzing modularity based on connectivity patterns. Models such as the "preferential attachment" model by Barabási and Albert (1999) and the evolving modularity model by Valverde and Solé (2007) have helped to understand how modular architectures can emerge in biological networks. These suggest that modularity can arise as a result of selection for both robustness and evolvability.

Collectively, all these minimal models have provided valuable insights into the mechanisms and principles underlying the emergence and evolution of modularity and hierarchy in biological systems. However, many challenges lie ahead, such as integrating these principles into more realistic modeling frameworks that capture the complexity of biological systems at multiple scales and the empirical validation of these theoretical predictions.

In summary, the theoretical foundations for describing the origins and properties of hierarchical modularity in biological systems have been explored from different perspectives, including complex systems theory, evolutionary biology, and network theory. These efforts have revealed the emergence of modularity at multiple scales as a fundamental organizational principle that can confer key evolutionary advantages to biological systems, such as adaptability, robustness, and efficiency.

3 Advantages of modularity

To advance towards a universal theory of the role of modularity in the development of complex life forms, we must first identify the evolutionary advantages (EAs) that this type of structure may provide, regardless of the specific features or scale of the system. By considering various theoretical and experimental realizations of modularity, we propose here a general classification of the key EAs of multiscale modularity into four classes that can be identified in a variety of biological systems. These EAs can be briefly listed as follows:

EA 1 The reuse and recombination of modular components facilitate the evolution of new functions and rapid adaptation of organisms to changing environments (Patthy, 1999; Bashton and Chothia, 2007).

EA 2 Modularity enhances the robustness of biological systems by limiting the propagation of perturbations and allowing for the independent evolution of sub-systems (Wagner et al., 2007; Samal and Jain, 2008).

EA 3 Hierarchical modularity enables the efficient processing and integration of information across multiple scales of biological organization (Barabási et al., 2003; Meunier et al., 2009; Maier et al., 2019).

EA 4 Modularity enables the integration of simpler components into more complex systems, providing a pathway for the evolution of biological complexity, the division of labor, and the emergence of novel functions (Baldwin and Clark, 2000).

These advantages play a crucial role in the emergence of modular organization across multiple scales. By facilitating adaptability, robustness, efficient information processing, and the integration of simple elements into more complex components, modularity allows for the evolution and survival of increasingly complex living systems. This process can develop iteratively, with modules at one level serving as building blocks for higher-level modules, leading to the formation of multiple nested hierarchies of modular structures at larger and larger scales.

4 Biological examples

To illustrate the evolutionary advantages of modularity presented in the previous section, we will briefly describe a series of examples that demonstrate how the key benefits of modularity manifest themselves in concrete biological systems, providing evidence for the central role of modularity in shaping the self-organization of structure and function in living systems.

At the molecular level, the modular architecture of proteins allows for the recombination of functional domains, facilitating the evolution of new functionalities, which corresponds to an advantage of type EA 1. For instance, the shuffling of protein domains through mechanisms such as exon shuffling and gene duplication has been a major driver of protein evolution (Patthy, 1999). This modular organization enables proteins to adapt rapidly to new challenges without the need to evolve entirely new structures from scratch.

Gene regulatory networks provide another example of a type EA 1 benefit of modularity. The lac operon in *E. coli*, for instance, is a modular regulatory system composed of a promoter, an operator, and structural genes that control the expression of enzymes involved in lactose metabolism. This modular structure facilitates the efficient control of gene expressions and has been found to regulate different metabolic processes in other bacterial species, thus showing that it can be reused and adapted to control diverse functions (Browning et al., 2019). Similarly, the eukaryotic cell cycle is regulated by a modular network of interacting proteins (cyclins and cyclin-dependent kinases), with each protein complex forming a functional module that drives a specific phase of the cycle (Schulze-Gahmen et al., 1995). The modular organization of these regulatory networks enables the reuse and recombination of regulatory modules, facilitating the emergence of new functionalities and the adaptation to diverse environmental conditions.

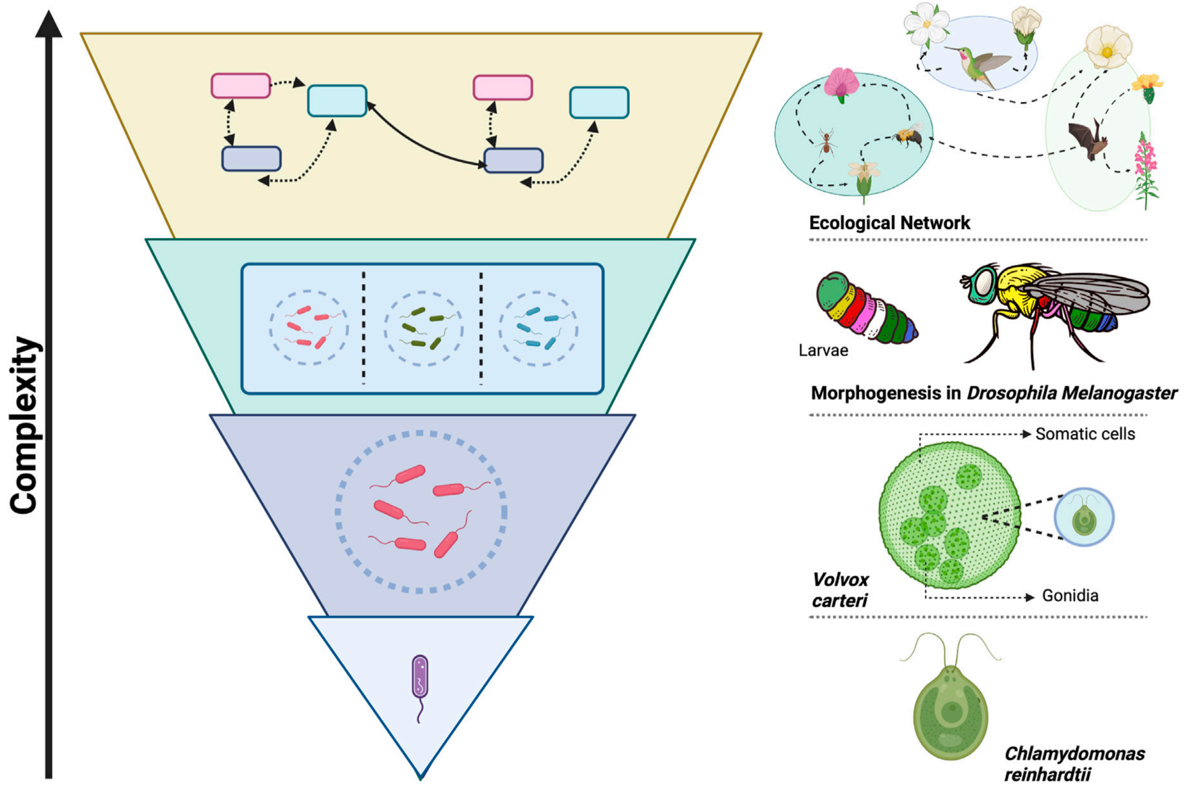


FIGURE 1
Modularity as a Path to Complexity in Biological Systems. The figure illustrates the role of modularity as a universal organizing principle, observed across multiple scales and biological contexts, that enables the evolution of greater complexity. This complexity arises from the integration of interacting modules, which give rise to new functions and emergent properties at each hierarchical level (Wolf et al., 2018). On the left, a schematic diagram shows how biological systems self-organize modularly at different levels, highlighting their hierarchical nature, where each level is composed of modular subsystems that integrate at higher levels. On the right, specific examples demonstrate this principle across various biological contexts and scales. At the unicellular level, *Chlamydomonas reinhardtii* can form colonies like *Volvox carteri*, an organism in transition towards multicellularity. In these colonies, cells organize into modules specialized in reproduction (gonidia) and motility (somatic cells), improving efficiency and division of labor (Herron, 2016). At the multicellular level, modular organization is observed in various processes, such as morphogenesis in *Drosophila melanogaster*, where Hox genes facilitate the formation of specialized modules and complex structures for diverse physiological functions (Hubert and Wellik, 2023). At the ecosystem level, networks of interactions between species also exhibit modularity, with groups of species interacting more closely with each other, contributing to ecosystem stability and resilience (Olesen et al., 2007). This framework provides an integrative perspective for understanding the role of modularity in the evolution of biological complexity. This image was created with BioRender.com

We can also identify the benefits of modularity in the very different context of cognitive processes. In this case, modularity allows the brain to efficiently process complex information by integrating specialized modules that operate in a relatively autonomous manner (Sperber, 2002; Carruthers, 2006), which corresponds to a type EA 3 case. This organization enables the coexistence of functional specialization and integration, as exemplified by language processing, which involves the coordination of multiple specialized modules, such as phonological, syntactic, and semantic processing units (Fodor, 1983; Robbins, 2009). The modular structure of brain networks is hierarchically organized, with smaller, more specialized modules nested within larger, more integrative modules (Meunier et al., 2009). This hierarchical modularity allows for efficient information processing within specialized domains while also enabling the emergence of higher-level cognitive functions through the integration of these modules. It can thus be characterized as conferring not only type EA 3 but also type EA 4 advantages.

In yet a different context, at the ecosystem level, it has been shown that modularity contributes to stability and resilience by compartmentalizing interactions between species, which corresponds to a type EA 2 benefit. In this case, modular ecosystems are characterized by groups of species that interact more strongly within modules than between modules (Olesen et al., 2007). This compartmentalization can limit the spread of perturbations and prevent cascading failures across the entire ecosystem (Stouffer and Bascompte, 2011), thereby enhancing robustness.

Finally, an example of a type EA 4 advantage of modularity can be found in the modular organization of metabolic networks, where the integration of simpler modules allows for the generation of more complex metabolic capabilities. Photosynthesis, for instance, comprises distinct modules, such as light-harvesting complexes and electron transport chains, which integrate to convert light into chemical energy (in the form of ATP and NADPH) (Stirbet et al., 2020). Similarly, the citric acid cycle consists of a modular assembly of enzymatic subunits that form an integrated functional module, which enables the evolution of novel metabolic functions through the recombination of existing modules. In both cases, modularity enables the hierarchical integration of simpler modules into more complex metabolic systems, facilitating the emergence of novel functionalities. For example, photosynthesis can further integrate with other modules (such as the carbon fixation pathway) to enable plants to synthesize glucose from CO2, whereas the citric acid cycle can couple with other metabolic pathways to generate energy and precursors for biosynthesis (Akram, 2014).

The examples presented above illustrate how the key evolutionary advantages of modularity can be identified in biological systems across different scales and levels of complexity, showing that the general properties of biological modularity go beyond the specificities of a given system realization.

5 Discussion

The ubiquity of modular organization across biological scales, from molecular networks to ecosystems, shows the fundamental importance of this organizing principle in the emergence and evolution of complex life forms. As we have shown above, by compartmentalizing biological systems into relatively autonomous, functionally specialized sub-systems, modularity allows for the reuse and recombination of existing modules to support new functions, enhance robustness, enable efficient information processing, and facilitate the evolution of biological complexity.

Understanding modularity as a fundamental principle of organization across scales could unveil its power as a unifying concept, placing it among the few universal principles proposed to explain the remarkable tendency of evolution to generate increasingly complex systems. Figure 1 illustrates this principle, showcasing modularity's role in biological complexity through specific examples at different levels of organization. Another such principle is criticality, which refers to the state of a system at the boundary between order and chaos, where it exhibits a balance between robustness and adaptability (Munoz, 2018). Robustness refers to a system's ability to maintain its functionality while facing perturbations, while adaptability refers to its capacity to adjust to changing conditions (Wagner, 2005; Whitacre, 2012). Notably, modularity and criticality share essential features that enhance robustness and adaptability. For example, modularity contributes to robustness by localizing perturbations within modules, and it supports adaptability by enabling the recombination of evolved modules as a faster way to adjust to new conditions, rather than having to develop entirely new solutions (Kashtan and Alon, 2005; Clune et al., 2013).

This striking convergence of modularity and criticality raises thought-provoking questions: Could these principles be deeply interconnected, representing complementary facets of a more fundamental organizational framework? Might the modular architecture of biological systems facilitate their self-organization towards critical states, thereby unlocking the adaptive advantages associated with criticality (Irani and Alderson, 2023)? The intriguing parallels between modularity and criticality invite us to explore the interplay between these properties, potentially uncovering a more comprehensive understanding of the principles that shape the structure and dynamics of complex biological systems across scales.

Despite the significant progress made in understanding the modular organization of biological systems, many challenges and open questions remain. The development of more advanced computational tools for detecting and analyzing modularity across scales could provide deeper insights into the structure and function of complex biological networks. Furthermore, exploring the interplay between modularity and other organizational principles, such as hierarchy and criticality, could provide novel design principles for engineered systems.

The emerging era of cell engineering harnesses the modularity of cells to program complex biological functions, paving the way for transformative advances in biotechnology and medicine (Lim and Pawson, 2010; Lim, 2022). By unraveling the mechanisms that enable the integration of lower-level modules into increasingly complex hierarchies, we may gain a deeper understanding of the processes that gave rise to the first living organisms and the subsequent evolution of biological complexity (Ruiz-Mirazo et al., 2017).

The perspective that we present here highlights the importance of modularity and hierarchical organization as fundamental principles in the design and function of living systems across multiple scales. By identifying the key evolutionary advantages conferred by modular organization, we provide a unifying lens for understanding the emergence of modular hierarchical structures in biology and the mechanisms underlying the resilience, adaptability, and evolvability of living systems. This knowledge not only improves our fundamental understanding of biology but also provides opportunities for applications in a variety of fields, from bioengineering to the design of complex adaptive systems.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

SH: Conceptualization, Writing–original draft, Writing–review and editing. CH: Conceptualization, Writing–original draft, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This project was made possible through the support of Grant 62213 from the John Templeton Foundation. This funding supported the authors' research and writing of this article.

Conflict of interest

Author CH was employed by CHuepe Labs.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Akram, M. (2014). Citric acid cycle and role of its intermediates in metabolism. *Cell. Biochem. biophysics* 68, 475–478. doi:10.1007/s12013-013-9750-1

Alon, U. (2007). Network motifs: theory and experimental approaches. *Nat. Rev. Genet.* 8, 450–461. doi:10.1038/nrg2102

Baldwin, C. Y., and Clark, K. B. (2000). *Design rules, the power of modularity*. Cambridge: MIT press.

Barabási, A.-L., and Albert, R. (1999). Emergence of scaling in random networks. *science* 286, 509–512. doi:10.1126/science.286.5439.509

Barabási, A.-L., Dezső, Z., Ravasz, E., Yook, S.-H., and Oltvai, Z. (2003). Scale-free and hierarchical structures in complex networks. *AIP Conf. Proc.* 661, 1–16. doi:10.1063/1.1571285

Barabasi, A.-L., and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5, 101–113. doi:10.1038/nrg1272

Bashton, M., and Chothia, C. (2007). The generation of new protein functions by the combination of domains. *Structure* 15, 85–99. doi:10.1016/j.str.2006.11.009

Bonner, J. T. (1988). *The evolution of complexity by means of natural selection*. New Jersey: Princeton University Press.

Bordenstein, S. R., and Theis, K. R. (2015). Host biology in light of the microbiome: ten principles of holobionts and hologenomes. *PLoS Biol.* 13, e1002226. doi:10.1371/journal.pbio.1002226

Browning, D. F., Godfrey, R. E., Richards, K. L., Robinson, C., and Busby, S. J. (2019). Exploitation of the escherichia coli lac operon promoter for controlled recombinant protein production. *Biochem. Soc. Trans.* 47, 755–763. doi:10.1042/BST20190059

Carruthers, P. (2006). *The architecture of the mind*. Oxford: Oxford University Press.

Cheverud, J., Schlosser, G., and Wagner, G. (2004). *Modularity in development and evolution*.

Clune, J., Mouret, J.-B., and Lipson, H. (2013). The evolutionary origins of modularity. *Proc. R. Soc. b Biol. Sci.* 280, 20122863. doi:10.1098/rspb.2012.2863

Felice, R. N., Randau, M., and Goswami, A. (2018). A fly in a tube: macroevolutionary expectations for integrated phenotypes. *Evolution* 72, 2580–2594. doi:10.1111/evo.13608

Fodor, J. A. (1983). *The modularity of mind*. Cambridge: MIT press.

Gray, M. W. (2017). Lynn margulis and the endosymbiont hypothesis: 50 years later. *Mol. Biol. Cell.* 28, 1285–1287. doi:10.1091/mbc.E16-07-0509

Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999). From molecular to modular cell biology. *Nature* 402, C47–C52. doi:10.1038/35011540

Herron, M. D. (2016). *Origins of multicellular complexity: Volvox and the volvocine algae*.

Holldobler, B., and Wilson, E. O. (2009). *The superorganism: the beauty elegance and strangeness of insect societies*. New York City: WW Norton and Company.

Hubert, K. A., and Wellik, D. M. (2023). Hox genes in development and beyond. *Development* 150, dev192476. doi:10.1242/dev.192476

Huitzil, S., Huepe, C., Aldana, M., and Frank, A. (2023). The missing link: how the holobiont concept provides a genetic framework for rapid evolution and the inheritance of acquired characteristics. *Front. Ecol. Evol.* 11, 1279938. doi:10.3389/fevo.2023.1279938

Huitzil, S., Sandoval-Motta, S., Frank, A., and Aldana, M. (2018). Modeling the role of the microbiome in evolution. *Front. physiology* 9, 1836. doi:10.3389/fphys.2018.01836

Huitzil, S., Sandoval-Motta, S., Frank, A., and Aldana, M. (2020). Phenotype heritability in holobionts: an evolutionary model. *Symbiosis Cell. Mol. Med. Evol. Aspects* 69, 199–223. doi:10.1007/978-3-030-51849-3_7

Irani, M., and Alderson, T. H. (2023). Tuning criticality through modularity in biological neural networks. *J. Neurosci.* 43, 5881–5882. doi:10.1523/JNEUROSCI.0865-23.2023

Kadelka, C., Wheeler, M., Veliz-Cuba, A., Murrugarra, D., and Laubenbacher, R. (2023). Modularity of biological systems: a link between structure and function. *J. R. Soc. Interface* 20, 20230505. doi:10.1098/rsif.2023.0505

Kashtan, N., and Alon, U. (2005). Spontaneous evolution of modularity and network motifs. *Proc. Natl. Acad. Sci.* 102, 13773–13778. doi:10.1073/pnas.0503610102

Kirk, D. L. (2005). A twelve-step program for evolving multicellularity and a division of labor. *BioEssays* 27, 299–310. doi:10.1002/bies.20197

Klingenberg, C. P., Mebus, K., and Auffray, J.-C. (2003). Developmental integration in a complex morphological structure: how distinct are the modules in the mouse mandible? *Evol. Dev.* 5, 522–531. doi:10.1046/j.1525-142x.2003.03057.x

Lim, W. A. (2022). The emerging era of cell engineering: harnessing the modularity of cells to program complex biological function. *Science* 378, 848–852. doi:10.1126/science.add9665

Lim, W. A., and Pawson, T. (2010). Phosphotyrosine signaling: evolving a new cellular communication system. *Cell.* 142, 661–667. doi:10.1016/j.cell.2010.08.023

Lobo, I. (2008). Biological complexity and integrative levels of organization. *Nat. Educ.* 1, 141.

Lorenz, D. M., Jeng, A., and Deem, M. W. (2011). The emergence of modularity in biological systems. *Phys. life Rev.* 8, 129–160. doi:10.1016/j.plrev.2011.02.003

Maier, B. F., Huepe, C., and Brockmann, D. (2019). Modular hierarchical and power-law small-world networks bear structural optima for minimal first passage times and cover time. *J. Complex Netw.* 7, 865–895. doi:10.1093/comnet/cnz010

Melo, D., Porto, A., Cheverud, J. M., and Marroig, G. (2016). Modularity: genes, development, and evolution. *Annu. Rev. Ecol. Evol. Syst.* 47, 463–486. doi:10.1146/annurev-ecolsys-121415-032409

Mengistu, H., Huizinga, J., Mouret, J.-B., and Clune, J. (2016). The evolutionary origins of hierarchy. *PLoS Comput. Biol.* 12, e1004829. doi:10.1371/journal.pcbi.1004829

Meunier, D., Lambiotte, R., Fornito, A., Ersche, K., and Bullmore, E. T. (2009). Hierarchical modularity in human brain functional networks. *Front. neuroinformatics* 3, 37. doi:10.3389/neuro.11.037.2009

Munoz, M. A. (2018). Colloquium: criticality and dynamical scaling in living systems. *Rev. Mod. Phys.* 90, 031001. doi:10.1103/revmodphys.90.031001

Olesen, J. M., Bascompte, J., Dupont, Y. L., and Jordano, P. (2007). The modularity of pollination networks. *Proc. Natl. Acad. Sci.* 104, 19891–19896. doi:10.1073/pnas.0706375104

Patthy, L. (1999). Genome evolution and the evolution of exon-shuffling—a review. *Gene* 238, 103–114. doi:10.1016/s0378-1119(99)00228-0

Pimm, S. L. (1991). *The balance of nature? ecological issues in the conservation of species and communities*. USA: University of Chicago Press.

Raff, R. A. (1996). *The shape of life*, 544.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabási, A.-L. (2002). Hierarchical organization of modularity in metabolic networks. *science* 297, 1551–1555. doi:10.1126/science.1073374

Robbins, P. (2009). *Modularity of mind*.

Rosenberg, E., and Zilber-Rosenberg, I. (2018). The hologenome concept of evolution after 10 years. *Microbiome* 6, 78–14. doi:10.1186/s40168-018-0457-9

Ruiz-Mirazo, K., Briones, C., and de la Escosura, A. (2017). Chemical roots of biological evolution: the origins of life as a process of development of autonomous functional systems. *Open Biol.* 7, 170050. doi:10.1098/rsob.170050

Sagan, L. (1967). On the origin of mitosing cells. *J. Theor. Biol.* 14, 255–274. doi:10.1016/0022-5193(67)90079-3

Samal, A., and Jain, S. (2008). The regulatory network of e. coli metabolism as a boolean dynamical system exhibits both homeostasis and flexibility of response. *BMC Syst. Biol.* 2, 21–18. doi:10.1186/1752-0509-2-21

Schaffer, L. V., and Ideker, T. (2021). Mapping the multiscale structure of biological systems. *Cell. Syst.* 12, 622–635. doi:10.1016/j.cels.2021.05.012

Schliwa, M., and van Blerkom, J. (1981). Structural interaction of cytoskeletal components. *J. Cell. Biol.* 90, 222–235. doi:10.1083/jcb.90.1.222

Schulze-Gahmen, U., Brandsen, J., Jones, H. D., Morgan, D. O., Meijer, L., Vesely, J., et al. (1995). Multiple modes of ligand recognition: crystal structures of cyclin-dependent protein kinase 2 in complex with atp and two inhibitors, olomoucine and isopentenyladenine. *Proteins Struct. Funct. Bioinforma.* 22, 378–391. doi:10.1002/prot.340220408

Simon, H. A. (1962). The architecture of complexity. *Proc. Am. philosophical Soc.* 106, 467–482.

Smith, J. M., and Szathmary, E. (1997). *The major transitions in evolution*. Oxford: OUP.

Sole, R. V., and Montoya, M. (2001). Complexity and fragility in ecological networks. *Proc. R. Soc. Lond. Ser. B Biol. Sci.* 268, 2039–2045. doi:10.1098/rspb.2001.1767

Solé, R. V., and Valverde, S. (2008). Spontaneous emergence of modularity in cellular networks. *J. R. Soc. Interface* 5, 129–133. doi:10.1098/rsif.2007.1108

Sperber, D. (2002). in *Defense of massive modularity*. doi:10.7551/mitpress/4108.003.0008

Stirbet, A., Lazár, D., Guo, Y., and Govindjee, G. (2020). Photosynthesis: basics, history and modelling. *Ann. Bot.* 126, 511–537. doi:10.1093/aob/mcz171

Stouffer, D. B., and Bascompte, J. (2011). Compartmentalization increases food-web persistence. *Proc. Natl. Acad. Sci.* 108, 3648–3652. doi:10.1073/pnas.1014353108

Valverde, S., and Solé, R. V. (2007). Self-organization versus hierarchy in open-source social networks. *Phys. Rev. E* 76, 046118. doi:10.1103/PhysRevE.76.046118

Wagner, A. (2005). Robustness, evolvability, and neutrality. *FEBS Lett.* 579, 1772–1778. doi:10.1016/j.febslet.2005.01.063

Wagner, G. P., and Altenberg, L. (1996). Perspective: complex adaptations and the evolution of evolvability. *Evolution* 50, 967–976. doi:10.1111/j.1558-5646.1996.tb02339.x

Wagner, G. P., Pavlicev, M., and Cheverud, J. M. (2007). The road to modularity. *Nat. Rev. Genet.* 8, 921–931. doi:10.1038/nrg2267

Whitacre, J. M. (2012). Biological robustness: paradigms, mechanisms, and systems principles. *Front. Genet.* 3, 67. doi:10.3389/fgene.2012.00067

Wolf, Y. I., Katsnelson, M. I., and Koonin, E. V. (2018). Physical foundations of biological complexity. *Proc. Natl. Acad. Sci.* 115, E8678-E8687–E8687. doi:10.1073/pnas.1807890115

Zelditch, M. L., and Goswami, A. (2021). What does modularity mean? *Evol. Dev.* 23, 377–403. doi:10.1111/ede.12390

TYPE Perspective
PUBLISHED 24 July 2024
DOI 10.3389/fsysb.2024.1394084

OPEN ACCESS

EDITED BY
Edoardo Saccenti,
Wageningen University and Research,
Netherlands

REVIEWED BY
Julia Koblitz,
German Collection of Microorganisms and Cell Cultures GmbH (DSMZ), Germany
Maria Suarez-Diez,
Wageningen University and Research,
Netherlands

*CORRESPONDENCE
John Casey,
✉ casey26@llnl.gov
Ali Navid,
✉ navid1@llnl.gov

RECEIVED 01 March 2024
ACCEPTED 13 May 2024
PUBLISHED 24 July 2024

CITATION
Casey J, Bennion B, D'haeseleer P, Kimbrel J, Marschmann G and Navid A (2024), Transporter annotations are holding up progress in metabolic modeling.
Front. Syst. Biol. 4:1394084.
doi: 10.3389/fsysb.2024.1394084

Transporter annotations are holding up progress in metabolic modeling

John Casey[1]*, Brian Bennion[1], Patrik D'haeseleer[2], Jeffrey Kimbrel[2], Gianna Marschmann[3] and Ali Navid[1]*

[1]Biochemical and Biophysical Systems Group, Lawrence Livermore National Laboratory, Livermore, CA, United States, [2]Systems and Synthetic Biology Group, Lawrence Livermore National Laboratory, Livermore, CA, United States, [3]Earth and Environmental Sciences, Lawrence Berkeley National Laboratory, Berkeley, CA, United States

Mechanistic, constraint-based models of microbial isolates or communities are a staple in the metabolic analysis toolbox, but predictions about microbe-microbe and microbe-environment interactions are only as good as the accuracy of transporter annotations. A number of hurdles stand in the way of comprehensive functional assignments for membrane transporters. These include general or non-specific substrate assignments, ambiguity in the localization, directionality and reversibility of a transporter, and the many-to-many mapping of substrates, transporters and genes. In this perspective, we summarize progress in both experimental and computational approaches used to determine the function of transporters and consider paths forward that integrate both. Investment in accurate, high-throughput functional characterization is needed to train the next-generation of predictive tools toward genome-scale metabolic network reconstructions that better predict phenotypes and interactions. More reliable predictions in this domain will benefit fields ranging from personalized medicine to metabolic engineering to microbial ecology.

KEYWORDS

metabolic modeling, transporter annotation, microbial community modeling, flux balance analysis, functional genomics

1 Introduction

Living systems interact with their surroundings. They acquire resources from their environment; co-operate, steal from, compete against, or kill their neighbors. Molecular compounds are the primary effectors of such interactions and thus the extent of these behaviors depend on the specialized transport proteins that move substances across membrane interfaces, into and out of cellular compartments. Microbes have designed transporters to access an incredible diversity of chemical species, enabling them to harbor pathways that generate cytotoxic byproducts (e.g., photorespiratory phosphoglycolate; Bauwe et al., 2012), to survive in harsh environments (e.g., acid mine drainage; Baker and Banfield, 2003), to harvest scarce resources (e.g., Lake Vostok, buried beneath 4 km of ice; Karl et al., 1999), to communicate with one another (e.g., quorum sensing in *Vibrio*; Hammer and Bassler, 2003), to attack one another (e.g., antibiotic production in soils; Chandra and Kumar, 2017), and to maintain a delicate balance of redox couples (Falkowski et al., 2008). For those interested in mechanistic modeling of such systems, knowing the full repertoire of microbial transport processes is crucial to predicting their dynamics in different habitats. This article describes the origins, state-of-the-art, challenges and

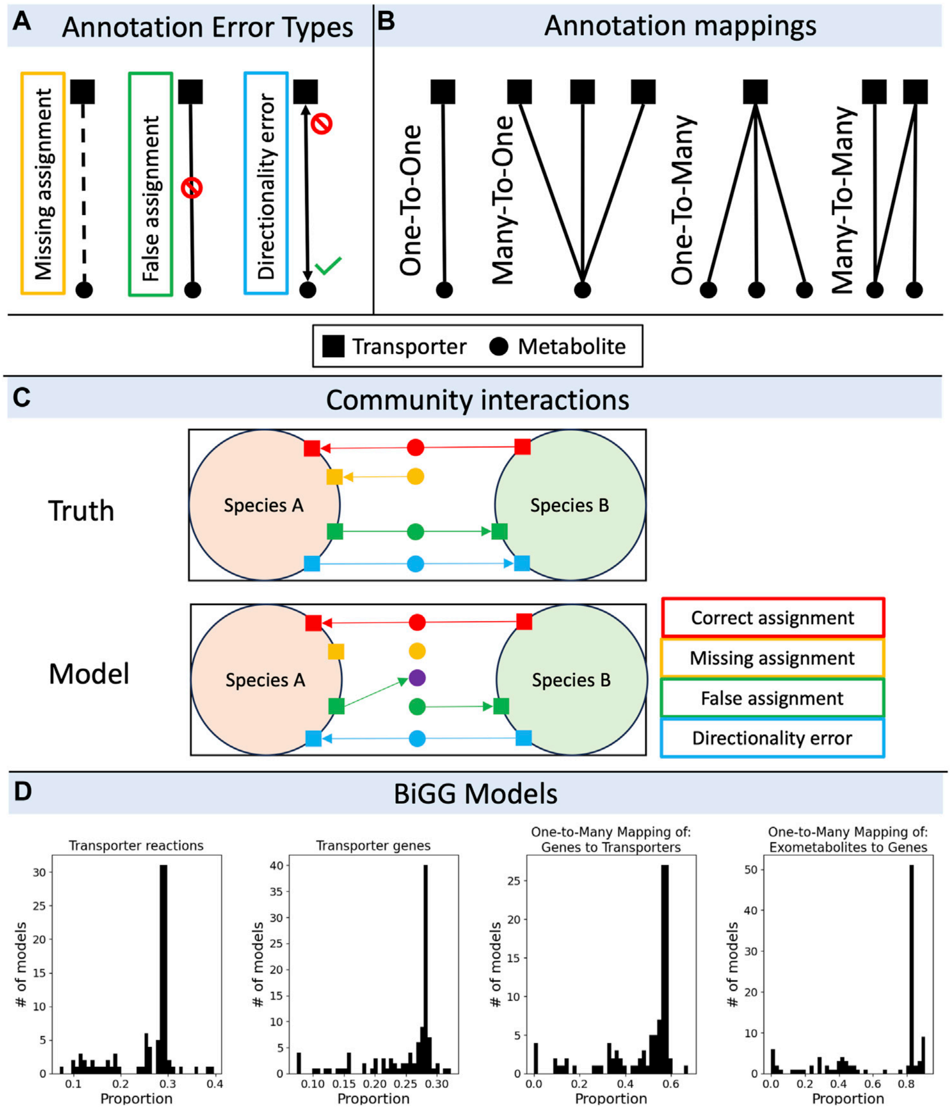


FIGURE 1
The pitfalls of transporter annotations in community metabolic modeling. **(A)** Types of errors encountered when assigning a single putative transporter to a single substrate. An annotation may miss an assignment where there should be one, may create an assignment where there should not, or may get the direction(s) of transport wrong (either due to an incorrect orientation of an irreversible process, or due to a reversibility error). **(B)** Mappings from transporter genes to substrates are non-unique. A single gene may map to a single substrate or multiple substrates, a single gene may be a part of a complex with multiple genes which map to a single substrate or multiple substrates. **(C)** Microbial interactions are variously affected by transporter annotation errors. For example, a species might not grow with missing assignment errors, the community might accumulate or deplete extracellular metabolites by false assignment errors, or a mutualism might be broken by directionality errors. **(D)** Analysis of transport mappings in BiGG models (n = 108 models). Histograms showing the proportion of transporter reactions to total reactions (left), the proportion of transporter genes to total genes (second from left), the proportion of one-to-many gene-to-transporter mappings to total transporter genes (second from right), and the proportion of one-to-many exometabolite-to-transporter gene mappings to total exometabolites (right). The large peaks correspond, mostly, to models of *Escherichia coli*, which are overrepresented in the BiGG database.

future prospects of transporter functional annotation that we hope will serve as a "call to arms" for doubling efforts in both computational and experimental approaches.

Mechanistic, constraint-based modeling in systems biology has benefitted immensely from standardization of the model reconstruction process (Thiele and Palsson, 2010; Heirendt et al., 2019), testing and reporting the quality of models (MEMOTE; Lieven et al., 2020), consolidation of new algorithms and software into just a few dominant software platforms (overwhelmingly COBRA; Ebrahim et al., 2013; Heirendt et al., 2019), and sharing in just a few dominant formats (overwhelmingly SBML; Keating et al., 2020). That coordination has paved the way for an ever-growing and active community of software developers, engineers, systems biologists and computational biologists working to relax many of the rigid assumptions of the first generation of flux balanced models (Varma and Palsson, 1994). While the software and protocols are fairly thorough, there are several aspects of model reconstruction that are a bit flimsy, including what to do about polymers, quinones, and, as we discuss in detail here, transporters. Some authors may take the effort to report what those decisions were and why they were made, but there is certainly space for our community to weigh in on these persistent concerns.

The accuracy of genome-scale metabolic model (GEM) predictions are strongly correlated to the quality and completeness of the metabolic network reconstructions (Bernstein et al., 2023). The availability of transport mechanisms for import of nutrients greatly influences choice of gap-filled reactions in both automatically generated and curated models. This issue is further complicated by the "moonlighting" nature of some proteins (Jeffrey, 2018) where under different conditions they assume different functional roles. Many proteins also exhibit weak promiscuous activities for a variety of metabolites which leads to an "underground metabolism" that plays a major role in the fitness of organisms (Noterbaart et al., 2018). Not accurately accounting for presence of some imported metabolites will lead to exclusion of these reactions from the final network reconstruction and could lead to errors in assessing the robustness of a system to various types of perturbation. In previous work we have shown that functional annotation tools generate metabolic annotations that are incomplete and inconsistent with each other, and that the same is true for transporter annotations, with typically less than half the transporter annotation tools having substrate predictions that are sufficiently detailed to be incorporated in a metabolic model (Griesemer et al., 2018).

2 Discussion

2.1 Transporter annotations: what could go wrong?

Pitfalls in matching transporters to their substrates come in a variety of flavors. We define three elemental error types—missing assignments, false assignments, and directionality errors (Figure 1A). There may be a fourth, somewhat more esoteric error type not included in the figure that applies to the case of a transporter that modifies a substrate during import (e.g., the phosphotransferase complex). These are likely rare and we have not encountered one, but an error in the annotation of the substrate modification or choice of cofactor (e.g., symporters) could conceivably occur. The frequency of different error types is likely variable for different species and for different annotation tools, but for some approximate context we quantified these errors in the model organism *E. coli* K12 MG1655, comparing an extensively curated GEM (iML1515; Monk et al., 2017) against an automatically generated GEM for the same genome using CarveMe (v1.5.2; Machado et al., 2018). Although transporter annotations in iML1515 may be updated in the future, we consider it a high-quality benchmark for evaluating error rates in automatically generated GEMs. In the CarveMe draft model, missing assignments accounted for 8.9%, false assignments accounted for 16.2%, and directionality errors accounted for 4.5% of the total transport reactions. Thus, nearly a third of annotated transporter functions were in error; because this strain is massively overrepresented in the BiGG database (King et al., 2016) that CarveMe references, we should treat these error rates as an underestimate of the error rate expected for non-model organisms using the same method. Griesemer and others showed that genome coverage by metabolic annotation tools, and discrepancies in annotation across different tools are significantly worse for organisms that are more phylogenetically distant from well-studied model organisms such as *E. coli* and *B. subtilis*, and we expect the same to be true for transporter annotations (Griesemer et al., 2018).

Each error type applies in GEMs to four types of gene-protein-reaction (GPR) mappings—one-to-one, one-to-many, many-to-one, and many-to-many (Figure 1B). Non-unique mappings between transporter genes, transporter proteins, and substrates arise from the possibility that individual transporters have more (one-to-one) or less (one-to-many) specificity in binding or selective permeability, and that individual substrates may bind or pass through one (one-to-one) or more (many-to-one) transporters. An analysis of all manually curated models in the BiGG database (King et al., 2016) revealed a wide range of unique mapping frequencies, with 36% ± 29% (range 0%–91%) of exometabolites mapping uniquely to a single transporter gene (n = 108 models; Figure 1D). As an added layer of complexity, gene products may be associated with more than one transporter complex (e.g., the GLUT1 subunit is present in multiple sugar transporters), which themselves may have broad substrate specificity (many-to-many) or serve as a common structural protein for various transporters. As we explore sources for the different error types and how those errors propagate through non-unique mappings in more detail (Figure 1C), it is worth reviewing the current state-of-the-art in automated functional transporter annotation tools and the databases they reference to address these pitfalls.

2.2 Transporter annotation tools and databases

Besides the major sequence repositories, there are currently two primary online database resources dedicated to transporters, and several more niche databases which focus on specific taxonomic groups or transporter types (Table 1). With two decades of development and curation, the Transporter Classification Database (TCDB; Saier, 2006; Saier et al., 2009; Saier et al., 2014; Saier et al., 2016; Saier et al., 2021) remains a central clearinghouse for transporter structures, bioinformatics tools, and is the official home of the Transporter Classification (TC) system ontology, a scheme based on mechanism, energy source, taxonomy and substrate. Since 2001, the International Union of Biochemistry and Molecular Biology (IUBMB) has designated the TC system as the formally recognized ontology for membrane transporters across all domains of life (Busch and Saier, 2003). Each entry in TCDB is manually curated and often accompanied by a detailed summary of the literature, and is maintained by a well-known authority on transporters. Surprisingly, Kroll and others reported that more than half of TCDB entries scored poorly (2 or below, on a scale from 1 to 5) on the UniProt annotation scale, and instead opted to rely on GO and UniProt entries (only those with a score of 5; Kroll et al., 2023). TransportDB (now in version 2.0; Elbourne et al., 2017) is another popular resource for systems biologists which builds on the TCDB and NCBI datasets, with entries currently available for 2,761 organisms (predominantly bacteria, though there are some eukaryotes and archaea) through a graphical and convenient web-portal. Entries in TransportDB are computationally derived with their accompanying annotation tool called TransAAP.

TABLE 1 Databases dedicated to transporters. NA, URL not maintained.

Database	Description	URL	Reference
ABCdb	Prokaryotic ATP binding cassettes. Curated and computational partitions	www-abcdb.biotoul.fr/	Fichant et al. (2006)
ARAMEMNON	Plant membrane proteins. Computational	aramemnon.botanik.uni-koeln.de/	Schwacke et al. (2003), Schwacke and Flügge (2018)
TCDB	All transporters. Curated	www.tcdb.org/	Saier (2006), Saier et al. (2009), Saier et al. (2014), Saier et al. (2016), Saier et al. (2021)
YTPdb	Yeast membrane proteins. Curated	NA	Brohée et al. (2010)
TransportDB 2.0	All transporters. Computational	http://www.membranetransport.org	Elbourne et al. (2017)

TABLE 2 Annotation tools dedicated to transporters. Note that some portals appear to no longer be maintained (NA), while others have changed URLs since publication.

Name	Notes	URL	Reference
TransAAP	Integrated with TransportDB	www.membranetransport.org/	Elbourne et al. (2023)
TIP	Integrated with PathwayTools; parses existing text-based annotations	bioinformatics.ai.sri.com/ptools/	Lee et al. (2008), Karp et al. (2020)
TrSSP	Standalone, SVM annotation	www.zhaolab.org/TrSSP/	Mishra et al. (2014)
TRIAGE	Formerly the annotation tool for Merlin	NA	Dias et al. (2017)
TransATH	Standalone, automated pipeline based on Saier's protocol	NA	Aplop and Butler (2017)
TranCEP	Standalone, combined homology and SVM annotation	github.com/bioinformatics-group/TranCEP	Alballa et al. (2020)
TranSyt	Successor to TRIAGE, standalone and integrated with Merlin, KBASE	transyt.bio.di.uminho.pt/	Cunha et al. (2023)
TransportTP	Standalone, combined homology and SVM annotation	NA	Li et al. (2009)
PortPred	Standalone. Combined DL-based protein embeddings and ML classification	github.com/MarcoAnteghini/PortPred	Anteghini et al. (2023)
SPOT	Standalone. DL using Transformer Networks for classification of transporter-substrate vector pairs	github.com/AlexanderKroll/SPOT	Kroll et al. (2023)

A chronology of transporter annotation tools, their various approaches, and a summary of their performance is available elsewhere (Alballa et al., 2020; Cunha et al., 2023), and we simply provide a convenient lookup table with short descriptions and URLs for reference (Table 2). Recently, the TranSyT tool (Cunha et al., 2023) has emerged as a front-runner alongside TransAAP. In the spirit of integration and ease of use, TranSyT can be implemented as a standalone app to generate a SBML file of transport reactions, or within popular automated GEM reconstruction pipelines like Merlin (Capela et al., 2022) and the ModelSEED reconstruction tools in KBase (Faria et al., 2023). TranSyT also scores annotations, a feature which may be leveraged for merging multiple annotation sources (Henry et al., 2010; Greisemer et al., 2018) or for generating ensemble GEM reconstructions.

2.3 Modeling microbial community interactions

Genome scale models have been used in simulating microbial interactions for nearly two decades (reviewed by Heinken et al., 2021), and numerous algorithms have tackled the problem from different angles (reviewed by Biggs et al., 2015; Bauer and Thiele, 2018; Deiner and Gibbons, 2023; Scott et al., 2023). The architecture of community models, whether they ought to be compartmentalized or pooled into a "super-organism," and whether one should attempt to sample the combinatorial interactions with flux balance analysis or to isolate the elementary modes of exchanges was pondered early on (Taffs et al., 2009; Perez-Garcia et al., 2016). Common to most of the more recent attempts is a compartmentalized approach with either stationary or dynamic flux balance analysis, wherein each strain-specific model interacts through an extracellular "compartment" through the exchange of metabolites. Intuitively (and formally; Klitgord and Segre, 2010), the compartmentalization of pathways, or parts of pathways, or of entire metabolic networks strongly influences predicted flux distributions and interactions. For example, a non-compartmentalized model might regenerate ATP from ADP in the absence of a proton motive force. Thus, an accurate accounting of which substrates, which products, and which reactions are where is vital to constraining fluxes and identifying modes of species-species interactions within a community.

Automated reconstruction of draft GEMs has improved considerably over the past decade (Machado et al., 2018; Wang et al., 2018; Heirendt et al., 2019; Faria et al., 2023), making great strides in closing the gap with curated models from genome information alone, but a recent analysis of automated and non-gapfilled draft GEMs showed dismal performance in predicting substrate utilization (Gralka et al., 2023). While there is still no substitute for manual curation by a skilled hand, draft GEM quality could be markedly improved through more comprehensive transporter annotations (Zuniga et al., 2021). Expansion from monoculture simulations to more complex communities likely amplifies these errors, resulting in poor agreement between predicted and actual growth rates in a gut community using three of the latest community FBA algorithms (Pearson correlations of 0.07, at best; Joseph et al., 2024). Special attention to microbial interactions (Sung et al., 2017) was given in the AGORA bacteria reconstructions (Magnúsdóttir et al., 2017; Heinken et al., 2023) and for the human host (the number of extracellular transport reactions ballooned from 537 in Recon1 to 1,537 in Recon2; Sahoo et al., 2014), but clearly there is room for more accurate and comprehensive representation of transport processes to improve growth and interaction predictions.

2.4 Challenges for transporter annotation databases and tools

Guiding principles from the larger systems biology community of shared access, integration and formatting, consistent with the FAIR principles (Barker et al., 2022), should be adopted when building relational databases and the tools that draw from them. This includes providing persistent link identifiers for genes, proteins, and substrates to common resources (e.g., NCBI, PubChem, BRENDA, RHEA) wherever possible, providing documented API's for user access, adhering to community standard formats like SBML and JSON, in the case of tools, working with other developers to integrate with community standard reconstruction pipelines like COBRA and KBase. As we look to the next-generation of transporter annotation tools, especially those that build from emerging methods in machine learning and artificial intelligence, databases that prioritize these principles will be more readily accessed and leveraged.

Database and tool developers should also seek to provide, wherever possible, a minimal set of functional attributes of transporter gene annotations required for GEM reconstruction. We have identified five such attributes: membrane localization, membrane orientation (inward vs. outward facing), binding reversibility, substrate specificity, and reaction stoichiometry. We will discuss the current approaches and challenges in assigning these attributes.

2.4.1 Membrane localization

With the exception of a few exceptionally well-studied model organisms, protein localization across an entire proteome, or even a substantial portion, is typically unknown *a priori*. A number of predictive tools are based on homology to manually curated databases of proteins of known localization (e.g., PSORT; Yu et al., 2010) or based on identification of transmembrane domains and their orientation (e.g., TMPred; Cuthbertson et al., 2005). Today, 77 protein subcellular localization prediction tools are now listed in bio.tools (reviewed in Li et al., 2023), with the newest generation (e.g., TmAlphaFold; Dobson et al., 2023) taking advantage of recent advances in structural prediction. Several are tailored to specific model organisms, while others draw from a broader taxonomic resolution. In the absence of sanity-checks for each compartmentalized reaction during the reconstruction process for a particular species, and given the importance of assigning transporters to the correct membrane, it may be wise to consider a consensus localization (e.g., COMPARTMENTS; Binder et al., 2014) from a collection of the most relevant sorting tools and other sources.

2.4.2 Transporter orientation and reversibility

Secondary-active transporters like ion symporters and antiporters are typically reversible, but are often practically irreversible under physiological conditions. However, a famous counter-example is the oxygen-dependent transport of glutamate into and out of nerve cells (Szatkowski et al., 1990). Even in this non-canonical case, forward and reverse kinetics may be radically different for inward- and outward-facing protein orientations (Zhang et al., 2007). Primary-active transporters are, to our knowledge, strictly irreversible. Because of its functional classification scheme, annotation to the TC ontology should cover all but the most egregious cases of reversibility.

2.4.3 Substrate specificity

Because assigning substrates to transporters is the crux of the matter, we conducted an analysis of TransportDB 2.0 (Elbourne et al., 2017), the most extensive database of transporter annotations currently available. The dataset comprised 2,661 unique substrate names associated with 940,581 substrate-transporter pairs, distributed among 2,745 organisms. Substrates link identifiers were unavailable, and a single substrate often appeared with multiple names (e.g., "sodium ion" vs. "Na+"), making an estimate of the true number of unique substrate-transporter pairs difficult. For a subset of the unique substrate names (for practical reasons, those which appeared in more than 8 organisms), we manually assigned substrates into four categories: known (e.g., "Oxalate"), putative (containing a "?"; e.g., "Oxalate?"), ambiguous ("a carboxylic acid"), and unknown (e.g., "metabolite"). From this categorization across all organisms, we found that 52% ± 9% were known, 9% ± 4% were putative, 31% ± 8% were ambiguous, and 9% ± 6% were unknown (Figure 2A). Although the full 5-level TC system ontology terms are returned with TransAAP, the datasets available through TransportDB 2.0 contain only the first three levels (194 unique terms). From this coarse resolution, we found that only 5 ontology terms represented a majority (66% ± 9%) of all transporter annotations across all organisms, with a single term (3.A.1; ATP binding cassettes) representing nearly half (45% ± 11%; Figure 2B).

A single transporter may have similar affinity for multiple compounds, or even entire classes of compounds. This means that in some cases, a transporter might be annotated to an ambiguous level of substrate specificity (e.g., "a dicarboxylate") not because of a lack of knowledge of the appropriate dicarboxylate molecule it transports (annotation is a missing one-to-one mapping), but rather because it has broad specificity for multiple dicarboxylate molecules (annotation is truly a one-to-many mapping); perhaps even with comparable kinetic properties. Modest changes of just one or two residues in transporter binding domains

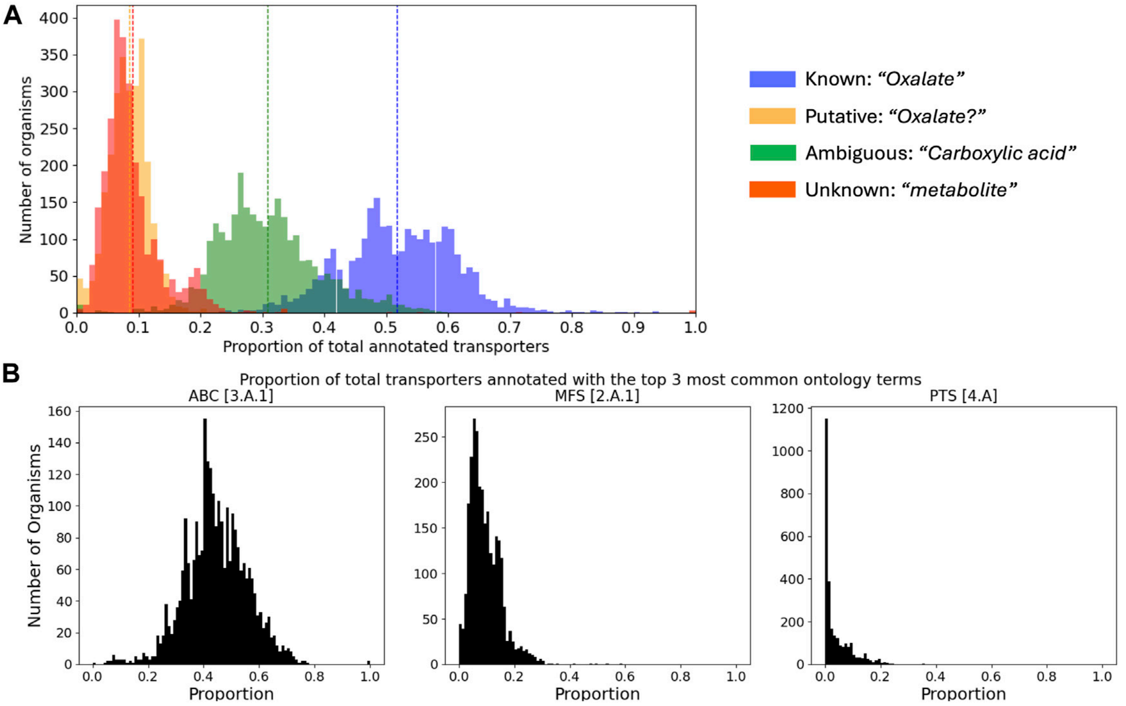


FIGURE 2
Summary of transporter annotations retrieved from TransportDB 2.0. **(A)** Distributions of the proportion of transporters annotated to different levels of specificity across all organisms. Vertical dashed lines correspond to the mean of each distribution, and an example of each category is provided. **(B)** Distributions of the proportion of transporters of the top 3 most abundant [super-] families across all organisms. ABC–ATP binding cassette; MFS–major facilitator superfamily; PTS–phosphotransfer-driven group translocators.

can affect substrate specificity and even stoichiometry, as is the case for the cation/proton antiporters (Masrati et al., 2018), so degeneracy in substrate specificity might be unfortunately necessary.

2.5 The trouble with diffusion

Although the selective permeability of membrane lipids with different lipid compositions have been described in great detail (Hannesschlaeger et al., 2019), diffusion reactions beyond the gasses and a few waste products are rarely included in GEM reconstructions. This may partly be due to the arbitrary nature of delineating the broad spectrum of diffusion rates, from fast (order 10^{-2} m^2 s^{-1}; e.g., oxygen) to slow (10^{-10} m^2 s^{-1}; e.g., high molecular weight polar compounds) diffusing molecules. In general, phosphorylated metabolites might be considered slow, eliminating a sizable portion of the total intracellular metabolites, but the line becomes blurry when considering small nonpolar metabolites like fatty acids, alkanes or alcohols. To make matters worse, the decision to include a diffusive reaction for a metabolite which is also actively transported would result in an underestimate of energy costs in standard FBA. In addition to specificity in transmembrane permeability, diffusive transport across other intracellular compartments, like the shell proteins of cyanobacterial carboxysomes which show preference for negatively charged ions (Mahinthichaichan et al., 2018), should be represented. Knowledge of the localization of pathways, or parts of pathways within, can aid in filtering the list of candidate diffusive reactions into and out of subcellular compartments, but this area is ripe for progress.

2.6 Prospects for computational approaches to transporter functional annotation

The state-of-the-art in transporter annotation brings together sequence alignment, systems biology ontologies, and structure analysis to make predictions about whether a gene product is a transporter, where it might be located, its orientation, and what substrates it might bind. Nevertheless, we find that many transporters lack sufficient coverage in one or more of the required attributes. A leap forward will address gene-protein-reaction specificity first.

We propose a concept for a computational pipeline built on existing tools to progressively narrow the search space of potential transporter-substrate binding pairs. By limiting the number of candidate substrates for each predicted transporter structure, one can devise a strategy to limit compute resources and alleviate some of the scalability problem for downstream experimental validation. The pipeline (Figure 3), makes parallel use of bioinformatics, systems biology tools and molecular dynamics simulations to generate a short-list of substrates with relatively high predicted ligand binding affinities. The workflow begins with homology search against the TCDB to annotate genes to the lowest level of ontology, given some threshold alignment. Although the TC System is not phylogenetically structured *per se*, an analogous approach to "Lowest Common Ancestor" (e.g., MEGAN; Huson et al., 2007) could be used to assign ontology terms at a threshold confidence level. In this scheme, a gene with close sequence similarity to a transporter gene in the TCDB is annotated to level 5 (e.g., 2.A.1.1.1), whereas another with weaker alignment is annotated to level 3 (e.g., 2.A.1). Structuring the depth of annotation is a conservative strategy to

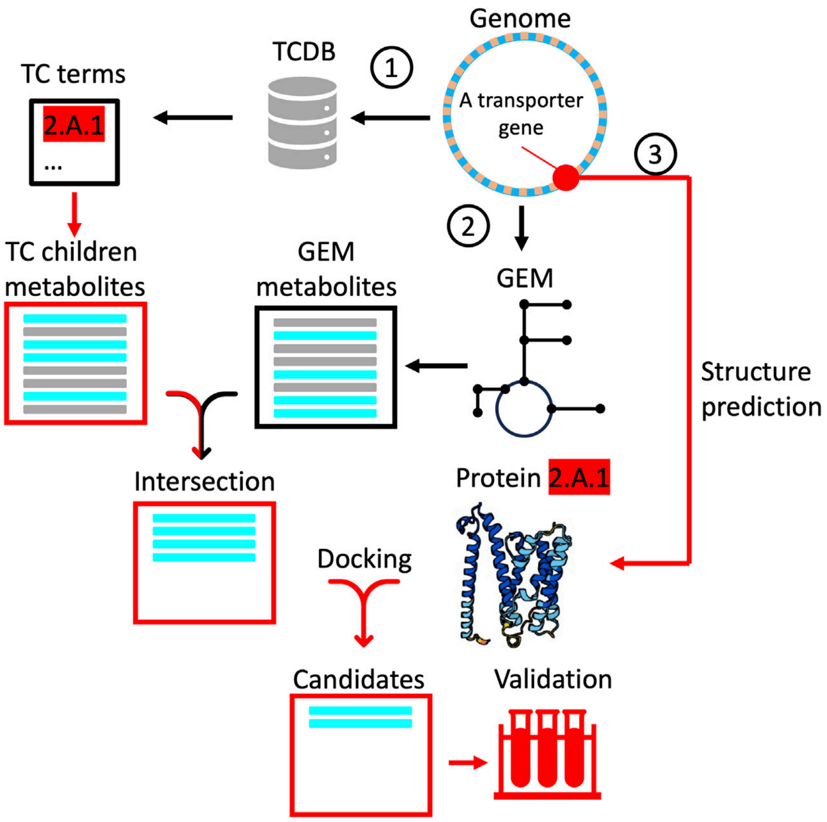


FIGURE 3
A proposed computational workflow to progressively narrow the search space for experimental validation of transporter functional annotations. Red lines correspond to paths followed for a single transporter and are repeated for all un-annotated transporters, while black lines correspond to paths taken (once) for the whole genome. The pipeline begins (1) with alignment of transporter genes to the TCDB, retrieving a list (horizonal bars) of all children metabolites associated with the lowest common ancestor ontology term. In another path (2), a draft GEM is reconstructed to generate a list of all intracellular metabolites synthesized or degraded in the metabolic network. The intersection of both lists (cyan bars) is passed to a third path (3) as candidates for docking simulations using the predicted protein structure. Predicted binding affinities that exceed some threshold are finally passed as candidates for experimental validation.

generate a list of children substrates that the query structure could possibly transport (i.e., all substrates beneath 2.A.1). In a parallel step, a draft GEM is reconstructed, returning the full set of intracellular metabolites. By taking the intersection of these two lists, we pare down the candidate substrates to only those which the organism could conceivably take up or secrete. More stringent approaches exist at this step, including an analysis of uptake and secretion potential given the free exchange of all intracellular metabolites across the system boundary using flux variability analysis (Gudmundsson and Thiele, 2010), but the concept remains the same. Finally, from the intersection set, predictions of ligand binding affinity are used to generate a ranking of candidates. This step takes advantage of advances in structure prediction (e.g., AlphaFold; Jumper et al., 2021; RoseTTAFold; Baek et al., 2021), binding site inference, docking and molecular dynamics simulations (e.g., Ohnuki et al., 2023). One approach here is to infer transporter binding sites from homologus ligands and their cognate binding pockets already in the PDB databank (PDBspheres; Zemla et al., 2022). Fusion Docking-ML calculation can then be performed to determine the most favorable ligand poses in the transporter (Jones et al., 2021). If increased fidelity is desired, various versions of molecular dynamics simulations can be performed to qualitatively and/or quantitatively predict favorable dynamical protein-ligand interactions and associated binding constants (Sohraby and Nunes-Alves, 2023). This approach benefits from high throughput, with each simulation taking approximately 0.01 s/ligand (Zhang et al., 2014), but may suffer from the lack of sensitivity for low molecular weight ligands (less than 4 carbons) and metals, although progress is being made (c.f., zinc; Wang, 2023). An exciting development in this area is quantum docking simulations (Heifetz, 2020), which would, in principle, allow quantitation of binding affinities for these small molecules. The drawback with this quantum docking is throughput, with simulations taking on the order of minutes to hours depending on the size of the binding pocket, each. At this stage, depending on one's objectives and the resources available, one might either submit the best candidates for experimental validation or simply apply a threshold affinity for annotation.

2.7 Prospects for transporter functional genomics

With the advent of reliable protein structure prediction tools such as AlphaFold (Jumper et al., 2021), we will likely see many of our current sequence-to-function annotation tools replaced by a whole new generation of sequence-to-structure-to-function tools over the next decade, both for enzyme annotation and for substrate-specific transporter annotation. However, the availability of large-scale substrate specificity data to train such tools will likely continue to be a bottleneck. While computational methods can pare down the search space of transporter-ligand binding candidates, evidence for transporter annotations should come from experimental validation, preferably *in vivo* (David et al., 2019). Recent advances in laboratory automation and mass spectrometry are dramatically increasing the throughput of functional and phenotypic screening (Coutant et al., 2019), and there is potential for functional genomics guided by mechanistic models. For instance, dynamic FBA can be used to identify target genes to generate smaller, metabolic process-specific deletion libraries for subsequent phenotyping (Brunnsåker et al., 2023). To our knowledge, these approaches have not yet been applied to transporters but could be easily adapted using Biolog-like screens (Bochner et al., 2001) or exometabolomics (Jenkins Sánchez et al., 2022). One high-throughput approach involves the use of a substrate-selective riboswitch as biosensors (Genee et al., 2016). When expressed along with metagenomic DNA fragments, transformants could be screened for their ability to grow on the substrate, and in so doing, the authors could assign function to uncharacterized transporters and identified numerous transporter annotations in error for multiple substrates. Another exciting recent development is Boundary Flux Analysis (reviewed in Lewis, 2024), a method to link changes in metabolite concentrations in growth media to constraints on uptake or secretion rates in GEMs. This approach appears scalable and holds great promise for screening deletion libraries.

3 Conclusion

Errors in transporter annotation arise from a variety of sources, most often resulting in missing or false assignments to substrates. Because of the non-unique mapping of genes to transporters to substrates, these errors metastasize, contributing to horrendous performance in the genotype-phenotype mapping of automated GEM reconstructions based on genome annotation alone. Mischaracterization of species-environment interactions is compounded when inferring microbial interactions in community models, leading to further expansion of spurious and false interaction predictions, and therefore poor fidelity to observations. To complement the progress enjoyed by other aspects of GEM reconstruction, we need to pursue new computational and experimental approaches to the transporter annotation problem. We offer a strawman workflow combining hierarchical ontology filtering with molecular dynamics simulations, and look to emerging high-throughput screening methods to validate predictions. Until the larger systems biology community and sponsors prioritize this challenge, we can continue to expect diminishing returns on advances in microbiome modeling.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: BIGG Models (http://bigg.ucsd.edu/) and Transport DB2 (http://www.membranetransport.org/).

Author contributions

JC: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Visualization, Writing–original draft, Writing–review and editing. BB: Writing–review and editing. PD'h: Writing–review and editing. JK: Writing–review and editing. GM: Writing–review and editing. AN: Conceptualization, Funding acquisition, Project administration, Supervision, Writing–original draft.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was funded as part of the LLNL μBiospheres Scientific Focus Area by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomic Sciences program under FWP SCW1039.

Acknowledgments

We appreciate helpful feedback from Peter Karp, Emanuel Cunha, Oscar Dias, Liam Elbourne, Bruno Andre, and Almut Heinken. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344, and by Lawrence Berkeley National Laboratory under Contract DE-AC02-05CH11231. LLNL-JRNL-860975.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Alballa, M., Aplop, F., and Butler, G. (2020). TranCEP: predicting the substrate class of transmembrane transport proteins using compositional, evolutionary, and positional information. *PLoS ONE* 15, e0227683. doi:10.1371/journal.pone.0227683

Anteghini, M., Santos, V. A. M. D., and Saccenti, E. (2023). PortPred: exploiting deep learning embeddings of amino acid sequences for the identification of transporter proteins and their substrates. *J. Cell. Biochem.* 124, 1803–1824. doi:10.1002/jcb.30490

Aplop, F., and Butler, G. (2017). TransATH: transporter prediction via annotation transfer by homology. *ARPN J. Eng. Appl. Sci.* 12, 317–324.

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 373, 871–876.

Baker, B. J., and Banfield, J. F. (2003). Microbial communities in acid mine drainage. *FEMS Microbiol. Ecol.* 44, 139–152. doi:10.1016/S0168-6496(03)00028-X

Barker, M., Chue Hong, N. P., Katz, D. S., Lamprecht, A.-L., Martinez-Ortiz, C., Psomopoulos, F., et al. (2022). Introducing the FAIR Principles for research software. *Sci. Data* 9, 622. doi:10.1038/s41597-022-01710-x

Bauer, E., and Thiele, I. (2018). From metagenomic data to personalized *in silico* microbiotas: predicting dietary supplements for Crohn's disease. *npj Syst. Biol. Appl.* 4, 27. doi:10.1038/s41540-018-0063-2

Bauwe, H., Hagemann, M., Kern, R., and Timm, S. (2012). Photorespiration has a dual origin and manifold links to central metabolism. *Curr. Opin. Plant Biol.* 15, 269–275. doi:10.1016/j.pbi.2012.01.008

Bernstein, D. B., Akkas, B., Price, M. N., and Arkin, A. P. (2023). Evaluating *E. coli* genome-scale metabolic model accuracy with high-throughput mutant fitness data. *Mol. Syst. Biol.* 19, e11566. doi:10.15252/msb.202311566

Biggs, M. B., Medlock, G. L., Kolling, G. L., and Papin, J. A. (2015). Metabolic network modeling of microbial communities. *WIREs Syst. Biol. Mech. Dis.* 7, 317–334. doi:10.1002/wsbm.1308

Binder, J. X., Pletscher-Frankild, S., Tsafou, K., Stolte, C., O'Donoghue, S. I., Schneider, R., et al. (2014). COMPARTMENTS: unification and visualization of protein subcellular localization evidence. *Database* 2014, bau012. doi:10.1093/database/bau012

Bochner, B. R., Gadzinski, P., and Panomitros, E. (2001). Phenotype MicroArrays for high-throughput phenotypic testing and assay of gene function. *Genome Res.* 11, 1246–1255. doi:10.1101/gr.186501

Brohée, S., Barriot, R., Moreau, Y., and André, B. (2010). YTPdb: a wiki database of yeast membrane transporters. *Biochimica Biophysica Acta (BBA) - Biomembr.* 1798, 1908–1912. doi:10.1016/j.bbamem.2010.06.008

Brunnsåker, D., Reder, G. K., Soni, N. K., Savolainen, O. I., Gower, A. H., Tiukova, I. A., et al. (2023). High-throughput metabolomics for the design and validation of a diauxic shift model. *Syst. Biol. Appl.* 9, 11. doi:10.1038/s41540-023-00274-9

Busch, W., Saier, M. H., and International Union of Biochemistry and Molecular Biology IUBMB (2004). The IUBMB-endorsed transporter classification system. *Mol. Biotechnol.* 27, 253–262. doi:10.1385/mb:27:3:253

Capela, J., Lagoa, D., Rodrigues, R., Cunha, E., Cruz, F., Barbosa, A., et al. (2022). merlin, an improved framework for the reconstruction of high-quality genome-scale metabolic models. *Nucleic Acids Res.* 50, 6052–6066. doi:10.1093/nar/gkac459

Chandra, N., and Kumar, S. (2017). "Antibiotics producing soil microorganisms," in *Antibiotics and antibiotics resistance genes in soils: toxicity, risk assessment and management* (Berlin, Germany: Springer International Publishing), 1–18.

Coutant, A., Roper, K., Trejo-Banos, D., Bouthinon, D., Carpenter, M., Grzebyta, J., et al. (2019). Closed-loop cycles of experiment design, execution, and learning accelerate systems biology model development in yeast. *Proc. Natl. Acad. Sci. U. S. A.* 116, 18142–18147. doi:10.1073/pnas.1900548116

Cunha, E., Lagao, D., Faria, J. P., Liu, F., Henry, C. S., and Dias, O. (2023). *TranSyT*, an innovative framework for identifying transport systems. *Bioinformatics* 39. doi:10.1093/bioinformatics/btad466

Cuthbertson, J. M., Doyle, D. A., and Sansom, M. S. P. (2005). Transmembrane helix prediction: a comparative evaluation and analysis. *Protein Eng. Des. Sel.* 18, 295–308. doi:10.1093/protein/gzi032

David, R., Byrt, C. S., Tyerman, S. D., Gilliham, M., and Wege, S. (2019). Roles of membrane transporters: connecting the dots from sequence to phenotype. *Ann. Bot.* 124, 201–208. doi:10.1093/aob/mcz066

Dias, O., Gomes, D., Vilaca, P., Cardoso, J., Rocha, M., Ferreira, E. C., et al. (2017). Genome-wide semi-automated annotation of transporter systems. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 14, 443–456. doi:10.1109/TCBB.2016.2527647

Diener, C., and Gibbons, S. M. (2023). More is different: metabolic modeling of diverse microbial communities. *mSystems* 8, e0127022–22. doi:10.1128/msystems.01270-22

Dobson, L., Szekeres, L. I., Gerdán, C., Langó, T., Zeke, A., and Tusnády, G. E. (2023). TmAlphaFold database: membrane localization and evaluation of AlphaFold2 predicted alpha-helical transmembrane protein structures. *Nucleic Acids Res.* 51, D517–D522. doi:10.1093/nar/gkac928

Ebrahim, A., Lerman, J. A., Palsson, B. O., and Hyduke, D. R. (2013). COBRApy: COnstraints-based reconstruction and analysis for Python. *BMC Syst. Biol.* 7, 74. doi:10.1186/1752-0509-7-74

Elbourne, L. D. H., Tetu, S. G., Hassan, K. A., and Paulsen, I. T. (2017). TransportDB 2.0: a database for exploring membrane transporters in sequenced genomes from all domains of life. *Nucleic Acids Res.* 45, D320–D324. doi:10.1093/nar/gkw1068

Elbourne, L. D. H., Wilson-Mortier, B., Ren, Q., Hassan, K. A., Tetu, S. G., and Paulsen, I. T. (2023). TransAAP: an automated annotation pipeline for membrane transporter prediction in bacterial genomes. *Microb. Genomics* 9, mgen000927. doi:10.1099/mgen.0.000927

Falkowski, P. G., Fenchel, T., and Delong, E. F. (2008). The microbial engines that drive earth's biogeochemical cycles. *Science* 320, 1034–1039. doi:10.1126/science.1153213

Faria, J. P., Liu, F., Edirisinghe, J. N., Gupta, N., Seaver, S. M. D., Freiburger, A. P., et al. (2023). ModelSEED v2: high-throughput genome-scale metabolic model reconstruction with enhanced energy biosynthesis pathway prediction (preprint). *Syst. Biol.* doi:10.1101/2023.10.04.556561

Fichant, G., Basse, M.-J., and Quentin, Y. (2006). ABCdb: an online resource for ABC transporter repertories from sequenced archaeal and bacterial genomes. *FEMS Microbiol. Lett.* 256, 333–339. doi:10.1111/j.1574-6968.2006.00139.x

Genee, H. J., Bali, A. P., Petersen, S. D., Siedler, S., Bonde, M. T., Gronenberg, L. S., et al. (2016). Functional mining of transporters using synthetic selections. *Nat. Chem. Biol.* 12, 1015–1022. doi:10.1038/nchembio.2189

Gralka, M., Pollak, S., and Cordero, O. X. (2023). Genome content predicts the carbon catabolic preferences of heterotrophic bacteria. *Nat. Microbiol.* 8, 1799–1808. doi:10.1038/s41564-023-01458-z

Griesemer, M., Kimbrel, J. A., Zhou, C. E., Navid, A., and D'haeseleer, P. (2018). Combining multiple functional annotation tools increases coverage of metabolic annotation. *BMC Genomics* 19, 948. doi:10.1186/s12864-018-5221-9

Gudmundsson, S., and Thiele, I. (2010). Computationally efficient flux variability analysis. *BMC Bioinforma.* 11, 489–494. doi:10.1186/1471-2105-11-489

Hammer, B. K., and Bassler, B. L. (2003). *Quorum* sensing controls biofilm formation in *Vibrio cholerae*. *Mol. Microbiol.* 50, 101–104. doi:10.1046/j.1365-2958.2003.03688.x

Hannesschlaeger, C., Horner, A., and Pohl, P. (2019). Intrinsic membrane permeability to small molecules. *Chem. Rev.* 119, 5922–5953. doi:10.1021/acs.chemrev.8b00560

Heifetz, A. (2020). Quantum mechanics in drug discovery, methods in molecular biology (New York, NY: Springer US).doi:10.1007/978-1-0716-0282-9

Heinken, A., Basile, A., and Thiele, I. (2021). Advances in constraint-based modelling of microbial communities. *Curr. Opin. Syst. Biol.* 27, 100346. doi:10.1016/j.coisb.2021.05.007

Heinken, A., Hertel, J., Acharya, G., Ravcheev, D. A., Nyga, M., Okpala, O. E., et al. (2023). Genome-scale metabolic reconstruction of 7,302 human microorganisms for personalized medicine. *Nat. Biotechnol.* 41, 1320–1331. doi:10.1038/s41587-022-01628-0

Heirendt, L., Arreckx, S., Pfau, T., Mendoza, S. N., Richelle, A., Heinken, A., et al. (2019). Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. *Nat. Protoc.* 14, 639–702. doi:10.1038/s41596-018-0098-2

Henry, C. S., DeJongh, M., Best, A. A., Frybarger, P. M., Linsay, B., and Stevens, R. L. (2010). High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat. Biotechnol.* 28, 977–982. doi:10.1038/nbt.1672

Huson, D. H., Auch, A. F., Qi, J., and Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Res.* 17, 377–386. doi:10.1101/gr.5969107

Jeffery, C. J. (2018). Protein moonlighting: what is it, and why is it important? *Philisophical Trans. R. Soc. B* 373, 20160523. doi:10.1098/rstb.2016.0523

Jenkins Sánchez, L. R., Claus, S., Muth, L. T., Salvador López, J. M., and Van Bogaert, I. (2022). Force in numbers: high-throughput screening approaches to unlock microbial transport. *Curr. Opin. Biotechnol.* 74, 204–210. doi:10.1016/j.copbio.2021.11.012

Jones, D., Kim, H., Zhang, X., Zemla, A., Stevenson, G., Bennett, W. F. D., et al. (2021). Improved protein–ligand binding affinity prediction with structure-based deep fusion inference. *J. Chem. Inf. Model.* 61, 1583–1592. doi:10.1021/acs.jcim.0c01306

Joseph, C., Zafeiropoulos, H., Bernaerts, K., and Faust, K. (2024). Predicting microbial interactions with approaches based on flux balance analysis: an evaluation. *BMC Bioinforma.* 25, 36. doi:10.1186/s12859-024-05651-7

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589. doi:10.1038/s41586-021-03819-2

Karl, D. M., Bird, D. F., Björkman, K., Houlihan, T., Shackelford, R., and Tupas, L. (1999). Microorganisms in the accreted ice of Lake Vostok, Antarctica. *Science* 286, 2144–2147. doi:10.1126/science.286.5447.2144

Karp, P. D., Paley, S. M., Midford, P. E., Krummenacker, M., Billington, R., Kothari, A., et al. (2020). Pathway Tools version 24.0: integrated software for pathway/genome informatics and systems biology. ArXiv.

Keating, S. M., Waltemath, D., König, M., Zhang, F., Dräger, A., Chaouiya, C., et al. (2020). SBML Level 3: an extensible format for the exchange and reuse of biological models. *Mol. Syst. Biol.* 16, e9110. doi:10.15252/msb.20199110

King, Z. A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J. A., et al. (2016). BiGG Models: a platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res.* 44, D515–D522. doi:10.1093/nar/gkv1049

Klitgord, N., and Segre, D. (2010). The importance of compartmentalization in metabolic flux models: yeast as an ecosystem of organelles. *Genome Inf.* 22, 41–55. PMID:20238418. doi:10.1142/9781848165786_0005

Kroll, A., Niebuhr, N., Butler, G., and Lercher, M. J. (2023). A general prediction model for substrates of transport proteins (preprint). *Bioinformatics*. doi:10.1101/2023.10.31.564943

Lee, T. J., Paulsen, I., and Karp, P. (2008). Annotation-based inference of transporter function. *Bioinformatics* 24, i259–i267. doi:10.1093/bioinformatics/btn180

Lewis, I. A. (2024). Boundary flux analysis: an emerging strategy for investigating metabolic pathway activity in large cohorts. *Curr. Opin. Biotechnol.* 85, 103027. doi:10.1016/j.copbio.2023.103027

Li, H., Benedito, V. A., Udvardi, M. K., and Zhao, P. X. (2009). TransportTP: a two-phase classification approach for membrane transporter prediction and characterization. *BMC Bioinforma.* 10, 418. doi:10.1186/1471-2105-10-418

Li, J., Zou, Q., and Yuan, L. (2023). A review from biological mapping to computation-based subcellular localization. *Mol. Ther. - Nucleic Acids* 32, 507–521. doi:10.1016/j.omtn.2023.04.015

Lieven, C., Beber, M. E., Olivier, B. G., Bergmann, F. T., Ataman, M., Babaei, P., et al. (2020). MEMOTE for standardized genome-scale metabolic model testing. *Nat. Biotechnol* 38, 272–276. doi:10.1038/s41587-020-0446-y

Machado, D., Andrejev, S., Tramontano, M., and Patil, K. R. (2018). Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic Acids Res.* 46, 7542–7553. doi:10.1093/nar/gky537

Magnúsdóttir, S., Heinken, A., Kutt, L., Ravcheev, D. A., Bauer, E., Noronha, A., et al. (2017). Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nat. Biotechnol.* 35, 81–89. doi:10.1038/nbt.3703

Mahinthichaichan, P., Morris, D. M., Wang, Y., Jensen, G. J., and Tajkhorshid, E. (2018). Selective permeability of carboxysome shell pores to anionic molecules. *J. Phys. Chem. B* 122, 9110–9118. doi:10.1021/acs.jpcb.8b06822

Masrati, G., Dwivedi, M., Rimon, A., Gluck-Margolin, Y., Kessel, A., Ashkenazy, H., et al. (2018). Broad phylogenetic analysis of cation/proton antiporters reveals transport determinants. *Nat. Commun.* 9, 4205. doi:10.1038/s41467-018-06770-5

Mishra, N. K., Chang, J., and Zhao, P. X. (2014). Prediction of membrane transport proteins and their substrate specificities using primary sequence information. *PLoS ONE* 9, e100278. doi:10.1371/journal.pone.0100278

Monk, J. M., Lloyd, C. J., Brunk, E., Mih, N., Sastry, A., King, Z., et al. (2017). iML1515, a knowledgebase that computes *Escherichia coli* traits. *Nat. Biotechnol.* 35, 904–908. doi:10.1038/nbt.3956

Notebaart, R. A., Kintses, B., Feist, A. M., and Papp, B. (2018). Underground metabolism: network-level perspective and biotechnological potential. *Curr. Opin. Biotechnol.* 49, 108–114. doi:10.1016/j.copbio.2017.07.015

Ohnuki, J., Jaunet-Lahary, T., Yamashita, A., and Okazaki, K. (2023). Accelerated molecular dynamics and AlphaFold uncover a missing conformational state of transporter protein OxlT. BioRxiv.

Perez-Garcia, O., Lear, G., and Singhal, N. (2016). Metabolic network modeling of microbial interactions in natural and engineered environmental systems. *Front. Microbiol.* 7, 673. doi:10.3389/fmicb.2016.00673

Sahoo, S., Aurich, M. K., Jonsson, J. J., and Thiele, I. (2014). Membrane transporters in a human genome-scale metabolic knowledgebase and their implications for disease. *Front. Physiology* 5, 91. doi:10.3389/fphys.2014.00091

Saier, M. H., Reddy, V. S., Moreno-Hagelsieb, G., Hendargo, K. J., Zhang, Y., Iddamsetty, V., et al. (2021). The transporter classification database (TCDB): 2021 update. *Nucleic Acids Res.* 49, D461–D467. doi:10.1093/nar/gkaa1004

Saier, M. H., Reddy, V. S., Tamang, D. G., and Västermark, Å. (2014). The transporter classification database. *Nucleic Acids Res.* 42, D251–D258. doi:10.1093/nar/gkt1097

Saier, M. H., Reddy, V. S., Tsu, B. V., Ahmed, M. S., Li, C., and Moreno-Hagelsieb, G. (2016). The transporter classification database (TCDB): recent advances. *Nucleic Acids Res.* 44, D372–D379. doi:10.1093/nar/gkv1103

Saier, M. H., Tran, C. V., and Barabote, R. D. (2006). TCDB: the Transporter Classification Database for membrane transport protein analyses and information. *Nucleic Acids Res.* 34, D181–D186. doi:10.1093/nar/gkj001

Saier, M. H., Yen, M. R., Noto, K., Tamang, D. G., and Elkan, C. (2009). The transporter classification database: recent advances. *Nucleic Acids Res.* 37, D274–D278. doi:10.1093/nar/gkn862

Schwacke, R., and Flügge, U.-I. (2018). "Identification and characterization of plant membrane proteins using ARAMEMNON," in *Plant membrane proteomics, methods in molecular biology*. Editors H.-P. Mock, A. Matros, and K. Witzel (New York, New York, NY: Springer), 249–259. doi:10.1007/978-1-4939-7411-5_17

Schwacke, R., Schneider, A., Van Der Graaff, E., Fischer, K., Catoni, E., Desimone, M., et al. (2003). ARAMEMNON, a novel database for Arabidopsis integral membrane proteins. *Plant Physiol.* 131, 16–26. doi:10.1104/pp.011577

Scott, W. T., Benito-Vaquerizo, S., Zimmermann, J., Bajić, D., Heinken, A., Suarez-Diez, M., et al. (2023). A structured evaluation of genome-scale constraint-based modeling tools for microbial consortia. *PLoS Comput. Biol.* 19, e1011363. doi:10.1371/journal.pcbi.1011363

Sohraby, F., and Nunes-Alves, A. (2023). Advances in computational methods for ligand binding kinetics. *Trends Biochem. Sci.* 48, 437–449. doi:10.1016/j.tibs.2022.11.003

Sung, J., Kim, S., Cabatbat, J. J. T., Jang, S., Jin, Y.-S., Jung, G. Y., et al. (2017). Global metabolic interaction network of the human gut microbiota for context-specific community-scale analysis. *Nat. Commun.* 8, 15393. doi:10.1038/ncomms15393

Szatkowski, M., Barbour, B., and Attwell, D. (1990). Non-vesicular release of glutamate from glial cells by reversed electrogenic glutamate uptake. *Nature* 348, 443–446. doi:10.1038/348443a0

Taffs, R., Aston, J. E., Brileya, K., Jay, Z., Klatt, C. G., McGlynn, S., et al. (2009). *In silico* approaches to study mass and energy flows in microbial consortia: a syntrophic case study. *BMC Syst. Biol.* 3, 114–116. doi:10.1186/1752-0509-3-114

Thiele, I., and Palsson, B. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.* 5, 93–121. doi:10.1038/nprot.2009.203

Varma, A., and Palsson, B. O. (1994). Metabolic flux balancing: basic concepts, scientific and practical use. *Nat. Biotechnol.* 12, 994–998. doi:10.1038/nbt1094-994

Wang, H., Marcišauskas, S., Sánchez, B. J., Domenzain, I., Hermansson, D., Agren, R., et al. (2018). RAVEN 2.0: a versatile toolbox for metabolic network reconstruction and a case study on Streptomyces coelicolor. *PLoS Comput. Biol.* 14, 10065411–e1006617. doi:10.1371/journal.pcbi.1006541

Wang, K. (2023). GPDOCK: highly accurate docking strategy for metalloproteins based on geometric probability. *Briefings Bioinforma.* 24, bbac620. doi:10.1093/bib/bbac620

Yu, N. Y., Wagner, J. R., Laird, M. R., Melli, G., Rey, S., Lo, R., et al. (2010). PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* 26, 1608–1615. doi:10.1093/bioinformatics/btq249

Zemla, A. T., Allen, J. E., Kirshner, D., and Lightstone, F. C. (2022). PDBspheres: a method for finding 3D similarities in local regions in proteins. *NAR Genomics Bioinforma.* 4, lqac078. doi:10.1093/nargab/lqac078

Zhang, X., Wong, S. E., and Lightstone, F. C. (2014). Toward fully automated high performance computing drug discovery: a massively parallel virtual screening pipeline for docking and molecular mechanics/generalized born surface area rescoring to improve enrichment. *J. Chem. Inf. Model.* 54, 324–337. doi:10.1021/ci4005145

Zhang, Z., Tao, Z., Gameiro, A., Barcelona, S., Braams, S., Rauen, T., et al. (2007). Transport direction determines the kinetics of substrate transport by the glutamate transporter EAAC1. *Proc. Natl. Acad. Sci. U. S. A.* 104, 18025–18030. doi:10.1073/pnas.0704570104

Zuniga, C., Tibocha-Bonilla, J. D., and Betenbaugh, M. J. (2021). Kinetic, metabolic, and statistical analytics: addressing metabolic transport limitations among organelles and microbial communities. *Curr. Opin. Biotechnol.* 71, 91–97. doi:10.1016/j.copbio.2021.06.024

frontiers | Frontiers in Systems Biology

TYPE Review
PUBLISHED 02 August 2024
DOI 10.3389/fsysb.2024.1407994

OPEN ACCESS

EDITED BY
Rahuman S. Malik-Sheriff,
European Bioinformatics Institute (EMBL-EBI),
United Kingdom

REVIEWED BY
Subash Balsamy,
King Abdullah University of Science and Technology, Saudi Arabia
Dilan Pathirana,
University of Bonn, Germany

*CORRESPONDENCE
Robert W. Smith,
robert1.smith@wur.nl

RECEIVED 27 March 2024
ACCEPTED 03 July 2024
PUBLISHED 02 August 2024

CITATION
Noordijk B, Garcia Gomez ML, ten Tusscher KHWJ, de Ridder D, van Dijk ADJ and Smith RW (2024), The rise of scientific machine learning: a perspective on combining mechanistic modelling with machine learning for systems biology.
Front. Syst. Biol. 4:1407994.
doi: 10.3389/fsysb.2024.1407994

The rise of scientific machine learning: a perspective on combining mechanistic modelling with machine learning for systems biology

Ben Noordijk[1,2], Monica L. Garcia Gomez[2,3], Kirsten H. W. J. ten Tusscher[2,3], Dick de Ridder[1,2], Aalt D. J. van Dijk[2,4] and Robert W. Smith[5]*

[1]Bioinformatics Group, Wageningen University and Research, Wageningen, Netherlands, [2]CropXR Institute, Utrecht, Netherlands, [3]Experimental and Computational Plant Development, Institute of Environmental Biology; Theoretical Biology, Institute of Biodynamics and Biocomplexity, Department of Biology, Utrecht University, Utrecht, Netherlands, [4]Biosystems Data Analysis, Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, Netherlands, [5]Laboratory of Systems and Synthetic Biology, Wageningen University and Research, Wageningen, Netherlands

Both machine learning and mechanistic modelling approaches have been used independently with great success in systems biology. Machine learning excels in deriving statistical relationships and quantitative prediction from data, while mechanistic modelling is a powerful approach to capture knowledge and infer causal mechanisms underpinning biological phenomena. Importantly, the strengths of one are the weaknesses of the other, which suggests that substantial gains can be made by combining machine learning with mechanistic modelling, a field referred to as Scientific Machine Learning (SciML). In this review we discuss recent advances in combining these two approaches for systems biology, and point out future avenues for its application in the biological sciences.

KEYWORDS

machine learning, mechanistic models, scientific machine learning (SciML), ordinary differential equations, system identification, parameter estimation, biology-informed neural network (BINN)

1 Introduction

Classically, systems biology has primarily focused on the use of dynamic mechanistic models to elucidate the underpinnings of natural phenomena. Popular model formalisms applied include ordinary and partial differential equations (ODEs and PDEs, respectively), Boolean networks, Petri nets, cellular automata, individual-based models, and combinations of these. Properties of mechanistic models—including the type of equation or rules, initial conditions, or parameter values—depend on the field, question of interest, and expertise of the researchers involved and are often determined or constrained by the limited availability and quality of experimental data. While classic, low-dimensional models can fit a range of concentration-, time-, and space-dependent datasets (Michaelis and Menten, 1913; Lotka, 1920; Volterra, 1926; Hodgkin and Huxley, 1952), for larger, high-dimensional biological systems such models can be difficult to construct due to the so-called curse of dimensionality (Bellman, 1957): as many variables and hence model parameters are necessary to describe a high-dimensional system, it is virtually impossible to generate sufficient experimental measurements to properly estimate these parameters. Only if many existing parameters are known *a priori* (e.g., reaction rates from experimental measurements), they can be used to construct a quantitative mechanistic model that overcomes the curse of dimensionality (Karr et al., 2012). Alternatively, coarser models such as Flux Balance Analysis and Boolean models are typically applied to large metabolic or regulatory networks, as their assumptions lead to simpler models (Xiao, 2009; Orth et al., 2010). Mechanistic models have been indispensable tools to test if our current understanding of biology is necessary and sufficient to describe experimental data, all while having interpretable inner workings. Nevertheless, a gap exists whereby high-throughput time- or space-dependent data is not yet readily used to construct detailed, large mechanistic models.

More recently, state-of-the art machine learning (ML) algorithms have been developed and applied to the increasing wealth of biological data. Since these are data-driven methods that are built to infer patterns from large, high-dimensional datasets, they have enabled high accuracy in applications such as protein structure and function prediction (Jumper et al., 2021; Kulmanov et al., 2018), single-cell transcriptomics modelling (Lopez et al., 2018), and more (see Baker et al., 2018; Sapoval et al., 2022). However, many of these ML methods have limited biological interpretability, and do not elucidate underlying biological mechanisms in the way that mechanistic models can.

Given their complementary strengths and weaknesses, integration between ML and mechanistic models, also called SciML, is a promising new field, which has already gained popularity in scientific disciplines such as engineering (Willard et al., 2022), crop modelling (Maestrini et al., 2022), and physics (Karniadakis et al., 2021). Indeed, there is a great interest in combining these two approaches and their application in diverse fields (Legaard et al., 2023; Tong et al., 2020; von Rueden et al., 2021). In this review, we discuss the latest advances in combining ML and mechanistic modelling approaches—particularly in the form of ODEs or PDEs—applied to systems biology. Notably, while similar reviews for fields like biomedical multiscale models exist (Alber et al., 2019), and reviews such as Gazestani and Lewis (2019) concentrate solely on deep learning—a subset of ML—our focus is on innovative approaches in merging biological knowledge with various ML approaches within the systems biology domain. Here, we aim to provide a perspective on the use of SciML for the study of biological systems, and thus we do not explicitly focus on performing the modelling in practice. For more information on SciML-related software packages and best practices, please refer to the Supplementary Material.

We first describe methods leveraging prior biological knowledge or mechanistic models to augment the interpretability and accuracy of ML models. Subsequently, we explore how ML techniques can contribute to the development and simulation of mechanistic models. Next, we review models that intrinsically merge mechanistic models with ML, and the synergy this provides. Finally, we provide a perspective on potential new avenues for integration of ML and mechanistic models. A brief overview of all categories of models that we discuss is given in Table 1, where we highlight what mechanistic model and ML building blocks they are built of, and for what goal they are integrated.

2 Combining ML with prior knowledge

2.1 Constraining ML model structure

Machine learning is concerned with computational methods that learn (i.e., are trained) to perform a certain task based on example data. A wide range of methods are available, each differing primarily in the assumptions they impose on a problem. This results in a trade-off between the model's complexity and its ability to learn any given problem, known as the bias-variance trade-off (Geman et al., 1992). As a major subfield of ML, neural networks (NNs, more recently called deep learning, DL) consist of simple functions ("units" or "nodes") that calculate a weighted combination of their inputs and then apply a non-linear transformation to produce an output. By combining several layers of such units, given a dataset of examples of input x and desired output y, sufficiently large NNs can in principle be trained to approximate any function (Hornik et al., 1989)

$$\hat{y} = \mathrm{NN}(x, w, b) \tag{1}$$

where w and b represent the internal weights and biases of the NN, respectively. For readability, subsequent equations will omit explicit mention of these parameters.

NNs have shown great potential in systems biology (Sapoval et al., 2022) to, for example, relate multi-omics data to drug response (Sharifi-Noghabi et al., 2019). Nevertheless, the broad deployment and practical utility of NNs is still limited by a number of factors. First, NNs can be hard to generalise to different biological contexts as they easily overfit the specific training data available. Second, as highly parameterised universal approximation methods, NNs suffer from a lack of interpretability. Therefore, it makes sense to inform NNs with existing biological knowledge to constrain their complexity, a task for which NNs are well-suited. Conventionally, such approaches start from an existing NN architecture (e.g., a multi-layer perceptron, MLP, or a recurrent NN, RNN) and limit some of its internal connections based on biological data or prior knowledge, thus reducing the number of parameters to be estimated. In some cases, this allows certain elements of the NN to take on a mechanistic meaning, which "opens up the black box." Here we discuss methods where NN performance and/or interpretability has been aided by inclusion of established biological insights.

A first way to enforce biological prior knowledge is by creating a sparsely connected MLP, where each node represents a biological entity (e.g., a gene, protein complex, or full cell organelle) and nodes are only connected if they are known to interact based on experimental or computational biological evidence (Elmarakeby et al., 2021). Such a sparse MLP has been applied to cell growth models, where connections were informed by Gene Ontology (GO) terms (Ma et al., 2018) and to modelling signalling and transcriptional regulation, where each connection is based on known interactions between genes, proteins, and their pathway

TABLE 1 Overview of the SciML approaches covered in this review, the models they merge, and the goal of integration (NN, neural network, MM, mechanistic model, ML, machine learning, ODE, ordinary differential equation).

Section	Name	Starting point	Combine with	Goal
2.1	Constraining ML model structure	Standard fully connected NN	Dataset of (predicted) biological interactions, only connect nodes in NN if there is evidence for an interaction	Make nodes and edges take on meaning; increase interpretability
2.2	Mechanistic model simulations as input for ML	Existing MM	NN to make predictions based on MM output	Perform task that MM could not do in isolation
3.1	Selecting from a library of candidate terms	Terms from which ODEs could be constructed	ML to select key terms from the library	Identify ODEs that fit dataset using a small number of candidate terms
3.2	Finding hidden mechanisms	ODEs with some terms (i.e., mechanisms) already known	NN to fit unknown terms	ODE model with increased performance; potentially information about what terms should be added to the ODEs
3.3	No candidate terms are known	ODEs missing terms that are needed to explain rate of change	NN that predicts the rate of change of each element (e.g., gene), based on all other elements in the system	Accurate, but hard to interpret method to predict temporal patterns
3.4	NN to enhance model simulations	Parameterised ODEs	NN that predicts the solution of the ODEs	Faster solving of the ODE system
4.1	ML to aid in fitting sparse, noisy data	ODEs that should be fit to noisy and/or sparse data	NN to interpolate the data while adhering to the limits that the ODEs provide	Interpolate data (without overfitting) for finding parameters of ODEs
4.2	Parametrisation of metabolic systems	High-dimensional system of ODEs with yet unknown parameters	NN that predicts a set of parameters, and NN that can classify if parameters are good or not	Find parameters for large system of ODEs that make it consistent with experimental data

membership (Fortelny and Bock, 2020; Hartman et al., 2023). Overall, these studies find that such biologically-constrained MLPs outperform existing predictive models, suffer less from overfitting compared to their fully connected counterparts, and allow for meaningful biological interpretability. However, there is no agreed upon best method yet to extract biological insights from these sparse MLPs.

MLPs are not the only NN architecture that can be used as a blueprint for biology-informed ML. For example, in a recurrent neural network (RNN), the matrix governing the calculation of the hidden state from the previous time point's hidden state can be likened to an interaction matrix (graph) between molecules in a signalling network (Nilsson et al., 2022). Therefore, this matrix can be constrained to only include known interactions, which prevents overfitting, and enables genome-scale modelling of intracellular signalling. Moreover, this matrix can be further constrained by existing knowledge of dynamical systems, e.g., by restraining the system's largest eigenvalue to be smaller than one, as this ensures that the RNN always converges to a steady state or equilibrium. Other architectures, such as convolutional neural networks (CNN), have also been constrained with prior knowledge in fields such as physics (Zhang Z. et al., 2023). However, in the field of systems biology we were unable to find examples of such applications yet, even though CNNs could be used to study, e.g., spatial cell-cell interactions.

Overall, this highlights the potential for constructing biologically-constrained NNs by starting with existing NN architectures that effectively align with the structure of the biological problem being addressed. Nevertheless, not all prior biological knowledge naturally lends itself to this, and the most insightful way to extract meaning from the internal workings of an NN remains to be elucidated.

2.2 Mechanistic model simulations as input

An alternative way to make use of biological knowledge is to use the output of mechanistic models (defined more in depth in Section 3) as "input" to an ML model (Gelbach et al., 2022; Myers et al., 2023). Note that this should be distinguished from "integrated models," where part of the system is modelled using ODEs and another part using ML; here, we focus on cases where multiple ODE simulations are performed to generate data to train the ML model.

One classic approach is so-called simulation-based inference, which refers to a suite of techniques for inferring model parameters when the likelihood function is not tractable (Cranmer et al., 2020). A likelihood function quantifies the probability of observing a set of data given a specific set of parameter values in a model. Parameter values can then be optimised by maximising this likelihood. Classical approaches for simulation-based inference include, e.g., approximate Bayesian computation (ABC), where parameters are repeatedly drawn from a prior distribution, a simulation is run with those parameters, and the parameter values are retained as a sample of the posterior distribution if the simulated data is sufficiently close to the observed data. This yields a probability distribution for parameter values given a model structure and a dataset. The approach is case-based, in the sense that for a new set of observations, the entire estimation procedure must be run again.

A second approach is to create a model for the likelihood by estimating the distribution of simulated data with, e.g., kernel density estimation. Compared to ABC, it has the advantage of spreading the costs of the initial investment in simulation across various analyses or parameter estimates: new data points can be evaluated more efficiently. Here, recent developments that use NNs now allow density estimation to scale to high-dimensional data. An example is normalising flows, in which variables described by, e.g., a multivariate Gaussian are transformed through a parameterised invertible transformation. Several such steps can be stacked, and the parameters of the transformations are trained by maximising the likelihood of the observed data. A recent example of such an approach is Bayesflow (Radev et al., 2020), which trains two neural networks on simulated data: i) a summary network, which reduces a set of observations to learned summary statistics (for time-series, typically a long-short-term memory (LSTM) network is used, which is a variant of the above-mentioned RNN); and ii) an inference network, which learns the posterior given these summary statistics. The latter is implemented as a normalising flow. Bayesflow has been used for systems biology problems in Arruda et al. (2023) to consider measurements for different cells or patients, and simulate a heterogeneous cell population using a non-linear mixed-effects model of (single-cell) translation.

An alternative to simulation-based inference is to use transfer learning (Przedborski et al., 2021). This leverages features and representations learned by solving one problem to help solve a related but different problem. After pretraining a model on a large dataset, it can be transferred and fine-tuned for a new task with smaller datasets, accelerating learning and improving performance. This approach is especially useful when labelled data for the target task is limited or expensive to obtain. In the specific example of Przedborski et al. (2021), simulated clinical trial data was obtained from an already calibrated ODE model for immunotherapy, describing time evolution of various cell types based on molecular interactions. Note that this existing model was not directly aimed at distinguishing between patients responding and non-responding to treatment. To do so, an additional classification model was developed. Relevant features for distinguishing response from non-response were selected from the initial conditions and kinetic parameters of the ODE model simulations. These features were then used as inputs to an NN, which was pretrained on the simulated data to classify virtual patients as responders or non-responders. Subsequently, transfer learning was used to fine-tune the model on real clinical data.

Both biologically-constrained MLPs and ODE-input ML have typically been applied to datasets where the final output is static (i.e. a state that does not change). For dynamic outputs, it may be better to start with a mechanistic model and enhance it using ML, as discussed in the next section.

3 ML to enhance mechanistic models

Ordinary differential equation (ODE) models are a commonly used framework to model biological dynamical systems. As the affordability and accessibility of many experimental methods have increased, and the scale of data generation has grown dramatically, mechanistic models have become larger (Fröhlich et al., 2018), more detailed, and less abstract. This leads to a need for both new methods for model construction (i.e., identifying the unknown terms in an equation), and for improved numerical algorithms to address the high computational requirements of ODE solving. Here, we discuss four ways in which ML can support the construction and simulation of mechanistic models: i) if potential terms in the ODE are already known and a subset should be selected, ii) if some terms are still unknown, iii) if all candidate terms are unknown, and iv) if ODE solving should be enhanced.

3.1 Selecting from a library of candidate terms

The first step of any mechanistic modelling study is to define the equations of the model based on prior knowledge of the biological system. These equations describe the rate at which a variable changes over time and/or space, and how it depends on other variables in the system and parameters/reaction rates. The mathematical notation for such a system generally reads

$$\frac{dx}{dt} = f(x, p, t) \tag{2}$$

where dx/dt is the rate of change of species or variables x over time, which is determined by reactions f with parameters (or rate constants) p. These reactions may be influenced by time t. In systems biology, the functions f could represent defined chemical reactions between variables, e.g., conversions between different states or enzyme-catalyzed Michaelis-Menten reactions, that depend on parameters p with clear biological definitions, e.g., transcription, translation, complex formation, (de)phosphorylation, dilution, degradation, and diffusion rates. Consequently, many systems biology models are constructed from the same set of mathematical terms, or building blocks, with a direct biological interpretation (Ingalls, 2013; Klipp et al., 2016).

Another factor to consider is the size of the model, i.e. the number of variables and/or parameters. This is often constrained by the data availability, namely which system species and rates have been measured. In the process of model construction, a key question for the modeller is then whether a model needs to be complete—in the sense that all known variables x need to be contained within the model—or whether a smaller, abstract model is sufficient to explain the available data. This is referred to as *model parsimony* and measures such as the Akaike Information Criterion can be used to compare model structures (Portet, 2020). In practice, this means that systems biologists often search for models with a limited number of "hidden," or unmeasured, variables to reduce the uncertainty in predictions made for measured variables.

Both considerations above—equation formulation and model size—can be biased by the researchers' preferences and prior knowledge. To avoid this, ML has recently been applied to construct models based on data in an unbiased manner. For example, Erdem and Birtwistle (2023) utilised ML to infer gene networks from integrated -omics data and used these connections to expand an existing mechanistic model (Erdem et al., 2022; 2023). Alternatively, when a library of potential terms in f is already known, the SINDy (sparse identification of non-linear dynamics) family of symbolic regression methods has been developed to select the most relevant terms from this library (Brunton et al., 2016; Champion et al., 2019; Massonis et al., 2023). The SINDy method (Brunton et al., 2016) rewrites an ODE, as in Eq. 2, into

$$f(x) = \Theta(\mathbf{X})\Xi \tag{3}$$

where $\Theta(\mathbf{X})$ is a time-dependent matrix containing a library of candidate mathematical terms for the ODE (e.g., $\cos(x(t))$, $x^2(t), \ldots$), and Ξ is a sparse matrix containing parameters detailing the rates of each associated mathematical term in the equation. To obtain the matrix Ξ from data, we can minimise a loss function

$$\mathcal{L} = \left(\frac{dx_d}{dt} - \Theta(\mathbf{X})\Xi\right)^2 \tag{4}$$

where dx_d/dt is the numerically approximated time-derivative of time-dependent measurements. When the loss function $\mathcal{L}$ approaches zero, the predicted ODEs produce solutions that match the time-dependent measurements of variables. To prevent complex models being obtained, this optimisation problem is solved with sparse regression methods, such that Ξ is a sparse vector containing as many zeros as possible (Brunton et al., 2016). Test cases in the literature encompass a variety of oscillatory systems, including Lorenz attractors, swinging pendulums (which have recently been related to cell cycle models (Dragoi et al., 2024)), spatial patterning, and glycoloysis pathways in yeast. Moreover this SINDy methodology has since been extended to model non-linear dynamics using implicit functions (Kaheman et al., 2020) and to create structurally identifiable models (Massonis et al., 2023). One recent extension of the SINDy method used autoencoder NNs to reduce the dimensionality of data x to a smaller set of "intrinsic coordinates" z, which can be modelled and used to reproduce the observations seen in the larger system (Champion et al., 2019). In this instance the neural network calculates

$$z = \mathrm{NN}(x) \tag{5}$$

where $|z| < |x|$, and dz/dt provides knowledge about the larger system dx/dt. Compared to linear dimension reduction approaches such as principal component analysis or dynamic mode decomposition, this nonlinear approach may lead to poor interpretability of the dynamic variables, but it allows for more complex models to be simplified and analytically explored.

3.2 Finding hidden mechanisms

In a second, less constrained, modelling approach, universal ordinary differential equations incorporate NNs into the differential equations themselves. In this case, the mathematical definition of a reaction or relationship between model variables may be unknown, and a neural network is trained to determine the time-dependent rate of change. An example universal ordinary differential equation would then take the form

$$\frac{dx}{dt} = f(x) + \mathrm{NN}(x,t) \tag{6}$$

where $f(x)$ models known relations, whilst $\mathrm{NN}(x,t)$ is a time-dependent NN that represents unknown interactions. The equations are then fit to data as part of training the NN. Such methods have been applied to ODEs (such as the oscillatory Lotka-Volterra system), PDEs for describing spatio-temporal biological phenomena (Rackauckas et al., 2021), and chemical master equations describing stochastic kinetics of small genetic networks including feedback loops (Jiang et al., 2021). Hence, they have proven to be very convenient when commonly used mathematical functions do not provide a model with a good fit to data. Bringing universal ordinary differential equations together with SINDy provides the opportunity, as in Rackauckas et al. (2021), to determine an unknown time-dependent reaction rate, followed by approximating the best mathematical definition of the reaction rate using SINDy. This would allow models to simultaneously be constructed directly from data whilst building on pre-existing knowledge (contained in $f(x)$).

In a complementary approach, one can use the output of the NNs (e.g., a plot of $\mathrm{NN}(x)$ vs. x) to estimate the precise mathematical expression (functional form) that describes an unknown term (Lagergren et al., 2020; Daryakenari et al., 2024). Lagergren et al. (2020) showed that MLPs could be used to estimate cell growth and diffusion terms in a PDE model describing scratch assay experiments where cells repopulate available space on a surface. From this analysis, explicit mathematical functions could be approximated to create a phenomenological that then showed these two terms were not sufficient for a fully accurate MLP fit. Based on this discrepancy, the authors also added a time-delay term which yielded a better model fit, even when taking into account the increased number of parameters. This methodology was demonstrated on both simulated and *in vitro* data.

3.3 No candidate terms are known

As a third approach, neural ODEs (nODEs) (Chen et al., 2019) can be used to estimate the rate of change of the system. Here, no underlying assumptions about the functional form of the dynamics are made, and the neural network outputs the rate of change of x,

$$\frac{dx}{dt} = \mathrm{NN}(x,t). \tag{7}$$

nODEs have been applied for transcriptomic forecasting (i.e., predicting gene expression over time) (Erbe et al., 2023), but provide limited biological interpretability. To enhance interpretability and integrate biological insights, Hossain et al. (2024) incorporated prior knowledge into the neural network architecture, specifically by adding soft constraints which steer the nODE connections to putative transcription factor-gene interactions obtained through transcription factor binding site enrichment (comparable to Section 2.1). The methodology was performed to model gene expression changes in yeast cell cycles, breast cancer progression, and B cell dynamics from ChIP-seq and RNA-Seq datasets. This approach increased performance, led to a sparser NN, and could be used to reconstruct underlying gene regulatory networks. Potentially, this gene regulatory network could be used as a starting point for a more insightful mechanistic model, built up using some of the aforementioned methods. For single-cell transcriptomics, Chen et al. (2022) and Zhang J. et al. (2023) used an autoencoder to predict RNA velocities or expressions, respectively. To gain biological insights into the workings of the autoencoder, the latent layer could be probed for biological insights.

Nevertheless, elucidating the inner workings of nODEs remains a challenge compared to more traditional ODE/PDE models. Moreover, their predictive performance can still be improved, especially for sparse, noisy biological data measurements.

3.4 Neural networks to enhance model simulations

Once model equations have successfully been obtained, the next step in model construction is to define parameters and simulate the system. During parameter optimisation (i.e. data fitting), a differential equation model is solved many tens of thousands of times with different sets of parameter values before the output simulations are compared with experimental data. In the absence of extensive parallelisation, the computational cost of numerically solving the model often leads to long run times for parameter optimisation. Since traditional ODE solvers are computationally demanding, researchers have considered the use of NNs to output the solution of an ODE given time t as an input. The NN is then trained to minimise a loss function that ensures the NN's output adheres to the underlying ODE (Grossmann et al., 2023).

This approach can be extended to PDEs, providing the NN with time and spatial coordinates as has been done by Han et al. (2018), Nabian and Meidani (2019), and Wang and Wang (2024) for high-dimensional systems consisting of 50–100 equations. In these examples, the spatial coordinates of the PDE are modelled using a stochastic time-dependent processes and used as inputs into an NN to predict the evolution of system components over space and time.

Comparisons between this NN-based ODE/PDE solving method and traditional approaches, such as finite element methods, reveal two key insights (Han et al., 2018; Nabian and Meidani, 2019; Grossmann et al., 2023; Stiasny and Chatzivasileiadis, 2023; Wang and Wang, 2024). First, there is debate as to whether NNs can predict solutions to differential equations with similarly high accuracy as their finite-element counterparts. For example, Grossmann et al. (2023) show that their methodology provides PDE solutions with higher relative error compared to finite-element methods. Notably, the relative errors found in Grossmann et al. (2023) are comparable with those for high-dimensional systems (Han et al., 2018; Wang and Wang, 2024). Second, the evaluation time of differential equation systems using NNs does not change with the accuracy of solutions, in contrast to finite element methods which take longer when higher accuracy is required (Grossmann et al., 2023). This hints to the possibility that parallelisation of NN evaluation could dramatically speed up large-scale model simulations at the cost of slightly decreased accuracy of numerical approximations. To the best of our knowledge, researchers have not yet been able to bridge the gap in relative error between NN solutions and solutions obtained using finite-element methods.

In summation, the examples above illustrate how ML methods can be applied to differential equation models to identify what terms should be used in equations, predict novel terms in equations, and speed up numerical approximation of complex models.

4 Integrating mechanistic models and ML

4.1 ML to aid in fitting sparse, noisy data

Many of the methods discussed above require numerous time point measurements with minimal noise, which is often difficult to achieve for biological problems. Hence, generating an estimation of the experimental data at unmeasured time points can greatly assist in mechanistic model fitting and provide insight into the underlying biological dynamics:

$$\hat{x} = \mathrm{NN}(t). \tag{8}$$

However, since MLPs commonly contain thousands of parameters, they are prone to overfitting the training data and may not generalise well to out-of-sample scenarios (Willard et al., 2022). Such function-estimating NNs can be made robust by constraining them using known ODEs, i.e., making these models physics-informed neural networks (PINNs) (Raissi et al., 2019). A first approach is to make their derivative be as close as possible to *a priori* ODE/PDEs that describe (aspects of) the known underlying biological system. Such an approach was demonstrated by Yazdani et al. (2020) on three biological datasets, and was implemented through the loss function:

$$\mathcal{L} = \underbrace{(\hat{x} - x)^2}_{\text{Data loss}} + \underbrace{\left(\frac{d}{dt}\hat{x} - f(\hat{x}, t)\right)^2}_{\text{ODE loss}} \tag{9}$$

The first term ensures a close match between the NN-interpolated data $\hat{x}$ and the experimental data x, while the second term keeps the MLP derivatives in agreement with the *a priori* ODEs f. $\frac{d}{dt}\hat{x}$ is found by automatic differentiation through the NN. Minimising this loss function not only allows the NN to more robustly fit the noisy training data, but also allows for simultaneous fitting of parameters in the *a priori* ODEs f. All in all, this demonstrates that the unidirectional interactions discussed so far can be integrated, where mechanistic models inform ML, and vice-versa.

On simulated datasets, Yazdani et al. (2020) demonstrate that this approach successfully estimates practically identifiable parameters (i.e., those that can be uniquely determined from experimental data) for oscillatory or adaptive models with 5–20 unknown parameters and 5–10 system variables. It would be interesting to determine how successful the methodology is with sparser experimental datasets than those used in this study. Note that this approach only works if the complete ODE equations are known *a priori*; if parts are unknown, methods as described in Section 3.2 could be used, as shown by Lagergren et al. (2020).

In this nascent field, researchers integrating NNs with biological knowledge use some ambiguous nomenclature for models, where similar methods have been given different names, and different methods have been given similar names. Table 2 provides an overview (not aiming to be complete) striving to disambiguate terminology.

TABLE 2 Nomenclature for integration of neural networks with biological knowledge.

Info		Characteristics				
Study	**Name for approach**	**Underlying ML structure**	ODE/PDE in loss function	ML-structure constrained by biological knowledge	MLP as term in ODE/ PDE	ODE as input to ML (no simultaneous fitting)
Lagergren et al. (2020)	Biologically informed neural network (BINN)	MLP (fully connected) with PDE	Yes	No	Yes	No
Elmarakeby et al. (2021)	Biologically informed neural network (BINN)	MLP (sparse)	No	Yes	No	No
Hartman et al. (2023)	Biologically informed neural network (BINN)	MLP (sparse)	No	Yes	No	No
Yazdani et al. (2020)	Systems biology informed neural network (SBINN)	MLP (fully connected) with ODE	Yes	No	No	No
Przedborski et al. (2021)	Systems biology informed neural network (SBINN)	MLP (fully connected)	No	No	No	Yes
Ma et al. (2018)	*Visible neural network* (VNN)	MLP (sparse)	No	Yes	No	No
Fortelny and Bock. (2020)	Knowledge primed neural network (KPNN)	MLP (sparse)	No	Yes	No	No
Nilsson et al. (2022)	Large-scale knowledge-EMBedded Artificial Signaling-networks (LEMBAS)	RNN (sparse)	No	Yes	No	No

4.2 Parametrisation of metabolic systems

The use of system features alongside simulated or real data has also been applied to NNs evaluating parameters of metabolic systems, such as catalytic rates or maximal rate velocities and Michaelis constants. Choudhury et al. (2022, 2023) present REKINDLE and RENNAISANCE, that apply generative adversarial neural networks (GANs) to find sets of metabolic enzyme parameters that recapitulate metabolic profiles of *E. coli* in steady state conditions. Such mathematical models incorporate tens of state variables and hundreds of model parameters. In REKINDLE (Choudhury et al., 2022), a generator NN is trained to produce model parameter sets with such accuracy that a discriminator NN cannot predict whether they are real or fake when compared with "ground-truth" parameter sets. In RENNAISANCE (Choudhury et al., 2023), several GANs are optimised by a genetic algorithm to produce parameter sets that lead to a model consistent with experimentally determined metabolic responses (e.g., speed at which metabolic pathways reach steady state, system stability, etc.), an approach that foregoes the need for comparison with "ground-truth" parameter sets. In the initial generation of the genetic algorithm, many GANs are created and compared for their ability to produce relevant parameter sets that yield accurate steady state levels of metabolic concentrations. Following generations are then populated with GANs that are perturbed versions of the previous generations best-performing network. Over time, a population of highly performing GANs are then obtained and allows users to analyse variability of model parameters and dynamics for metabolic pathways. The output of both REKINDLE and RENAISSANCE can be used to simulate metabolic systems under different experimental conditions (at steady state or within dynamic bioreactors), compare predicted metabolic parameters with experimentally determined counterparts (and use experimentally measured parameter values to further constrain optimisation solutions), and to predict how metabolic reactions change between physiological states.

Finally, Sukys et al. (2022) have created Nessie, an NN that takes a time-point and model parameters as input and predicts probability distributions of single cell mRNA or protein copy numbers. By then comparing the distributions of system variables with experimentally-determined copy number distributions, the method allows for the back-calculation and estimation of single cell parameter distributions. The authors applied this idea to genetic feedback loops, toggle switches, and kinase pathways. The NN approach made analysis of relationships between parameters and system properties—e.g., the parameters responsible for bimodality in a simple autoregulatory feedback loop—approximately ten thousand times faster.

In summary, recent developments propose a seamless integration of NNs with mechanistic models, and we envision that further progress in this research direction will enable models with increased applicability, interpretability, and performance.

5 Prospective applications: from gene regulatory networks to whole organisms

In the previous sections we reviewed existing work, where mechanistic modelling constrains or informs ML methods, where ML helps construct mechanistic models, and methodologies where

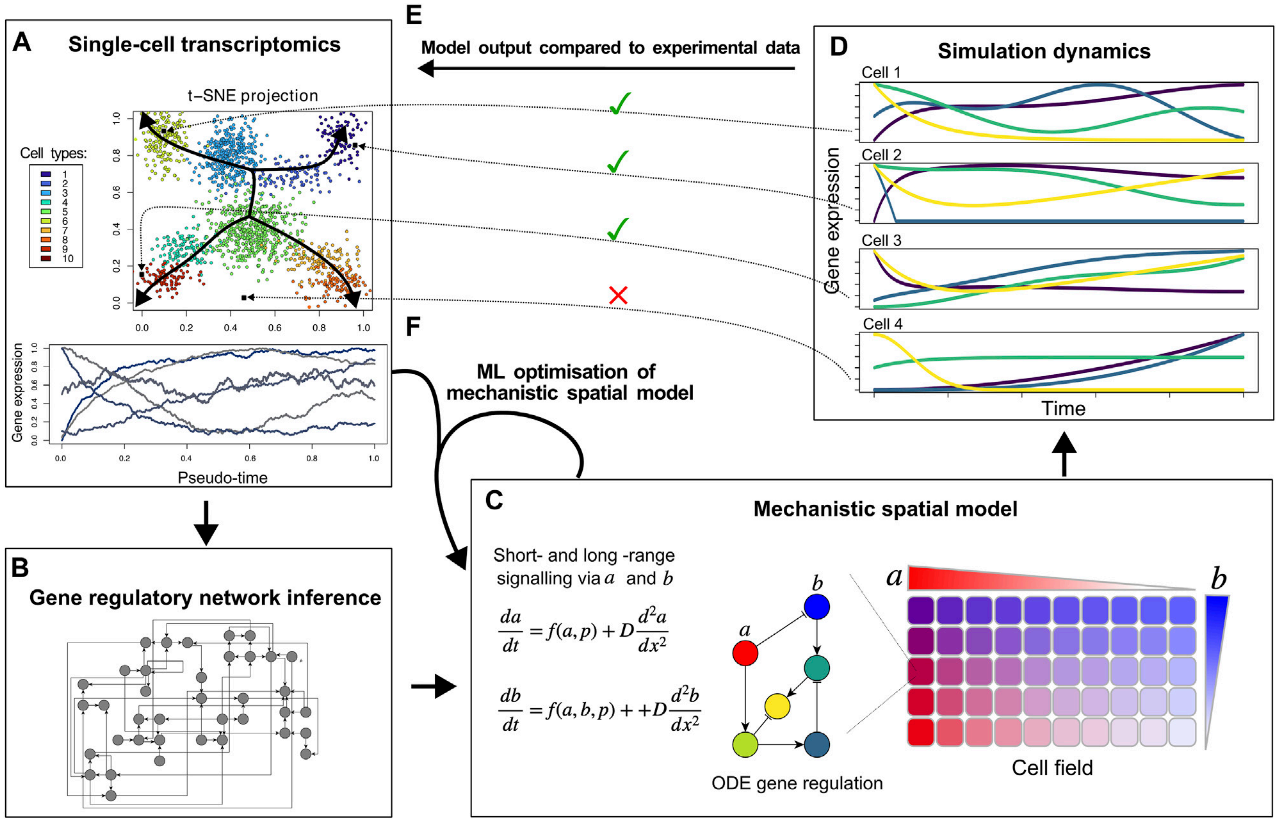

A
Single-cell transcriptomics
t–SNE projection
Cell types:
1
2
3
4
5
6
7
8
9
10
0.0
0.2
0.4
0.6
0.8
1.0
Gene expression
Pseudo-time
E
Model output compared to experimental data
✓
✓
✓
✗
D
Simulation dynamics
Cell 1
Cell 2
Cell 3
Cell 4
Gene expression
Time
F
ML optimisation of mechanistic spatial model
B
Gene regulatory network inference
C
Mechanistic spatial model
Short- and long -range signalling via a and b
$\frac{da}{dt} = f(a,p) + D\frac{d^2a}{dx^2}$
$\frac{db}{dt} = f(a,b,p) + +D\frac{d^2b}{dx^2}$
a
b
ODE gene regulation
a
b
Cell field

FIGURE 1
Proposed hybrid mechanistic-ML models for developmental tissue patterning. Based on single-cell transcriptomic data **(A)**, ML methods can infer a regulatory network **(B)**, that can be used as a building block of a mechanistic spatial model incorporating known and hypothesised details of cell-cell signalling and morphogen gradients **(C)**. By comparing the cell differentiation trajectories produced by the model **(D)** to the actual expression data and cell fate clusters **(E and A)**, an iterative approach can identify missing genes, short-range cell signalling, and/or morphogen gradients to optimise the hybrid model **(F)**.

these two start to become intertwined. Clearly, exploiting the synergy between ML and mechanistic models can lead to more accurate, better interpretable models in systems biology, which will enhance our capacity to modify the behaviour and performance of biological systems in an informed way. Although the balance between ML and mechanistic modelling within integrated approaches may be a matter of taste, expertise of the scientist, and the availability of data and prior knowledge or models, mechanistic models in the end are most easily interpreted. In this last part we therefore turn our focus to how we envision the integration of (multiple) ML techniques could lead to the improvement and expansion of mechanistic models. Additionally, we suggest how ML methods can model residual components to improve predictive power.

5.1 Potential for hybrid approaches to understand tissue developmental patterning

As an illustrative example, in developmental biology the aim is to decipher how cells with identical genetic make up decide which genes to express when and where, in order to produce a patterned specialised tissue consisting of a variety of distinct cell types. In recent years, single-cell transcriptomics combined with ML dimensionality reduction approaches such as tSNE and UMAP (van der Maaten and Hinton, 2008; McInnes et al., 2020) are increasingly used to identify gene expression clusters corresponding to the distinct cell fates occurring in the tissue under study. Subsequently, a pseudotime-based ordering of these cell states enables the reconstruction of temporal trajectories describing cell fate development and transitions (Trapnell, 2015; Saelens et al., 2019) (Figure 1A). Thus far, these methods have mostly been used to identify novel cell types, including the gene expression profiles uniquely identifying these. Frequently, novel cell states are identified that are intermediates of previously known cell types (Jo et al., 2021; Gan et al., 2022), increasing our knowledge of the gene expression changes that cells experience on their path to differentiation. Additionally, subdivisions of previously known cell fates into distinct categories or rare novel cell types are frequently detected (Grün et al., 2015; Tang et al., 2017; Krenkel et al., 2019; Fu et al., 2020). This fine-grained level of understanding has only been possible through the combination of single-cell sequencing with ML methods.

Other ML approaches have been applied to infer gene regulatory networks from single-cell transcriptomics data, identify potential regulatory links between genes, and find the specific cell types in which these regulatory interactions take place (Aibar et al., 2017; Pratapa et al., 2020; Kamimoto et al., 2023) (Figure 1B). Still, it is highly non-trivial to determine whether the recovered regulatory interactions offer a full explanation for the observed cell fate dynamics. In fact, this may be unlikely given that single-cell sequencing is technically limited in the number of transcripts sampled for each cell, with absence of transcripts—particularly lowly-expressed transcription factors—not necessarily meaning absence of expression (Ke et al., 2023). Thus it appears an interesting research direction to combine these methods with spatially explicit mechanistic models of cell fate dynamics that can not only incorporate gene regulatory dynamics but also direct short range cell-cell signalling, longer range morphogen gradient based signalling, transcription factor complex formation, and protein stability regulation, (Figure 1C). While recently ML methods have also emerged aimed at inferring cell-cell interactions from single-cell sequencing data, this has thus far been limited to leveraging known ligand-receptor pairs (Jin et al., 2021; Wilk et al., 2023).

To construct such a mechanistic model for cell fate patterning, the regulatory network inferred by ML can serve as input into the mechanistic network model (Figures 1B,C). Likely, the ML-inferred network is large and different networks may be recovered depending on the specific inference algorithm used, potentially necessitating taking an ensemble approach (Marbach et al., 2012; De Clercq et al., 2021). Network complexity could be reduced by scoring regulatory interactions based on how frequently they are recovered by different algorithms, the integration of transcription factor binding measurements, and known transcription factor-promoter interactions. Additionally, network pruning approaches derived from NN pruning methods could be used to reduce complexity of these regulatory networks (Yeom et al., 2021).

Through simulating a mechanistic model of the multicellular tissue (cell field) that incorporates the inferred gene regulatory network, cell-cell signalling, and the role of morphogens (Figure 1C), *in silico* gene expression dynamics across the tissue can be generated (Figure 1D). Similar to the actual *in vivo* measurements, such *in silico* dynamics can be clustered into cell fates and organised according to their temporal dynamics, enabling a direct comparison with the *in vivo* data (Figure 1E). Mismatches between these simulated and actual cell fates and their dynamics can then be used to further improve and complete the mechanistic model (Figure 1F). This model optimisation should likely involve ML-based optimisation of parameters not present in the experimental data. Examples of these are protein stability, types of cell-cell signalling and their downstream effects, and/or cellular division dynamics. Finally, the integration of the mechanistic and the ML models might include the incorporation of additional relevant genes and interactions based on correlations with already modelled genes or with the phenotype aimed to be described.

Eventually, this could result in an interpretable mechanistic-ML model that reproduces ML-derived cell types, dynamics of cell fates, and inferred cell-cell signalling. We envision that iterating between model learning and adaptive weighting and pruning/sparsifying of inferred networks will help create models which balance explanatory power and model complexity.

5.2 Whole organism studies as a potential scenario for a hybrid mechanistic-ML model

In organisms, both local and systemic responses occur. These responses involve a wide range of spatial and temporal scales, as well as complex interactions between different organs. Here, we use plants as an example of such a multi-scale process, in which the growth and development of organs occurs throughout their lifetime and is regulated by environmental conditions like nutrient stress, drought, high temperatures, shading, or diseases. Ultimately, the organism's performance depends on the coordination of all its parts, necessitating or the development of organism-level models that account for the dynamic processes occurring in each organ. Mechanistic models are typically limited in the number of temporal and spatial scales that can be covered within a single modelling framework, as well as in the number of relevant variables that can be considered. As an example of a modelling framework to study whole organism models, Functional Structural Plant (FSP) models integrate processes at the individual leaf and root level, overall shoot and root level, and entire plant level. In theory, FSP models can include molecular details on how each organ is regulated, e.g., root growth, even if not resolved to the level of individual cells. Still, they tend to be biased towards heavily studied adaptive responses with a clear morphological phenotype, such as preferential foraging towards high nutrient patches, stomatal closure and root elongation under drought, shoot elongation and more upright posture of leaves under high temperature and shading, and reduction in growth to redirect energy to defence under disease pressure (Ruffel et al., 2011; Huot et al., 2014; Pierik and Testerink, 2014; Quint et al., 2016; Buti et al., 2020). In contrast, transcriptomic data reveal that next to these processes with a clear observable output, a large range of metabolic and physiological responses are set in motion by stresses as well. These include changes in nitrate and carbon metabolism, membrane composition, osmotic regulation, and overall rewiring of protein translation. There are missing regulatory layers that are also important to explain an organism's responses. The lack of detailed description of the regulation and temporal dynamics of many of these processes suggest these could be more suited for ML rather than mechanistic modelling, yet still require integration within a single model.

As an example, let us assume our overall organism model contains several functional submodules governing specific morphological and physiological responses in individual organs. For a plant this will represent, e.g., root growth, hypocotyl (stem) growth, or stomatal aperture in leaves (Figure 2A). For stomatal aperture and hypocotyl elongation, key molecular players and interactions have been identified experimentally, enabling the construction of mechanistic models and explaining how they regulate plant development (Figure 2B, top part of each panel). However, many more relevant players and interactions are likely to be discovered. A promising approach to fill knowledge gaps would be to simulate these submodules using the existing mechanistic models, and compare simulated gene expression with transcriptomics measurements to determine how much of the observed dynamics of known key regulatory genes is already explained by the model, and how much "residual" is not explained yet. ML could then be used to infer which genes missing from the mechanistic model could explain these residuals (Figure 2B bottom part, c), potentially under the condition that their

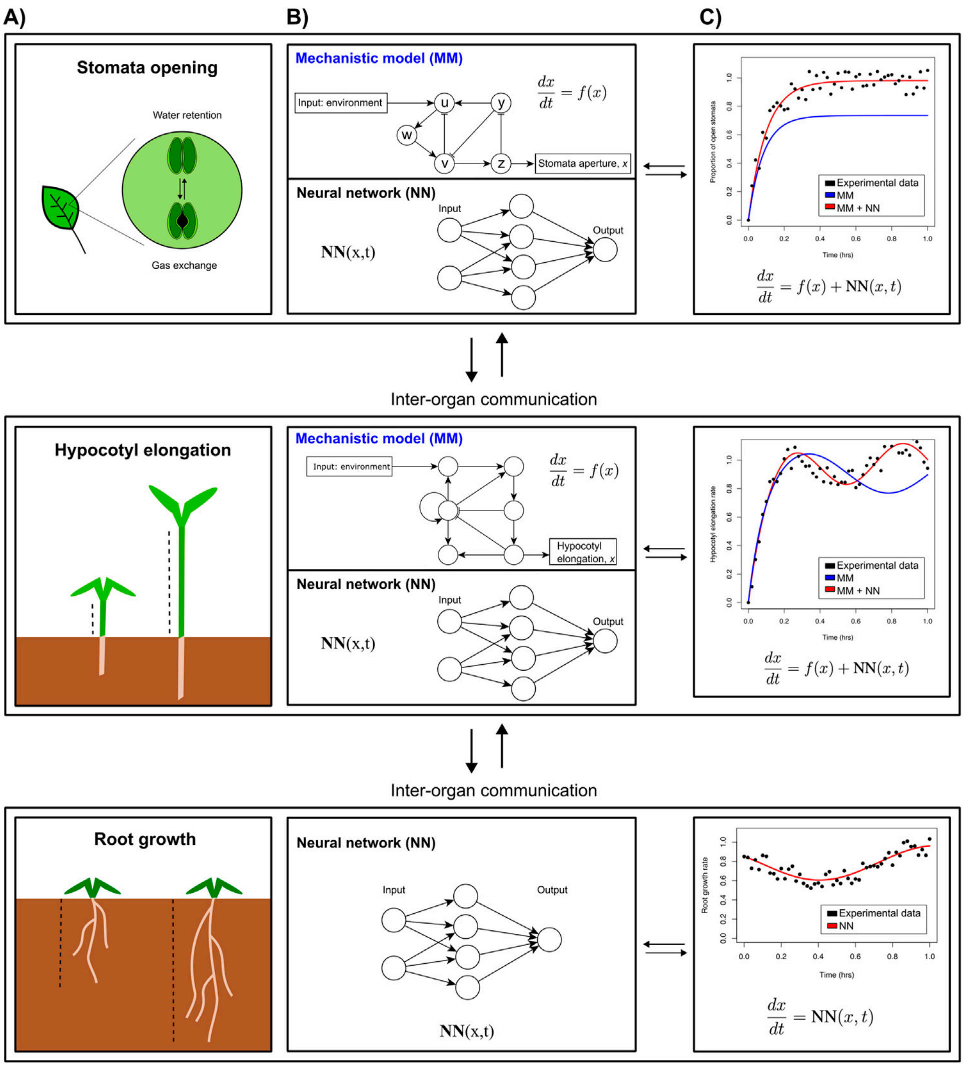


FIGURE 2
Multiscale whole organism model that models various phenotypes. **(A)**. Envisioned iterative strategy integrating mechanistic models (MMs) and neural networks (NNs) **(B)**, that in turn can be used to yield more accurate predictions **(C)**. The hybrid models developed for individual parts of an organism can then be connected to account for inter-organ communication through exchange of molecular regulators and/or nutrients.

regulatory connections to the genes in the mechanistic model can be determined or inferred. The accuracy of the fit between the mechanistic module response and observations can then be improved by iteratively incorporating these novel genes into the mechanistic model, while ensuring high model quality measures that balance accuracy and model complexity (such as the Bayesian or Akaike information criteria, BIC and AIC). Finally, any dynamics that are still not explained by the mechanistic model—including additional genes—can be integrated through an NN term, generating a partly hybrid mechanistic ML module (Figure 2C).

A second possible application of integrated mechanistic-ML modelling would be in the many responses that are not yet properly understood or identified, but do impact the organism's performance. Firstly, ML approaches could be developed to predict a particular phenotype, e.g., plant weight, given a number of morphological, transcriptional, and physiological responses. Feature importance assigned by the ML model would support the parametrisation of the organism-level mechanistic model. Secondly, ML approaches could be used to model the behaviour of still poorly understood response modules for which no mechanistic models can be formulated, (e.g., root growth in Figure 2B). Finally, the functional modules need to be connected (because of reciprocal dependencies or shared regulatory genes), as do different parts or organs of an organism, based on reciprocal exchange of molecular information. For plants, some root-shoot and shoot-root signals have been identified to date, yet many more likely remain to be discovered. ML-based approaches can help predict such missing connections between the different functional modules as well as distinct plant parts.

It should be noted that even though this particular section discusses plants, the foreseen approaches are equally applicable to different fields of research and other organisms, for example, in modelling a virtual human with mechanistic modules for certain well-studied organ systems, supplemented with ML modules for less well-studied parts and supported by ML-based predictors.

6 Conclusion

As discussed, mechanistic models are knowledge-driven approaches that offer insights into underlying biological mechanisms, but are hard to scale up to high dimensions in terms of compute time, parametrisation, and interpretability. On the other hand, ML is data-driven, allowing it to make accurate predictions using large amounts of high-dimensional data, yet it often allows for limited insight into the dynamic mechanisms underlying biological functions. Thus, the strengths of one method are the weaknesses of the other, implying that their integration would be a promising means to achieve both mechanistic understanding and accurate predictions in systems biology.

In our review, we have discussed methods which have either successfully integrated biological knowledge or mechanistic modelling into ML; used ML to help build, fit, or speed up mechanistic models; or fully integrated both approaches. Especially developments in this last category are promising; they allow each step of the procedure to be informed by its influence on the final result and help us overcome typical research challenges such as sparse and/or noisy data, unknown contributing factors, or lack of biological interpretabilty. We end with a vision on how iteratively applying several ML approaches to inform mechanistic modelling may aid in developing quantitatively detailed yet mechanistically tractable models for fields such as developmental patterning or whole organism physiology. This integrative approach promises to yield hybrid models with accurate yet biologically interpretable outputs. Such models can then be used to guide in an informed way the selection of desired behaviours of the biological system under study.

The ability to extract meaningful biological insight from SciML approaches is likely to remain a major focus for future research. Only by "opening up the black box" can we illuminate the complexities of biological processes, which are essential towards deepening our scientific understanding of mechanisms that govern the life we find all around us. Iteratively combining ML with mechanistic modelling is one of several powerful means to achieve this goal.

Author contributions

BN: Visualization, Writing–original draft, Writing–review and editing. MG: Visualization, Writing–original draft, Writing–review and editing. KT: Writing–original draft, Writing–review and editing, Visualization. DR: Writing–original draft, Writing–review and editing. AD: Writing–original draft, Writing–review and editing. RS: Conceptualization, Writing–original draft, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. BN, MG, KT, DR, and AV are supported by the long-term program PlantXR: A new generation of breeding tools for extra-resilient crops (KICH3. LTP.20.005) which is financed by the Dutch Research Council (NWO), the Foundation for Food and Agriculture Research (FFAR), companies in the plant breeding and processing industry, and Dutch universities. These parties collaborate in the CropXR Institute (www.cropxr.org) that is founded through the National Growth Fund (NGF) of the Netherlands. RS is supported by NWO-M project 17336 ("Systems biology analysis of infection structure development in a plant pathogenic fungus").

Acknowledgments

BN used ChatGPT 3.5 to help rephrase a limited number of sentences in the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fsysb.2024.1407994/full#supplementary-material

References

Aibar, S., González-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., et al. (2017). SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* 14, 1083–1086. doi:10.1038/nmeth.4463

Alber, M., Buganza Tepole, A., Cannon, W. R., De, S., Dura-Bernal, S., Garikipati, K., et al. (2019). Integrating machine learning and multiscale modeling—perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences. *npj Digit. Med.* 2, 115–211. doi:10.1038/s41746-019-0193-y

Arruda, J., Schälte, Y., Peiter, C., Teplytska, O., Jaehde, U., and Hasenauer, J. (2023). An amortized approach to non-linear mixed-effects modeling based on neural posterior estimation. doi:10.1101/2023.08.22.554273

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

Bellman, R. (1957). *Dynamic programming*. New Jersey: Princeton University Press.

Brunton, S. L., Proctor, J. L., and Kutz, J. N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Natl. Acad. Sci.* 113, 3932–3937. doi:10.1073/pnas.1517384113

Buti, S., Hayes, S., and Pierik, R. (2020). The bHLH network underlying plant shade-avoidance. *Physiol. Plant.* 169, 312–324. doi:10.1111/ppl.13074

Champion, K., Lusch, B., Kutz, J. N., and Brunton, S. L. (2019). Data-driven discovery of coordinates and governing equations. *Proc. Natl. Acad. Sci.* 116, 22445–22451. doi:10.1073/pnas.1906995116

Chen, R. T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. (2019). Neural ordinary differential equations. doi:10.48550/arXiv.1806.07366

Chen, Z., King, W. C., Hwang, A., Gerstein, M., and Zhang, J. (2022). DeepVelo: single-cell transcriptomic deep velocity field learning with neural ordinary differential equations. *Sci. Adv.* 8, eabq3745. doi:10.1126/sciadv.abq3745

Choudhury, S., Moret, M., Salvy, P., Weilandt, D., Hatzimanikatis, V., and Miskovic, L. (2022). Reconstructing kinetic models for dynamical studies of metabolism using generative adversarial networks. *Nat. Mach. Intell.* 4, 710–719. doi:10.1038/s42256-022-00519-y

Choudhury, S., Narayanan, B., Moret, M., Hatzimanikatis, V., and Miskovic, L. (2023). Generative machine learning produces kinetic models that accurately characterize intracellular metabolic states. doi:10.1101/2023.02.21.529387

Cranmer, K., Brehmer, J., and Louppe, G. (2020). The frontier of simulation-based inference. *Proc. Natl. Acad. Sci.* 117, 30055–30062. doi:10.1073/pnas.1912789117

Daryakenari, N. A., Florio, M. D., Shukla, K., and Karniadakis, G. E. (2024). AI-Aristotle: a physics-informed framework for systems biology gray-box identification. *PLOS Comput. Biol.* 20, e1011916. doi:10.1371/journal.pcbi.1011916

De Clercq, I., Van de Velde, J., Luo, X., Liu, L., Storme, V., Van Bel, M., et al. (2021). Integrative inference of transcriptional networks in Arabidopsis yields novel ROS signalling regulators. *Nat. Plants* 7, 500–513. doi:10.1038/s41477-021-00894-1

Dragoi, C.-M., Tyson, J. J., and Novák, B. (2024). Newton's cradle: cell cycle regulation by two mutually inhibitory oscillators. doi:10.1101/2024.05.18.594803

Elmarakeby, H. A., Hwang, J., Arafeh, R., Crowdis, J., Gang, S., Liu, D., et al. (2021). Biologically informed deep neural network for prostate cancer discovery. *Nature* 598, 348–352. doi:10.1038/s41586-021-03922-4

Erbe, R., Stein-O'Brien, G., and Fertig, E. J. (2023). Transcriptomic forecasting with neural ordinary differential equations. *Patterns* 4, 100793. doi:10.1016/j.patter.2023.100793

Erdem, C., and Birtwistle, M. R. (2023). MEMMAL: a tool for expanding large-scale mechanistic models with machine learned associations and big datasets. *Front. Syst. Biol.* 3, 1099413. doi:10.3389/fsysb.2023.1099413

Erdem, C., Gross, S. M., Heiser, L. M., and Birtwistle, M. R. (2023). MOBILE pipeline enables identification of context-specific networks and regulatory mechanisms. *Nat. Commun.* 14, 3991. doi:10.1038/s41467-023-39729-2

Erdem, C., Mutsuddy, A., Bensman, E. M., Dodd, W. B., Saint-Antoine, M. M., Bouhaddou, M., et al. (2022). A scalable, open-source implementation of a large-scale mechanistic model for single cell proliferation and death signaling. *Nat. Commun.* 13, 3555. doi:10.1038/s41467-022-31138-1

Fortelny, N., and Bock, C. (2020). Knowledge-primed neural networks enable biologically interpretable deep learning on single-cell sequencing data. *Genome Biol.* 21, 190. doi:10.1186/s13059-020-02100-5

Fröhlich, F., Kessler, T., Weindl, D., Shadrin, A., Schmiester, L., Hache, H., et al. (2018). Efficient parameter estimation enables the prediction of drug response using a mechanistic pan-cancer pathway model. *Cell. Syst.* 7, 567–579. doi:10.1016/j.cels.2018.10.013

Fu, Y., Huang, X., Zhang, P., van de Leemput, J., and Han, Z. (2020). Single-cell RNA sequencing identifies novel cell types in Drosophila blood. *J. Genet. Genomics = Yi Chuan Xue Bao* 47, 175–186. doi:10.1016/j.jgg.2020.02.004

Gan, Y., Guo, C., Guo, W., Xu, G., and Zou, G. (2022). Entropy-based inference of transition states and cellular trajectory for single-cell transcriptomics. *Briefings Bioinforma.* 23, bbac225. doi:10.1093/bib/bbac225

Gazestani, V. H., and Lewis, N. E. (2019). From genotype to phenotype: augmenting deep learning with networks and systems biology. *Curr. Opin. Syst. Biol.* 15, 68–73. doi:10.1016/j.coisb.2019.04.001

Gelbach, P. E., Zheng, D., Fraser, S. E., White, K. L., Graham, N. A., and Finley, S. D. (2022). Kinetic and data-driven modeling of pancreatic β-cell central carbon metabolism and insulin secretion. *PLOS Comput. Biol.* 18, e1010555. doi:10.1371/journal.pcbi.1010555

Geman, S., Bienenstock, E., and Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Comput.* 4, 1–58. doi:10.1162/neco.1992.4.1.1

Grossmann, T. G., Komorowska, U. J., Latz, J., and Schönlieb, C.-B. (2023). Can physics-informed neural networks beat the finite element method? doi:10.48550/arXiv.2302.04107

Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., et al. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* 525, 251–255. doi:10.1038/nature14966

Han, J., Jentzen, A., and E, W. (2018). Solving high-dimensional partial differential equations using deep learning. *Proc. Natl. Acad. Sci.* 115, 8505–8510. doi:10.1073/pnas.1718942115

Hartman, E., Scott, A. M., Karlsson, C., Mohanty, T., Vaara, S. T., Linder, A., et al. (2023). Interpreting biologically informed neural networks for enhanced proteomic biomarker discovery and pathway analysis. *Nat. Commun.* 14, 5359. doi:10.1038/s41467-023-41146-4

Hodgkin, A. L., and Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiology* 117, 500–544. doi:10.1113/jphysiol.1952.sp004764

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Netw.* 2, 359–366. doi:10.1016/0893-6080(89)90020-8

Hossain, I., Fanfani, V., Fischer, J., Quackenbush, J., and Burkholz, R. (2024). Biologically informed NeuralODEs for genome-wide regulatory dynamics. *bioRxiv.*, 529835. doi:10.1101/2023.02.24.529835

Huot, B., Yao, J., Montgomery, B. L., and He, S. Y. (2014). Growth-defense tradeoffs in plants: a balancing act to optimize fitness. *Mol. Plant* 7, 1267–1287. doi:10.1093/mp/ssu049

Ingalls, B. P. (2013). *Mathematical modeling in systems biology: an introduction*. 1 edn. Cambridge (Mass.): The MIT Press.

Jiang, Q., Fu, X., Yan, S., Li, R., Du, W., Cao, Z., et al. (2021). Neural network aided approximation and parameter inference of non-Markovian models of gene expression. *Nat. Commun.* 12, 2618. doi:10.1038/s41467-021-22919-1

Jin, S., Guerrero-Juarez, C. F., Zhang, L., Chang, I., Ramos, R., Kuan, C.-H., et al. (2021). Inference and analysis of cell-cell communication using CellChat. *Nat. Commun.* 12, 1088. doi:10.1038/s41467-021-21246-9

Jo, K., Sung, I., Lee, D., Jang, H., and Kim, S. (2021). Inferring transcriptomic cell states and transitions only from time series transcriptome data. *Sci. Rep.* 11, 12566. doi:10.1038/s41598-021-91752-9

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589. doi:10.1038/s41586-021-03819-2

Kaheman, K., Kutz, J. N., and Brunton, S. L. (2020). SINDy-PI: a robust algorithm for parallel implicit sparse identification of nonlinear dynamics. *Proc. R. Soc. A Math. Phys. Eng. Sci.* 476, 20200279. doi:10.1098/rspa.2020.0279

Kamimoto, K., Stringa, B., Hoffmann, C. M., Jindal, K., Solnica-Krezel, L., and Morris, S. A. (2023). Dissecting cell identity via network inference and *in silico* gene perturbation. *Nature* 614, 742–751. doi:10.1038/s41586-022-05688-9

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., and Yang, L. (2021). Physics-informed machine learning. *Nat. Rev. Phys.* 3, 422–440. doi:10.1038/s42254-021-00314-5

Karr, J. R., Sanghvi, J. C., Macklin, D. N., Gutschow, M. V., Jacobs, J. M., Bolival, B., et al. (2012). A whole-cell computational model predicts phenotype from genotype. *Cell.* 150, 389–401. doi:10.1016/j.cell.2012.05.044

Ke, Y., Minne, M., Eekhout, T., and De Rybel, B. (2023). "Single cell RNA-sequencing in arabidopsis root tissues," in *Plant gene regulatory networks: methods and protocols*. Editors K. Kaufmann and K. Vandepoele (New York, NY: Springer US), 41–56. doi:10.1007/978-1-0716-3354-0_4

Klipp, E., Liebermeister, W., Wierling, C., and Kowald, A. (2016). *Systems biology: a textbook*. USA: John Wiley and Sons.

Krenkel, O., Hundertmark, J., Ritz, T. P., Weiskirchen, R., and Tacke, F. (2019). Single cell RNA sequencing identifies subsets of hepatic stellate cells and myofibroblasts in liver fibrosis. *Cells* 8, 503. doi:10.3390/cells8050503

Kulmanov, M., Khan, M. A., Hoehndorf, R., and Wren, J. (2018). DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics* 34, 660–668. doi:10.1093/bioinformatics/btx624

Lagergren, J. H., Nardini, J. T., Baker, R. E., Simpson, M. J., and Flores, K. B. (2020). Biologically-informed neural networks guide mechanistic modeling from sparse experimental data. *PLOS Comput. Biol.* 16, e1008462. doi:10.1371/journal.pcbi.1008462

Legaard, C., Schranz, T., Schweiger, G., Drgoňa, J., Falay, B., Gomes, C., et al. (2023). Constructing neural network based models for simulating dynamical systems. *ACM Comput. Surv.* 55 (236), 1–34. doi:10.1145/3567591

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nat. Methods* 15, 1053–1058. doi:10.1038/s41592-018-0229-2

Lotka, A. J. (1920). Analytical note on certain rhythmic relations in organic systems. *Proc. Natl. Acad. Sci.* 6, 410–415. doi:10.1073/pnas.6.7.410

Ma, J., Yu, M. K., Fong, S., Ono, K., Sage, E., Demchak, B., et al. (2018). Using deep learning to model the hierarchical structure and function of a cell. *Nat. Methods* 15, 290–298. doi:10.1038/nmeth.4627

Maestrini, B., Mimić, G., van Oort, P. A. J., Jindo, K., Brdar, S., Athanasiadis, I. N., et al. (2022). Mixing process-based and data-driven approaches in yield prediction. *Eur. J. Agron.* 139, 126569. doi:10.1016/j.eja.2022.126569

Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., et al. (2012). Wisdom of crowds for robust gene network inference. *Nat. Methods* 9, 796–804. doi:10.1038/nmeth.2016

Massonis, G., Villaverde, A. F., and Banga, J. R. (2023). Distilling identifiable and interpretable dynamic models from biological data. *PLOS Comput. Biol.* 19, e1011014. doi:10.1371/journal.pcbi.1011014

McInnes, L., Healy, J., and Melville, J. (2020). UMAP: uniform manifold approximation and projection for dimension reduction. doi:10.48550/arXiv.1802.03426

Michaelis, L., and Menten, M. L. (1913). Die kinetik der invertinwirkung. *Biochem. z* 49, 352.

Myers, P. J., Lee, S. H., and Lazzara, M. J. (2023). An integrated mechanistic and data-driven computational model predicts cell responses to high- and low-affinity EGFR ligands. *bioRxiv.*, 543329. doi:10.1101/2023.06.25.543329

Nabian, M. A., and Meidani, H. (2019). A deep learning solution approach for high-dimensional random differential equations. *Probabilistic Eng. Mech.* 57, 14–25. doi:10.1016/j.probengmech.2019.05.001

Nilsson, A., Peters, J. M., Meimetis, N., Bryson, B., and Lauffenburger, D. A. (2022). Artificial neural networks enable genome-scale simulations of intracellular signaling. *Nat. Commun.* 13, 3069. doi:10.1038/s41467-022-30684-y

Orth, J. D., Thiele, I., and Palsson, B. Ø. (2010). What is flux balance analysis? *Nat. Biotechnol.* 28, 245–248. doi:10.1038/nbt.1614

Pierik, R., and Testerink, C. (2014). The art of being flexible: how to escape from shade, salt, and drought. *Plant Physiol.* 166, 5–22. doi:10.1104/pp.114.239160

Portet, S. (2020). A primer on model selection using the Akaike Information Criterion. *Infect. Dis. Model.* 5, 111–128. doi:10.1016/j.idm.2019.12.010

Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A., and Murali, T. M. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat. Methods* 17, 147–154. doi:10.1038/s41592-019-0690-6

Przedborski, M., Smalley, M., Thiyagarajan, S., Goldman, A., and Kohandel, M. (2021). Systems biology informed neural networks (SBINN) predict response and novel combinations for PD-1 checkpoint blockade. *Commun. Biol.* 4, 877–915. doi:10.1038/s42003-021-02393-7

Quint, M., Delker, C., Franklin, K. A., Wigge, P. A., Halliday, K. J., and van Zanten, M. (2016). Molecular and genetic control of plant thermomorphogenesis. *Nat. Plants* 2, 15190. doi:10.1038/nplants.2015.190

Rackauckas, C., Ma, Y., Martensen, J., Warner, C., Zubov, K., Supekar, R., et al. (2021). Universal differential equations for scientific machine learning. doi:10.48550/arXiv.2001.04385

Radev, S. T., Mertens, U. K., Voss, A., Ardizzone, L., and Köthe, U. (2020). BayesFlow: learning complex stochastic models with invertible neural networks. doi:10.48550/arXiv.2003.06281

Raissi, M., Perdikaris, P., and Karniadakis, G. E. (2019). Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* 378, 686–707. doi:10.1016/j.jcp.2018.10.045

Ruffel, S., Krouk, G., Ristova, D., Shasha, D., Birnbaum, K. D., and Coruzzi, G. M. (2011). Nitrogen economics of root foraging: transitive closure of the nitrate-cytokinin relay and distinct systemic signaling for N supply vs. demand. *Proc. Natl. Acad. Sci. U. S. A.* 108, 18524–18529. doi:10.1073/pnas.1108684108

Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* 37, 547–554. doi:10.1038/s41587-019-0071-9

Sapoval, N., Aghazadeh, A., Nute, M. G., Antunes, D. A., Balaji, A., Baraniuk, R., et al. (2022). Current progress and open challenges for applying deep learning across the biosciences. *Nat. Commun.* 13, 1728. doi:10.1038/s41467-022-29268-7

Sharifi-Noghabi, H., Zolotareva, O., Collins, C. C., and Ester, M. (2019). MOLI: multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics* 35, i501–i509. doi:10.1093/bioinformatics/btz318

Stiasny, J., and Chatzivasileiadis, S. (2023). Physics-informed neural networks for time-domain simulations: accuracy, computational cost, and flexibility. *Electr. Power Syst. Res.* 224, 109748. doi:10.1016/j.epsr.2023.109748

Sukys, A., Öcal, K., and Grima, R. (2022). Approximating solutions of the Chemical Master equation using neural networks. *iScience* 25, 105010. doi:10.1016/j.isci.2022.105010

Tang, Q., Iyer, S., Lobbardi, R., Moore, J. C., Chen, H., Lareau, C., et al. (2017). Dissecting hematopoietic and renal cell heterogeneity in adult zebrafish at single-cell resolution using RNA sequencing. *J. Exp. Med.* 214, 2875–2887. doi:10.1084/jem.20170976

Tong, A., van Dijk, D., Stanley III, J. S., Amodio, M., Yim, K., Muhle, R., et al. (2020). "Interpretable neuron structuring with graph spectral regularization," in *Advances in intelligent data analysis XVIII.* Editors M. R. Berthold, A. Feelders, and G. Krempl (Cham: Springer International Publishing), 509–521. doi:10.1007/978-3-030-44584-3_40

Trapnell, C. (2015). Defining cell types and states with single-cell genomics. *Genome Res.* 25, 1491–1498. doi:10.1101/gr.190595.115

van der Maaten, L., and Hinton, G. (2008). Visualizing Data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.

Volterra, V. (1926). Fluctuations in the abundance of a species considered Mathematically1. *Nature* 118, 558–560. doi:10.1038/118558a0

von Rueden, L., Mayer, S., Beckh, K., Georgiev, B., Giesselbach, S., Heese, R., et al. (2021). Informed machine learning - a taxonomy and survey of integrating prior knowledge into learning systems. *IEEE Trans. Knowl. Data Eng.*, 1–1doi. doi:10.1109/TKDE.2021.3079836

Wang, M., and Wang, X. (2024). Hybrid neural networks for solving fully coupled, high-dimensional forward–backward stochastic differential equations. *Mathematics* 12, 1081. doi:10.3390/math12071081

Wilk, A. J., Shalek, A. K., Holmes, S., and Blish, C. A. (2023). Comparative analysis of cell–cell communication at single-cell resolution. *Nat. Biotechnol.* 42, 470–483. doi:10.1038/s41587-023-01782-z

Willard, J., Jia, X., Xu, S., Steinbach, M., and Kumar, V. (2022). Integrating scientific knowledge with machine learning for engineering and environmental systems. *ACM Comput. Surv.* 55, 1–37. doi:10.1145/3514228

Xiao, Y. (2009). A tutorial on analysis and simulation of boolean gene regulatory network models. *Curr. Genomics* 10, 511–525. doi:10.2174/138920209789208237

Yazdani, A., Lu, L., Raissi, M., and Karniadakis, G. E. (2020). Systems biology informed deep learning for inferring parameters and hidden dynamics. *PLOS Comput. Biol.* 16, e1007575. doi:10.1371/journal.pcbi.1007575

Yeom, S.-K., Seegerer, P., Lapuschkin, S., Binder, A., Wiedemann, S., Müller, K.-R., et al. (2021). Pruning by explaining: a novel criterion for deep neural network pruning. *Pattern Recognit.* 115, 107899. doi:10.1016/j.patcog.2021.107899

Zhang, J., Larschan, E., Bigness, J., and Singh, R. (2023a). scNODE: generative model for temporal single cell transcriptomic data prediction. doi:10.1101/2023.11.22.568346

Zhang, Z., Yan, X., Liu, P., Zhang, K., Han, R., and Wang, S. (2023b). A physics-informed convolutional neural network for the simulation and prediction of two-phase Darcy flows in heterogeneous porous media. *J. Comput. Phys.* 477, 111919. doi:10.1016/j.jcp.2023.111919

frontiers | Frontiers in Systems Biology

TYPE Original Research
PUBLISHED 29 August 2024
DOI 10.3389/fsysb.2024.1380018

OPEN ACCESS

EDITED BY
Shayn Peirce-Cottler,
University of Virginia, United States

REVIEWED BY
Elisa Domínguez-Hüttinger,
National Autonomous University of Mexico, Mexico
Rada Amin,
University of Nebraska-Lincoln, United States

*CORRESPONDENCE
Sarah Minucci,
✉ sarah.minucci@appliedbiomath.com

RECEIVED 31 January 2024
ACCEPTED 15 August 2024
PUBLISHED 29 August 2024

CITATION
Minucci S, Gruver S, Subramanian K and Renardy M (2024) A multi-scale semi-mechanistic CK/PD model for CAR T-cell therapy.
Front. Syst. Biol. 4:1380018.
doi: 10.3389/fsysb.2024.1380018

A multi-scale semi-mechanistic CK/PD model for CAR T-cell therapy

Sarah Minucci[1]*, Scott Gruver[1], Kalyanasundaram Subramanian[2] and Marissa Renardy[1]

[1]Applied BioMath, Concord, MA, United States, [2]Differentia Biotech, San Francisco, CA, United States

Chimeric antigen receptor T (CAR T) cell therapy has shown remarkable success in treating various leukemias and lymphomas. Cellular kinetic (CK) and pharmacodynamic (PD) behavior of CAR T cell therapy is distinct from other therapies due to its living nature. CAR T CK is typically characterized by an exponential expansion driven by target binding, fast initial decline (contraction), and slow long-term decline (persistence). Due to the dependence of CK on target binding, CK and PD of CAR T therapies are inherently and bidirectionally linked. In this work, we develop a semi-mechanistic model of CAR T CK/PD, incorporating molecular-scale binding, T cell dynamics with multiple phenotypes, and tumor growth and killing. We calibrate this model to published CK and PD data for a CD19-targeting CAR T cell therapy. Using sensitivity analysis, we explore variability in response due to patient- and drug-specific properties. We further explore the impact of tumor characteristics on CAR T-cell expansion and efficacy through individual- and population-level parameter scans.

KEYWORDS

CAR T-cell therapy, mechanistic modeling, quantitative systems pharmacology, sensitivity analysis, cellular kinetics

1 Introduction

Chimeric antigen receptor (CAR) T-cells are T-cells engineered to produce CARs which recognize and bind to a tumor antigen. In CAR T-cell therapy, a patient's T-cells are extracted and isolated, re-engineered to express a specific CAR, expanded *ex vivo*, and then infused back into the patient. Six such therapies have been approved for treating a variety of blood cancers (Chen et al., 2023). These therapies have been shown to produce long-lasting response and superior response rates to alternative treatments (Melenhorst et al., 2022; Sermer et al., 2020). As a result of the individualized nature of CAR T manufacturing, the contents of the dosed product will vary from patient to patient. Further, CAR T cellular kinetic behavior is distinct from other therapies due to its "living" nature; it is typically characterized by an exponential expansion, fast initial decline (contraction), and slow long-term decline (persistence). Additionally, cellular kinetics (CK) is not as well-studied as pharmacokinetics for more traditional drugs Chaudhury et al. (2020). Interactions between CAR T-cells and tumor cells are complex since tumor expansion has a significant impact on CAR T-cell expansion. Furthermore, much is still unknown about the workings of CAR T-cells in the body and there is not a standard monitoring process. Modeling can shed light on CAR-T cell CK/PD and inform future studies by mechanistically linking CAR T-cell doses to tumor cell growth and determining optimal drug properties to achieve efficacy and safety. Furthermore, patient characteristics can be incorporated into the model to provide individualized dose predictions and guide patient and indication selection.

Modeling and simulation has been used to understand CAR T-cell dynamics and efficacy (see, for example, reviews by Chaudhury et al. (2020); Nukala et al. (2021)) and the impact of preconditioning (Owens and Bozic, 2021). Until recently, the three distinct phases of CAR T cellular kinetics and the impact of different CD4$^+$ and CD8$^+$ T cell phenotypes had not been mechanistically described. Previous modeling work had captured CAR T cellular kinetics either empirically (Stein et al., 2019), mechanistically but without multiple phenotypes, (Singh et al., 2020), or mechanistically with effector/memory phenotypes but without separating CD4$^+$ and CD8$^+$ T cells (Hardiansyah and Ng, 2019). Recent work by Salem et al. (2023) has incorporated all of these features, developing a mechanistic model incorporating binding-driven CAR T-cell expansion and activity for multiple CD4$^+$ and CD8$^+$ T-cell phenotypes to match clinical data from multiple trials. Further analysis of such models will be useful to understand system behavior, inform engineering of CAR T-cells, and understand variability in patient populations. In particular, sensitivity analysis provides understanding of the key mechanisms driving expansion and efficacy.

Here, we present a semi-mechanistic cellular kinetic-pharmacodynamic (CK-PD) model for CAR T-cell therapy of B-cell non-Hodgkin lymphoma (NHL). Our model includes CD8$^+$ and CD4$^+$ naive, effector, and memory T-cell phenotypes, binding of CARs to their target antigen CD19, binding-driven activation and expansion of T-cells, T-cell death and conversion to memory cells, and binding-driven killing of B cells by CD8$^+$ effector cells. We demonstrate the ability of the model to capture published human CAR T-cell CK and PD data, and perform sensitivity analysis to understand key model features and predict the impact of variability in patient, tumor, and drug characteristics.

2 Methods

2.1 Data

The model was informed by and benchmarked to published human CAR T-cellular kinetics, B cell percentage, and clinical response data from a phase I clinical trial with IM19 CAR T-cells for 13 relapsed or refractory NHL patients (Ying et al., 2021). The CK data and the B cell aplasia data were both digitized using WebPlotDigitizer (Rohatgi, 2022). Two days prior to CAR T-cell infusion, patients were pre-treated with fludarabine and cyclophosphamide for 3 days to deplete endogenous lymphocytes. IM19 CAR T-cells were dosed by weight at 3 × 10^5, 1 × 10^6, or 3 × 10^6 cells per kg. The CD4:CD8 ratio of the infused CAR T-cells was reported for each of the patients.

2.2 Model structure

The model consists of a single compartment representing the blood. CAR T-cell and B cell populations and their corresponding receptor burdens are modeled explicitly. CAR T-cells, a fraction of which are CD8$^+$ and the remainder CD4$^+$, are dosed directly into the blood. All CAR T-cells are assumed to be naive at the time of dosing. CARs on both CD8$^+$ and CD4$^+$ CAR T-cells can bind to CD19. CD8$^+$ and CD4$^+$ naive CAR T-cells are activated at a rate proportional to the fraction of CAR that is bound to CD19. Activated CAR T-cells then proliferate and become effector cells. CD8$^+$ effector CAR T-cells can then kill B cells at a rate proportional to the fraction of CARs on CD8$^+$ effector T cells that are bound to CD19. We assume that CD4$^+$ effector CAR T-cells do not kill B cells as we focus only on direct effects (Alizadeh et al., 2023). Effector CAR T-cells either die or become memory cells. Memory CAR T-cells have a longer lifespan than effector cells, but do not participate in B cell killing. The model is intended to describe the initial response to CAR T therapy and therefore does not include any mechanisms for re-activation of memory cells. A diagram of the model reactions is shown in Figure 1. A more detailed description of the model equations is given below, where all states are in units of nmol.

2.3 Cell state equations

Infused CAR T cells are dosed directly into the blood ($T^{CD8}_{inf}(0) = fCD8T \times Dose$, where $fCD8T$ is the fraction of dosed CAR T-cells that are CD8$^+$). These cells can then become activated at a rate k_{act} or die at a rate $k_{death,inf}$. Activated CAR T cells divide at a rate $k_{div} = (2^{ndiv} - 1)/\tau$ to form effector cells at a rate $k_{diff} = 2^{ndiv}/\tau$, where $ndiv$ is the average number of divisions per activated cell and τ is the division time. At a rate of $k_{death,eff}$, a fraction f_{mem} of effector cells become memory cells and the remainder die. Memory T cells die at a rate of $k_{death,mem}$. This leads to the following equation for CD8$^+$ CAR T cells, and similarly for CD4$^+$ CAR T cells.

$$\frac{dT^{CD8}_{inf}}{dt} = -k^{CD8}_{act} T^{CD8}_{inf} - k^{CD8}_{death,inf} T^{CD8}_{inf}$$
$$\frac{dT^{CD8}_{act}}{dt} = k^{CD8}_{act} T^{CD8}_{inf} + k^{CD8}_{div} T^{CD8}_{act} - k^{CD8}_{diff} T^{CD8}_{act}$$
$$\frac{dT^{CD8}_{eff}}{dt} = k^{CD8}_{diff} T^{CD8}_{act} - k^{CD8}_{death,eff} T^{CD8}_{eff}$$
$$\frac{dT^{CD8}_{mem}}{dt} = k^{CD8}_{death,eff} f^{CD8}_{mem} T^{CD8}_{eff} - k^{CD8}_{death,mem} T^{CD8}_{mem}$$

Tumor cells are able to divide at a rate k^{tum}_{div} and be killed by CD8$^+$ effector CAR T cells at a rate $k_{kill} = f_{bound} k_{maxkill}$, where f_{bound} is the fraction of CD8$^+$ CAR that is bound to CD19.

$$\frac{dTumor}{dt} = k^{tum}_{div} Tumor - k_{kill} * Tumor * T^{CD8}_{eff}$$

Endogenous lymphocytes are produced at a zeroth order rate k_{prod} and die at a first order rate $k_{death,endo}$, resulting in the following equation.

$$\frac{dEndo}{dt} = k_{prod} Endo - k_{death,endo} Endo$$

2.4 Receptor equations

In addition to the cell-scale dynamics described above, molecular-scale dynamics are explicitly accounted for in the

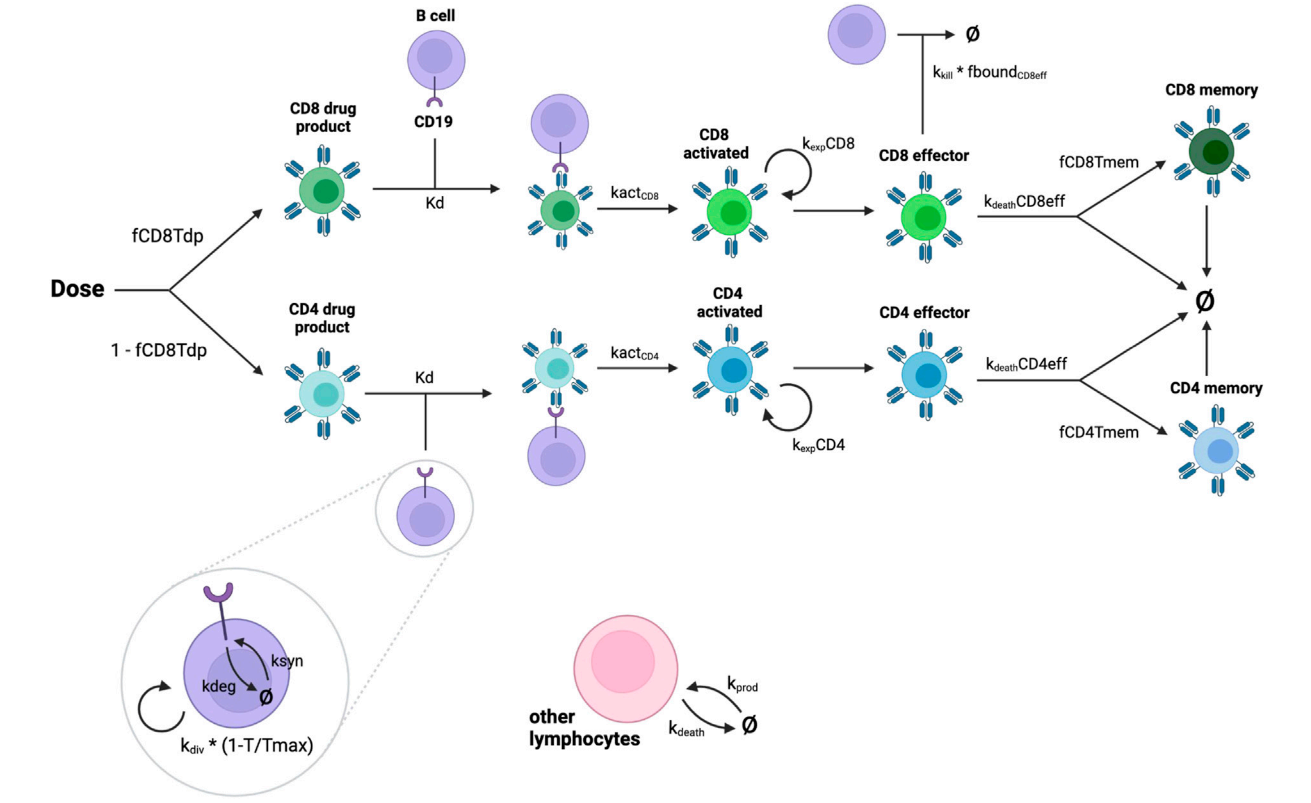


FIGURE 1
Diagram showing interactions represented in the model. CAR-T cells are dosed as part CD8⁺, part CD4⁺. Drug product cells are activated by binding to CD19 on malignant B cells. Activated cells replicate and become effector cells. CD8⁺ effector cells kill B cells. Effector CAR T-cells either die or become memory cells.

model. Receptor equations for CAR and CD19 are written such that the total receptor densities (CAR per T cell and CD19 per tumor cell) remain constant, as determined by a receptor per cell (RPC) parameter. Receptors are synthesized at a rate k_{syn} and internalized at a rate k_{int}. The equation for CD19 is written as follows, accounting for tumor cell division, synthesis and internalization of free CD19, binding/unbinding to CAR on different types of T cells, release from CAR:CD19 complex that is internalized with CAR, release from CAR:CD19 complex when a CAR T cell dies, and tumor cell death.

$$\begin{aligned}\frac{dmAg}{dt} &= k_{div}^{tum} * Tumor * RPC_{CD19} \big/ \left(N_{Av}/1e9\right) + k_{syn}^{CD19} Tumor - k_{int}^{CD19} CD19\\ &\quad - \frac{k_{on}}{V}\left(CAR_{inf}^{CD8} + CAR_{inf}^{CD4} + CAR_{eff}^{CD8} + CAR_{eff}^{CD4} + CAR_{mem}^{CD8} + CAR_{mem}^{CD4}\right)CD19\\ &\quad + k_{off}\left(CAR_{inf}^{CD8}: CD19 + CAR_{inf}^{CD4}: CD19 + CAR_{eff}^{CD8}: CD19\right.\\ &\quad \left. + CAR_{eff}^{CD4}: CD19 + \quad CAR_{mem}^{CD8}: CD19 + CAR_{mem}^{CD4}: CD19\right)\\ &\quad + k_{int}^{CAR}\left(CAR_{inf}^{CD8}: CD19 + CAR_{inf}^{CD4}: CD19 + CAR_{eff}^{CD8}: CD19\right.\\ &\quad \left. + CAR_{eff}^{CD4}: CD19 + \quad CAR_{mem}^{CD8}: CD19 + CAR_{mem}^{CD4}: CD19\right)\\ &\quad + k_{death,inf}^{CD8} CAR_{inf}^{CD8}: CD19 + k_{death,inf}^{CD4} CAR_{inf}^{CD4}: CD19\\ &\quad + k_{death,eff}^{CD8} CAR_{eff}^{CD8}: CD19 + k_{death,eff}^{CD4} CAR_{eff}^{CD4}: CD19\\ &\quad + k_{death,mem}^{CD8} CAR_{mem}^{CD8}: CD19 + k_{death,mem}^{CD4} CAR_{mem}^{CD4}: CD19 - k_{kill} T_{eff}^{CD8} CD19\end{aligned}$$

CARs on infused CAR T cells undergo synthesis and internalization, binding/unbinding with CD19, conversion to an activated state, loss from cell death, and release from CAR:CD19 complex when a tumor cell dies. The equations for CD8⁺ infused CAR and CAR: CD19 complex are shown below; equations for CD4⁺ CAR are similar.

$$\begin{aligned}\frac{dCAR_{inf}^{CD8}}{dt} &= k_{syn}^{CAR,CD8} T_{inf}^{CD8} - k_{int}^{CAR} CAR_{inf}^{CD8} - \frac{k_{on}}{V} CAR_{inf}^{CD8} CD19\\ &\quad + k_{off} CAR_{inf}^{CD8}: CD19 - k_{act}^{CD8} CAR_{inf}^{CD8} + k_{kill} T_{eff}^{CD8} CAR_{inf}^{CD8}: CD19\\ &\quad - k_{death,inf}^{CD8} CAR_{inf}^{CD8}\end{aligned}$$

$$\begin{aligned}\frac{dCAR_{inf}^{CD8}: CD19}{dt} &= \frac{k_{on}}{V} CAR_{inf}^{CD8} CD19 - k_{off} CAR_{inf}^{CD8}: CD19 - k_{int}^{CAR} CAR_{inf}^{CD8}: CD19\\ &\quad - k_{act}^{CD8} CAR_{inf}^{CD8}: CD19 - k_{kill} T_{eff}^{CD8} CAR_{inf}^{CD8}: CD19\\ &\quad - k_{death,inf}^{CD8} CAR_{inf}^{CD8}: CD19\end{aligned}$$

Since activated CAR T cells in the model are an intermediate state for the purposes of expansion and do not interact with the tumor, we do not include binding of activated CARs to CD19 in the model. CARs on activated CAR T cells follow the cellular kinetics (i.e., activation of infused CAR, division, and differentiation) so that CAR per T cell remains constant.

$$\begin{aligned}\frac{dCAR_{act}^{CD8}}{dt} &= k_{act}^{CD8}\left(CAR_{inf}^{CD8} + CAR_{inf}^{CD8}: mAg\right) + k_{div}^{CD8} CAR_{act}^{CD8}\\ &\quad - k_{diff}^{CD8} CAR_{act}^{CD8}\end{aligned}$$

CARs on effector and memory CAR T cells are synthesized and internalized, bind/unbind with CD19, undergo effector-to-memory conversion, are lost through T cell death, and are released from CAR: CD19 complex when a tumor cell dies. These equations are given below for CD8⁺ effector and memory CAR; equations for CD4⁺ CARs are similar.

$$
\begin{aligned}
\frac{dCAR_{eff}^{CD8}}{dt} &= k_{diff}^{CD8} CAR_{act}^{CD8} + k_{syn}^{CAR,CD8} T_{eff}^{CD8} - k_{int}^{CAR} CAR_{eff}^{CD8} - \frac{k_{on}}{V} CAR_{eff}^{CD8} CD19 \\
&+ k_{off} CAR_{eff}^{CD8}: CD19 + k_{kill} T_{eff}^{CD8} CAR_{eff}^{CD8}: CD19 - k_{death,eff}^{CD8} CAR_{eff}^{CD8} \\
\frac{dCAR_{eff}^{CD8}: CD19}{dt} &= \frac{k_{on}}{V} CAR_{eff}^{CD8} CD19 - k_{off} CAR_{eff}^{CD8}: CD19 - k_{int}^{CAR} CAR_{eff}^{CD8}: CD19 \\
&- k_{kill} T_{eff}^{CD8} CAR_{eff}^{CD8}: CD19 \\
\frac{dCAR_{mem}^{CD8}}{dt} &= k_{syn}^{CAR,CD8} T_{mem}^{CD8} - k_{int}^{CAR} CAR_{mem}^{CD8} - \frac{k_{on}}{V} CAR_{mem}^{CD8} CD19 \\
&+ k_{off} CAR_{mem}^{CD8}: CD19 + k_{kill} T_{eff}^{CD8} CAR_{mem}^{CD8}: CD19 \\
&+ f_{mem}^{CD8} k_{death,eff}^{CD8} CAR_{eff}^{CD8} - k_{death,mem}^{CD8} CAR_{mem}^{CD8} \\
\frac{dCAR_{mem}^{CD8}: CD19}{dt} &= \frac{k_{on}}{V} CAR_{mem}^{CD8} CD19 - k_{off} CAR_{mem}^{CD8}: CD19 - k_{int}^{CAR} CAR_{mem}^{CD8}: CD19 \\
&- k_{kill} T_{eff}^{CD8} CAR_{mem}^{CD8}: CD19
\end{aligned}
$$

2.5 Model parameterization

Values for the majority of the model parameters were inferred from literature as described below. The rest of the parameters were fit to individual patient data from Ying et al. (2021), described below.

2.3.1 Patient and tumor properties

The total blood volume was estimated to be 5L, based on the average human adult (Sharma and Sharma, 2018). The concentration of endogenous lymphocytes was assumed to be 10^9 per L. Endogenous lymphocytes were estimated to have an average lifespan of 30 days based on a steady-state assumption and benchmarking to observed T-cell recovery following autologous transplant (Hakim et al., 2005). We assume that 90% of endogenous lymphocytes are depleted by chemotherapy pretreatment prior to CAR T-cell infusion (Ying et al., 2019). The carrying capacity for the number of tumor cells was estimated to be 7×10^{12} based on the maximum tumor volume reported in Press et al. (Press et al., 1993), assuming an average cell diameter of 10μm (Das et al., 1991) and dividing the tumor volume by average cell volume to obtain a maximum number of cells. CD19 expression was estimated to be 5,000 receptors per B cell based on published values for patients with different types of lymphoma (D'Arena et al., 2000; Malik-Chaudhry et al., 2021; Spiegel et al., 2021). The internalization half-life of CD19 was estimated to be 4 h; published data indicates the internalization half-life can be as fast as 30 min in human B-cell lymphoma cell lines (Du et al., 2008) but as slow as 12+ hours in B-cell chronic lymphocytic leukemia patient samples (Sieber et al., 2003).

2.3.2 CAR T-cell properties

The CAR internalization half-life was estimated to be 6 h based on *in vitro* measurements for other CD19-targeting CAR T-cells Li et al. (2020). The mean activation time (that is, the time between binding to antigen and the start of cell proliferation) was estimated to be 18 h for CD8$^+$ CAR T-cells (Henrickson et al., 2008; Cui and Kaech, 2010) and 36 h for CD4$^+$ CAR T-cells (Kaech et al., 2002). Average lifespans for memory CAR T-cells were estimated to be 180 days for CD8$^+$ and 240 days for CD4$^+$ (Borghans et al., 2018). The CD4:CD8 ratio of the CAR T-cells for each patient were taken from Ying et al. (2021), and all infused CAR T-cells were assumed to be viable. We assumed expression levels of 12,700 CARs per T-cell for both CD8$^+$ and CD4$^+$ cells based on a published average estimate for a HMW-MAA-specific CAR on CD8$^+$ T cells (Anikeeva et al., 2021). We assumed that CARs bind to CD19 with an affinity of 1 nM based on reported affinities for high affinity CAR T variants (Jayaraman et al., 2020), with a binding on-rate of 0.001/nM/s.

Remaining model parameters, namely, the number of divisions per T-cell, time per T-cell division, drug product and effector cell lifespan, memory cell fraction, and initial tumor burden, were fit to data as described in the following subsection.

2.3.3 Calibration and benchmarking

Considerable variability in CAR T-cell expansion and efficacy is present in the data. To describe individual variability in CK, the following parameters were fit to individual CK trajectories: initial tumor burden, the number of divisions for activated T cells, and the fraction of effector cells that become memory cells. The time per T cell division and drug product and effector cell life spans were fit globally to all patient data. Optimization was performed using a Python-based trust region optimization method. Additionally, the percentage of B cells out of total cells in the model was calibrated to B cell aplasia data by tuning the number of endogenous lymphocytes in the model within a small, biologically reasonable range such that the mean and range of model outputs captured the general trend observed in the data. The rate of tumor cell division was also tuned to match the observed rebound in B cell aplasia data.

2.4 Model simulation and analysis

The model was implemented and simulations were performed with Applied BioMath's proprietary QSP modeling platform. Analysis and plotting were performed with Python version 3.11.8.

Global sensitivity analysis (GSA) was evaluated using two methods: Sobol indices estimated via the Fourier Amplitude Sensitivity Test (FAST), implemented using SALib (Herman and Usher, 2017; Iwanaga et al., 2022), and partial rank correlation coefficients (PRCC), implemented using Pingouin (Vallat, 2018).

In the GSA, model parameters for which we had individual data or fitted values (body weight, fraction of CD8$^+$ CAR T cells, initial tumor burden, number of CAR T cell divisions, and fraction of memory cells) were varied across the full range of individual values. Where possible, published ranges for individual parameters were used. CD19 expression was varied from 1,500 to 16,825 receptors per cell based on a published range for mantle cell lymphoma (D'Arena et al., 2000). Tumor doubling time was varied from 24 h to 30 days, based on the range reported in Roesch et al. (2014). Binding affinity was varied from 0.32 to 14.3 nM based on the range of values for CD19 CARs reported in Jayaraman et al. (2020). Remaining model parameters were varied 2-fold up and down nominal values. All parameters were sampled from a log-uniform distribution within their respective ranges, with a sample size of 5,000. Simulations were initialized with a 10^6 cells/kg dose. Model outputs considered in the sensitivity analysis were the peak concentration of CAR T-cells (Cmax) and the tumor burden at day 20.

To explore temporal dynamics and explore the impact of tumor characteristics, we performed one-at-a-time scans of tumor division time and CD19 expression per cell. The model was simulated for specific patients as well as for the full patient population using different values of these parameters, while keeping other model parameters fixed.

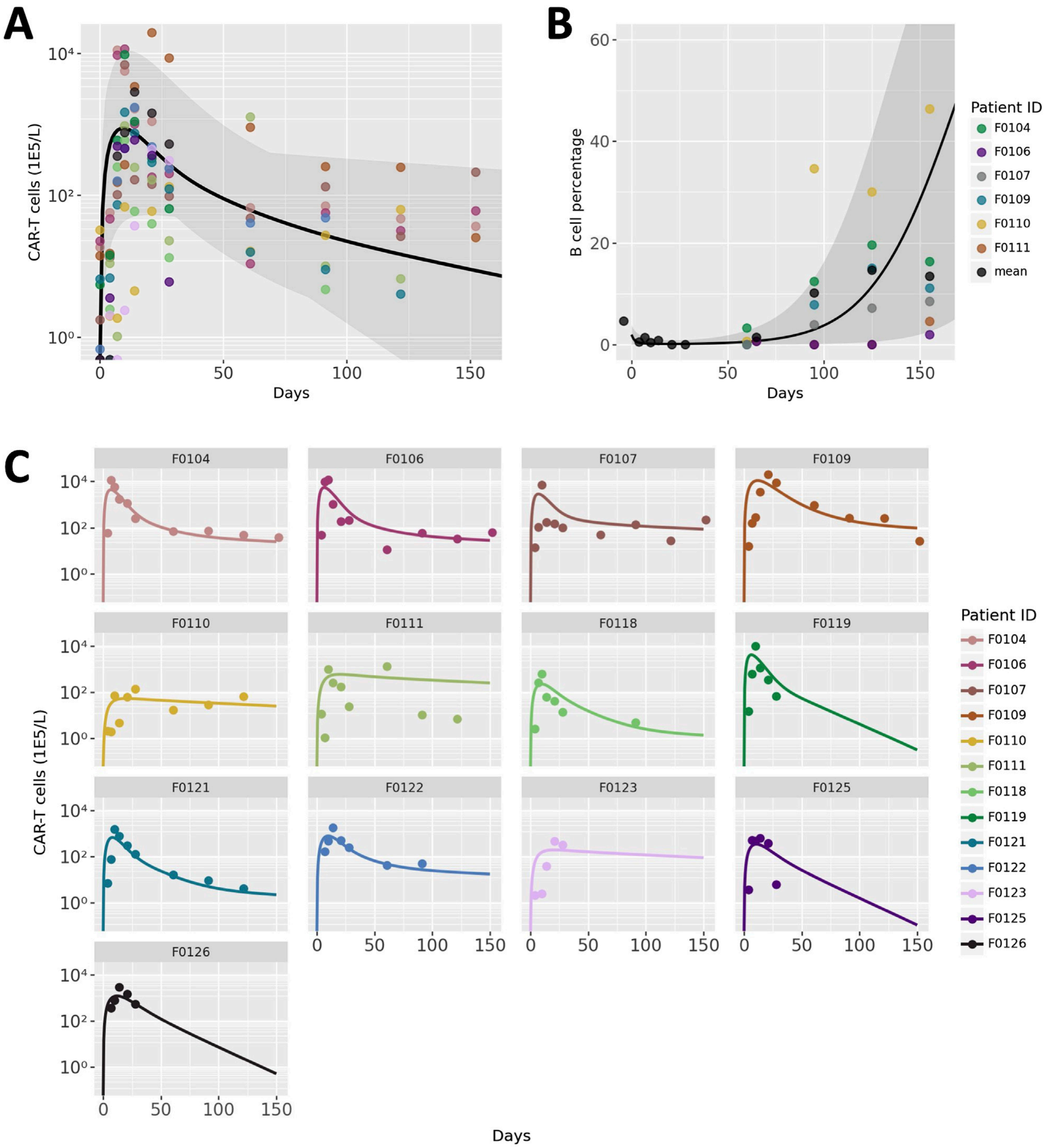


FIGURE 2
Model calibration and benchmarking results. **(A)** Patient population simulations for CAR T-cell CK. Black line indicates average model fit and shaded region represent the full range of individual trajectories. Points represent data, with colors representing different patients. **(B)** Patient population simulations for B cell aplasia. **(C)** Individual patient CAR T-cell CK data and simulations. Each panel represents an individual patient, the ID of which is labeled at the top of each panel.

3 Results

3.1 CK fitting and PD benchmarking

CAR T-cell trajectories vary widely from patient to patient. Our model was developed and calibrated to capture the typical phases of CAR T-cell CK as well as the variability between patients through fitting a combination of patient-specific and global parameter values. Results of optimization of CAR T-cell concentration to clinical data are shown in Figure 2. Figure 2A shows the average trajectory and full range across all 13 patients and Figure 2C shows each fitted patient simulation and data. The model adequately describes the

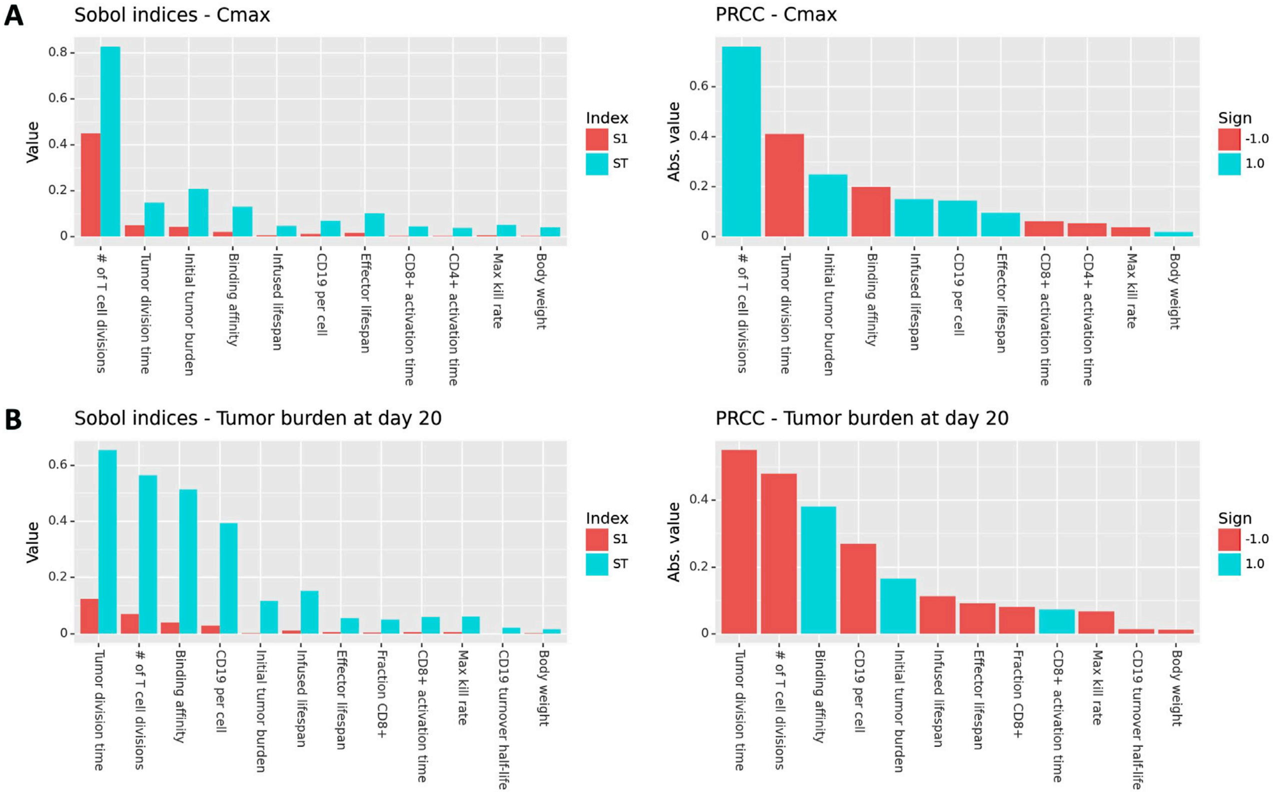


FIGURE 3
Global sensitivity results (Sobol indices and PRCC) for Cmax and tumor burden at day 20 post-treatment. Only parameters with a *p*-value less than 0.05 and that rank in the top ten for at least one measure of sensitivity are shown.

overall behavior of the data despite the significant variability between patients as well as within each patient data set. The full table of final parameter values can be found in the Supplementary Material.

Endogenous lymphocyte concentration and tumor doubling time were hand-tuned to a small degree to match measurements of B cells as a percentage of total lymphocytes, a measure of the efficacy of the CAR T-cells. Due to challenges with digitization, B cell aplasia data from only 6 of the 13 patients were distinguishable and are shown in Figure 2. The average and range of model simulations for all patient parameterizations capture the general trend of the data well and spans the variability between patients.

3.2 Sensitivity analysis

To explore the impact of model parameters on Cmax and efficacy, we performed Global Sensitivity Analysis (GSA). Results for Cmax are shown in Figure 3A. Sobol indices (first order and total order) and PRCC values are shown for all model parameters that had a *p*-value less than 0.05 and ranked in the top ten parameters for at least one measure of global sensitivity. The number of CAR T-cell divisions upon activation contributes to more than 80% of the variability in Cmax, which is a far greater contribution than any of the other parameters. The next most influential parameters are tumor growth rate, initial tumor burden, and mAb-CD19 binding affinity, which drive expansion through CAR-antigen interactions. CAR T cell life spans and CD19 expression are also influential. The ordering of parameters is roughly consistent between first order Sobol index, total order Sobol index, and PRCC. However, total order Sobol indices are generally at least two-fold larger than first order Sobol indices, indicating that there are interactions between parameters.

GSA results for tumor burden at day 20 are shown in Figure 3B. The most influential parameters are the tumor division time, number of T cell divisions, binding affinity, initial tumor burden, and CD19 expression is also influential. Tumor- and binding-related parameters are comparably influential on efficacy as the number of T cell divisions. This is in contrast to the results for Cmax, where the number of T cell divisions was by far the most influential parameter. This indicates that while expansion and efficacy are often correlated, patient properties such as tumor growth rate, initial tumor burden, and CD19 expression are more important for driving efficacy than they are for driving expansion. This is because CAR-CD19 interactions are required for both expansion and tumor cell killing.

3.3 Effects of tumor properties on CAR T expansion and efficacy

To investigate potential mechanisms related to patient to patient variability in response, we evaluated the effects of B cell division time and CD19 expression on B cells. These two parameters, which were informed by literature and not varied in Figure 2C were shown to be influential parameters by the GSA. We first focus on two patient

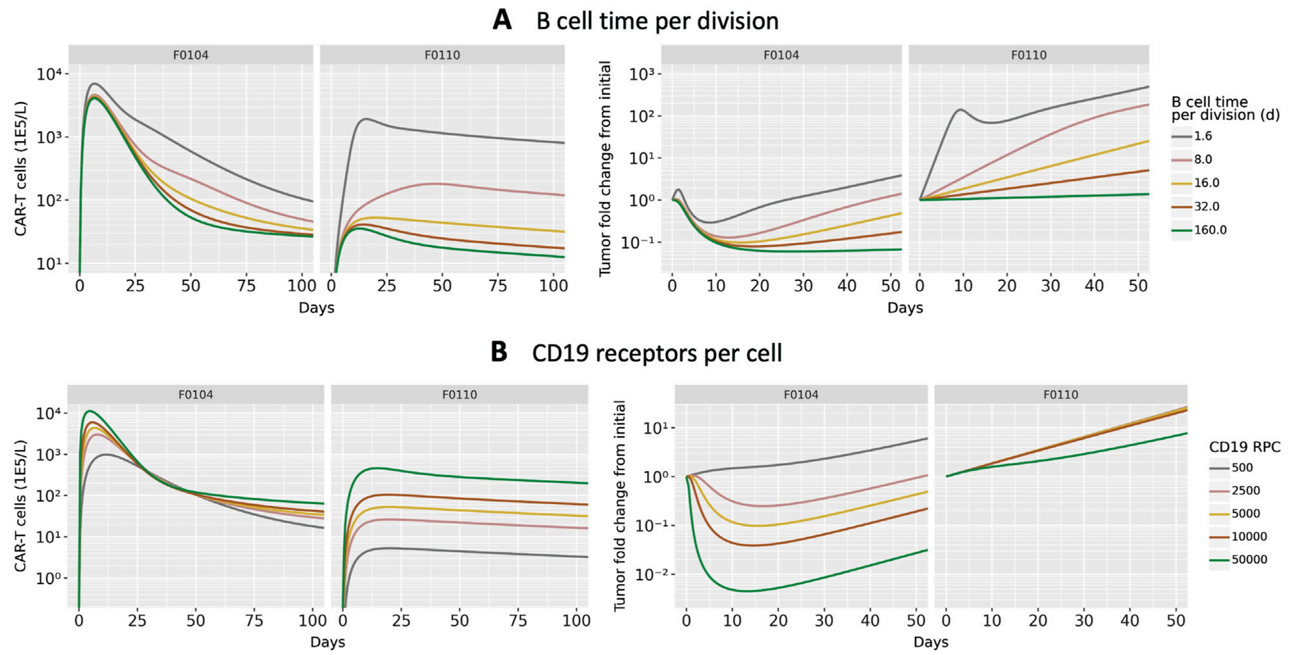


FIGURE 4
Results of scanning key model parameters representing patient characteristics for two individual patient parameterization. Parameters were scanned up to ~10x above and below their nominal parameterization for patient parameterizations F0104 and F0110. Simulations of CAR T-cell concentration and total B cell fold change from initial are shown for scans of **(A)** B cell time per division and **(B)** CD19 receptors per cell.

parameterizations (F0104 and F0110) which had distinct CK and tumor growth profiles. F0104 has a typical CK profile consisting of expansion, contraction, and persistence, paired with a clear reduction in tumor growth, while F0110 had continued tumor growth and less defined expansion and contraction phases. Figure 4 shows simulations of CAR T-cell concentration and tumor dynamics for these two patients, scanning over both parameters. Parameters are varied 10-fold up and down from nominal values to explore a wide range of system behaviors.

Scanning over B cell division time, shown in Figure 4A, revealed qualitatively different behavior between the two patients. For patient F0104, the model predicts that a B cell division time corresponds to a more gradual contraction, resulting in a greater concentration of CAR T-cells over time. The division time does not significantly impact the Cmax. However, for patient F0110, the slope of the contraction phase is relatively consistent across division times but the Cmax increases with faster division times. For both patients, the greater expansion of CAR T-cells is not sufficient to reverse tumor cell growth. A faster B cell division time results in more tumor growth regardless of CAR T-cell concentration for the parameter range scanned. Within the first 10–15 days, there is an acute reduction in tumor cells in response to initial CAR T-cell expansion for the fastest tumor cell division time, 1.6 days. However, this effect is transient and the faster B cell division time results in faster rebound of the tumor. For the slower tumor cell division times, CAR T expansion does reduce the tumor size; this combined with the generally slower tumor growth results in slow tumor growth in the longer term.

Figure 4B shows the results of varying CD19 expression on B cells. Higher CD19 expression leads to additional binding to CAR T-cells and subsequent activation, increasing CAR T-cell expansion. This looks different for each of the patient parameterizations; simulated CK for patient F0110 shows greater sensitivity to CD19 expression compared to that of patient F0104. For F0104, higher CD19 expression leads to faster expansion and faster contraction, causing a sharper peak in the CAR T CK. Higher CD19 expression also leads to greater long-term persistence of CAR T-cells. Patient F0110 exhibits greater expansion and persistence with varying levels of CD19 expression, with no evident contraction phase.

Examining the individually fit parameters for F0104 and F0110 sheds light on the unique behaviors of both the CAR T-cells and tumor cells between patient simulations. Patient F0104 has a smaller fraction of effector CAR T-cells that become memory cells, a larger initial tumor burden, and a slightly higher number of CAR T-cell divisions upon activation compared to patient F0110. This leads to greater expansion (and therefore greater efficacy) of the CAR T-cells for patient F0104, but potentially less persistence. For patient F0110, the lower expansion and greater memory cell formation leads to no clear contraction phase in the CK. The corresponding tumor growth curves show no impact of treatment except a small reduction in tumor growth rate at the highest receptor expression level scanned.

To assess the behavior of CAR T-cells and tumor growth on a population level, the parameter value for either B cell division time or CD19 expression was updated one at a time for each patient. These parameters were varied ranges described in literature: Roesch et al. (2014) report NHL doubling times

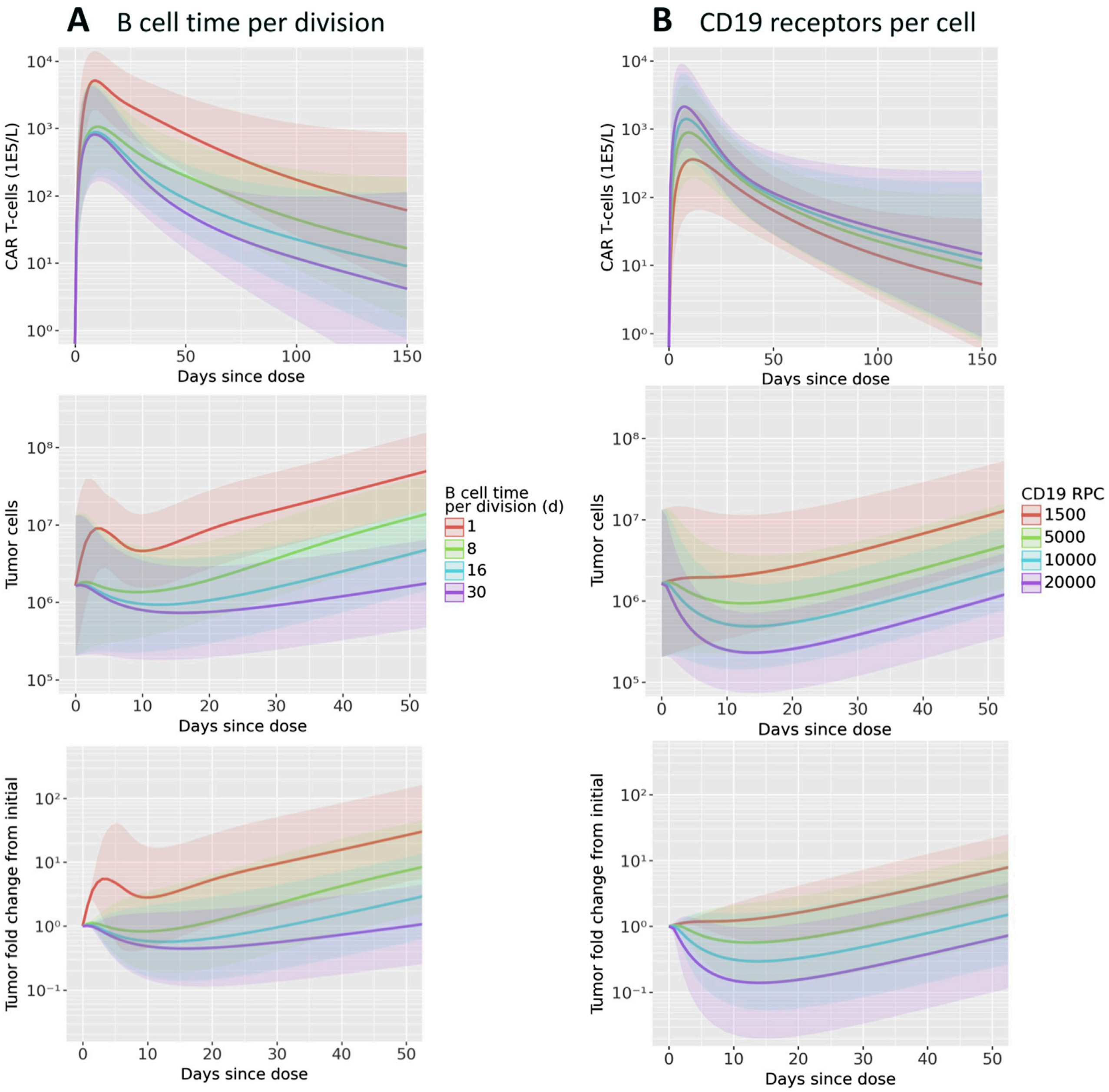


FIGURE 5
Results of scanning key model parameters representing patient characteristics across all patient parameterizations. Parameters were scanned across ranges consistent with values reported in the literature. Simulations of CAR T-cell concentration, total B cells, and B cell fold change from initial are shown for scans of **(A)** B cell time per division and **(B)** CD19 receptors per cell. Mean and one standard deviation of all patient results are shown.

from 24 h to 30 days, and D'Arena et al. (2000) report a standard deviation for CD19 expression of about 1/3 the mean value. The minimum value for CD19 expression reported was much lower (10-fold lower than the mean value), which we also include in the parameter scan. Figure 5 shows the mean and standard deviation across all patients for CK, tumor cell count, and tumor fold change. Overall, the same patterns described above in the patient-specific scans hold true: faster B cell division times yield more CAR T-cell expansion and greater tumor growth, and higher CD19 expression leads to more CAR T-cells and improved tumor cell killing. Within the physiological ranges tested, B cell division time has an impact on both CK and tumor cell growth by close to an order of magnitude, on average. Notably, the rate of tumor regrowth is similar for all tumor doubling times, indicating that the increased persistence of CAR T-cells does counteract the increased tumor growth.

The range of reported CD19 expression is quite varied, and the model predicts that this parameter could have a significant impact on treatment efficacy. Between the maximum and minimum values scanned, within the range of reported values, there is about an order of magnitude difference in the CAR T cell Cmax. Furthermore, for the lowest CD19 RPC, there is essentially no tumor growth inhibition. The three higher RPC values do show inhibition, with a reduction from baseline of up to 10x.

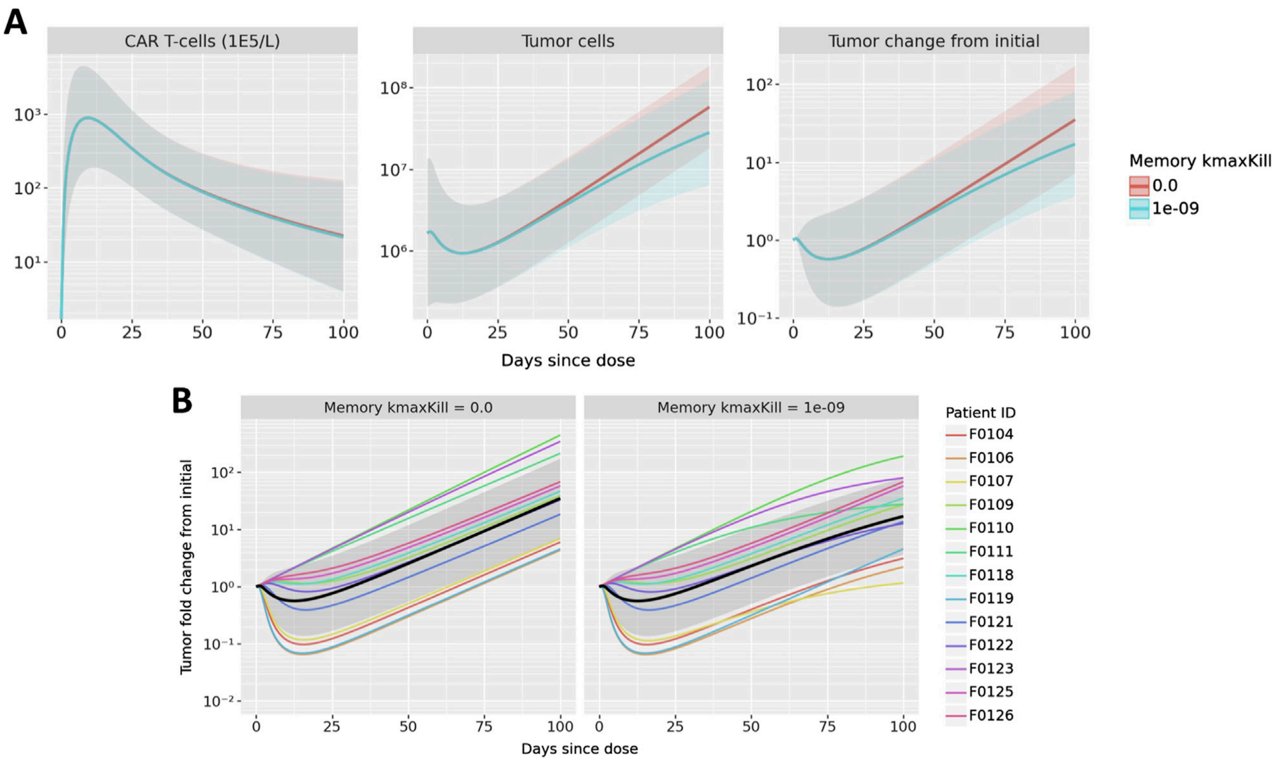


FIGURE 6
Results of exploring the impact of memory cell killing. Simulations in which memory cells have the same killing capacity as active cells are compared to the nominal case in which memory cells do not kill tumor cells. Simulations of CAR T-cell concentration, total B cells, and B cell fold change from initial are shown in **(A)** where mean and one standard deviation of all patient results are compared. B cell fold change from initial is shown in **(B)**; multicolored lines represent individual patient trajectories, black curve represents the mean across all patients, and the gray region represents one standard deviation.

3.4 Exploratory analysis: memory cell killing

To show how the model can explore questions about both individual and population-level dynamics, we performed simulations to understand the potential impact of memory cell killing. In the nominal simulations, we assume that memory cells do not kill tumor cells. For this analysis, we compare the nominal simulations against those in which memory cell killing has the same killing capacity as activated cells. Figure 6 shows the results for both the population level and individual trajectories.

First observing the population-level dynamics in Figure 6A, the impact on memory cell killing is observed only in the tumor, not in CAR T CK. Furthermore, the model predicts that any difference is observed after 50 days. This makes sense due to the delayed appearance of memory cells and the subsequent growth of the memory cell population - thus, memory cell killing is predicted to have a small overall impact on reducing tumor growth during the terminal phase of CAR T expansion. Since the exact memory cell populations may vary from patient to patient, the impact of memory cell killing may also be observed on a patient level, shown in Figure 6B. It is evident that some patients show little impact of memory cell killing, sch as F0125 and F0126. On the other hand, tumor growth in patients F0107, F0111, F0123 nearly plateaus as compared to the nominal parameterization which has linear growth.

Although there is insufficient data to inform the true activity of memory CAR-T cells in the model, this hypothetical analysis shows the ability of the model to differentiate the impact of treatment on individual patients as compared to a population-level impact. In these simulations, a moderate population-level effect was the result of an aggregated variety of patient effects, from no impact to a signficiant impact. Furthermore, the model shows in what populations and at what times the impact of these changes might be observed.

4 Discussion

Our mechanistic modeling approach incorporating molecular-scale and cell-scale dynamics successfully captured CAR T CK-PD and revealed key system behaviors. Mechanistic modeling is necessary to capture the interplay of target engagement, T cell expansion, and tumor cell killing. CAR T-cell therapy is distinct from other therapeutics in that CK and PD are inter-dependent. This dependency is demonstrated in our model by the sensitivity of Cmax to tumor and binding parameters.

Global sensitivity analysis revealed that both drug-specific and patient-specific properties can potentially explain variability in response to CAR T therapy. The most influential drug-specific properties are the number of divisions per activated CAR T-cell and the binding of the CAR for CD19. The number of divisions for activated cells is the most influential factor for peak CAR T-cell expansion, and was also highly influential for tumor killing. This number of divisions could potentially be increased through further engineering or refining of manufacturing processes for the CAR T product, for example, through selection for naive cells (Arcangeli et al., 2022). Importantly, while we classify this as a drug-specific property, this could be variable across patients since the CAR T-cells are manufactured from the patients' own cells. Thus, individual variability in the number of T cell divisions can also contribute to observed variability in CK and efficacy. Binding affinity also impacted both CAR T-cell expansion and efficacy, which could be improved in engineering of the CAR.

While our modeling suggests that CAR T expansion is driven primarily by number of divisions, global sensitivity analysis shows that tumor properties such as CD19 expression and growth rate are comparatively more influential in driving efficacy. Tumor growth rate was also highly influential on CK. This demonstrates two things: (1) while expansion often correlates with efficacy, expansion itself is not necessarily sufficient for tumor shrinkage, and (2) variability in patient characteristics will lead to significant variability in both exposure and response. Modeling provides insight into this variability and can be used to inform patient, target, and indication selection.

In individual- and population-level model simulations, we observed that although a faster tumor growth rate corresponds to increased CAR T-cell expansion and distinct CK profiles, this increased expansion is often not enough to control the faster-growing tumor. This implies that drug characteristics may need to be modified in order to target more aggressive tumors. Notably, while there is little to no predicted tumor shrinkage for this CAR T with faster growing tumors for most patients, treatment is still effective in slowing tumor growth both short- and long-term, providing a benefit to patients. Increased target expression drives both increased expansion and stronger tumor killing. Patients with low target expression may be poor candidates for this type of treatment due to poor expansion and little anti-tumor activity, leading to lack of response. This also suggests that target expression should be a key consideration in both target and indication selection, while balancing toxicity concerns. Furthermore, the power of individualized parameterizations of the model was demonstrated in the memory cell killing exploratory analysis. Although the population-level simulation showed a small overall reduction in tumor growth, some patient trajectories showed signficant reduction while others showed nearly no impact. Although this was a hypothetical exploration due insufficient data, these simulations demonstrate that modeling can have a large impact on understanding individual patient dynamics.

In the future, this model and analysis could help drive decisions in CAR T-cell design, manufacturing, patient selection, patient-specific dose selection, and efficacious dose selection for novel CAR Ts. This model could be further refined by adding other T cell phenotypes, cytokines, immune cells types, and additional reactions such as re-activation of memory cells. One key limitation of the current work is lack of direct measurements of tumor burden over time to inform efficacy. Rather, we relied of B cell aplasia data and assumptions about the native immune population to estimate tumor reduction. Additional efficacy data would help to better constrain the model and may allow for individualized efficacy modeling. Additional patient-specific data such as CD19 expression could enable individualized predictions of efficacy/response through a digital twin approach. Another limitation of this model is that it does not account for effects of CD4$^+$ T cells on tumor cell killing. This model could also be extended to study other targets and indications, including solid tumors for which there are currently no approved CAR T-cell therapies.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7518381/.

Ethics statement

Ethical approval was not required for the study involving humans in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

Author contributions

SM: Conceptualization, Formal Analysis, Visualization, Writing–original draft. SG: Data curation, Investigation, Writing–original draft. KS: Conceptualization, Supervision, Writing–review and editing. MR: Formal Analysis, Methodology, Supervision, Visualization, Writing–original draft.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

We thank Raibatak Das for developing the first iteration of this model.

Conflict of interest

Authors SM, SG, MR are employed by Applied BioMath. Author KS is employed by Differentia Bio.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fsysb.2024.1380018/full#supplementary-material

References

Alizadeh, D., Wang, D., and Brown, C. E. (2023). Uncovering the role of cd4+ car t cells in cancer immunotherapy. *Cancer Res.* 83, 2813–2815. doi:10.1158/0008-5472.CAN-23-1948

Anikeeva, N., Panteleev, S., Mazzanti, N., Terai, M., Sato, T., and Sykulev, Y. (2021). Efficient killing of tumor cells by CAR-T cells requires greater number of engaged CARs than TCRs. *J. Biol. Chem.* 297, 101033. doi:10.1016/j.jbc.2021.101033

Arcangeli, S., Bove, C., Mezzanotte, C., Camisa, B., Falcone, L., Manfredi, F., et al. (2022). Car t cell manufacturing from naive/stem memory t lymphocytes enhances antitumor responses while curtailing cytokine release syndrome. *J. Clin. investigation* 132, e150807. doi:10.1172/JCI150807

Borghans, J. A., Tesselaar, K., and de Boer, R. J. (2018). Current best estimates for the average lifespans of mouse and human leukocytes: reviewing two decades of deuterium-labeling experiments. *Immunol. Rev.* 285, 233–248. doi:10.1111/imr.12693

Chaudhury, A., Zhu, X., Chu, L., Goliaei, A., June, C. H., Kearns, J. D., et al. (2020). Chimeric antigen receptor T cell therapies: a review of cellular kinetic-pharmacodynamic modeling approaches. *J. Clin. Pharmacol.* 60, S147-S159–S159. doi:10.1002/jcph.1691

Chen, Y.-J., Abila, B., and Mostafa Kamel, Y. (2023). CAR-T: what is next? *Cancers (Basel)* 15, 663. doi:10.3390/cancers15030663

Cui, W., and Kaech, S. M. (2010). Generation of effector CD8+ T cells and their conversion to memory T cells. *Immunol. Rev.* 236, 151–166. doi:10.1111/j.1600-065X.2010.00926.x

D'Arena, G., Musto, P., Cascavilla, N., Dell'Olio, M., Di Renzo, N., and Carotenuto, M. (2000). Quantitative flow cytometry for the differential diagnosis of leukemic B-cell chronic lymphoproliferative disorders. *Am. J. Hematol.* 64, 275–281. doi:10.1002/1096-8652(200008)64:4<275::aid-ajh7>3.0.co;2-y

Das, D. K., Gupta, S. K., Datta, B. N., and Sharma, S. C. (1991). Working formulation of the non-Hodgkin's lymphomas: a study of cell size and mitotic indices in cytologic subtypes. *Diagn. Cytopathol.* 7, 499–503. doi:10.1002/dc.2840070511

Du, X., Beers, R., FitzGerald, D. J., and Pastan, I. (2008). Differential cellular internalization of anti-CD19 and -CD22 immunotoxins results in different cytotoxic activity. *Cancer Res.* 68, 6300–6305. doi:10.1158/0008-5472.CAN-08-0461

Hakim, F. T., Memon, S. A., Cepeda, R., Jones, E. C., Chow, C. K., Kasten-Sportes, C., et al. (2005). Age-dependent incidence, time course, and consequences of thymic renewal in adults. *J. Clin. investigation* 115, 930–939. doi:10.1172/JCI22492

Hardiansyah, D., and Ng, C. M. (2019). Quantitative systems pharmacology model of chimeric antigen receptor T-cell therapy. *Clin. Transl. Sci.* 12, 343–349. doi:10.1111/cts.12636

Henrickson, S. E., Mempel, T. R., Mazo, I. B., Liu, B., Artyomov, M. N., Zheng, H., et al. (2008). T cell sensing of antigen dose governs interactive behavior with dendritic cells and sets a threshold for T cell activation. *Nat. Immunol.* 9, 282–291. doi:10.1038/ni1559

Herman, J., and Usher, W. (2017). SALib: an open-source python library for sensitivity analysis. *J. Open Source Softw.* 2, 97. doi:10.21105/joss.00097

Iwanaga, T., Usher, W., and Herman, J. (2022). Toward SALib 2.0: advancing the accessibility and interpretability of global sensitivity analyses. *Socio-Environmental Syst. Model.* 4, 18155. doi:10.18174/sesmo.18155

Jayaraman, J., Mellody, M. P., Hou, A. J., Desai, R. P., Fung, A. W., Pham, A. H. T., et al. (2020). CAR-T design: elements and their synergistic function. *EBioMedicine* 58, 102931. doi:10.1016/j.ebiom.2020.102931

Kaech, S. M., Wherry, E. J., and Ahmed, R. (2002). Effector and memory T-cell differentiation: implications for vaccine development. *Nat. Rev. Immunol.* 2, 251–262. doi:10.1038/nri778

Li, W., Qiu, S., Chen, J., Jiang, S., Chen, W., Jiang, J., et al. (2020). Chimeric antigen receptor designed to prevent ubiquitination and downregulation showed durable antitumor efficacy. *Immunity* 53, 456–470. doi:10.1016/j.immuni.2020.07.011

Malik-Chaudhry, H. K., Prabhakar, K., Ugamraj, H. S., Boudreau, A. A., Buelow, B., Dang, K., et al. (2021). TNB-486 induces potent tumor cell cytotoxicity coupled with low cytokine release in preclinical models of B-NHL. *MAbs.* 13 (1), e1890411-1–e1890411-15. doi:10.1080/19420862.2021.1890411

Melenhorst, J. J., Chen, G. M., Wang, M., Porter, D. L., Chen, C., Collins, M. A., et al. (2022). Decade-long leukaemia remissions with persistence of CD4+ CAR T cells. *Nature* 602, 503–509. doi:10.1038/s41586-021-04390-6

Nukala, U., Rodriguez Messan, M., Yogurtcu, O. N., Wang, X., and Yang, H. (2021). A systematic review of the efforts and hindrances of modeling and simulation of CAR T-cell therapy. *AAPS J.* 23, 52–20. doi:10.1208/s12248-021-00579-9

Owens, K., and Bozic, I. (2021). Modeling CAR T-cell therapy with patient preconditioning. *Bull. Math. Biol.* 83, 42–36. doi:10.1007/s11538-021-00869-5

Press, O. W., Eary, J. F., Appelbaum, F. R., Martin, P. J., Badger, C. C., Nelp, W. B., et al. (1993). Radiolabeled-antibody therapy of B-cell lymphoma with autologous bone marrow support. *N. Engl. J. Med.* 329, 1219–1224. doi:10.1056/NEJM199310213291702

Roesch, K., Hasenclever, D., and Scholz, M. (2014). Modelling lymphoma therapy and outcome. *Bull. Math. Biol.* 76, 401–430. doi:10.1007/s11538-013-9925-3

Rohatgi, A. (2022). *Webplotdigitizer*.

Salem, A. M., Mugundu, G. M., and Singh, A. P. (2023). Development of a multiscale mechanistic modeling framework integrating differential cellular kinetics of car t-cell subsets and immunophenotypes in cancer patients. *CPT Pharmacometrics and Syst. Pharmacol.* 12, 1285–1304. doi:10.1002/psp4.13009

Sermer, D., Batlevi, C., Palomba, M. L., Shah, G., Lin, R. J., Perales, M.-A., et al. (2020). Outcomes in patients with DLBCL treated with commercial CAR T cells compared with alternate therapies. *Blood Adv.* 4, 4669–4678. doi:10.1182/bloodadvances.2020002118

Sharma, R., and Sharma, S. (2018). Physiology, blood volume. *StatPearls*.

Sieber, T., Schoeler, D., Ringel, F., Pascu, M., and Schriever, F. (2003). Selective internalization of monoclonal antibodies by B-cell chronic lymphocytic leukaemia cells. *Br. J. Haematol.* 121, 458–461. doi:10.1046/j.1365-2141.2003.04305.x

Singh, A. P., Zheng, X., Lin-Schmidt, X., Chen, W., Carpenter, T. J., Zong, A., et al. (2020). Development of a quantitative relationship between CAR-affinity, antigen abundance, tumor cell depletion and CAR-T cell expansion using a multiscale systems PK-PD model. *MAbs.* 12 (1), e1688616-1–e1688616-21. doi:10.1080/19420862.2019.1688616

Spiegel, J. Y., Patel, S., Muffly, L., Hossain, N. M., Oak, J., Baird, J. H., et al. (2021). Car t cells with dual targeting of cd19 and cd22 in adult patients with recurrent or refractory b cell malignancies: a phase 1 trial. *Nat. Med.* 27, 1419–1431. doi:10.1038/s41591-021-01436-0

Stein, A. M., Grupp, S. A., Levine, J. E., Laetsch, T. W., Pulsipher, M. A., Boyer, M. W., et al. (2019). Tisagenlecleucel model-based cellular kinetic analysis of chimeric antigen receptor–T cells. *CPT pharmacometrics and Syst. Pharmacol.* 8, 285–295. doi:10.1002/psp4.12388

Vallat, R. (2018). Pingouin: statistics in python. *J. Open Source Softw.* 3, 1026. doi:10.21105/joss.01026

Ying, Z., He, T., Wang, X., Zheng, W., Lin, N., Tu, M., et al. (2021). Distribution of chimeric antigen receptor-modified T cells against CD19 in B-cell malignancies. *BMC cancer* 21, 198–207. doi:10.1186/s12885-021-07934-1

Ying, Z., Huang, X. F., Xiang, X., Liu, Y., Kang, X., Song, Y., et al. (2019). A safe and potent anti-CD19 CAR T cell therapy. *Nat. Med.* 25, 947–953. doi:10.1038/s41591-019-0421-7

TYPE Methods
PUBLISHED 12 September 2024
DOI 10.3389/fsysb.2024.1444912

Building virtual patients using simulation-based inference

Nathalie Paul[1], Venetia Karamitsou[2], Clemens Giegerich[2], Afshin Sadeghi[1], Moritz Lücke[1], Britta Wagenhuber[2], Alexander Kister[1,3] and Markus Rehberg[2]*

[1]Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS, Sankt Augustin, Germany, [2]Sanofi R&D, Translational Disease Modeling, Frankfurt, Germany, [3]eScience Division (S.3), Federal Institute for Materials Research and Testing, Berlin, Germany

Check for updates

OPEN ACCESS

EDITED BY
Gary An,
University of Vermont, United States

REVIEWED BY
David Phillip Nickerson,
The University of Auckland, New Zealand
Wilhelm Huisinga,
University of Potsdam, Germany
Thomas Kerbusch,
Certara, United States

*CORRESPONDENCE
Markus Rehberg,
✉ markus.rehberg@sanofi.com

RECEIVED 06 June 2024
ACCEPTED 29 August 2024
PUBLISHED 12 September 2024

CITATION
Paul N, Karamitsou V, Giegerich C, Sadeghi A, Lücke M, Wagenhuber B, Kister A and Rehberg M (2024) Building virtual patients using simulation-based inference.
Front. Syst. Biol. 4:1444912.
doi: 10.3389/fsysb.2024.1444912

In the context of *in silico* clinical trials, mechanistic computer models for pathophysiology and pharmacology (here Quantitative Systems Pharmacology models, QSP) can greatly support the decision making for drug candidates and elucidate the (potential) response of patients to existing and novel treatments. These models are built on disease mechanisms and then parametrized using (clinical study) data. Clinical variability among patients is represented by alternative model parameterizations, called virtual patients. Despite the complexity of disease modeling itself, using individual patient data to build these virtual patients is particularly challenging given the high-dimensional, potentially sparse and noisy clinical trial data. In this work, we investigate the applicability of simulation-based inference (SBI), an advanced probabilistic machine learning approach, for virtual patient generation from individual patient data and we develop and evaluate the concept of nearest patient fits (SBI NPF), which further enhances the fitting performance. At the example of rheumatoid arthritis where prediction of treatment response is notoriously difficult, our experiments demonstrate that the SBI approaches can capture large inter-patient variability in clinical data and can compete with standard fitting methods in the field. Moreover, since SBI learns a probability distribution over the virtual patient parametrization, it naturally provides the probability for alternative parametrizations. The learned distributions allow us to generate highly probable alternative virtual patient populations for rheumatoid arthritis, which could potentially enhance the assessment of drug candidates if used for *in silico* trials.

KEYWORDS

virtual patients, QSP modeling, individual patient fitting, machine learning, artificial intelligence, simulation-based inference

1 Introduction

Quantitative Systems Pharmacology (QSP) models provide mechanistic insights into the dynamic interactions between complex pathophysiological reactions and pharmacological interventions, which yield dynamic responses of protein biomarkers and clinical endpoints (Bradshaw et al., 2019; Sorger et al., 2011). Different model parameterizations can represent variability in disease mechanisms and thereby capture a large range of patients and endotypes. An individual parameter set θ for the QSP model is here denoted as a virtual patient and determines its biomarker and disease score response to a specific treatment ($QSP(\theta)$). Finding and identifying these parameterizations θ within the disease biology network allows us to model and assess virtual patients individually and predict their disease progression and treatment response to novel drugs.

The generation of virtual patients is either driven by hypothesis, to capture, for example, high-level features of responses observed in the clinic where no data is available (Friedrich, 2016), or driven by collected clinical outcome data. Often, such data comes as summary statistics over the patient population and requires the use of parameter searches and parameter weighting methods (e.g., prevalence weighting (Klinke, 2008; Howell et al., 2012; Schmidt et al., 2013; Allen et al., 2016)). Ideally, the clinical data includes individual patient-level data which makes an explicit fit of real patients possible (Björnsson et al., 2019; Allen et al., 2021; Luo et al., 2022). The latter approach requires data preparation, high performance fitting algorithms and efficient computation pipelines to achieve a robust quantitative representation of several hundreds of patients given noisy, locally sparse and high-dimensional individual clinical data. Since the results of the individual patient fits are used to guide drug development decisions, we here seek a broad understanding of virtual patients in terms of how likely it is that they indeed describe the real patient data.

Integrating machine learning (ML) approaches to QSP modeling is a powerful strategy to tackle the computational challenges associated with mechanistic modeling of such complex biological systems (reviewed extensively in (Aghamiri et al., 2022) and (Zhang et al., 2022)). ML has been successfully implemented in parameter estimation (Wajima et al., 2009), model-order reduction (Derbalah et al., 2022), virtual patient generation (Rieger et al., 2018; Parikh et al., 2022) and the assessment of stochastic effects (McComb et al., 2022).

Here, we investigate the applicability of a novel ML approach for building virtual patients. We use simulation-based inference (SBI) that has, to the best of our knowledge, not been applied to such large QSP models yet. As an example, we use a proprietary QSP model for rheumatoid arthritis and fit it to individual patient data where patients have been treated with an anti-TNF drug. SBI approaches are advanced ML techniques for inferring a parameterization of a simulator given prior knowledge and empirical data (Lueckmann et al., 2021). While classic fitting algorithms output a point estimate for a parametrization (Byrd et al., 2000; Egea et al., 2009), SBI produces a probability distribution over the parametrization space, yielding a much more informative result. Prior knowledge in terms of an expert-designed reference patient parametrization is used to build an initial belief about the desired probability distribution. The belief then gets updated based on clinical data observations. The resulting learned probability distribution provides the probability of specific patient parameterizations and thus technically makes it possible to not only discover a single patient parameterization of high probability but multiple ones. The probability distribution could hence be used to generate new realistic virtual patients during *in silico* trials that may participate in future studies.

In a second step, we propose to leverage knowledge from already built virtual patients (from the same population) to enhance the performance of the algorithm. Instead of using the reference parametrization as prior knowledge for a new patient fit, we use an already learned parametrization of a similar patient. The so-called nearest patient fit (SBI NPF) thus starts from an improved initial belief. We expect a more consistent fit among patients of similar type, which would support an easier identification of virtual patient subgroups. To identify a similar patient, we define a vicinity criterion on the clinical data.

2 Methods

2.1 Clinical data

The individual patient data was taken from the MONARCH study [NCT02332590, anti-TNF study arm: n = 155 (Burmester et al., 2017; Gabay et al., 2017; Gaby et al., 2020)]. A total of 133 patients were used for individual patient fitting. Individual patients were fitted amongst others to cell counts (lymphocytes, macrophages), blood protein biomarkers (CRP, MMP-3, RANKL, OPG, OC, CXCL13, sICAM-1 and IL6) as well as clinical readouts (SJC28, TJC28, DAS28-CRP). The data was taken at baseline until 24 weeks of treatment with up to eight measurement time points. Population statistics of the data is available at https://zenodo.org/doi/10.5281/zenodo.12808208.

2.2 QSP model and simulation

The QSP model (built in SimBiology®, https://mathworks.com/products/simbiology.html) contains 96 ordinary differential equations (ODE) definitions, 260 reactions, 100 initial and repeated assignments and over 1,000 literature references for parameterization of 450 parameters. For simulation in Julia (version 1.8.3, https://julialang.org/), the Julia Package Sundials (package that interfaces SUNDIALS 5.2.0 library, https://github.com/SciML/Sundials.jl) with the solver CVODE_BDF() and absolute and relative tolerances of 1E-6 were used to solve the ODE system. The QSP model is shown in Supplementary Figure S1 (supplement).

The reference parametrization of the QSP model is a pre-implemented solution to an anti-TNF treatment based on various clinical, *in vitro* and animal *in vivo* experiments ranging from mechanistic to clinical outcome data (Biesemann et al., 2023).

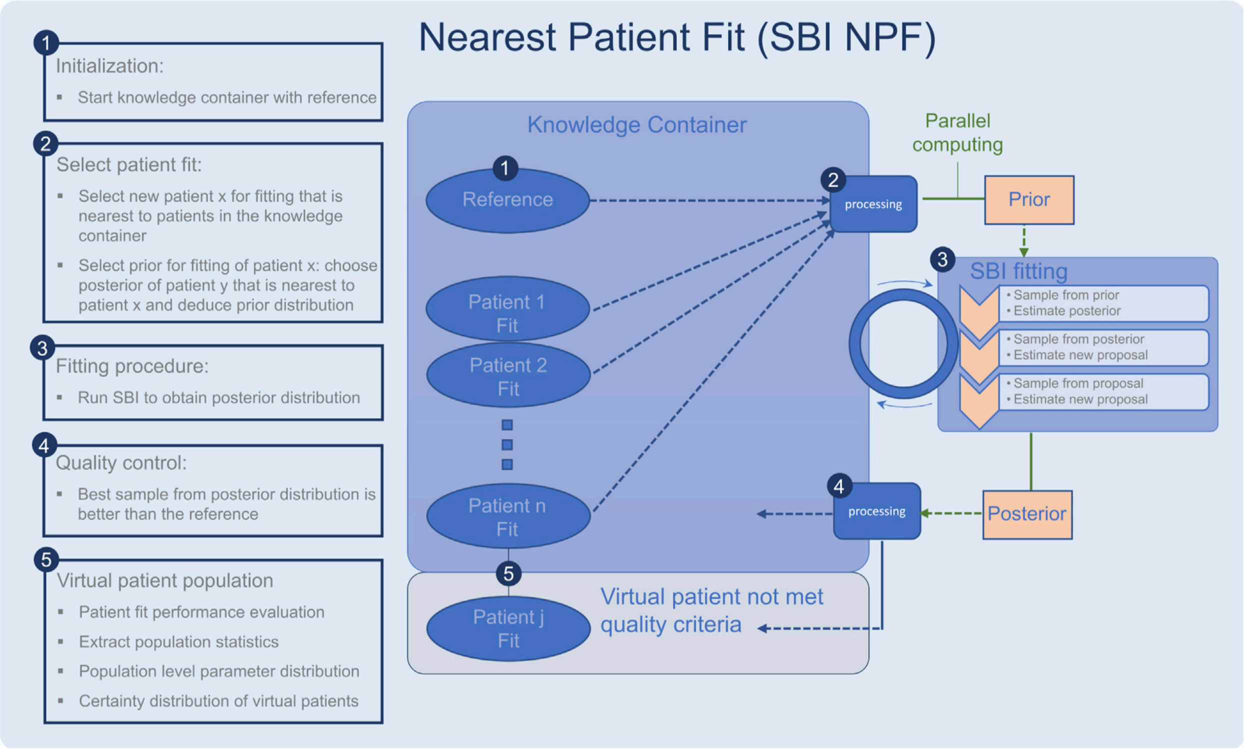
2.3 Global sensitivity analysis

Global sensitivity analysis allows us to determine the importance of QSP parameters on relevant simulation outputs. The analysis was performed during drug treatment, since this is the for the parameter optimization relevant scenario. We defined the parameter ranges by a ±30% interval around the reference parametrization and used Saltelli's sampling scheme (provided by the Python SALib module https://salib.readthedocs.io/en/latest/api.html#sobol-sensitivity-analysis, version 1.3.12). For a given parameter θ_i and a relevant QSP output variable X_j we calculated the total order sensitivity index S_{X_j,θ_i} following the Sobol procedure (Sobol, 2001). To deduce a single sensitivity value for each parameter θ_i, we aggregated the total order sensitivity S_{X_j,θ_i} over the relevant output variables weighted by their variance as

$$S_{\theta_i}^{agg} = \frac{\sum_{j=1}^{n} S_{X_j,\theta_i} Var\left[X_j\right]}{\sum_{j=1}^{n} Var\left[X_j\right]} \tag{1}$$

2.4 Simulation-based inference

Simulation-based inference (SBI) is a class of methods which apply statistical inference to learn the parameters of stochastic simulators (Lueckmann et al., 2021), and hence are applicable for



Nearest Patient Fit (SBI NPF)

1 Initialization:
- Start knowledge container with reference

2 Select patient fit:
- Select new patient x for fitting that is nearest to patients in the knowledge container
- Select prior for fitting of patient x: choose posterior of patient y that is nearest to patient x and deduce prior distribution

3 Fitting procedure:
- Run SBI to obtain posterior distribution

4 Quality control:
- Best sample from posterior distribution is better than the reference

5 Virtual patient population
- Patient fit performance evaluation
- Extract population statistics
- Population level parameter distribution
- Certainty distribution of virtual patients

FIGURE 1
Workflow of the nearest patient fit pipeline (SBI NPF), steps 1–5. The SBI fitting procedure is depicted in step 3. Ellipses represent patient fits and boxes represent processing steps.

learning parameters of QSP models. Statistical inference combines a prior distribution with empirical observations to conclude a posterior distribution. More precisely, given a prior probability distribution $p(\theta)$ over a parametrization $\theta \in \mathbb{R}^n$ and observed data $x_o \in \mathbb{R}^d$, it deduces the posterior probability distribution $p(\theta \mid x_o)$. Following Bayes theorem (Lee, 1989), the posterior is calculated based on the likelihood function $p(x|\theta)$. Since the analytical or numerical computation of the likelihood function is often intractable for complex simulations (Cranmer et al., 2020), SBI estimates the posterior in a "likelihood-free" manner, only relying on samples of the simulator $x \sim sim(\theta)$.

In this work, we evaluate an SBI approach which learns the posterior distribution with a density estimation neural network (*neural posterior estimation*). More precisely, the desired posterior $p(\theta \mid x)$ is assumed to be a member of a family of probability densities q_κ parametrized by κ that can be of various not-predefined shapes (e.g., multimodal). The distribution parameters κ are learned with a neural network $F(x, w)$, where w denotes the adjustable weights of the neural network and x denotes its input, i.e., $p(\theta|x) \approx q_{F(x,w)}(\theta)$. The weights of the neural network are trained by minimizing the loss function $L(w) = \sum_{i=1}^{M} - \log q_{F(x_i,w)}(\theta_i)$ over generated training samples $\{(\theta_i, x_i)\}_i$ where the parameters θ_i are sampled from the prior $\theta_i \sim p(\theta)$ and the corresponding simulation results x_i are sampled from the QSP model $x_i \sim QSP(\theta_i)$. Since QSP simulations are expensive, we use the sample efficient algorithm *sequential neural posterior estimation* (Greenberg et al., 2019). Only those training samples (θ_i, x_i) are considered relevant, where the simulation result x_i is close to the clinical data x_o of the patient to be fitted. Such training samples are generated by drawing parametrizations θ_i from a sequentially refined posterior estimate $\tilde{p}(\theta|x)$ which is called proposal posterior, *cf.* Figure 1, point 3. Since the posterior under a proposal does not coincide with the desired posterior under the prior, the authors in (Greenberg et al., 2019) present a re-parameterization of the problem to automatically transform between estimates of the proposal posterior $\tilde{p}(\theta|x)$ and the true desired posterior $p(\theta \mid x)$. The sequential procedure leads to more informative and thus overall fewer training samples from the simulator.

2.4.1 Usage for individual patient fitting

To run the selected SBI approach, a variety of hyperparameters must be configured which are problem specific. First, to reduce the complexity of the optimization task, we selected an appropriate subset of the QSP parameters for fitting using global sensitivity analysis (Section 2.3) and expert knowledge. Second, we chose a prior distribution over the fitting parameters. Third, we selected the neural network-based density estimator F which models the posterior, the number of rounds in the sequential procedure of the algorithm as well as the number of samples drawn per round used to produce a posterior estimate.

Additional tuning of data and simulation outcome was applied: To handle the measurement noise of the patient data, we introduced (multiplicative lognormal) noise to the QSP simulation output during training leading to a stochastic simulator. We considered the scale parameter of the lognormally distributed noise as a fitting parameter which allowed us to regulate and learn the appropriate amount of noise per patient.

Due to the high dimensionality of the patient data, we reduced it to summary statistics for fitting. More precisely, we represented each biomarker timeseries by its median and the difference between its 0.9- and 0.1-quantile, indicating the rate with which a biomarker increases or decreases. For 16 biomarkers, this resulted in a 32-dimensional representation of the clinical data.

2.4.2 Nearest patient fit (SBI NPF)

The described SBI algorithm fits each patient individually and independently. We investigated an additional approach for boosting the performance by leveraging knowledge of an already learned similar patient.

Since the raw biomarker time series data in the clinical study is sparse, we used the introduced representation of the clinical data in terms of a set of statistical features and computed the similarity between patients as the Euclidean distance in the normalized feature space. More precisely, we used the Euclidean metric proposed in (Dixon, 1979) which is designed for the presence of missing data since not all patients have measurements for all 16 biomarkers:

$$d\left(P_i, P_j\right) = \sqrt{\sum_{b \in CB} \frac{2|CB|}{32}\left(\left(median\left(P_{i,b}\right) - median\left(P_{j,b}\right)\right)^2 + \left(Q_{0.9}\left(P_{i,b}\right) - Q_{0.1}\left(P_{j,b}\right)\right)^2\right)} \tag{2}$$

where CB denotes the set of common biomarkers of patient P_i and patient P_j, $median\ (P_{i,b})$ is the normalized median of biomarker b for patient P_i, and $Q_x\ (P_{i,b})$ for $x \in (0, 1)$ is the normalized x-quantile.

To implement the suggested nearest patient fit approach, we considered the fitting process of the patient cohort as a sequential procedure, *cf.* Figure 1. In each step, we fit a batch of patients in parallel and the procedure terminates when all patients are fitted. Throughout the process, the knowledge we gain from successful patient fits is collected in a so-called knowledge container which makes the knowledge available for the subsequent patient fitting experiments. The developed pipeline is described in detail in the following:

- The knowledge container is initialized with the reference patient, which is generated with the reference parametrization in the QSP model (Figure 1, step 1).
- A processing module selects a batch of patients for fitting (Figure 1, step 2), which are nearest to the current patients in the container according to our similarity metric (Equation 2). The prior for each patient fit is defined based on the learned parametrization of its most similar patient in the knowledge container.
- Each selected patient is fitted with the SBI algorithm (Figure 1, step 3).
- The quality of each resulting patient-specific posterior distribution is assessed by a processing module (Figure 1, step 4). If the learned parameterization is better than the reference parametrization according to the loss function in Section 2.6, the patient fit is put into the knowledge container. If not, its knowledge is not reused for the subsequent SBI experiments, but it is still part of our learned virtual patient population (Figure 1, step 5).

2.4.3 Implementation

We chose a cross-platform implementation to combine fast and robust ordinary differential equation solvers from Julia with high performance SBI methods from Python (version 3.9.12, https://www.python.org/). We used the Python SBI implementation provided by (Lueckmann et al., 2021) and customized the simulation and patient data handling as described above. Information exchange between the SBI algorithm and the QSP model was handled using hdf5-files (in Python: https://pypi.org/project/h5py/, version 3.6.0, in Julia: https://juliaio.github.io/HDF5.jl/stable/, version v0.16.16). The fitting experiments were performed on a Linux server with Intel(R) Xeon(R) Gold 6226R 65 core CPU that has 775 GB memory available, resulting in fitting times of approximately 4 h per patient.

2.5 Benchmarks

Scatter search for MATLAB (SSm, Release 2014A) developed by (Egea et al., 2009) and a gradient-based method (fmincon developed by Mathworks) was used as benchmark on a Windows machine (11th Gen Intel(R) Core(TM) i7-11850H) using MATLAB R2021b and Simbiology version 6.2. Parameter bounds have been set twofold around the reference parametrization. The computation time for a single patient fit was set to 4 h, which met the convergence criterion.

2.6 Evaluation metrics

An individual patient fit yields a QSP parametrization θ. The quality of the parametrization was assessed by comparing the corresponding QSP output to the clinical data c as

$$L(\theta, c) = \sqrt{\frac{1}{\sum_{b=1}^{B} T_b} \sum\nolimits_{b=1}^{B} \sum\nolimits_{t=1}^{T_b} \left(\frac{QSP_{b,t}(\theta) - c_{b,t}}{\max\ (c_{b,1}, \ldots, c_{b,T_b})}\right)^2} \tag{3}$$

where B denotes the number of biomarkers, T_b the number of clinical measurement time points of biomarker b, $QSP_{b,t}\ (\theta)$ the QSP output for biomarker b at time t when parametrized with θ, and $c_{b,t}$ the clinical observation of biomarker b at time t. As biomarker values may be on different scales, we used a maximum-scaling for equal weighting. Since all considered fitting algorithms (SBI, SSm, fmincon) start from the reference parametrization θ_{ref}, we evaluated their performance against the reference parameterization in terms of relative loss reduction as

$$gap(\theta, c) = \frac{L\left(\theta_{ref}, c\right) - L(\theta, c)}{L\left(\theta_{ref}, c\right)} \tag{4}$$

where $\boldsymbol{\theta}$ denotes the parametrization determined by the respective fitting algorithm. $\boldsymbol{gap} < \mathbf{0}$ depicts worse data fits than the reference parameterization while $\boldsymbol{gap} > \mathbf{0}$ depicts improved data fits over the reference parameterization with $\boldsymbol{gap} = \mathbf{1}$ as the best possible case. $\boldsymbol{gap} = \mathbf{0}$ depicts no improvement over the reference parameterization.

For SBI, which produces a probability distribution over the parametrization, we defined the ultimate parametrization $\boldsymbol{\theta_{sbi}}$ as the best one out of 100 samples drawn from the posterior. To evaluate the quality of the posterior distribution, we also reported the fraction of samples which are better than the reference parametrization,

$$\boldsymbol{frac}\left(\boldsymbol{\mathcal{D}_{post}}, \boldsymbol{c}\right) = \frac{\sum_{k=1}^{100} \mathbf{1}_{\{L(\theta_{k,sbi}, c) < L(\theta_{ref}, c)\}}}{\mathbf{100}} \tag{5}$$

where $\boldsymbol{\theta_{k,sbi}}$ denotes the k-th drawn sample from the learned posterior distribution $\boldsymbol{\mathcal{D}_{post}}$ and $\mathbf{1}$ is the indicator function.

TABLE 1 Table shows results of hyperparameter tuning.

Hyperparameter	Value
Training procedure	
Number of rounds	50
Number of simulations per round	100
Prior	
Distribution	Lognormal
Prior scale	0.25
Prior loc	Reference parametrization + for noise: 0.2
Density estimator	
Neural network	“made”
Hidden features	100
Number of atoms	25

3 Results

3.1 Selected hyperparameter values

The hyperparameters which control the sequential training procedure (e.g., number of rounds) as well as the architecture of the density estimation neural network, were optimized with grid search (see Table 1 for an overview of the determined hyperparameter values). A relevant subset of 25 QSP parameters was selected for fitting based on biological expert knowledge, which is often a reasonable first step (Cheng et al., 2022), and global sensitivity analysis results. Figure 2 shows thirteen parameters identified as key determinants of the model output by expert priority (A), as well as the twelve most sensitive parameters (of the remaining ones) identified by global sensitivity analysis (B) (Equation 1). The parameters selected by expert priority were categorized into “immune cell numbers in blood”, “sensitivity of immune processes to cytokine levels” and “simulation of immune cells”. Variability in the expert priority parameters across virtual patients leads to variability in cell populations that play a significant role in disease pathophysiology and response to treatment. Note that the aggregated Sobol indices of the expert priority parameters are comparable to those of the high sensitivity parameters.

For the 25 fitting parameters we chose a lognormal prior distribution $\boldsymbol{LogNormal}(\boldsymbol{loc}, \boldsymbol{scale})$ centered around the reference parametrization with parameters $\boldsymbol{loc} = \log(\boldsymbol{\theta_{ref}})$ and $\boldsymbol{scale} = 0.25$. As the reference parametrization $\boldsymbol{\theta_{ref}}$ simulates a typical patient, the prior can be an informed starting point for an individual patient fit. The lognormal distribution was chosen to keep the range of parameter values positive. Moreover, it covers the different scales of the parameters with a single *scale* value since by definition the amount of variance caused by the *scale* parameter also depends on the parameter ***loc*** (the higher ***loc***, the higher the variance) For SBI NPF we derived the prior from the knowledge container (see Section 3.4) by centering the lognormal distribution around the parametrization of the patient’s nearest container patient (and using the same *scale* value as above).

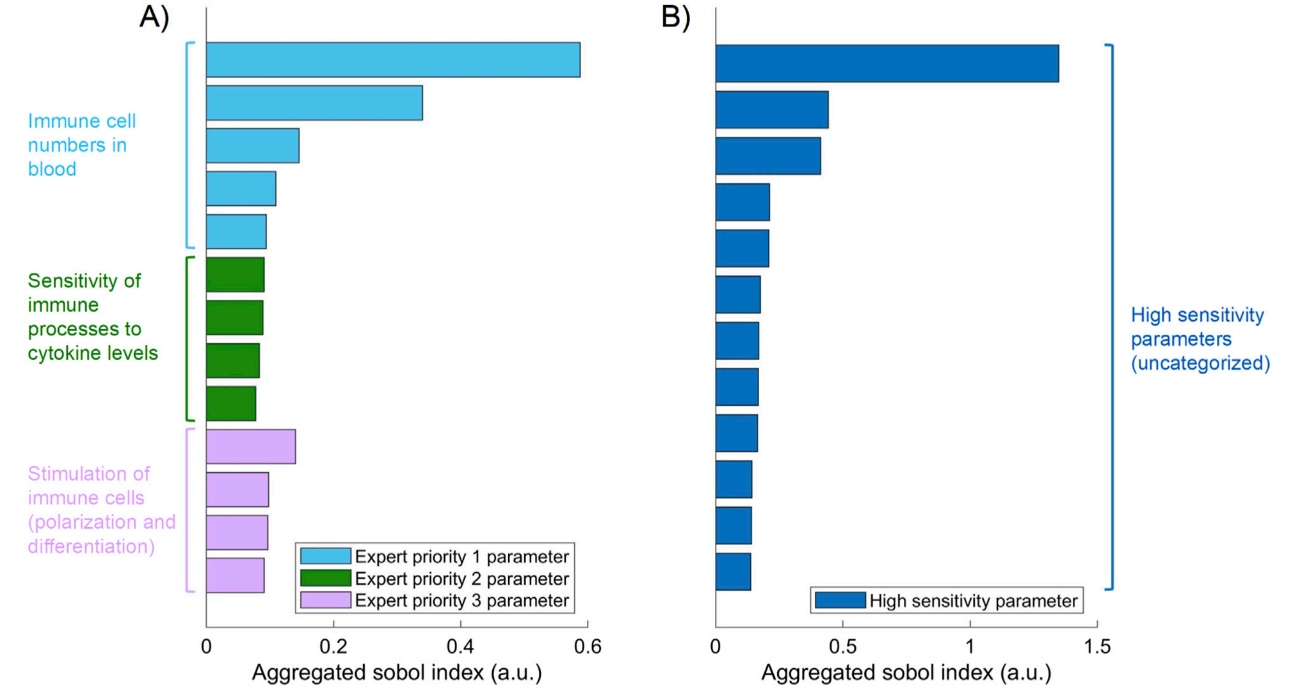


FIGURE 2
Aggregated Sobol indices for the 12 most sensitive parameters **(B)** and 13 expert priority parameters **(A)** selected by their role in the QSP RA model. Parameters are grouped by the corresponding category (color).

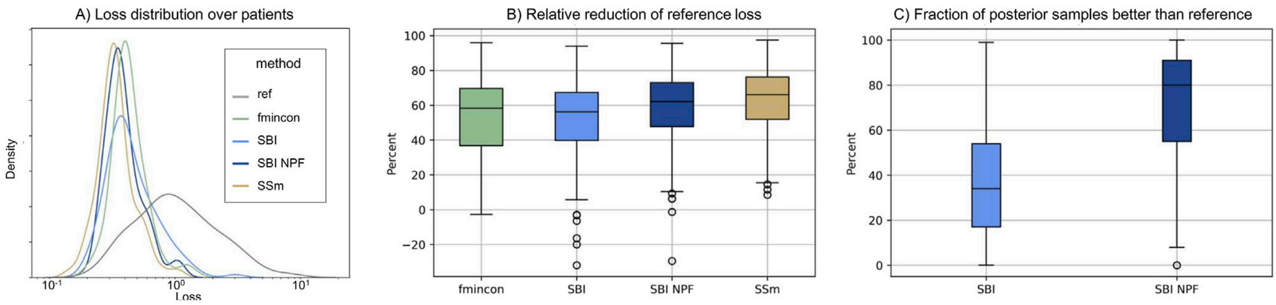


FIGURE 3
(A) Distribution of loss values (Equation 3) over the patient population (n = 133) for the different methods. **(B)** Distribution of the relative reduction of the reference loss over the patient population (n = 133) shown for the different methods calculated using the gap function (Equation 4). Boxes represent interquartile-ranges with a line at the median, whiskers extend to the last data point up to 1.5-fold of the interquartile range and circles represent outliers. **(C)** Distribution of the fraction of posterior samples which outperform the reference fit (ref) for both SBI approaches (SBI and SBI NPF) calculated from Equation 5.

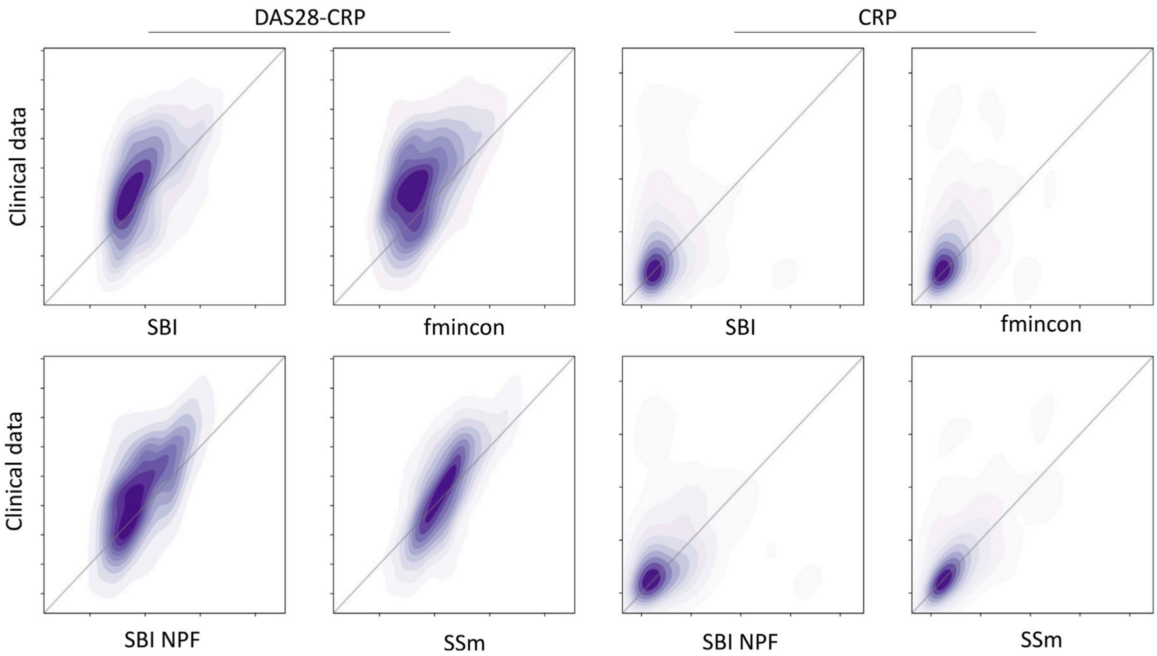


FIGURE 4
Correlation between all patient's observations from the clinical data (*y*-axis) and the respective simulation results (*x*-axis) depicted as a density plot for a blood biomarker (CRP on the left) and a disease score (DAS28-CRP on the right). Simulation results were generated using the individual parameter estimates from the four different algorithms (SBI, fmincon, SBI NPF and SSm). Dark-shaded areas indicate high density while soft-shaded areas indicate low density.

3.2 Virtual patient generation

Individual patient fits were performed for the presented SBI approaches as well as for the selected benchmarks (SSm, fmincon) and evaluated according to the loss function in Equation 3. For each method, the distribution of the loss over all patients is depicted in Figure 3A. All fitting algorithms lead to loss curves with smaller mean loss values and smaller variance compared to the reference. For each patient fit, the relative reduction of the reference loss is shown in Figure 3B as a distribution over the population. The fitting performance range is spanned by the two benchmarks fmincon and SSm. While the performance distribution of SBI resembles fmincon, SBI NPF yields a clear improvement which is similar to the SSm performance distribution. For both SBI approaches there are a few outlier patients, for which the reference is better than the respective SBI result (i.e., negative relative reduction in Figure 3B). When comparing the losses patient-wise, SBI NPF improves over SBI for 82% of the patients. SBI NPF does not only outperform SBI in terms of the best posterior sample but also in terms of the whole learned posterior distribution, *cf.* Figure 3C. It shows that for SBI typically 34% out of 100 posterior samples are better than the reference parametrization, while for SBI NPF this number is around 80%. Visual predictive checks on a biomarker level are presented in Figures 4, 5 for the c-reactive protein(CRP) and a disease score

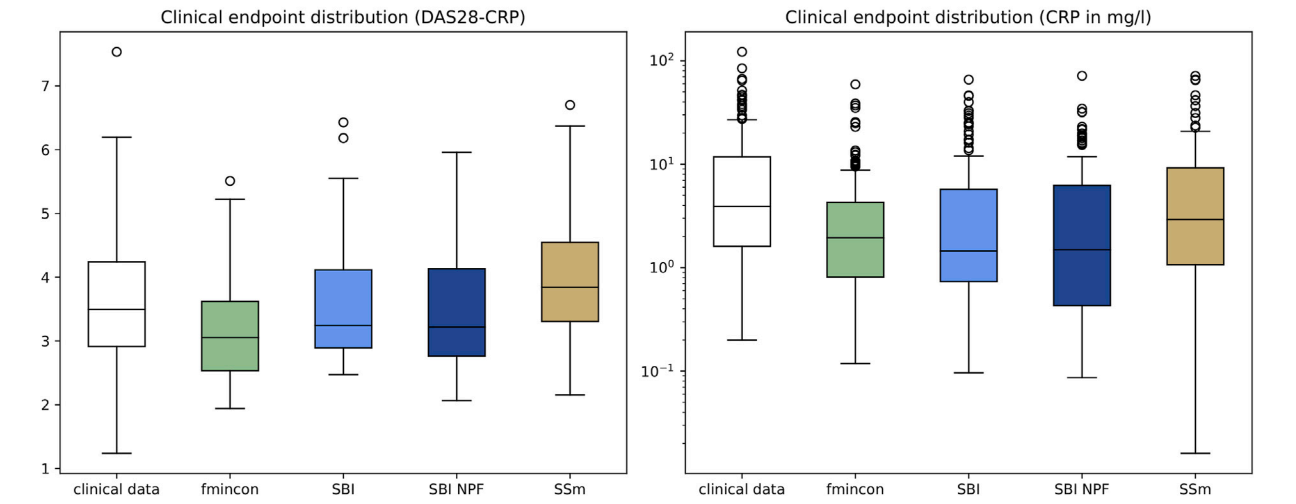


FIGURE 5
Depiction of clinical endpoints and corresponding simulation results as distributions over the patient population (n = 133) after 24 weeks of treatment for DAS28-CRP (left) and CRP (right). Simulation results were generated using the individual parameter estimates from the four different algorithms. Boxes represent interquartile-ranges with a line at the median, whiskers extend to the last data point up to 1.5-fold of the interquartile range and circles represent outliers.

(DAS28-CRP). Figure 4 compares the clinical biomarker observations of all patients (y-axis) at all time points to the corresponding simulations of the model with the parameter sets of the respective fitting algorithm (x-axis). This density correlation plot illustrates that, similar to the benchmarks, the SBI approaches are overall able to describe the clinical data sufficiently well. The visual predictive checks also reflect that SBI NPF leads to better fits than SBI. In Figure 5 we depict the distribution of the clinical data and the obtained simulation results (after 24 weeks of treatment) over the patient population. Inter-patient variability is large in the clinical data endpoints and the fitting methods are generally able to capture this variability under the chosen parameter bounds. An example of an individual fit obtained by SBI is shown in the Supplementary Figure S2 for the CRP data. In summary, the empirical evaluations demonstrate that the SBI approaches can compete with classic fitting methods in the field in terms of fitting quality and fitting speed. Moreover, the suggested SBI NPF pipeline significantly improved over SBI.

3.3 Comparison of virtual patients

For each patient, SBI produces a posterior probability distribution over the considered 25-dimensional parameter space. Exemplary one-dimensional marginal posteriors are depicted in Figure 6 for three different parameters. One column depicts the marginal distribution for a specific QSP model parameter for three different patients which all started from the same prior (grey). For each parameter (column), the three learned patient-individual posteriors (blue) differ significantly from each other. While a learned posterior can have moved far away from the prior, i.e., the reference parametrization, they can also resemble each other, at least in the one dimension depicted in this figure (similarity of the here depicted one-dimensional marginal prior and posterior does not imply similarity of the 25-dimensional prior and posterior distributions). Overall, we observe multiple shapes of the marginal posteriors, which range from very concentrated distributions to broader and flat ones.

Note that the sampled parameter sets from a patient posterior distribution contain between-parameter relationships (example given in the supplement as parallel coordinate plot, see Supplementary Figure S3) and can be used to explore correlations (example given in the supplement as correlation matrix, see Supplementary Figure S4).

4 Discussion

4.1 Concept: Generation of virtual patients by fitting individual patient data

QSP models are typically built in several steps. Individual mechanistic parameters, such as binding or dissociation as well as mechanistic pathway modules are first calibrated based on *in vitro* and *in vivo* experiments and, in the final step, are then fitted to clinical study data such as biomarker concentrations and disease activity endpoints (Cheng et al., 2022).

Often this clinical data is only available as summary statistics, which requires weighting methods to ensure a proper distribution of the inferred parameter sets (Klinke, 2008; Schmidt et al., 2013). It requires difficult assumptions on which patients may exist in the real world and has consequences for prediction of drug efficacy.

Fitting of individual clinical data circumvents these assumptions but is limited, in good cases, to only a few hundred patients where the individual data is often provided without uncertainty statistics (such as standard deviation). The lack of uncertainty statistics denies the use of sophisticated approaches for generating alternative parameterizations for a single patient, such as bootstrapping (Tibshirani and Efron, 1986).

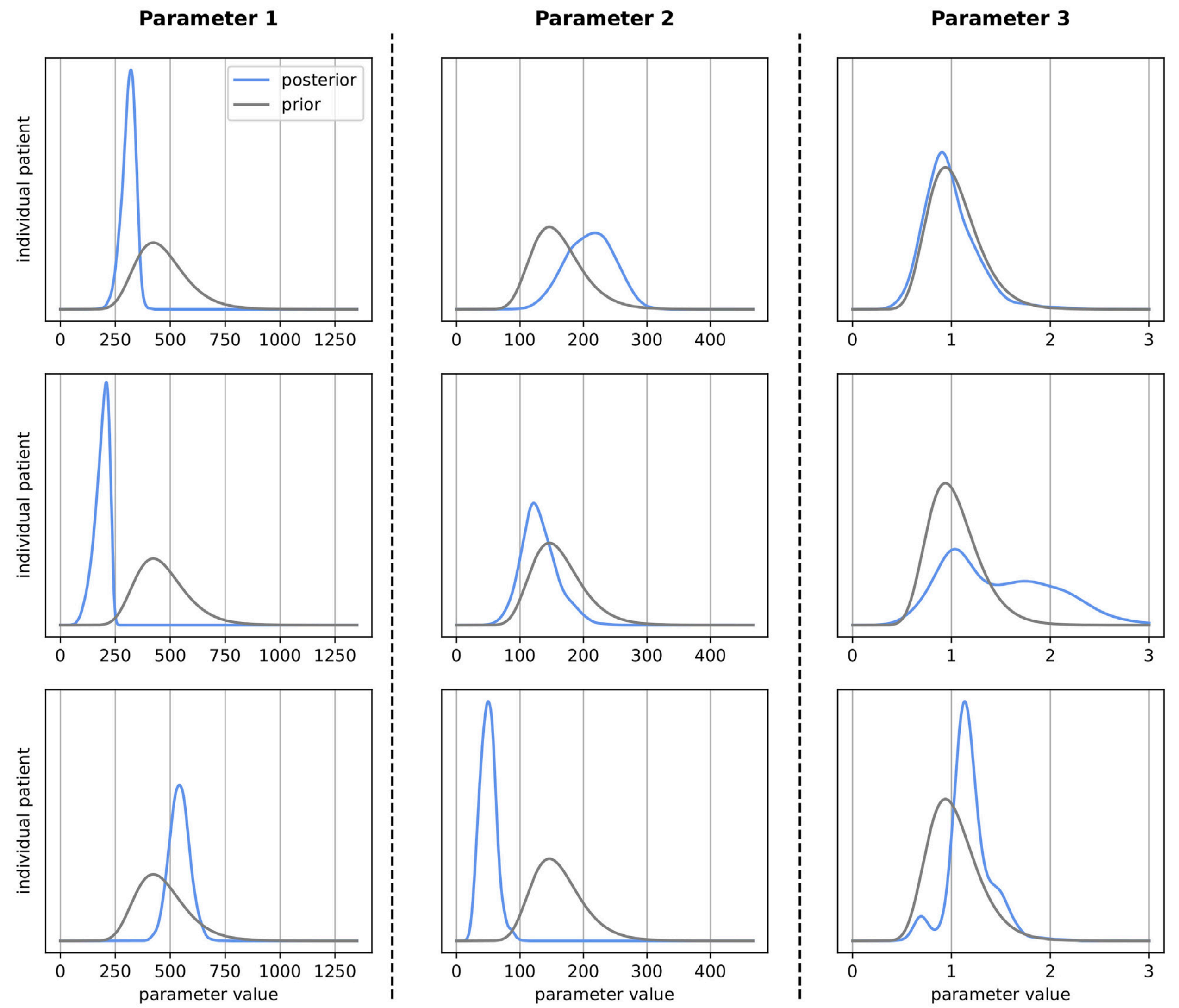


FIGURE 6
Prior (grey) vs. selected patient-individual posterior (blue) one-dimensional marginal distributions for three model parameters. Every subplot stands for an individual patient. Each column represents one specific parameter. *X*-axes represent the parameter value used in the QSP model (parameter-specific units) and *y*-axes represent the density.

By applying a simulation-based inference method, we generated parameter probability distributions during the patient fit, directly providing alternative parametrizations for real patients. More precisely, sampling from the probability distributions yield different highly likely parametrizations for an individual patient, which can then be used to achieve a larger virtual patient population. Thus, the above-mentioned limitations of individual patient data have been overcome and the generated virtual population is based on real patients, which is advantageous compared to weighting methods and their assumptions.

The subsequent validation of the generated virtual population, either from individual patient fitting or from hypothesis-based methods, is usually achieved by predicting the population outcome of other studies, for example, drugs with different mode of action or different dosing schemes, under consideration of the baseline characteristics of the study population.

4.2 SBI for fitting individual patient data

In this work, we employ SBI to learn a distribution over the QSP model parametrization for an individual rheumatoid arthritis patient and build a virtual patient from it. The goal is to identify regions in the parameter space which best explain the patient observations, i.e., where the corresponding simulated biomarker values match the patient's clinical data.

The approach is particularly interesting for the described setup since there may exist multiple optimal QSP parametrizations to model the patient data. The learned probability distribution in the parameter space naturally provides the probability of certain parameterizations and can be used to explore alternative parameterizations. Another benefit of SBI is that it treats the simulation as a black box, similar to SSm and fmincon.

There exists a variety of SBI algorithms in the literature, see (Lueckmann et al., 2021) for a detailed overview, from which we chose the sample-efficient algorithm *sequential neural posterior estimation*. The choice of the SBI approach as well as of the stochastic global and deterministic local approach can yield differences in the benchmarking as their performance needs to be considered as partially problem specific (Egea et al., 2009). Furthermore, the applied data statistics and the data noise handling can influence the result performance.

4.3 Choice of hyperparameters

Within this challenging optimization problem, algorithms and settings of hyperparameters are an impactful choice that is based on the underlying optimization criterion and performance assumptions. Alternative hyperparameter settings may yield similar or better results and can be subject of further analysis.

To reduce the complexity of the optimization problem and to achieve high quality model fits, we selected the most relevant parameters for model fitting by assessing the parameter influence on biomarker-related model outputs through global sensitivity analysis (Sobol, 2001). In addition, expert priority parameters have been included in the parameter estimation (Cheng et al., 2022). The quantitative choice of 25 parameters seems arbitrary but alternative parameter numbers did not improve the result of the parameter estimation.

4.4 Performance of virtual patient generation

The results of this work demonstrate that fitting of individual patients can yield virtual patients that each outperform the reference and that the model parameterizations can represent the variability in clinical response typically seen in the data. The variability in the patient data was very high, *cf.* Figure 5, which is expected for rheumatoid arthritis as heterogeneous disease, and poses a real challenge for individual patient fitting but also for predicting response (Rehberg et al., 2021). Obviously, the inter-patient variability is a consequence of phenotypic differences and measurement noise. As noise cannot be explained biologically with the mechanistic QSP model, a perfect correlation between clinical data and model predictions in Figure 4 is difficult to achieve (see also (Schmidt et al., 2013)). Yet the discussed algorithms show a different fitting performance with fmincon performing worst, SSm performing best and SBI being in between. Fmincon generally is less suited for our optimization task than the others as it searches for a local and not necessarily global optimum. While fmincon and SSm provide only point estimates, SBI provides a distribution, i.e., multiple parameter estimates with corresponding probabilities. We note that the fitting approach with SBI uses summary statistics of the clinical data and not its raw observations like the benchmarks, which could be a disadvantage. Yet overall, the SBI approaches get reasonably close to SSm. Our results also illustrate that SBI can handle a high-dimensional parameter space of 25 parameters and make them suited for such kind of QSP problems. For comparison, SBI approaches in the literature focused, so far, on setups of only 2–10 parameters (Lueckmann et al., 2021; Reza et al., 2022; Boelts et al., 2023; Boyali et al., 2021). The fact that SBI could be improved with SBI NPF for 82% of the patients demonstrates a high potential of the nearest patient fit pipeline developed in this work. It showcases the influence and necessity of good prior estimates for SBI algorithms. However, 18% of the patients were better fitted with SBI, which starts from a presumably less appropriate prior distribution. While SBI approaches are inherently stochastic, the impact on fitting quality was minor in repetitive experiments. We must assume that the SBI NPF pipeline has room for improvement in defining the patient vicinity criteria and/or that patient vicinity is not always of benefit, as a QSP model may require very different parametrizations to produce similar outputs (Duffull and Gulati, 2020). To conclude on the SBI NPF pipeline, the developed concept of nearest patient fits is not specific to SBI but represents a generic contribution that can be transferred to any fitting algorithm which considers initial solutions.

4.5 Comparison of virtual patients

The patient-specific posterior marginal distributions show that very diverse QSP model parametrizations can be necessary to describe individual patients well, which SBI was able to learn. The different shapes of the marginal posteriors indicate the flexibility of the chosen SBI approach (sequential neural posterior estimation) in modelling probability distributions. While concentrated distributions can indicate a high certainty in the virtual patient parametrizations, flat distributions may point towards those that are uncertain. One advantage of the learned distributions is that alternative virtual patient parametrizations can directly be generated through sampling. I.e., new highly-probable patient fits can be easily generated without re-running the optimization solver or using other metrics and assumptions such as prevalence weighting. These alternate parameterizations of a virtual patient may describe the fitted data equally well and may represent differences in the disease mechanisms. Exploring alternate parametrizations is fundamental to assess the range of treatment outcomes of an individual patient.

On the population level, aggregation of the given patient-specific posterior distributions may allow the application of population statistics for assessment of subgroups, patient differences and population spread.

4.6 General conclusion

In this work, we find SBI approaches to be powerful tools in creating virtual patients using individual patient data. SBI achieved the same performance in patient fits compared to benchmark algorithms and provides parameter probability distributions, which can be used to explore alternative parameterizations for real patients to create more confidence in predicting clinical outcomes for *in silico* trials. Furthermore, leveraging patient similarities observed in the clinical data, improved the performance and may be suited as a generalizable strategy in generating virtual patients.

Data availability statement

The original contributions presented in the study are included in the Supplementary Material, further inquiries can be directed to the corresponding author.

Ethics statement

Ethical approval was not required for the study involving humans in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

Author contributions

NP: Conceptualization, Data curation, Investigation, Methodology, Supervision, Writing–original draft, Writing–review and editing. VK: Data curation, Investigation, Validation, Writing–original draft, Writing–review and editing. CG: Data curation, Investigation, Methodology, Writing–original draft, Writing–review and editing. AS: Investigation, Methodology, Writing–original draft, Writing–review and editing. ML: Investigation, Methodology, Writing–original draft, Writing–review and editing. BW: Conceptualization, Supervision, Writing–original draft, Writing–review and editing. AK: Conceptualization, Investigation, Methodology, Supervision, Writing–original draft, Writing–review and editing. MR: Conceptualization, Investigation, Supervision, Writing–original draft, Writing–review and editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

The authors thank Tim Wirtz, Anastasios Siokis, Hans-Christoph Schneider and Rebekka Fendt for helpful discussions and acknowledge their contributions.

Conflict of interest

VK, CG, BW and MR were employed by Sanofi R&D.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fsysb.2024.1444912/full#supplementary-material

SUPPLEMENTARY FIGURE S1

Schematic representation of key interactions in the QSP RA model. The model is composed of a blood compartment (from which immune cells are recruited) and a synovial tissue compartment (side of inflammation). Thickness of connection dots illustrate the influence strength. Dots represent positive influence, bars represent negative influence (only for Tregs). Abbreviations: TNF: tumor necrosis factor, FLS: fibroblast-like synoviocytes IL6R: interleukin-6 receptor, CRP: c-reactive protein, DAS28-CRP: disease activity score 28 with CRP, TH: T helper cells, MAC: macrophages, Treg: T regulatory cells, B: B cells.

SUPPLEMENTARY FIGURE S2

Individual patient fit of QSP model to c-reactive protein (CRP) data obtained by SBI. The clinical patient data is depicted as circles. Data before treatment start shows baseline characteristics of that individual patient and the drop in CRP shows response to the treatment (treatment time: 24 weeks). Each dashed blue line represents a fit obtained by SBI. More precisely, it represents the QSP simulation result when parametrized with a sample from the learned patient posterior. Note that the depicted fitting result was obtained by fitting 16 clinical biomarkers and endpoints from that patient simultaneously (including CRP).

SUPPLEMENTARY FIGURE S3

Parallel coordinate plot of parameter sets sampled from a patient posterior distribution obtained by SBI. The 25 fitting parameters are depicted as p1, . . ., p25 as columns, each equipped with a *y*-axis showing the respective parameter value range. Each line (from p1 to p25) is one parameter set derived from SBI for the given patient, obtained by sampling from the learned patient posterior distribution. A line's color represents the quality of that parameter set in terms of the resulting fitting loss (Equation 3). Parameter sets in dark green color illustrate low loss values while orange parameter sets have higher loss values. A total of 100 parameter sets is shown.

SUPPLEMENTARY FIGURE S4

Parameter correlation calculated for the parameter sets shown in Supplementary Figure S3: The heatmap shows the correlation between the 25 parameters shown as p1, . . ., p25 obtained from the 100 parameter sets depicted in Supplementary Figure S3. Numbers are Pearson correlation coefficients and are highlighted in red for positive correlation and in blue for negative correlation.

References

Aghamiri, S. S., Amin, R., and Helikar, T. (2022). Recent applications of quantitative systems pharmacology and machine learning models across diseases. *J. Pharmacokinet. Pharmacodyn.* 49 (1), 19–37. doi:10.1007/s10928-021-09790-9

Allen, A., Siefkas, A., Pellegrini, E., Burdick, H., Barnes, G., Calvert, J., et al. (2021). A digital twins machine learning model for forecasting disease progression in stroke patients. *Appl. Sci.* 11 (12), 5576. doi:10.3390/app11125576

Allen, R. J., Rieger, T. R., and Musante, C. J. (2016). Efficient generation and selection of virtual populations in quantitative systems pharmacology models. *CPT Pharmacometrics Syst. Pharmacol.* 5 (3), 140–146. doi:10.1002/psp4.12063

Biesemann, N., Margerie, D., Asbrand, C., Rehberg, M., Savova, V., Agueusop, I., et al. (2023). Additive efficacy of a bispecific anti–TNF/IL-6 nanobody compound in translational models of rheumatoid arthritis. *Sci. Transl. Med.* 15 (15), eabq4419. doi:10.1126/scitranslmed.abq4419

Björnsson, B., Borrebaeck, C., Elander, N., Gasslander, T., Gawel, D. R., Gustafsson, M., et al. (2019). Digital twins to personalize medicine. *Genome Med.* 12 (1), 4. doi:10.1186/s13073-019-0701-3

Boelts, J., Harth, P., Gao, R., Udvary, D., Yanez, F., Baum, D., et al. (2023). Simulation-based inference for efficient identification of generative models in connectomics. *bioRxiv*. doi:10.1371/journal.pcbi.1011406

Boyali, A., Thompson, S., and Wong, D. R. (2021). "Identification of vehicle dynamics parameters using simulation-based inference," in *2021 IEEE intelligent vehicles symposium workshops (IV workshops)*.

Bradshaw, E. L., Spilker, M. E., Zang, R., Bansal, L., He, H., Jones, R. D. O., et al. (2019). Applications of quantitative systems pharmacology in model-informed drug discovery: perspective on impact and opportunities. *CPT Pharmacometrics Syst. Pharmacol.* 8 (11), 777–791. doi:10.1002/psp4.12463

Burmester, G. R., Lin, Y., Patel, R., van Adelsberg, J., Mangan, E. K., Graham, N. M., et al. (2017). Efficacy and safety of sarilumab monotherapy versus adalimumab monotherapy for the treatment of patients with active rheumatoid arthritis (MONARCH): a randomised, double-blind, parallel-group phase III trial. *Ann. Rheum. Dis.* 76 (5), 840–847. doi:10.1136/annrheumdis-2016-210310

Byrd, R. H., Gilbert, J. C., and Nocedal, J. (2000). A trust region method based on interior point techniques for nonlinear programming. *Math. Program.* 89, 149–185. doi:10.1007/pl00011391

Cheng, Y., Straube, R., Alnaif, A., Huang, L., Leil, T., and Schmidt, B. (2022). Virtual populations for quantitative systems pharmacology models. *Methods Mol. Biol.*, 2486: 129–179. doi:10.1007/978-1-0716-2265-0_8

Cranmer, K., Brehmer, J., and Louppe, G. (2020). The frontier of simulation-based inference. *Proc. Natl. Acad. Sci. U.S.A.* 117 (48), 30055–30062. doi:10.1073/pnas.1912789117

Derbalah, A., Al-Sallami, H., Hasegawa, C., Gulati, A., and Duffull, S. B. (2022). A framework for simplification of quantitative systems pharmacology models in clinical pharmacology. *Br. J. Clin. Pharmacol.* 88 (4), 1430–1440. doi:10.1111/bcp.14451

Dixon, J. K. (1979). Pattern recognition with partly missing data. *IEEE Trans. Syst. Man. Cybern. Syst.* 9 (10), 617–621. doi:10.1109/tsmc.1979.4310090

Duffull, S., and Gulati, A. (2020). Potential issues with virtual populations when applied to nonlinear quantitative systems pharmacology models. *CPT Pharmacometrics Syst. Pharmacol.* 9 (11), 613–616. doi:10.1002/psp4.12559

Egea, J. A., Vazquez, E., Banga, J. R., and Martí, R. (2009). Improved scatter search for the global optimization of computationally expensive dynamic models. *J. Glob. Optim.* 43, 175–190. doi:10.1007/s10898-007-9172-y

Friedrich, C. (2016). A model qualification method for mechanistic physiological QSP models to support model-informed drug development. *CPT Pharmacometrics and Syst. Pharmacol.* 5 (2), 43–53. doi:10.1002/psp4.12056

Gabay, C., Msihid, J., Paccard, C., Zilberstein, M., Graham, N. M., and Boyapati, A. (2017). FRI0227 Sarilumab significantly suppresses circulating biomarkers of bone resorption and cardiovascular risk compared with adalimumab: biomarker analysis from the phase 3 monarch study. *Ann. Rheum. Dis.* 76, 570. doi:10.1136/annrheumdis-2017-eular.4534

Gaby, C., Burmester, G., Strand, V., Msihid, J., Zilberstein, M., Kimura, T., et al. (2020). Sarilumab and adalimumab differential effects on bone remodelling and cardiovascular risk biomarkers, and predictions of treatment outcomes. *Arthritis Res. and Ther.* 22 (1), 70. doi:10.1186/s13075-020-02163-6

Greenberg, D., Nonnenmacher, M., and Macke, J. (2019). "Automatic posterior transformation for likelihood-free inference," in *International conference on machine learning*.

Howell, B. A., Yang, Y., Kumar, R., Woodhead, J., Harrill, A., Clewell, H. 3., et al. (2012). *In vitro* to *in vivo* extrapolation and species response comparisons for drug-induced liver injury (DILI) using DILIsym™: a mechanistic, mathematical model of DILI. *J. Pharmacokinet. Pharmacodyn.* 39 (5), 527–541. doi:10.1007/s10928-012-9266-0

Klinke, D. (2008). Integrating epidemiological data into a mechanistic model of type 2 diabetes: validating the prevalence of virtual patients. *Ann. Biomed. Eng.* 36 (2), 321–334. doi:10.1007/s10439-007-9410-y

Lee, P. M. (1989). *Bayesian statistics*. London: Oxford University Press.

Lueckmann, J.-M., Boelts, J., Greenberg, D., Goncalves, P., and Macke, J. (2021). "Benchmarking simulation-based inference," in *International conference on artificial intelligence and statistics*.

Luo, M. C., Nikolopoulou, E., and Gevertz, J. L. (2022). From fitting the average to fitting the individual: a cautionary tale for mathematical modelers. *Front. Oncol.* 12, 793908. doi:10.3389/fonc.2022.793908

McComb, M., Blair, R. H., Lysy, M., and Ramanathan, M. (2022). Machine learning-guided, big data-enabled, biomarker-based systems pharmacology: modeling the stochasticity of natural history and disease progression. *J. Pharmacokinet. Pharmacodyn.* 49 (1), 65–79. doi:10.1007/s10928-021-09786-5

Parikh, J., Rumbell, T., Butova, X., Myachina, T., Acero, J. C., Khamzin, S., et al. (2022). Generative adversarial networks for construction of virtual populations of mechanistic models: simulations to study Omecamtiv Mecarbil action. *J. Pharmacokinet. Pharmacodyn.* 49 (1), 51–64. doi:10.1007/s10928-021-09787-4

Rehberg, M., Giegerich, C., Praestgaard, A., van Hoogstraten, H., Iglesias-Rodriguez, M., Curtis, J. R., et al. (2021). Identification of a rule to predict response to sarilumab in patients with rheumatoid arthritis using machine learning and clinical trial data. *Rheumatol. Ther.* 8, 1661–1675. doi:10.1007/s40744-021-00361-5

Reza, M., Zhang, Y., Nord, B., Poh, J., Ciprijanovic, A., and Strigari, L. (2022). "Estimating cosmological constraints from galaxy cluster abundance using simulation-based inference," in *ICML 2022 workshop on machine learning for astrophysics*.

Rieger, T. R., Allen, R. J., Bystricky, L., Chen, Y., Colopy, G. W., Cui, Y., et al. (2018). Improving the generation and selection of virtual populations in quantitative systems pharmacology models. *Prog. Biophys. Molec Bio* 139, 15–22. doi:10.1016/j.pbiomolbio.2018.06.002

Schmidt, B. J., Casey, F. P., Paterson, T., and Chan, J. R. (2013). Alternate virtual populations elucidate the type I interferon signature predictive of the response to rituximab in rheumatoid arthritis. *BMC Bioinforma.* 14, 221. doi:10.1186/1471-2105-14-221

Sobol, I. (2001). Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Math. Comput. SIMULAT* 55 (1-3), 271–280. doi:10.1016/S0378-4754(00)00270-6

Sorger, P., Allerheiligen, S., Abernethy, D., Altman, R., Brouwer, K., Califano, A., et al. (2011). *Quantitative and systems pharmacology in the post-genomic era: new approaches to discovering drugs and understanding therapeutic*. Maharashtra, India: NIH White Paper by the QSP Workshop Group.

Tibshirani, B., and Efron, R. (1986). Bootstrap methods for standard errors, confidence. *Stat. Sci.* 27 (2), 54–77.

Wajima, T., Isbister, G. K., and Duffull, S. B. (2009). A comprehensive model for the humoral coagulation network in humans. *Clin. Pharmacol. Ther.* 86, 290–298. doi:10.1038/clpt.2009.87

Zhang, T., Androulakis, I. P., Bonate, P., Cheng, L., Helikar, T., Parikh, J., et al. (2022). Two heads are better than one: current landscape of integrating QSP and machine learning: an ISoP QSP SIG white paper by the working group on the integration of quantitative systems pharmacology and machine learning. *J. Pharmacokinet. Pharmacodyn.* 49 (1), 5–18. doi:10.1007/s10928-022-09805-zFebruary, 2022)

frontiers | Frontiers in Systems Biology

TYPE Methods
PUBLISHED 18 November 2024
DOI 10.3389/fsysb.2024.1460369

OPEN ACCESS

EDITED BY
Rongling Wu,
The Pennsylvania State University (PSU),
United States

REVIEWED BY
Jianrong Wang,
Michigan State University, United States
Tao He,
San Francisco State University, United States

*CORRESPONDENCE
Xiaoxi Shen,
✉ rcd67@txstate.edu

RECEIVED 05 July 2024
ACCEPTED 30 October 2024
PUBLISHED 18 November 2024

CITATION
Shen X and Wang X (2024) An exploration of testing genetic associations using goodness-of-fit statistics based on deep ReLU neural networks.
Front. Syst. Biol. 4:1460369.
doi: 10.3389/fsysb.2024.1460369

An exploration of testing genetic associations using goodness-of-fit statistics based on deep ReLU neural networks

Xiaoxi Shen* and Xiaoming Wang

Department of Mathematics, Texas State University, San Marcos, TX, United States

As a driving force of the fourth industrial revolution, deep neural networks are now widely used in various areas of science and technology. Despite the success of deep neural networks in making accurate predictions, their interpretability remains a mystery to researchers. From a statistical point of view, how to conduct statistical inference (e.g., hypothesis testing) based on deep neural networks is still unknown. In this paper, goodness-of-fit statistics are proposed based on commonly used ReLU neural networks, and their potential to test significant input features is explored. A simulation study demonstrates that the proposed test statistic has higher power compared to the commonly used t-test in linear regression when the underlying signal is nonlinear, while controlling the type I error at the desired level. The testing procedure is also applied to gene expression data from the Alzheimer's Disease Neuroimaging Initiative (ADNI).

KEYWORDS

deep neural networks, goodness-of-fit test, asymptotic normality, sample splitting, genetic association

Introduction

Since the creation of backpropagation, neural networks have regained their popularity, and deep neural networks are now the fundamental building blocks of sophisticated artificial intelligence. For instance, in computer vision, convolutional neural networks (CNNs) (LeCun, 1989) are commonly used for object detection, while recurrent neural networks (RNNs) (Rumelhart et al., 1988), or more recently, transformers (Vaswani et al., 2017) play vital roles in natural language processing.

One of the main reasons for the superior performance of deep learning models is that neural networks are universal approximators. In fact, in the early 1990s, various research established the universal approximation property for shallow neural networks, as well as their derivatives with squashing activation functions—functions that are monotonically increasing and approach 0 and 1 when the variable tends to negative and positive infinity, respectively (Cybenko, 1989; Hornik et al., 1989; Pinkus, 1999) showed that any neural network has the universal approximation property as long as the activation function is not a polynomial. Recently, similar results have also been established for deep neural networks with the Rectified Linear Unit (ReLU) activation function (Nair and Hinton, 2010). Another important characteristic of shallow neural networks is that the approximation rate to certain smooth functions is independent of the dimensionality of the input features (Barron, 1993), making neural networks a great candidate to avoid curse of dimensionality. For example (Shen et al., 2023; Braun et al., 2024), have shown that the rate of convergence of shallow neural networks is independent of the input dimension when the underlying function resides in the Barron space.

Such nice approximation properties provide deep neural networks with great potential for modeling complex genotype-phenotype relationships, and a lot of research has been done in this direction. For instance, a deep learning method known as DANN (Quang et al., 2014) was proposed to make predictions on the deleteriousness of genetic variants. In terms of predicting effects of the non-coding regions, DanQ (Quang and Xie, 2016) integrated CNNs and Bidirectional Long Short-Term Memory networks to capture different aspects of DNA sequences and outperformed other similar methods in various metrics. More recently (Zhou et al., 2023), used deep neural networks to model Alzheimer's disease (AD) polygenic risk and the deep learning methods outperform traditional methods such as weighted polygenic risk score model and LASSO (Tibshirani, 1996).

Despite empirical and theoretical evidence on the powerful prediction performance of deep neural networks, an overlooked problem in deep learning is the interpretability of these models. From a statistical perspective, the interpretability of deep learning models can be improved if we know how to conduct statistical inference using deep neural networks. In recent years, several works have been done in this direction. For example (Horel and Giesecke, 2019), proposed a significant test based on shallow neural network using empirical process theory. However, the asymptotic distribution of the test statistic is hard to compute. Recently, Shen et al. (2021) and Shen et al. (2022) proposed two testing procedures for shallow neural networks with sigmoid activation function. Both of these testing procedures are easier to implement and have better performance compared to *t*-test or *F* test in linear regression. Dai et al. (2024) also proposed a black box testing procedure to test conditional independence between features and response. Below we would like to point out several challenges one needs to conquer in order to develop hypotheses testing based on deep learning models:

1. Classical statistical hypothesis testing techniques in parametric models are difficult to apply in DNNs. One reason is that the parameters (weights and biases) are unidentifiable in general (Fukumizu, 2003), making them hard to interpret. For example, in linear regression, testing the significance of a covariate is equivalent to testing the coefficient attached to it is equal to 0 or not. However, in a DNN, there are many ways to make the covariate vanish in the model. As an example one can let all the weights directly attached to an input feature be 0 or one can also let all the weights for each hidden-to-output unit to be 0.
2. The number of tuning parameters to train a DNN is large. There is no general guideline on how to choose the number of layers and the number of hidden units in each layer to achieve desirable performance in a DNN. Additionally, in the training process, how to wisely select the learning rate and the number of iterations needed is also unclear. Without carefully choosing these tuning parameters, it is likely that the trained DNN will overfit the data. Although overfitting might be acceptable for prediction, it generally needs to be avoided when conducting statistical hypothesis testing.
3. There is lack of theoretical guarantees to ensure the performance of DNNs as tools in genetic association studies. Current theories on DNNs mainly focus on evaluating the generalization errors of DNNs. Many results available are based on the assumption of high-dimensional regime, where the sample size and the number of features are of the same order, or in the polynomial regime, where the sample size grows polynomially as the number of features (Mei et al., 2022; Mei and Montanari, 2022). These conditions are easily satisfied in tasks like image classification, where one can use the data augmentation strategy to manually generate new samples. In genetic studies, however, researchers usually face a limited sample size but a huge number of genetic variants, making those results less attractive in genetic studies.

In this paper, we proposed a goodness-of-fit test based on deep ReLU neural networks, extending the work of (Shen et al., 2021). The rest of the paper is organized as follows: Section 2 provides a brief introduction to deep neural networks, followed by the proposed goodness-of-fit test. Results from simulation studies and real data analyses are presented in Section 3, and conclusions are drawn in Section 4.

Methods

Deep neural networks (DNNs)

A perceptron (Rosenblatt, 1958) originated from mimicking the functionality of a neuron in the human brain. As shown in Figure 1A, the green node is the only computation unit in a perceptron, and it outputs a nonlinear transformation of the linear combination of input units. Such a transformation in a computation unit is often called an activation function. By stacking multiple perceptrons together, a shallow neural network, shown in Figure 1B, is obtained. The blue computation nodes in the middle are known as the hidden units. Each of them computes a nonlinear activation of a linear combination of the nodes in the input layer. The green nodes are known as output units, and each of them applies a linear or nonlinear activation to a linear combination of the outputs from the hidden units. When the number of hidden layers is more than one, as shown in Figure 1C, a deep neural network is obtained.

Throughout the remainder of the paper, we consider deep neural networks with only one output unit and linear activation is applied to the output unit. In particular, the output of a deep neural network with L hidden layer can be represented as

$$f(\boldsymbol{x}) = \boldsymbol{W}_{L+1}\sigma(\boldsymbol{W}_L\sigma(\cdots\boldsymbol{W}_2\sigma(\boldsymbol{W}_1\boldsymbol{x}))), \tag{1}$$

where $\boldsymbol{W}_l$ is an $n_l \times n_{l-1}$ matrix containing the weights between the (L-1)th layer and the lth layer. Here n_l is the number of nodes in the lth layer. By convention, the 0$^{\text{th}}$ layer represents the input layer, while the (l+1)th layer represents the output layer and therefore, $n_0 = p$ and $n_{L+1} = 1$ by our model assumption. $\sigma: \mathbb{R} \rightarrow \mathbb{R}$ is a nonlinear activation function and in this paper, we considered one of the most used nonlinear activation functions, the Rectified Linear Unit (ReLU) activation function (Nair and Hinton, 2010).

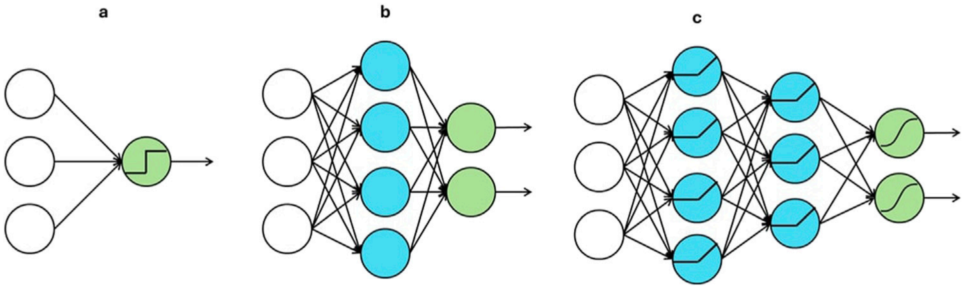


FIGURE 1
Architectures of **(A)** a perceptron, **(B)** a shallow neural network and **(C)** a deep neural network.

That is, $\sigma(x) = \max\{x, 0\}$. In (1), when σ is applied to a matrix or a vector, it is considered as an elementwise operation.

Goodness-of-fit test based on DNNs

We consider the following nonparametric regression model:

$$Y_i = f_0(\boldsymbol{X}_i) + \varepsilon_i, i = 1, \ldots, n$$

where $(\boldsymbol{X}_i, Y_i), i = 1, \ldots, n$ are i.i.d pairs of data points with $\boldsymbol{X}_i = [X_{i1}, \ldots, X_{ip}]^T \in \mathbb{R}^p$ being the vector of covariates for the ith individual and Y_i being the response for the ith individual. $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. random errors with mean 0 and variance σ^2. Moreover, f_0 is an underlying function to be estimated using deep neural networks through minimizing the squared error loss:

$$\hat{f}_n = argmin_{f \in \mathcal{F}_{DNN}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(\boldsymbol{X}_i))^2,$$

where $\mathcal{F}_{DNN}$ is the class of deep neural networks of the form Equation 1, that is,

$$\mathcal{F}_{DNN} = \left\{ f(\boldsymbol{x}) = \boldsymbol{W}_{L+1}\sigma(\boldsymbol{W}_L\sigma(\cdots\boldsymbol{W}_2\sigma(\boldsymbol{W}_1\boldsymbol{x}))) : \|f\|_\infty \le M \right\}.$$

In addition, we assume that $\boldsymbol{X}_i$ come from a continuous distribution, $Y_i \in [-M, M]$ for some $M > 0$ and the underlying function is bounded, that is $\|f_0\|_\infty \le M$. These assumptions are required to provide an upper bound for $\|\hat{f}_n - f_0\|_{L^2}$ as demonstrated in (Farrell et al., 2021).

Our goal is to develop a statistical hypothesis testing procedure to test whether certain covariates should be included in the model or not based on the deep neural network estimator $\hat{f}_n$. In other words, for $S \subset \{1, \ldots, p\}$, a subset of indices of covariates, the null hypothesis is H_0: $X_j, j \in S$ are not significant. To gain some insights of the testing procedure, recall that in multiple linear regression, testing the significance of a predictor is equivalent to testing whether its coefficient is zero or not. This is the well-known t-test procedure. However, due to the unidentifiability of neural network parameters, such a method cannot be easily applied to neural networks. On the other hand, such a t-test is equivalent to an F test by comparing the mean squared error under the full model where the predictor is involved and the reduced model where the predictor is excluded from the model. Our goodness-of-fit test for deep neural networks is constructed based on such an idea.

Following (Shen et al., 2021), we proposed to use a goodness-of-fit (GoF) type statistic for genetic association studies using DNNs. Here are the steps to construct the GoF test statistic.

1. Randomly partitioned the dataset into two parts. Denote $0 < \gamma \le 0.5$ to be the proportion of the first part among the total n data points. Also let $m = \lfloor \gamma n \rfloor$ be the number of data points in the first part so that $n - m$ is the number of data points in the second part. For simplicity, we denote $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_m, Y_m)$ to be the first part of the data and $(\boldsymbol{X}_{m+1}, Y_{m+1}), \ldots, (\boldsymbol{X}_n, Y_n)$ to be the second part of the data.
2. Use the first part is used to fit the data under the null hypothesis H_0 and this is done by training a deep neural network whose input layer only involves the covariates $X_j, j \notin S$. The second part is used to fit the data under the alternative hypothesis which is done by fitting a deep neural network using all the covariates. The mean squares errors of these two model fittings are given by

$$T_0 = \frac{1}{m} \sum_{i=1}^{m} \left(Y_i - \hat{f}_{H_0}(\boldsymbol{X}_i)\right)^2,$$

$$T_1 = \frac{1}{n-m} \sum_{i=m+1}^{n} \left(Y_i - \hat{f}_{H_1}(\boldsymbol{X}_i)\right)^2.$$

3. The asymptotic distribution of T_0 and T_1 can be obtained in a similar fashion as of (Shen et al., 2021). Combining Lemma 3 in (Shen et al., 2021) and Theorem 2 in (Farrell et al., 2021), it follows that under the null hypothesis H_0, both T_0 and T_1 are asymptotically standard normally distributed when $B_n L_n \log B_n \log n = o(n)$ where B_n is the number of parameters in the DNN and L_n is the number of hidden layers in the DNN. Therefore,

$$\left[\left(\frac{1}{m} + \frac{1}{n-m}\right)\kappa\right]^{-\frac{1}{2}} (T_0 - T_1) \xrightarrow{d} N(0, 1),$$

where $\kappa = \mathbb{E}(\varepsilon^4)$ is the fourth moment of the random error provided that $B_n L_n \log B_n \log n = o(n)$.

4. The GoF test statistic can be obtained by replacing κ by a consistent estimator:

$$T = \left[\left(\frac{1}{m} + \frac{1}{n-m}\right)\hat{\kappa}_n\right]^{-\frac{1}{2}} (T_0 - T_1),$$

As mentioned in (Yatchew, 1992), a possible choice for $\hat{\kappa}_n$ is

$$\hat{\kappa}_n = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{f}_{H_0}(\boldsymbol{X}_i)\right)^4 - \left(\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{f}_{H_0}(\boldsymbol{X}_i)\right)^2\right)^2.$$

5. The p-value of the test is then calculated the same way as in a two-sided Z-test. In other words, $p = \mathbb{P}(|T| > |t|)$, where t is the observed test statistic.

Network structures

A sufficient condition, as has been mentioned above, to ensure asymptotic normality is $B_n L_n \log B_n \log n = o(n)$. In fact, this condition provides some guidance on how to choose the network structure. Since B_n is the number of parameters in a DNN, $B_n \asymp n^{*2} L_n$, where $n^\star = \max\{n_1, \ldots, n_{L_n}\}$. Therefore, $B_n L_n \log B_n \log n \asymp n^{*2} L_n^2 \log(n^\star L_n) \log n$. Now we consider the following scenarios:

- If $L_n = O(1)$, such as a shallow ReLU neural network, then the sufficient condition is equivalent to $n^{*2} \log n^\star \log n = o(n)$. In this case, one can choose $n^\star = O(n^{\frac{1}{2}-\alpha})$ for some $0 < \alpha < \frac{1}{2}$.
- If $n^\star = O(1)$, i.e., each hidden layer has a bounded number of hidden units, then the sufficient condition is equivalent to $L_n^2 \log L_n \log n = o(n)$. In this case, one can choose $L_n = O(n^{\frac{1-\alpha}{2}})$ for some $0 < \alpha < 1$.
- If both $n^\star$ and L_n can increase with the sample size, then one can choose $n^\star = O(n^\alpha)$ and $L_n = O(n^\beta)$ as long as α and β satisfy $0 < \alpha + \beta < \frac{1}{2}$.

Results

Simulation 1

In this section, we conducted a simulation study to evaluate our proposed test's type I error and power. Since in genetic studies, linear models are the most used method to detect genetic associations, we compared our proposed test with the t-test in linear regression. Specifically, we generated the response variable via the following equation:

$$Y_i = f_0(X_{i1}) + \varepsilon_i, i = 1, \ldots, n,$$

where $\boldsymbol{X}_i = [X_{i0}, X_{i1}]^T, i = 1, \ldots, n$ are i.i.d. random vectors sampled from a uniform distribution on the square $[-1, 1]^2$. $\varepsilon_i, i = 1, \ldots, n$ are i.i.d. random variables sampled from a normal distribution $\mathcal{N}(0, 0.5^2)$. In the simulation, we consider two different functions f_0. One is the quadratic function $f_0(x) = x^2$ and the other one is a trigonometric function $f_0(x) = \cos(2\pi x)$.

Since the first component does not involve in the simulation equation, it was used to evaluate the performance of the type I error of the proposed test. The null hypothesis to be tested is H_0: X_0 is not significant, or equivalently, the index set for this null hypothesis is $S = \{0\}$. The second component in $\boldsymbol{X}_i$ was involved in generating the response, it was therefore to be used to evaluate the power of the proposed test. In this case, the null hypothesis to be tested is H_0: X_1 is not significant, or equivalently, the index set for this null hypothesis is $S = \{1\}$. To test significance of each component, we applied the testing procedure as mentioned above. We started by partitioning the data set into two parts with ratio $\gamma = 0.1$ and $\gamma = 0.5$. Then the majority of the data was used to train a shallow or a deep ReLU neural network under the alternative hypothesis while the minority of the data was used to calculate the mean squared error under the null hypothesis. When we trained the neural networks, the following three network structures were used:

- A shallow ReLU neural network with the number of hidden units being $\lfloor n^{1/3} \rfloor$.
- A deep ReLU neural network with the number of hidden layer being $\lfloor n^{1/3} \rfloor$ and each hidden layer has 18 hidden units.
- A deep ReLU neural network with $\lfloor n^{1/4} \rfloor$ hidden layers and each hidden layer has $\lfloor n^{1/4} \rfloor$ hidden units.

All the three network structures used here meet the requirement as mentioned in section 2.3. In the simulation, we considered sample sizes being 200, 500, 1,000 and 2000. The stochastic gradient descent algorithm was applied, and the batch size was determined so that 20 batches were used for each sample size. 200 epochs were used to run the stochastic gradient descent. To further alleviate the possible overfitting, we applied dropout to each hidden unit in the network with a dropout rate being 0.05. To obtain the empirical type I error and the empirical power, 1,000 Monte Carlo replications were conducted. Tables 1, 2 below summarize the simulation results.

Based on Tables 1, 2, it can be easily seen that linear models and the proposed GoF test can control the empirical type I error very well at level 0.05, except that the proposed GoF test is slightly conservative when the sample size is small for the quadratic signal for the split-ratio $\gamma = 0.1$, while the empirical type I error rate of the GoF test is slightly inflated for small sample size when the split ratio $\gamma = 0.5$. The empirical powers of proposed GoF test based on ReLU neural networks are consistently much higher compared to the t-test in linear model, which suggests that the proposed GoF test can outperform the t-test in linear model when the underlying signal is nonlinear. On the other hand, it is worth noting that when $\gamma = 0.1$, shallow ReLU neural networks achieve higher empirical power than deep ReLU neural networks in both cases, especially when the sample size is relatively large. On the contrary, when the underlying function is the cosine function and the sample size is 200, deep ReLU neural networks have higher power compared to the shallow ones. Similar situations can also be seen for $\gamma = 0.5$, but for the cosine signal, deep neural networks with structure 1 (growing number of hidden layers and fixed number of hidden units in each layer) achieve higher power compared to shallow neural networks. Therefore, we believe that these observations suggest that the rule of parsimony still applies in ReLU neural networks.

Simulation 2

In many situations, a response variable can be related to multiple causal variables. In this simulation, we investigated the performance of the proposed method under such a scenario. In particular, the response variable in this simulation was generated based on the following equation:

TABLE 1 Comparisons between linear model and goodness-of-fit test based on ReLU neural networks under quadratic signal.

		$\gamma = 0.1$				$\gamma = 0.5$			
Sample size		200	500	1,000	2,000	200	500	1,000	2,000
Type I Error	Linear Model	0.047	0.047	0.055	0.048	0.041	0.041	0.038	0.054
	Shallow ReLU NN	0.028	0.053	0.050	0.053	0.102	0.066	0.056	0.053
	Deep ReLU NN 1	0.030	0.054	0.049	0.052	0.108	0.066	0.053	0.050
	Deep ReLU NN 2	0.046	0.048	0.039	0.042	0.088	0.061	0.055	0.051
Power	Linear Model	0.058	0.071	0.068	0.076	0.073	0.068	0.058	0.063
	Shallow ReLU NN	0.152	0.367	0.580	0.858	0.484	0.736	0.955	1.000
	Deep ReLU NN 1	0.098	0.295	0.543	0.787	0.594	0.774	0.952	0.998
	Deep ReLU NN 2	0.056	0.176	0.448	0.738	0.273	0.513	0.830	0.944

TABLE 2 Comparisons between linear model and goodness-of-fit test based on ReLU neural networks under cosine signal.

		$\gamma = 0.1$				$\gamma = 0.5$			
Sample size		200	500	1,000	2,000	200	500	1,000	2,000
Type I Error	Linear Model	0.063	0.046	0.062	0.051	0.055	0.048	0.049	0.060
	Shallow ReLU NN	0.057	0.050	0.056	0.063	0.072	0.079	0.056	0.050
	Deep ReLU NN 1	0.054	0.048	0.056	0.059	0.081	0.075	0.048	0.050
	Deep ReLU NN 2	0.039	0.061	0.040	0.052	0.064	0.076	0.048	0.052
Power	Linear Model	0.051	0.058	0.061	0.055	0.062	0.050	0.043	0.068
	Shallow ReLU NN	0.106	0.483	0.876	0.952	0.551	0.858	0.966	0.996
	Deep ReLU NN 1	0.228	0.295	0.413	0.425	0.970	0.982	0.981	0.922
	Deep ReLU NN 2	0.042	0.083	0.262	0.622	0.218	0.541	0.789	0.911

$$Y_i = |X_{1i}| + 2X_{2i}^2 + \cos(2\pi X_{3i}) + \epsilon_i,$$

where all the covariates $X_{0i}, X_{1i}, X_{2i}, X_{3i}$ are i.i.d. random variables from Uniform[-1,1]. The random error term is sampled from $\mathcal{N}(0, 0.5^2)$. Similar to Simulation 1, the variable X_0 is not involved in the underlying function, so it was used to check type I error of the test, and the other three variables were used to evaluate the power of the test.

In this scenario, the hypotheses of interest are H_0: X_j is not significant for $j \in S$ with $S = \{0\}$ for type I error and $S = \{1\}, \{2\}, \{3\}$ respectively for the three variables used to evaluate power. We used the same deep neural network structures and the same choices of tuning parameters as we did in Simulation 1. Table 3 summarize the empirical type I error rates and the empirical power of the proposed method, linear model, and the black-box test under the sample sizes 200, 500, 1,000, and 2,000.

As we can see from Table 3, both linear model t-test and the proposed GoF test can control the type I error rate very well. Similar to what we saw from Simulation 1, even the underlying function contains multiple causal variables, the proposed GoF test can still detect the significance of the variables having nonlinear associations with the response variable.

Real data analyses

Alzheimer's disease (AD) is one of the most common neurodegenerative diseases with a substantial genetic component (Karch et al., 2014; Sims et al., 2020). Therefore, it is of great importance to have an efficient method to screen the genetic components that are associated with AD pathogenesis so that early treatments can be applied for disease management (Zissimopoulos et al., 2015). To investigate the performance of our proposed GoF test in identifying AD-related genes, we applied our proposed method to the gene expression data from Alzheimer's Disease Neuroimaging Initiative (ADNI).

The hippocampus region plays a vital role in memory (Mu and Gage, 2011) and the shrinkage of hippocampus volume is an early symptom of AD (Schuff et al., 2009). Therefore, we chose the hippocampus volume as the phenotype in the real data analysis. After removing individuals with missing values for hippocampus volume and merging data from individuals having both gene expression information and hippocampus volume, a total of 464 individuals and 15,837 gene expressions were obtained. We then regressed the scaled hippocampus volume onto some important predictors including age, gender and education status.

TABLE 3 Comparisons between linear model and goodness-of-fit test based on ReLU neural networks under multiple causal variables.

		$\gamma = 0.1$				$\gamma = 0.5$			
Sample size		200	500	1,000	2,000	200	500	1,000	2,000
Type I Error (X_0)	Linear Model	0.058	0.046	0.044	0.043	0.052	0.047	0.056	0.048
	Shallow ReLU NN	0.046	0.043	0.044	0.064	0.076	0.064	0.048	0.054
	Deep ReLU NN 1	0.044	0.044	0.045	0.065	0.071	0.061	0.046	0.055
	Deep ReLU NN 2	0.047	0.043	0.042	0.063	0.063	0.064	0.046	0.054
Power (X_1)	Linear Model	0.066	0.061	0.056	0.042	0.040	0.045	0.049	0.041
	Shallow ReLU NN	0.049	0.064	0.108	0.127	0.128	0.134	0.172	0.287
	Deep ReLU NN 1	0.050	0.068	0.070	0.078	0.130	0.131	0.136	0.181
	Deep ReLU NN 2	0.048	0.055	0.058	0.074	0.084	0.072	0.075	0.107
Power (X_2)	Linear Model	0.081	0.075	0.065	0.062	0.074	0.065	0.070	0.087
	Shallow ReLU NN	0.057	0.387	0.710	0.967	0.533	0.859	0.974	0.998
	Deep ReLU NN 1	0.076	0.106	0.119	0.146	0.514	0.777	0.912	0.952
	Deep ReLU NN 2	0.051	0.057	0.072	0.321	0.170	0.361	0.647	0.834
Power (X_3)	Linear Model	0.045	0.055	0.065	0.059	0.040	0.050	0.054	0.064
	Shallow ReLU NN	0.046	0.082	0.373	0.568	0.163	0.228	0.273	0.314
	Deep ReLU NN 1	0.054	0.093	0.203	0.263	0.404	0.633	0.749	0.666
	Deep ReLU NN 2	0.050	0.042	0.055	0.119	0.077	0.111	0.171	0.309

TABLE 4 Top 10 significant genes selected from *t*-test in linear model and the GoF tests based on different ReLU neural network structures.

Linear model	Shallow ReLU neural network	Deep ReLU neural network 1	Deep ReLU neural network 2
SNRNP40	*GRM2*	*GRM2*	*GRM2*
PPIH	*DGCR6*	*DGCR6*	*DGCR6*
GPR85	*GPRC5D*	*BRCA2*	*NDRG1*
DNAJB1	*SMARCB1*	*KIF1C*	*GPRC5D*
WDR70	*NDRG1*	*NDRG1*	*KIF1C*
CYP4F2	*KIF1C*	*GPRC5D*	*KLF13*
NOD2	*NUDT22*	*NUDT22*	*COX20*
MEGF9	*BRCA2*	*COX20*	*NUDT22*
CTBP1-AS2	*COX20*	*SMARCB1*	*OR4A5*
PHYKPL	*REG1A*	*STAG3L4*	*STAG3L4*

The residual obtained will be used as the response variable to train ReLU neural networks. The network structures and hyperparameters in the ReLU neural networks used in the real data analysis were the same as in the simulation studies. Table 4 summarizes the top 10 significant genes selected from *t*-test in linear model and the GoF tests based on ReLU neural networks.

As can be seen from Table 4, the significant genes selected from the GoF test do not overlap with the ones selected from the linear models, and different network structures picked out similar genes. On the other hand, in (Shen et al., 2022), the top 10 significant genes selected using a testing procedure based on shallow sigmoid neural networks have large overlap with the ones selected from the linear model. This indicates that ReLU neural networks may be able to detect different signals that are hard to detect when using linear models or shallow sigmoid neural networks. Among them, the gene *GRM2* is the top pick. Although the biological mechanism of the association between these genes and AD needs further validation, it is worth pointing out that a recent study has shown that the metabotropic glutamate receptor 2 (mGluR2), a protein encoded by the gene *GRM2* plays a role in the pathogenesis of AD (Srivastava et al., 2020).

Discussions and conclusion

In this paper, we have proposed a goodness-of-fit test based on ReLU neural networks. The proposed test can be used to detect the significance of a predictor. Once the network structures are suitably chosen, the test statistics have an asymptotically normal distribution, making it easy to implement in practice. Simulation results have demonstrated that the proposed method can detect nonlinear underlying signals, and real data analysis also showed the potential that ReLU neural networks may detect signals that are hard to identify from linear models or even shallow sigmoid neural networks.

On the other hand, although the theoretical framework of the GoF test was proposed in this paper, in practice, the performance of a deep ReLU neural network also depends on the optimization algorithm used and the hyperparameters (e.g., learning rate, number of epochs, etc.) selected. So, there is still a gap in how the DNN can be used to conduct statistical inference on detecting significant variables. This will be our future work. In addition, while we mainly focused on testing a single variable (such as a gene expression in the real data analysis) in this paper, it is worthwhile to also investigate the performance of our proposed method on a wider range of datasets to evaluate the performance of the GoF test when testing a set of variants in a genetic region, such as in a chromosome or in a pathway. In addition, various significant testing procedures based on neural networks nowadays and as a future work, we plan to conduct a comprehensive comparison on these methods.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/supplementary material.

Author contributions

XS: Conceptualization, Formal Analysis, Methodology, Project administration, Supervision, Writing–original draft, Writing–review and editing. XW: Formal Analysis, Investigation, Software, Writing–original draft, Writing–review and editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

ChatGPT 4o was used to correct grammatical mistakes.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inf. Theory* 39, 930–945. doi:10.1109/18.256500

Braun, A., Kohler, M., Langer, S., and Walk, H. (2024). Convergence rates for shallow neural networks learned by gradient descent. *Bernoulli* 30, 475–502. doi:10.3150/23-BEJ1605

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Math. Control Signal Syst.* 2, 303–314. doi:10.1007/BF02551274

Dai, B., Shen, X., and Pan, W. (2024). Significance tests of feature relevance for a black-box learner. *IEEE Trans. Neural Netw. Learn. Syst.* 35, 1898–1911. doi:10.1109/TNNLS.2022.3185742

Farrell, M. H., Liang, T., and Misra, S. (2021). Deep neural networks for estimation and inference. *Econometrica* 89, 181–213. doi:10.3982/ECTA16901

Fukumizu, K. (2003). Likelihood ratio of unidentifiable models and multilayer neural networks. *Ann. Statistics* 31, 833–851. doi:10.1214/aos/1056562464

Horel, E., and Giesecke, K., 2019. Towards explainable ai: significance tests for neural networks. arXiv preprint arXiv:1902.06021.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Netw.* 2, 359–366. doi:10.1016/0893-6080(89)90020-8

Karch, C. M., Cruchaga, C., and Goate, A. M. (2014). Alzheimer's disease genetics: from the bench to the clinic. *Neuron* 83, 11–26. doi:10.1016/j.neuron.2014.05.041

LeCun, Y. (1989). "Generalization and network design strategies," in *Connectionism in perspective*. Editors R. Pfeifer, Z. Schreter, F. Fogelman, and L. Steels

Mei, S., Misiakiewicz, T., and Montanari, A. (2022). Generalization error of random feature and kernel methods: hypercontractivity and kernel matrix concentration. *Appl. Comput. Harmon. Analysis, Special Issue Harmon. Analysis Mach. Learn.* 59, 3–84. doi:10.1016/j.acha.2021.12.003

Mei, S., and Montanari, A. (2022). The generalization error of random features regression: precise asymptotics and the double descent curve. *Commun. Pure Appl. Math.* 75, 667–766. doi:10.1002/cpa.22008

Mu, Y., and Gage, F. H. (2011). Adult hippocampal neurogenesis and its role in Alzheimer's disease. *Mol. Neurodegener.* 6, 85. doi:10.1186/1750-1326-6-85

Nair, V., and Hinton, G. E. (2010). "Rectified linear units improve restricted Boltzmann machines," in Proceedings of the 27th international conference on machine learning, Haifa, June 21, 2010, 807–814.

Pinkus, A. (1999). Approximation theory of the MLP model in neural networks. *Acta Numer.* 8, 143–195. doi:10.1017/S0962492900002919

Quang, D., Chen, Y., and Xie, X. (2014). DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* 31, 761–763. doi:10.1093/bioinformatics/btu703

Quang, D., and Xie, X. (2016). DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic acids Res.* 44, e107. doi:10.1093/nar/gkw226

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.* 65, 386–408. doi:10.1037/h0042519

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1988). Learning representations by back-propagating errors. *Cogn. Model.* 5, 1. doi:10.1038/323533a0

Schuff, N., Woerner, N., Boreta, L., Kornfield, T., Shaw, L. M., Trojanowski, J. Q., and The Alzheimer's; Disease Neuroimaging Initiative (2009). MRI of hippocampal volume loss in early Alzheimer's disease in relation to ApoE genotype and biomarkers. *Brain* 132, 1067–1077. doi:10.1093/brain/awp007

Shen, X., Jiang, C., Sakhanenko, L., and Lu, Q. (2021). A goodness-of-fit test based on neural network sieve estimators. *Statistics and Probab. Lett.* 174, 109100. doi:10.1016/j.spl.2021.109100

Shen, X., Jiang, C., Sakhanenko, L., and Lu, Q. 2022. A sieve quasi-likelihood ratio test for neural networks with applications to genetic association studies. doi:10.48550/arXiv.2212.08255

Shen, X., Jiang, C., Sakhanenko, L., and Lu, Q. (2023). Asymptotic properties of neural network sieve estimators. *J. Nonparametric Statistics* 35, 839–868. doi:10.1080/10485252.2023.2209218

Sims, R., Hill, M., and Williams, J. (2020). The multiplex model of the genetics of Alzheimer's disease. *Nat. Neurosci.* 23, 311–322. doi:10.1038/s41593-020-0599-5

Srivastava, A., Das, B., Yao, A. Y., and Yan, R. (2020). Metabotropic glutamate receptors in alzheimer's disease synaptic dysfunction: therapeutic opportunities and hope for the future. *J. Alzheimers Dis.* 78, 1345–1361. doi:10.3233/JAD-201146

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Methodol.* 58, 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in Advances in Neural Information Processing Systems 30: Annual Conferenceon Neural Information Processing Systems, Long Beach, CA, December 4–9, 2017, 5998–6008.

Yatchew, A. J. (1992). Nonparametric regression tests based on least squares. *Econ. Theory* 8, 435–451. doi:10.1017/S0266466600013153

Zhou, X., Chen, Yu, Ip, F. C. F., Jiang, Y., Cao, H., Lv, G., et al. (2023). Deep learning-based polygenic risk analysis for Alzheimer's disease prediction. *Commun. Med.* 3, 49–20. doi:10.1038/s43856-023-00269-x

Zissimopoulos, J., Crimmins, E., and St.Clair, P. (2015). The value of delaying alzheimer's disease onset. *Forum Health Econ. Policy* 18, 25–39. doi:10.1515/fhep-2014-0013

frontiers | Frontiers in Systems Biology

TYPE Review
PUBLISHED 07 January 2025
DOI 10.3389/fsysb.2024.1422384

A multi-omics strategy to understand PASC through the RECOVER cohorts: a paradigm for a systems biology approach to the study of chronic conditions

Jun Sun[1]*, Masanori Aikawa[2], Hassan Ashktorab[3], Noam D. Beckmann[4,5], Michael L. Enger[6], Joaquin M. Espinosa[7], Xiaowu Gai[8,9], Benjamin D. Horne[10,11], Paul Keim[12,13,14], Jessica Lasky-Su[15], Rebecca Letts[16], Cheryl L. Maier[17], Meisha Mandal[6], Lauren Nichols[16], Nadia R. Roan[18,19], Mark W. Russell[20], Jacqueline Rutter[16], George R. Saade[21,22], Kumar Sharma[23,24], Stephanie Shiau[25], Stephen N. Thibodeau[26], Samuel Yang[27], Lucio Miele[28]* and NIH Researching COVID to Enhance Recovery (RECOVER) Consortium

[1]Department of Medicine, Division of Gastroenterology and Hepatology, University of Illinois Chicago, Chicago, IL, United States, [2]Cardiovascular Division and Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, United States, [3]Department of Medicine, Howard University, Washington, DC, United States, [4]Department of Medicine, Division of Data Driven and Digital Medicine (D3M), New York, NY, United States, [5]Charles Bronfman Institute for Personalized Medicine, Mount Sinai Clinical Intelligence Center, Icahn School of Medicine at Mount Sinai, New York, NY, United States, [6]RTI International, Durham, NC, United States, [7]Linda Crnic Institute for Down Syndrome, University of Colorado Anschutz Medical Campus, Aurora, CO, United States, [8]Department of Pathology and Laboratory Medicine, Children's Hospital Los Angeles, Los Angeles, CA, United States, [9]Department of Pathology, Keck School of Medicine, University of Southern California, Los Angeles, CA, United States, [10]Intermountain Medical Center Heart Institute, Murray, UT, United States, [11]Department of Medicine, Division of Cardiovascular Medicine, Stanford University, Stanford, CA, United States, [12]Department of Biology, Northern Arizona University, Flagstaff, AZ, United States, [13]Pathogens Genomics Program, Translational Genomics Institute (TGen), Phoenix, AZ, United States, [14]Department of Biology, University of Oxford, Oxford, United Kingdom, [15]Channing Department of Network Medicine, Brigham and Women's Hospital, Harvard University, Boston, MA, United States, [16]RECOVER patient representative, Durham, NC, United States, [17]Department of Pathology and Laboratory Medicine, Emory University School of Medicine, Atlanta, GA, United States, [18]Gladstone Institute of Virology, Gladstone Institutes, San Francisco, CA, United States, [19]Department of Urology, University of California San Francisco, San Francisco, CA, United States, [20]Department of Pediatrics, Division of Pediatric Cardiology, University of Michigan, Ann Arbor, MI, United States, [21]Department of Obstetrics and Gynecology, University of Texas Medical Branch, Galveston, TX, United States, [22]Department of Obstetrics and Gynecology, Eastern Virginia Medical School, Norfolk, VA, United States, [23]Center for Precision Medicine, University of Texas San Antonio Health Sciences Center, San Antonio, TX, United States, [24]Department of Medicine, Division of Nephrology, University of Texas San Antonio Health Sciences Center, San Antonio, TX, United States, [25]Department of Biostatistics and Epidemiology, Rutgers School of Public Health, Piscataway, NJ, United States, [26]Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, United States, [27]Department of Emergency Medicine, Stanford University, Stanford, CA, United States, [28]Department of Genetics, School of Medicine, Louisiana State University Health Sciences, Center New Orleans, New Orleans, LA, United States

OPEN ACCESS

EDITED BY
Yoram Vodovotz,
University of Pittsburgh, United States

REVIEWED BY
Elebeoba E. May,
University of Wisconsin-Madison, United States
Song Feng,
Pacific Northwest National Laboratory (DOE),
United States

*CORRESPONDENCE
Lucio Miele,
✉ lmiele@lsuhsc.edu
Jun Sun,
✉ junsun7@uic.edu

RECEIVED 23 April 2024
ACCEPTED 14 October 2024
PUBLISHED 07 January 2025

CITATION
Sun J, Aikawa M, Ashktorab H, Beckmann ND, Enger ML, Espinosa JM, Gai X, Horne BD, Keim P, Lasky-Su J, Letts R, Maier CL, Mandal M, Nichols L, Roan NR, Russell MW, Rutter J, Saade GR, Sharma K, Shiau S, Thibodeau SN, Yang S, Miele L and NIH Researching COVID to Enhance Recovery (RECOVER) Consortium (2025) A multi-omics strategy to understand PASC through the RECOVER cohorts: a paradigm for a systems biology approach to the study of chronic conditions.
Front. Syst. Biol. 4:1422384.
doi: 10.3389/fsysb.2024.1422384

Post-Acute Sequelae of SARS-CoV-2 infection (PASC or "Long COVID"), includes numerous chronic conditions associated with widespread morbidity and rising healthcare costs. PASC has highly variable clinical presentations, and likely includes multiple molecular subtypes, but it remains poorly understood from a molecular and mechanistic standpoint. This hampers the development of rationally targeted therapeutic strategies. The NIH-sponsored "Researching COVID to Enhance Recovery" (RECOVER) initiative includes several retrospective/prospective observational cohort studies enrolling adult, pregnant adult and pediatric patients respectively. RECOVER formed an "OMICS" multidisciplinary task force, including clinicians, pathologists, laboratory scientists and data scientists, charged with developing recommendations to apply cutting-edge system biology technologies to achieve the goals of RECOVER. The task force met biweekly over 14 months, to evaluate published evidence, examine the possible contribution of each "omics" technique to the study of PASC and develop study design recommendations. The OMICS task force recommended an integrated, longitudinal, simultaneous systems biology study of participant biospecimens on the entire RECOVER cohorts through centralized laboratories, as opposed to multiple smaller studies using one or few analytical techniques. The resulting multi-dimensional molecular dataset should be correlated with the deep clinical phenotyping performed through RECOVER, as well as with information on demographics, comorbidities, social determinants of health, the exposome and lifestyle factors that may contribute to the clinical presentations of PASC. This approach will minimize lab-to-lab technical variability, maximize sample size for class discovery, and enable the incorporation of as many relevant variables as possible into statistical models. Many of our recommendations have already been considered by the NIH through the peer-review process, resulting in the creation of a systems biology panel that is currently designing the studies we proposed. This system biology strategy, coupled with modern data science approaches, will dramatically improve our prospects for accurate disease subtype identification, biomarker discovery and therapeutic target identification for precision treatment. The resulting dataset should be made available to the scientific community for secondary analyses. Analogous system biology approaches should be built into the study designs of large observational studies whenever possible.

KEYWORDS

COVID-19, PASC, RECOVER, systems biology, multi-omics

1 Introduction

The term Post-Acute Sequelae of SARS-CoV-2 infection (PASC), also known as "Long COVID", refers to numerous conditions associated with widespread morbidity and rising healthcare costs. PASC has highly variable clinical presentations, and likely includes multiple molecular subtypes (Thompson et al., 2023; Sherif et al., 2023). The NIH-sponsored "Researching COVID to Enhance Recovery" (RECOVER) initiative includes retrospective/prospective cohort studies including an adult cohort (Horwitz et al., 2023), a cohort of pregnant adults (Metz et al., 2023; Reel et al., 2021) and a pediatric cohort (Gross et al., 2024; Reel et al., 2021). These studies aim to enroll a total of 12,580 adult non-pregnant patients, 2,300 adult pregnant patients and 19,300 pediatric patients to rapidly improve our understanding of and ability to predict, treat, and prevent Post-Acute Sequelae of SARS-CoV-2 infection (PASC, or "Long COVID") through deep clinical phenotyping and laboratory studies. THE RECOVER "OMICS" Task Force was charged with developing recommendations based on published evidence and the experiences of its members, to incorporate multi-omics into the analysis of RECOVER results.

2 Methods

2.1 Objectives

The "OMICS" task force of the RECOVER study, a multi-disciplinary committee including clinicians, pathologists, laboratory scientists and data scientists, was charged with developing recommendations to apply cutting-edge system biology technologies to achieve the goals of RECOVER. The task force met biweekly over 14 months, to evaluate published evidence, examine the possible contribution of each "omics" technique to the study of PASC, as well as the potential limitations of each technique, and develop a consensus recommendation. The work was divided into two stages. During the first stage, sub-committees with specific expertise on an "omics" technique examined evidence supporting the use of that technique to study PASC, the type of data it could generate and the mechanistic questions it could answer, based on published evidence and the experiences of its members, to incorporate multi-omics into the analysis of RECOVER results. Each sub-committee presented to the entire task force. During the second stage, the task force combined the findings of each sub-committee into a comprehensive study design recommendation.

3 Results and discussion

The OMICS task force recommended that integrated, longitudinal, simultaneous multi-omics studies of participant biospecimens be performed on the entire RECOVER cohort through centralized laboratories, as opposed to multiple smaller studies using one or few analytical techniques.

The RECOVER adult protocol (Horwitz et al., 2023) includes multiple biospecimen collections: nasopharyngeal or nasal swab, 2 8.5 mL aliquots of blood in serum separation tubes, 4 × 8 ml aliquots of blood in cell preparation tubes, 2 × 2.7 mL aliquots of blood in sodium citrate tubes for plasma proteomics, 1 × 10 ml aliquot of blood in EDTA tube, 1 2.5 mL aliquot of blood in PAXgene RNA tube, 1 × 10 ml urine (no additives), 1 × 2mL aliquot of saliva in Oragene OGR 600 and 1 25 mL aliquot of stool. Of these, stool is sent by participants while the other samples are processed locally as per protocol specifications and shipped in batches to the central tissue bank. Participants who consent to biospecimen donation for future research are asked to provide blood and nasopharyngeal/nasal swab biospecimens at enrollment, 90 and 180 days after the index date (date of first infection or negative COVID test), and then annually (Horwitz et al., 2023). Saliva is collected upon enrollment for genetic analysis. Urine and stool are collected biannually. Additionally, a battery of clinical laboratory tests is performed in CLIA-certified laboratories at enrollment, 90 and 189 days after the index date, and thereafter, abnormal tests are repeated annually. Specific symptoms or abnormal study results trigger “Tier 2” or “Tier 3” assessments (see (Horwitz et al., 2023) for details). A SARS-CoV-2 PCR test is performed at enrollment for all “uninfected” participants, who are also tested for SARS-CoV-2 nucleocapsid antibodies spike protein antibodies for unvaccinated participants.

In addition to study visits, imaging and laboratory tests, participants complete multiple surveys, using validated survey instruments whenever possible, at 90-day intervals throughout the study. At enrollment, data are collected on demographics, social determinants of health (SDOH), disability, characteristics of the initial SARS-CoV-2 infection (if applicable), pregnancy (if applicable), vaccination status, comorbidities, medications, and PASC symptoms. Subsequently, at 90-day intervals, data are collected on interim infections, time-varying social determinants, vaccinations, comorbidities, medications and symptoms (Horwitz et al., 2023). The PASC symptom survey was developed for RECOVER and includes an overall quality of life instrument (PROMIS-10) and screening for core symptoms (43 for biological males and 46 for biological females) drawn from existing literature plus input from patient representatives and investigators. Questions about depression, anxiety, post-traumatic stress disorder (PTSD), and grief are also included. Report of a symptom may trigger additional questions about that symptom. Details of survey instruments are in the original reference (Horwitz et al., 2023).

The pregnancy study (Metz et al., 2023) follows a similar design to the non-pregnant adult study, enrolling participants with suspected, probable or confirmed SARS-CoV-2 infection during pregnancy, or documented lack of exposure to SARS-CoV-2 during pregnancy. Study procedures and biospecimen collections are analogous to those in the non-pregnant adult study (Metz et al., 2023), with modifications for breastfeeding or *postpartum* participants, and additional health and developmental assessments for babies exposed *in utero* to SARS-CoV-2.

The RECOVER pediatric study (Gross et al., 2024) has a similar design, with limitations due to the age range of participants. All pediatric participants complete a single Tier 1 visit including PROMIS global health measures and symptom screening. This visit includes a donation of saliva and capillary blood. Depending on infection status, clinical history, symptoms and probability of PASC, pediatric participants are promoted to Tier 2 or Tier 3, which include additional biospecimen donations during the acute and post-acute phase of PASC, as well as additional clinical assessments and surveys. The types, aliquot numbers, and cadence of biospecimen collections are described in detail in (Gross et al., 2024).

In summary, each RECOVER study will generate vast longitudinal datasets including clinical, demographic, medication, SDOH and lifestyle data for each participant, as well as sufficient types and numbers of biospecimen aliquots to permit a comprehensive, longitudinal multi-omics investigation. Potential environmental exposures can furthermore be estimated from census tract or ZIP code data.

The multi-dimensional molecular dataset generated by the multi-omics investigation should be correlated with the deep clinical phenotyping performed through RECOVER, as well as with information on demographics, comorbidities, social determinants of health, the exposome and lifestyle factors collected through RECOVER surveys, that may contribute to the clinical presentations of PASC. Data generation and analytical strategies should leverage integrative bioinformatics and machine learning.

A major advantage, and a potential challenge, of multi-omics approaches is that datasets derived from different analytical techniques and measured using different scales must be integrated. Approaches including multi-omics integration paired with ML have been gaining popularity in clinical and biomedical research (see (Reel et al., 2021; Niranjan et al., 2023)), though this field is rapidly evolving. An important advantage inherent in multi-dimensional measurements is that the extent to which different measurements agree with each other or not is potentially informative. For instance, transcriptomic data may or may not be reflected in the relative abundance of protein products, or quantitative differences in non-coding RNA expression may or may not translate into relative abundance of potential target mRNAs, the proteins they encode or the metabolites that these proteins may process. System biology approaches based on multi-omics have been used successfully in the study of cardiovascular disease (Joshi et al., 2021).

With respect to PASC, strategies similar to what we propose have been used on a smaller scale. ML has been used in the context of a multi-step analytical strategy to combine proteomic and metabolomic data to generate a multi-omics biomarker predictive of the risk of PASC (Wang et al., 2023) and give insights on the metabolic pathways altered during PASC. Dimensionality reduction was achieved through unsupervised cluster analysis followed by autoencoder (AE), using a three-layer neural network. Supervised ML was then used to identify the minimal number of molecules predictive of adverse clinical outcomes. This study, though very promising, was limited by small sample size (117, of whom 105 were

TABLE 1 Complementary data types captured by multi-omics assays.

Data type	Assays							
	SNP NGS	Epigenome	Bulk RNASeq	scRNASeq	Proteome	CyTOF	Metabolome	Microbiome
Genetic risk factors	+							
Epigenetic modifications		+	+					
mRNA/splice variants			+					
ncRNA			+					
Viral RNA			+					
Immune phenotyping			+	+	+	+		+
Antibodies					+			
Cytokines, chemokines					+			
Peptide hormones					+			
Coagulation factors					+			
Viral proteins					+			
Post-translational modifications					+	+		
Human Metabolites							+	
Bacterial metabolites							+	
Toxins/drugs							+	
Vitamins/hormones							+	
Bacterial diversity								+

TABLE 2 Approximate sample requirements for multi-omics assays.

Approximate amount of material	Assays							
	SNP NGS, GWAS	Methylome	Bulk RNASeq	scRNASeq	Proteome	CyTOF	Metabolome	Microbiome
	300–400 ng DNA (50 μL blood)	50–100 ng DNA	PBMC in 1 mL blood (250 ng RNA)	100,000 PBMC (20–30 μL blood)	400–500 μL EDTA plasma	106 cells (200–300 μL blood)	50–100 μL plasma	<500 mg stool

used as a training cohort for model development and 10% as a validation cohort), the severity of acute COVIDs in the patients enrolled and the absence of a vaccinated group. Despite these limitations, these results indicate that similar analytical strategies can be used successfully on a much larger sample with broader phenotyping, to discover predictive biomarkers, therapeutic targets and risk factors and to generate mechanistic hypotheses.

Within the RECOVER study, an unsupervised ML approach has been used to identify clinical subtypes of PASC, after symptoms differentiating infected from uninfected patients were identified using LASSO (least absolute shrinkage and selection operator) (Thaweethai et al., 2023).

Highly multiplexed “omics” approaches measure common clinical analytes and many more parameters (Table 1) at a fraction of the cost of traditional clinical tests, oftentimes using similar quantities of specimens (Table 2). In a multi-omics approach, analytes within each category (e.g., proteins, lipids, nucleic acids, metabolites, and microbes) are all measured simultaneously, generating high-content data that is more than the sum of its parts. This approach allows the discovery of new molecular signatures to enhance our understanding of complex disease pathophysiology. These signatures may occur within a single analyte category, but more likely cover more complex patterns that span multiple molecular layers, e.g., genomics,

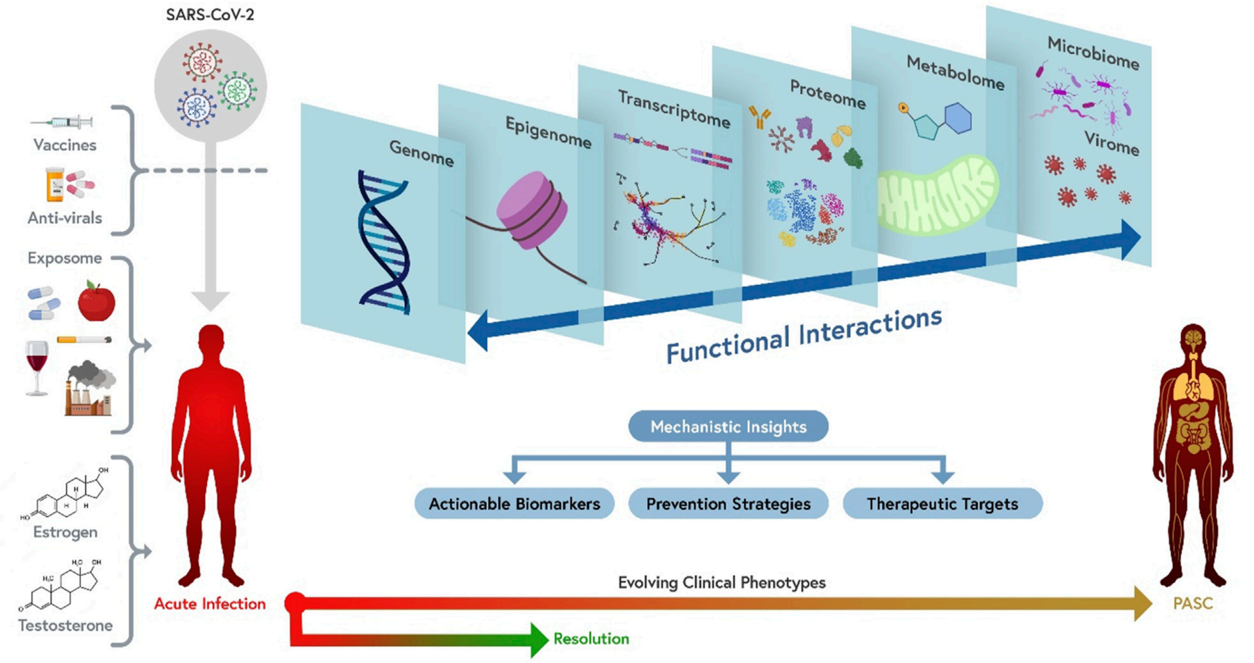


FIGURE 1
A comprehensive multi-omics approach to the mechanism(s) of PASC. From left to right: PASC is a consequence of infection with SARS-CoV-2. Different viral variants or sub-variants (represented in different colors) may have different probability of causing PASC or be associated with different presentations (e.g., due to different ability to cause persistent infection, to trigger pathogenic antibody responses, or to damage vascular endothelium). Vaccines and anti-viral agents can decrease the risk of PASC by interfering with viral persistence and replication. Multiple exposures, including diet, medications, tobacco, alcohol, environmental pollutants and co-morbidities, socioeconomic and psychosocial exposures, as well as sex hormones, can potentially affect the risk and clinical presentations of PASC. The combined effect of these factors results into evolving clinical phenotypes ranging from acute COVID-19 resolution to PASC through a number of mechanisms that can be best understood by simultaneously interrogating the multi-omics landscape of patients, including individual genomics, epigenomics, bulk and single-cell transcriptomics, plasma and cellular proteomics, metabolomics, and microbiome/virome. These different dimensions functionally interact with one another to determine pathogenetic mechanisms (e.g., persistent viral infection, modulated by individual genetics, triggers immune, inflammatory and metabolic changes that are in turn modulated by the intestinal and respiratory microbiomes and potentially by reactivation of other viruses). Insights generated by an integrated multi-omics investigation of patients with well-characterized clinical phenotypes are likely to identify actionable biomarkers (which may discriminate between PASC molecular subtypes), as well as therapeutic targets and prevention strategies. Orthogonal multi-omics tests repeated over time are the most informative approach to capture the pathogenesis of the different clinical presentations of PASC and their evolution over time.

epigenomics, transcriptomics, proteomics, lipidomics, metabolomics, and microbiomics (Figure 1). Deep multi-omics profiling will allow us to explore a broad spectrum of pathophysiological mechanisms (Table 3), define gene-environment interactions involved in the pathogenesis of PASC, identify molecular subtypes and candidate biomarkers and propose mechanism-based therapeutic strategies. The relative contributions of each “omics” we evaluated and considerations on data generation and analysis are described below.

3.1 Evidence supporting multi-omics technologies used in COVID-19 and PASC studies

3.1.1 Genomics

Genomics is an invaluable asset to understand disease risk, mechanism and etiology, and to serve as a backbone to allow for better modeling of multi-omics profiles in patient populations. Several genome-wide association studies (GWAS) have identified reproducible associations between specific loci and risk and outcomes of acute COVID-19 (Ferreira et al., 2022) with the most reproducible being with LZTFL1 and contiguous regions on 3q21.31 and ABO on 9q34.2. A recent GWAS study, currently in pre-print, detected an association between a locus near the FOXP4 gene and risk of developing PASC (Lammi et al., 2023). That study analyzed data from 24 studies conducted in 16 countries, totaling 6,450 PASC cases and 1,093,995 controls. However, most of the patients were of European ancestry, and this study should be replicated in a more diverse cohort. In GWAS studies, sample size and composition of study population (e.g., case/control ratio, ancestry, genetic admixture, *etc.*) are critical. FOXP4 is a broadly expressed transcription factor. Lammi et al. (Lammi et al., 2023) analyzed single-cell RNASeq data to confirm the expression of FOXP4 in surfactant-producing Type II alveolar cells and granulocytes. This correlation supports a possible mechanistic link, and demonstrates the importance of integrative multi-omics approaches. A recent computational study analyzed the evolution of predicted CD8 T-cell epitopes in SARS-CoV-2 variants and its correlation with clinical outcomes of acute COVID-19 in patients with different HLA genotypes, illustrating the importance of integrated analysis of viral and patient genomic data with clinical data (Kim et al., 2024). A similar approach could be used with PASC, and/or PASC clinical subtype, as an outcome. Beyond GWAS or

TABLE 3 Multi-omics assays generate information relevant to testing multiple mechanistic hypotheses for PASC.

Pathogenetic hypothesis	Assays							
	SNP lp-NGS, GWAS	Epigenome	Bulk RNASeq	scRNASeq	Proteome	CyTOF	Metabolome	Microbiome
Genetic predisposition	+							
Viral persistence			+		+			
Intra-patient viral evolution			+					
Non-SARS-CoV2 viral reactivation (EBV, others)			+		+			
Autoimmunity	+	+	+	+		+		+
Chronic inflammation		+	+	+	+	+	+	+
Endothelial damage					+	+		
Coagulation abnormalities					+			
Dysbiosis							+	+
Chronic stress		+	+	+	+		+	
Endocrine dysfunction					+		+	
Toxic exposures		+			+		+	

other genetic analyses, genotyping data can be used in conjunction with other multi-omics profiles to increase the likelihood of discovery. Identification of molecular quantitative trait loci (QTL) can be used to identify possible pathogenetic pathways (Debnath et al., 2020). Different technologies can be used to obtain genotyping information in PASC cases: high-density SNP-chips or low-pass sequencing are established platforms. Emerging technologies, such as nanopore long-read sequencing (Cuppen et al., 2022; Pervez et al., 2022), may also reduce the cost of whole-genome and whole-transcriptome sequencing.

3.1.2 Epigenomics

Epigenomics measure molecular events that regulate chromatin accessibility and expression, which can reflect long-term physiological states. Using methods such as chromatin immunoprecipitation sequencing (ChIP-seq), CUT&RUN, or assay for transposase-accessible chromatin using sequencing (ATAC-seq) (Yan et al., 2020; Sun et al., 2019), which can also be used in single-cell applications (Kashima et al., 2020), many epigenetic processes have been identified and associated with complex traits. One of the best characterized is DNA methylation (5 methyl-cytosine), which is altered in numerous human diseases. There is compelling evidence that changes in DNA methylation profiles are detectable in viral infections such as HIV (Bednarik et al., 1990; Zhang et al., 2016) and MERS (Menachery et al., 2018). In an epigenome-wide association study (EWAS) (Bhat and Jones, 2022), DNA methylation differences associated with a phenotype can be assessed at hundreds of thousands of cytosine-phosphate-guanine (CpG) sites across the epigenome. Several EWAS of COVID-19 in the literature found distinct patterns of DNA methylation associated with disease severity early in the disease course (Castro de Moura et al., 2021; Corley et al., 2021; Balnis et al., 2021; Zhou et al., 2021). EWAS in the ongoing Norwegian Corona Cohort Study also assessed whether there were differentially methylated CpGs between those with PASC (N = 41) compared to a remission group (N = 63), but did not find significant differences. However, the study was not longitudinal, and the authors point out that their sample size for PASC was small (Lee et al., 2022). The same study identified 3 differentially methylated sites associated with acute COVID-19 severity, including hypomethylation of IFI44L, an interferon response gene also associated with COVID-19 severity (Castro de Moura et al., 2021).

3.1.3 Transcriptomics

a) **Bulk Transcriptomics:** RNA transcripts act as intermediary components between genetic information and protein synthesis, and carry specific functions themselves. Transcripts are a regulation hub that responds to both environmental and genetic control, thus playing a major role in the molecular characterization of diseases. Non-coding RNAs fine-tune the expression levels of coding RNAs and their protein products, providing an additional level of regulation. Given its role as an ‘integration hub’ between genetic variation and environmental exposures, the transcriptome dataset is a key layer in multi-omics approaches. Bulk RNA sequencing (RNASeq) can measure the relative abundance of individual transcripts, and determine differences in mRNA splicing isoforms and RNA editing. Whole blood transcriptome analysis can accurately measure the expression levels of >16,000-20,000 RNA species, both protein-coding and non-coding, thus providing one of the most high-quality and high-content multi-omics datasets. Bulk transcriptomics integrates the effects of multiple key variables that can dynamically affect gene expression in blood cells (e.g., metabolic state, epigenetic variation, exposure to medications, stress *etc.*). Transcriptional signatures in the blood or cells of COVID-19 patients can help identify causal factors for acute or chronic complications as well as potential therapeutic targets (Jha et al., 2020; Asano et al., 2022). Recently, a long non-coding RNA-based ML model has been used to identify an RNA (LEF1-AS1) predictive of acute COVID-19 mortality in a ML-driven study of 1,286 patients in 15 institutions (Devaux et al., 2024). Additionally, a candidate signature of acute COVID-19 including 3 long non-coding RNA, 2 cytokines and 2 proteins in peripheral blood mononuclear cells (PMBCs) has been identified using a ML approach (Heydari et al., 2024). This study had a fairly small sample size (28 COVID-19 patients and 17 controls), but it illustrates the promise of multi-analyte biomarkers including RNAs in COVID-19. Current bioinformatics deconvolution approaches enable effective estimation of cell-type fraction and cell type specific gene expression in the peripheral immune system from bulk transcriptome data, offering a powerful tool for immune-phenotyping (Chen et al., 2018) that is complementary to plasma and cellular proteomics. Bulk samples employed for RNAseq can also be used for in-depth immune repertoire analyses (Galson et al., 2020). These data also allow prediction of physiological states, such as PANoptosis (Yang et al., 2024; Dai et al., 2023) or innate immunity activation (Karki and Kanneganti, 2022), and upstream regulators of these states (e.g., transcription factors, protein kinases, hormones), thus enabling the identification of potential therapeutic targets. Critically, transcriptomic data can also allow for the identification of circulating SARS-CoV-2 viral load from whole blood (including viral variant calling and, given sufficient sequence coverage, detection of intra-patient viral evolution), thus constituting an important tool to assess persistent viremia from sources such as vascular beds. Further virome/microbiome analyses of these data can capture other viruses/bacteria that may contribute to PASC pathophysiology (e.g., EBV). Several such transcriptome analyses have been completed for acute COVID-19 and PASC (Thompson et al., 2023; Hadjadj et al., 2020; Lucas et al., 2020; Sposito et al., 2021; Sullivan et al., 2021; Ziegler et al., 2021; Galbraith et al., 2022), but without integration with other omics. This supports the need for further transcriptome analyses in the RECOVER cohort in the context of a multi-omics approach.

b) **Single-cell transcriptomics:** Bulk transcriptomics measures RNA expression as an average of all cell types present in a sample. This can potentially mask the contribution of rare cell types or cellular states to the transcriptome. Single cell RNA sequencing (scRNAseq) can add further detail to immune phenotyping by measuring the transcriptomes of up to 20,000 individual cells simultaneously. This can provide highly detailed information, albeit at higher cost than bulk transcriptomics. scRNAseq protocols relevant to PASC can include Cellular Indexing of Transcriptomes and Epitopes by Sequencing (CITE-seq) and single cell VDJ sequencing (scVDJ) analyses, which can provide advanced immune phenotyping and T cell receptor/B cell receptor (TCR/BCR) repertoire data, respectively, on the same cell (Cadot et al., 2020; Kim et al., 2020; Lian et al., 2020; Saigusa and Ley, 2020; Frangieh et al., 2021; Mercatelli et al., 2021; Rodahl et al., 2021; Shi et al., 2021; Fan et al., 2022; Xu et al., 2022; He et al., 2022).

3.1.4 Proteomics

a) **Soluble proteins:** Protein-based biomarkers are commonly used for the diagnosis and management of myriad medical conditions and are likely to be useful for the prediction, diagnosis, prognosis and clinical management of PASC. Cytokines, chemokines, antibodies, coagulation factors, growth factors, complement cascade components, peptide hormones, and viral proteins can all be measured by high-content proteomic methods in plasma. Multiple technologies are now available to identify hundreds to thousands of individual proteins from very small volumes of serum, plasma, tissues, or cells, including peripheral blood mononuclear cells (PBMCs). These include mass spectrometry, SOMAscan® assays, Olink® proteomics, and PhIP-seq (phage immunoprecipitation sequencing), to name a few. Furthermore, some of these technologies (e.g., mass spectrometry conjugated with the newest search algorithms such as MSFragger (Kong et al., 2017)) enable the identification of protein isoforms and post-translational modifications, including novel ones. For example, while still under development, the latest SOMAscan® platform measures >7,000 proteins from a mere 125 µL of plasma or serum (Gold et al., 2012). Of critical importance for the study of autoimmunity in PASC, the PhIP-seq technology enables the identification of virtually all auto-antibodies produced by an individual (Mohan et al., 2018). A recent PhIP-seq study in a relatively small cohort identified a common autoreactive pattern in PASC patients and patients who had recovered from acute COVID-19 (Bodansky et al., 2023), raising important questions about the possible role of autoantibodies in PASC. Plasma proteomic biosignatures can inform on multiple pathophysiological processes at once, including but not restricted to various forms of inflammation (e.g., systemic, organ-specific, vascular), organ injury, vascular disorders, neurodegeneration, dysregulation of coagulation and fibrinolysis, and remote organ crosstalk *via* the blood. A wealth of proteomics data are already available for acute COVID-19 (Sullivan et al., 2021; Galbraith et al., 2022; Galbraith et al., 2021; D'Alessandro et al., 2020), as well as myriad auto-inflammatory conditions. In the context of an integrated multi-omics strategy, proteomics data could maximize the opportunities to discover mechanisms underlying PASC pathophysiology as well as molecular subtypes, clinically actionable biomarkers and treatment targets.

b) **Cellular proteomics-based immunophenotyping:** Immunophenotyping, which allows for the precise detection of membrane and intracellular proteins using antibodies, has identified signatures that are predictive of subsequent PASC (Peluso et al., 2021). Cellular proteomics-based immunophenotyping technologies can simultaneously quantify, at the single-cell level, expression levels of

40–50 surface and intracellular proteins of immune cells. These include high-parameter flow cytometry (e.g., BD X50), spectral flow cytometry (e.g., Cytek Aurora), and CyTOF (cytometry by time-of-flight, a.k.a. mass cytometry). CyTOF, the most common of these approaches, is a powerful high-dimensional immunophenotyping method that can, in a single specimen, quantify all major subsets of cells using its ~40 available channels. Alternatively, it can be used to deeply characterize one immune subset of interest (e.g., to interrogate phenotypes, homing properties, effector functions, and self-renewal capacities of T cells) (Neidleman et al., 2021a; Ma et al., 2021; Neidleman et al., 2021b; Neidleman et al., 2020). Phospho-CyTOF also enable analyses of signaling states of individual cells (Bendall et al., 2011). It can also be used to characterize the glycan features of immune cells at the single-cell level, informing on immune functions which are very much modulated by cell-surface glycosylation (Ma et al., 2022). As PASC is multifactorial and heterogeneous, approaches such as CyTOF which allow for broad and specific studies of immune subsets, will be key. Studies can, for example, examine how the global immune landscape is altered during PASC, as well as whether specific subsets implicated in COVID-19 disease progression or PASC (e.g., T cells, myeloid cells, neutrophils) exhibit subset-specific changes that can inform on mechanism of action. Studies that have begun to use CyTOF to explore immunological differences between fully recovered vs individuals with PASC have revealed a dysregulated adaptive immune response in the latter, e.g., global differences in T-cell subsets, sex-specific differences in cytolytic T-cells, increased frequency of T-cells migrating to inflamed tissues but also exhausted T-cells, as well as increased frequency of exhausted T-cells (Yin et al., 2023; Yin et al., 2024). Further studies using larger cohorts are warranted. From a practical standpoint, for the amount of data generated CyTOF is cost-effective and requires relatively few cells relative to if samples were to be analyzed using multiple low-parameter panels implemented in conventional flow cytometry.

3.1.5 Metabolomics

Metabolites are the end products of multiple pathways and often indicate the major phenotype(s) of metabolic and genetic disorders. From diabetes to inborn errors of metabolism, metabolites can often define the key pathways underlying complex diseases and serve as potential biomarkers. Metabolites may also mediate the downstream effects of genomic, epigenomic and transcriptomic processes, and in turn influence these processes to modify PASC phenotypes. As a measure of the status of hundreds of metabolic pathways, the overall metabolome and the lipidome represent biologically and mechanistically informative data streams. The endogenous metabolome captures a broad range of inflammatory processes, energy production, microbial metabolites, organ-specific biomarkers, lipids, carbohydrates, steroids, and amino acids, among other relevant information on physiologic processes. Furthermore, exogenous metabolites capture environmental exposures, including but not restricted to food and supplement intake, toxins (e.g., per-and polyfluoroalkylic substances, also known as PFAS, tobacco byproducts, illicit drugs), and medications (e.g., statins, ibuprofen, selective serotonin reuptake inhibitors), all of which may be important in the development or modification of PASC phenotypes, and cannot be easily predicted by other omics but can potentially impact the results of other omics tests. Importantly, these exogenous metabolites are not measurable by any other mechanisms. Microbial metabolites, also measured by metabolomics assays, may serve as important connectors to microbiome data. The interconnections between the metabolome and other multi-omics profiling illustrates an important aspect of multi-omics strategies: while metabolomics can provide crucial information as a single platform, it acts synergistically with other omics data in elucidating important functional relationships to PASC. Several small studies have demonstrated strong dysregulation of endogenous metabolites associated with particular PASC phenotypes (Valdés et al., 2022). For example, tryptophan metabolism was found to be dysregulated by several groups using metabolomic analyses in blood and urine studies (Bustamante et al., 2022; Dewulf et al., 2022), but the pathogenesis of this phenomenon is unclear. We believe that a comprehensive, longitudinal metabolomics investigation of PASC in the context of a multi-omics strategy in a sufficiently large cohort of patients with deep clinical phenotypes will help define and prioritize functional pathways.

3.1.6 Microbiome

The microbiome has multiple physiological roles in human health, including: i) extracting indigestible ingredients from food and synthesizing nutritional factors; ii) affecting host metabolism; iii) developing systemic and intestinal immunity; vi) providing signals for epithelial renewal and maintaining gut integrity; and iv) secreting anti-microbial products. Alterations of the microbiome may often be an initial disturbance with far-reaching ramifications on disease progression. The gut microbiomes of hospitalized COVID-19 patients were enriched with opportunistic pathogens such as *Clostridium hathewayi*, *Bacteroides nordii*, and *Actinomyces viscosus* (Zuo et al., 2020). In acute COVID-19, the gut microbiome is associated with immune responses and disease severity (Zhang et al., 2021; Maeda et al., 2022; Zuo et al., 2021) and also interacts with the lung microbiome (Zhu et al., 2022). Changes in the gut microbiome could influence respiratory tract infections through the common mucosal immune system. Conversely, respiratory tract dysbiosis and functional disorders due to COVID-19 also affect the digestive tract (Zhu et al., 2022). Studies have demonstrated SARS-CoV-2 interactions with host microbiome/virome communities, clotting/coagulation issues, dysfunctional brainstem/vagus nerve signaling, and immune cells (reviewed in (Proal and VanElzakker, 2021)). There is observational evidence of gut microbiome compositional alterations in patients with long-term complications of COVID-19 (Liu et al., 2022). However, the current studies have sample sizes varying from 8 to 130 patients and few studies followed patients beyond 6 months post-infection (reviewed in (Zhang et al., 2023)). A recent study (Xiong et al., 2023) using multi-omics of microbiome-host interactions identified phenotypic, intestinal microbial, and metabolic biomarkers for short-and long-term myalgic encephalomyelitis/chronic fatigue syndrome. Large amounts of microbiome data can be easily generated at low-cost in the RECOVER adult and pediatric cohorts. These data, when integrated with other multi-omics data, will allow for a better understanding of the virus-microbiome-host interactions and identifying microbial and metabolic biomarkers for PASC. Further studies are also needed to investigate whether microbiota modulation can prevent or facilitate the recovery from PASC.

3.2 Considerations on data generation and analysis

3.2.1 Data generation and randomization

Multi-omics data integration can generate valuable knowledge to understand disease pathogenesis. However, multi-omics data can often be burdened by large confounding signals that can prevent accurate modeling and successful discovery. It is therefore essential to appropriately design data collection and generation processes to ensure that such confounders are minimized, and that multi-omics data are amenable to address a large array of important biological questions aiming to characterize, understand and treat PASC. For example, it is usually better to reduce batch effects with a good study design that accurately accounts for them rather than attempting to correct for batch effects after the fact. One successful approach to minimize batch effects involves adequate randomization schemes that minimize risk of contamination of true signals by unwanted variation. Such an approach is powerful when biological questions are defined before data are collected, but often maximizes a distance metric between a measure of interest and the drivers of this unwanted variation at the cost of other potentially meaningful traits. In hypothesis-generating situations, other approaches, such as the inclusion of data generation controls (e.g., reference samples), profiled repeatedly within and across different multi-omics assays, have proven to be an important tool to control for unwanted variation, including confounding from technical variation. However, data generation controls can be complex to define, must contain enough material to be assayed repeatedly, need to capture the full range of biological variation in the multi-omics assessed, and depending on the number needed, can substantially increase data generation cost. These considerations are essential to design a successful multi-omics discovery effort, and it is therefore essential to include data generation experts as well as data scientists who know the biases of each multi-omics profiling technology in teams tasked with designing data generation strategies.

3.2.2 A multi-omics systems-biology approach to data analysis

Each of the omics assays generates vast datasets that require powerful analytical strategies (Gui et al., 2023; Martinez-Bartolome et al., 2018; Lee et al., 2019; Michelhaugh and Januzzi, 2023). Integrating data from multiple omics over time and with clinical, demographic and exposome data is the next level of analytical complexity. A multi-omics approach allows for the integration of multiple layers of information into systems biology models that capture the dynamic interplay between biological processes, allowing not only the study of the functional relationships between the molecular components of PASC, but the elucidation of their causal relationships (Beckmann et al., 2020; Kuijjer et al., 2019; Wang M. et al., 2021; Sonawane et al., 2022; Sonawane et al., 2019; Argelaguet et al., 2020). This approach is critical to understanding the pathogenesis of the clinical manifestations of PASC (Table 3). Integrating multi-omics data with the deep clinical and demographic phenotyping available *via* RECOVER will capture the most complete picture of disease pathophysiology, leading to more accurate identification and characterization of PASC subtypes as compared to individual omics studies of individual patient cohorts (Figure 1).

Multi-omics data are also important in substantiating and validating findings across individual omics platforms (e.g., a genetic polymorphism leading to transcriptomic, proteomic and metabolomics effects). Critical to this is the longitudinal capture of multi-omics data as the clinical presentations of PASC emerge and progress. Given the high-content nature of omics datasets, they support the development of machine learning (ML) class discovery approaches for identification of clinically relevant biosignatures. The rich datasets that will be produced as part of this effort will enable predictive and diagnostic algorithms to identify candidate biomarkers linked to disease outcomes. The thorough integration of these data into meaningful, queryable, and informative models is critical to understand the biological mechanisms, disease subtypes, progression and prognosis of PASC, to investigate the impact of modifiable risk factors and identify potential precision therapeutic approaches to PASC. Inherent in this, is the measure of these data at multiple time points throughout the disease process.

An example of the power of multi-omics approaches is the study of multisystem inflammatory syndrome in children (MIS-C), a serious complication of pediatric COVID-19. Longitudinal plasma bulk transcriptomics, combined with whole blood transcriptomics and plasma DNA epigenomics was recently used to develop multi-organ damage signatures indicative of MIS-C (Loy et al., 2022). This study complements previous genomic, proteomic and immunophenotyping investigations of MIS-C (Sacco et al., 2022; Gruber et al., 2020; Porritt et al., 2021; Carter et al., 2020) to delineate a clearer picture of its pathogenesis. We posit that a single comprehensive, integrated, longitudinal multi-omics approach would have reached similar conclusions as multiple consecutive studies focusing on 1-2 omics each. Such a comprehensive study, performed through centralized labs, would reduce lab-to-lab variability and pre-analytical variability, leveraging a large sample size with rich, highly standardized clinical phenotypes. Furthermore, it is difficult to predict ahead of time which omics would be the most consistent and/or most clinically informative, and which omics data are consistent with each other (e.g., a clinically informative RNA may predict the abundance of an enzyme that produces a metabolite, but if the protein abundance or the metabolite levels are not consistent with RNA abundance, perhaps because of short half-life of the protein or instability of the metabolite, that protein or its metabolite product would not be potential biomarkers or therapeutic targets).Another example of the importance of capturing multi-omics data across demographics and time points is the importance of sex and steroid hormones in the PASC population. Innate and adaptive, humoral and cell-mediated immune responses are impacted by hormones, and their dysregulation contributes to immune-mediated diseases including autoimmunity, a hallmark of PASC (Rojas et al., 2022; Moulton, 2018; Bereshchenko et al., 2018). Ovarian steroids recruit mast cells and T-regs to the uterus during pregnancy (Schumacher et al., 2014). Estradiol causes inflammasome activation in mast cells (Guo et al., 2021). Estradiol deficiency due to menopause and/or hypogonadism contributes to overactivity of the renin-angiotensin-aldosterone system (RAAS), while estrogen can contribute to mast cell activation syndrome (MCAS), which may contribute to the pathogenesis of PASC (O'Donnell et al., 2014; Szukiewicz et al., 2022; Arun et al., 2022). The SARS-CoV2 spike protein binds to and modulates both ACE2 and ERα receptors, and as sex hormones regulate the expression of ACE2 (Solis et al., 2022; Wang H. et al., 2021), the asymmetry in PASC development and clinical presentations between sexes - as well as across menstruation status and menstrual cycle time points - indicates that hormone measurements, which can be performed by metabolomics for non-peptide hormones and by proteomics for peptide hormones, are critical components of a multi-omics strategy.

3.2.3 Task force recommendations

We recommended the following strategy: germline whole genome sequencing (WGS) be performed on every RECOVER participant consented for genetic analysis to be used for GWAS studies. Epigenomics, bulk PBMC transcriptomics, plasma proteomics, plasma targeted metabolomics and stool proteomics should be performed on biospecimens from at least 2 time points per participant (baseline, 90 days and 180 days, or at a minimum baseline and 180 days) on biospecimens from as many participants as possible. Samples taken at later time points during the planned 4-year follow-up period may be analyzed as well in the future, particularly to investigate cases that persist long-term. However, the initial focus should be on the first 180 days post-enrollment, as the number of participants dropping out of the study or being lost to follow-up is likely to increase at later time points. This time window is likely to be long enough to compare COVID-19 cases that do result in PASC to cases that do not, which is one of the primary endpoints of the RECOVER studies, as well as PASC cases that resolve clinically within 180 days from cases that persist beyond that time, while maximizing sample size. Single-cell transcriptomics and/or single-cell immunophenotyping may be performed on subsets of participants from each arm of each cohort, to limit costs. Bioinformatics deconvolution of cellular populations based on bulk transcriptomics should be performed on all available PBMC biospecimens.

It must be pointed out that the adult and pregnant adult RECOVER cohorts include different arms: "acute infected" participants, who enroll within 30 days of a SARS-CoV-2 infection, "post-acute infected", who enroll after 30 days post-infection, "acute uninfected" enrolled within 30 days of a negative COVID-19 test and "post-acute uninfected", enrolled after 30 days post-negative test (Horwitz et al., 2023; Reel et al., 2021). This implies that baseline samples taken at enrollment are likely to reflect different pathobiological stages of disease in acute infected *versus* post-acute infected participants. It is also possible that a fraction of the "uninfected" participants will have experienced subclinical infections with SARS-CoV-2. Multi-omics analyses have the potential to identify these cases, particularly through proteomics-based identification of SARS-CoV-2 antibodies not detected by conventional tests. As there are significant differences in study design for the adult and pediatric cohorts (Horwitz et al., 2023; Reel et al., 2021), longitudinal biospecimens will only be available for Tier 2 pediatric participants, while baseline biospecimens will be available for all participants. Also, the amounts of blood/plasma available for the pediatric cohort will depend on the age of participants. With these considerations in mind, maximizing sample size should be the underlying principle. The main objective of this proposed multi-omics analysis is to generate a rich, multidimensional molecular profiling database to match the clinical, pathophysiological and socioeconomics data elements generated by the RECOVER studies. This data should be made available to the scientific community for secondary analyses.

4 Conclusion

Based on its analysis of the available evidence, the OMICS Task Force advocated an integrated "big data and systems biology" approach, using multi-omics to analyze biospecimens from the largest possible sample sizes in the RECOVER adult and pediatric cohorts, as opposed to single analyte assays or individual omics in multiple separate studies. This approach will maximize our ability to understand pathogenetic mechanisms in clinically defined patient subgroups, discover PASC molecular subtypes and guide precision therapeutic strategies. Centralized, streamlined omics analyses will limit potential inconsistencies associated with laboratory-to-laboratory and batch variations. In addition, multi-omics assays can capture most clinically assayed biomarkers at cheaper costs. Data generation on this scale can only be accomplished through highly multiplexed approaches, which will maximize opportunities to discover mechanisms underlying PASC pathophysiology.

PASC joins the number of poorly understood, chronic diseases that have been the bane of patients, healthcare providers and clinical researchers. While different clinical presentations of PASC have been described, traditional molecular approaches have thus far failed to produce a deep mechanistic understanding of the etiology, pathogenesis and molecular subtypes of PASC. Many of our recommendations have already been considered by the NIH through the peer-review process, resulting in the creation of a systems biology panel that is currently designing the studies we proposed. Currently, this panel is hammering down the details of the analytical strategies. The NIH RECOVER initiative offers an ideal opportunity to understand PASC in diverse populations, and can serve as a paradigm for the study of other complex, poorly-understood chronic diseases.

Author contributions

JS: Conceptualization, Writing–original draft, Writing–review and editing. MA: Conceptualization, Writing–original draft. HA: Conceptualization, Writing–original draft. NB: Conceptualization, Methodology, Visualization, Writing–original draft. ME: Project administration, Writing–original draft. JE: Conceptualization, Writing–original draft. XG: Conceptualization, Writing–original draft. BH: Conceptualization, Writing–original draft. PK: Conceptualization, Writing–original draft. JL-S: Conceptualization, Writing–original draft. RL: Conceptualization, Writing–original draft. CM: Conceptualization, Writing–original draft. MM: Conceptualization, Project administration, Writing–original draft. LN: Conceptualization, Writing–original

draft. NR: Conceptualization, Writing–original draft. MR: Conceptualization, Writing–original draft. JR: Conceptualization, Writing–original draft. GS: Conceptualization, Writing–original draft. KS: Conceptualization, Writing–original draft. SS: Conceptualization, Methodology, Writing–original draft. ST: Conceptualization, Writing–original draft. SY: Conceptualization, Writing–original draft. LM: Conceptualization, Visualization, Writing–original draft, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. The RECOVER studies and the RECOVER OMICS task force are supported by the National Institutes of Health, through the parent grants OT2HL161847, OT2HL161841, OT2HL156812.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J. C., et al. (2020). MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.* 21, 111. doi:10.1186/s13059-020-02015-1

Arun, S., Storan, A., and Myers, B. (2022). Mast cell activation syndrome and the link with long COVID. *Br. J. Hosp. Med. (Lond)* 83, 1–10. doi:10.12968/hmed.2022.0123

Asano, T., Chelvanambi, S., Decano, J. L., Whelan, M. C., Aikawa, E., and Aikawa, M. (2022). *In silico* drug screening approach using l1000-based connectivity map and its application to COVID-19. *Front. Cardiovasc Med.* 9, 842641. doi:10.3389/fcvm.2022.842641

Balnis, J., Madrid, A., Hogan, K. J., Drake, L. A., Chieng, H. C., Tiwari, A., et al. (2021). Blood DNA methylation and COVID-19 outcomes. *Clin. Epigenetics* 13, 118. doi:10.1186/s13148-021-01102-9

Beckmann, N. D., Lin, W. J., Wang, M., Cohain, A. T., Charney, A. W., Wang, P., et al. (2020). Multiscale causal networks identify VGF as a key regulator of Alzheimer's disease. *Nat. Commun.* 11, 3942. doi:10.1038/s41467-020-17405-z

Bednarik, D. P., Cook, J. A., and Pitha, P. M. (1990). Inactivation of the HIV LTR by DNA CpG methylation: evidence for a role in latency. *EMBO J.* 9, 1157–1164. doi:10.1002/j.1460-2075.1990.tb08222.x

Bendall, S. C., Simonds, E. F., Qiu, P., Amir el, A. D., Krutzik, P. O., Finck, R., et al. (2011). Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* 332, 687–696. doi:10.1126/science.1198704

Bereshchenko, O., Bruscoli, S., and Riccardi, C. (2018). Glucocorticoids, sex hormones, and immunity. *Front. Immunol.* 9, 1332. doi:10.3389/fimmu.2018.01332

Bhat, B., and Jones, G. T. (2022). Data analysis of DNA methylation epigenome-wide association studies (EWAS): a guide to the principles of best practice. *Methods Mol. Biol.* 2458, 23–45. doi:10.1007/978-1-0716-2140-0_2

Bodansky, A., Wang, C. Y., Saxena, A., Mitchell, A., Takahashi, S., Anglin, K., et al. (2023). *Autoantigen profiling reveals a shared post-COVID signature in fully recovered and Long COVID patients.* medRxiv.

Bustamante, S., Yau, Y., Boys, V., Chang, J., Paramsothy, S., Pudipeddi, A., et al. (2022). Tryptophan metabolism 'hub' gene expression associates with increased inflammation and severe disease outcomes in COVID-19 infection and inflammatory bowel disease. *Int. J. Mol. Sci.* 23, 14776. doi:10.3390/ijms232314776

Cadot, S., Valle, C., Tosolini, M., Pont, F., Largeaud, L., Laurent, C., et al. (2020). Longitudinal CITE-Seq profiling of chronic lymphocytic leukemia during ibrutinib treatment: evolution of leukemic and immune cells at relapse. *Biomark. Res.* 8, 72. doi:10.1186/s40364-020-00253-w

Carter, M. J., Fish, M., Jennings, A., Doores, K. J., Wellman, P., Seow, J., et al. (2020). Peripheral immunophenotypes in children with multisystem inflammatory syndrome associated with SARS-CoV-2 infection. *Nat. Med.* 26, 1701–1707. doi:10.1038/s41591-020-1054-6

Castro de Moura, M., Davalos, V., Planas-Serra, L., Alvarez-Errico, D., Arribas, C., Ruiz, M., et al. (2021). Epigenome-wide association study of COVID-19 severity with respiratory failure. *EBioMedicine* 66, 103339. doi:10.1016/j.ebiom.2021.103339

Chen, B., Khodadoust, M. S., Liu, C. L., Newman, A. M., and Alizadeh, A. A. (2018). Profiling tumor infiltrating immune cells with CIBERSORT. *Methods Mol. Biol.* 1711, 243–259. doi:10.1007/978-1-4939-7493-1_12

Corley, M. J., Pang, A. P. S., Dody, K., Mudd, P. A., Patterson, B. K., Seethamraju, H., et al. (2021). Genome-wide DNA methylation profiling of peripheral blood reveals an epigenetic signature associated with severe COVID-19. *J. Leukoc. Biol.* 110, 21–26. doi:10.1002/JLB.5HI0720-466R

Cuppen, E., Elemento, O., Rosenquist, R., Nikic, S., M, I. J., Zaleski, I. D., et al. (2022). Implementation of whole-genome and transcriptome sequencing into clinical cancer care. *JCO Precis. Oncol.* 6, e2200245. doi:10.1200/PO.22.00245

Dai, W., Zheng, P., Wu, J., Chen, S., Deng, M., Tong, X., et al. (2023). Integrated analysis of single-cell RNA-seq and chipset data unravels PANoptosis-related genes in sepsis. *Front. Immunol.* 14, 1247131. doi:10.3389/fimmu.2023.1247131

D'Alessandro, A., Thomas, T., Dzieciatkowska, M., Hill, R. C., Francis, R. O., Hudson, K. E., et al. (2020). Serum proteomics in COVID-19 patients: altered coagulation and complement status as a function of IL-6 level. *J. Proteome Res.* 19, 4417–4427. doi:10.1021/acs.jproteome.0c00365

Debnath, M., Banerjee, M., and Berk, M. (2020). Genetic gateways to COVID-19 infection: implications for risk, severity, and outcomes. *FASEB J.* 34, 8787–8795. doi:10.1096/fj.202001115R

Devaux, Y., Zhang, L., Lumley, A. I., Karaduzovic-Hadziabdic, K., Mooser, V., Rousseau, S., et al. (2024). Development of a long noncoding RNA-based machine learning model to predict COVID-19 in-hospital mortality. *Nat. Commun.* 15, 4259. doi:10.1038/s41467-024-47557-1

Dewulf, J. P., Martin, M., Marie, S., Oguz, F., Belkhir, L., De Greef, J., et al. (2022). Urine metabolomics links dysregulation of the tryptophan-kynurenine pathway to inflammation and severity of COVID-19. *Sci. Rep.* 12, 9959. doi:10.1038/s41598-022-14292-w

Fan, R., Liu, Y., DiStasio, M., Su, G., Asashima, H., Enninful, A., et al. (2022). Spatial-CITE-seq: spatially resolved high-plex protein and whole transcriptome co-mapping. *Res. Sq.* doi:10.21203/rs.3.rs-1499315/v1

Ferreira, L. C., Gomes, C. E. M., Rodrigues-Neto, J. F., and Jeronimo, S. M. B. (2022). Genome-wide association studies of COVID-19: connecting the dots. *Infect. Genet. Evol.* 106, 105379. doi:10.1016/j.meegid.2022.105379

Frangieh, C. J., Melms, J. C., Thakore, P. I., Geiger-Schuller, K. R., Ho, P., Luoma, A. M., et al. (2021). Multimodal pooled Perturb-CITE-seq screens in patient models define mechanisms of cancer immune evasion. *Nat. Genet.* 53, 332–341. doi:10.1038/s41588-021-00779-1

Galbraith, M. D., Kinning, K. T., Sullivan, K. D., Araya, P., Smith, K. P., Granrath, R. E., et al. (2022). Specialized interferon action in COVID-19. *Proc. Natl. Acad. Sci. U. S. A.* 119. doi:10.1073/pnas.2116730119

Galbraith, M. D., Kinning, K. T., Sullivan, K. D., Baxter, R., Araya, P., Jordan, K. R., et al. (2021). Seroconversion stages COVID19 into distinct pathophysiological states. *Elife* 10, e65508. doi:10.7554/eLife.65508

Galson, J. D., Schaetzle, S., Bashford-Rogers, R. J. M., Raybould, M. I. J., Kovaltsuk, A., Kilpatrick, G. J., et al. (2020). Deep sequencing of B cell receptor repertoires from COVID-19 patients reveals strong convergent immune signatures. *Front. Immunol.* 11, 605170. doi:10.3389/fimmu.2020.605170

Gold, L., Walker, J. J., Wilcox, S. K., and Williams, S. (2012). Advances in human proteomics at high scale with the SOMAscan proteomics platform. *N. Biotechnol.* 29, 543–549. doi:10.1016/j.nbt.2011.11.016

Gross, R. S., Thaweethai, T., Rosenzweig, E. B., Chan, J., Chibnik, L. B., Cicek, M. S., et al. (2024). Researching COVID to enhance recovery (RECOVER) pediatric study protocol: rationale, objectives and design. *PLoS One* 19, e0285635. doi:10.1371/journal.pone.0285635

Gruber, C. N., Patel, R. S., Trachtman, R., Lepow, L., Amanat, F., Krammer, F., et al. (2020). Mapping systemic inflammation and antibody responses in multisystem inflammatory syndrome in children (MIS-C). *Cell* 183, 982–995. doi:10.1016/j.cell.2020.09.034

Gui, Y., He, X., Yu, J., and Jing, J. (2023). Artificial intelligence-assisted transcriptomic analysis to advance cancer immunotherapy. *J. Clin. Med.* 12, 1279. doi:10.3390/jcm12041279

Guo, X., Xu, X., Li, T., Yu, Q., Wang, J., Chen, Y., et al. (2021). NLRP3 inflammasome activation of mast cells by estrogen via the nuclear-initiated signaling pathway contributes to the development of endometriosis. *Front. Immunol.* 12, 749979. doi:10.3389/fimmu.2021.749979

Hadjadj, J., Yatim, N., Barnabei, L., Corneau, A., Boussier, J., Smith, N., et al. (2020). Impaired type I interferon activity and inflammatory responses in severe COVID-19 patients. *Science* 369, 718–724. doi:10.1126/science.abc6027

He, H., Li, Z., Lu, J., Qiang, W., Jiang, S., Xu, Y., et al. (2022). Single-cell RNA-seq reveals clonal diversity and prognostic genes of relapsed multiple myeloma. *Clin. Transl. Med.* 12, e757. doi:10.1002/ctm2.757

Heydari, R., Tavassolifar, M. J., Fayazzadeh, S., Sadatpour, O., and Meyfour, A. (2024). Long non-coding RNAs in biomarking COVID-19: a machine learning-based approach. *Virol. J.* 21, 134. doi:10.1186/s12985-024-02408-9

Horwitz, L. I., Thaweethai, T., Brosnahan, S. B., Cicek, M. S., Fitzgerald, M. L., Goldman, J. D., et al. (2023). Researching COVID to Enhance Recovery (RECOVER) adult study protocol: rationale, objectives, and design. *PLoS One* 18, e0286297. doi:10.1371/journal.pone.0286297

Jha, P. K., Vijay, A., Halu, A., Uchida, S., and Aikawa, M. (2020). Gene expression profiling reveals the shared and distinct transcriptional signatures in human lung epithelial cells infected with SARS-CoV-2, MERS-CoV, or SARS-CoV: potential implications in cardiovascular complications of COVID-19. *Front. Cardiovasc Med.* 7, 623012. doi:10.3389/fcvm.2020.623012

Joshi, A., Rienks, M., Theofilatos, K., and Mayr, M. (2021). Systems biology in cardiovascular disease: a multiomics approach. *Nat. Rev. Cardiol.* 18, 313–330. doi:10.1038/s41569-020-00477-1

Karki, R., and Kanneganti, T. D. (2022). Innate immunity, cytokine storm, and inflammatory cell death in COVID-19. *J. Transl. Med.* 20, 542. doi:10.1186/s12967-022-03767-z

Kashima, Y., Sakamoto, Y., Kaneko, K., Seki, M., Suzuki, Y., and Suzuki, A. (2020). Single-cell sequencing techniques from individual to multiomics analyses. *Exp. Mol. Med.* 52, 1419–1427. doi:10.1038/s12276-020-00499-2

Kim, G. J., Elnaggar, J. H., Varnado, M., Feehan, A. K., Tauzier, D., Rose, R., et al. (2024). A bioinformatic analysis of T-cell epitope diversity in SARS-CoV-2 variants: association with COVID-19 clinical severity in the United States population. *Front. Sys. Biol.* 15, 1357731. in press. doi:10.3389/fimmu.2024.1357731

Kim, H. J., Lin, Y., Geddes, T. A., Yang, J. Y. H., and Yang, P. (2020). CiteFuse enables multi-modal analysis of CITE-seq data. *Bioinformatics* 36, 4137–4143. doi:10.1093/bioinformatics/btaa282

Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D., and Nesvizhskii, A. I. (2017). MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat. Methods* 14, 513–520. doi:10.1038/nmeth.4256

Kuijjer, M. L., Tung, M. G., Yuan, G., Quackenbush, J., and Glass, K. (2019). Estimating sample-specific regulatory networks. *iScience* 14, 226–240. doi:10.1016/j.isci.2019.03.021

Lammi, V., Nakanishi, T., Jones, S. E., Andrews, S. J., Karjalainen, J., Cortés, B., et al. (2023). *Genome-wide association study of long COVID* medRxiv 2023.06.29.23292056. doi:10.1101/2023.06.29.23292056

Lee, L. H., Halu, A., Morgan, S., Iwata, H., Aikawa, M., and Singh, S. A. (2019). XINA: a workflow for the integration of multiplexed proteomics kinetics data with network analysis. *J. Proteome Res.* 18, 775–781. doi:10.1021/acs.jproteome.8b00615

Lee, Y., Riskedal, E., Kalleberg, K. T., Istre, M., Lind, A., Lund-Johansen, F., et al. (2022). EWAS of post-COVID-19 patients shows methylation differences in the immune-response associated gene, IFI44L, three months after COVID-19 infection. *Sci. Rep.* 12, 11478. doi:10.1038/s41598-022-15467-1

Lian, Q., Xin, H., Ma, J., Konnikova, L., Chen, W., Gu, J., et al. (2020). Artificial-cell-type aware cell-type classification in CITE-seq. *Bioinformatics* 36, i542–i550. doi:10.1093/bioinformatics/btaa467

Liu, Q., Mak, J. W. Y., Su, Q., Yeoh, Y. K., Lui, G. C., Ng, S. S. S., et al. (2022). Gut microbiota dynamics in a prospective cohort of patients with post-acute COVID-19 syndrome. *Gut* 71, 544–552. doi:10.1136/gutjnl-2021-325989

Loy, C. J., Sotomayor-Gonzalez, A., Servellita, V., Nguyen, J., Lenz, J., Bhattacharya, S., et al. (2022). Nucleic acid biomarkers of immune response and cell and tissue damage in children with COVID-19 and MIS-C. *Cell Rep. Med.* 4, 101034. doi:10.1016/j.xcrm.2023.101034

Lucas, C., Wong, P., Klein, J., Castro, T. B. R., Silva, J., Sundaram, M., et al. (2020). Longitudinal analyses reveal immunological misfiring in severe COVID-19. *Nature* 584, 463–469. doi:10.1038/s41586-020-2588-y

Ma, T., McGregor, M., Giron, L., Xie, G., George, A. F., Abdel-Mohsen, M., et al. (2022). Single-cell glycomics analysis by CyTOF-Lec reveals glycan features defining cells differentially susceptible to HIV. *Elife* 11. doi:10.7554/elife.78870

Ma, T., Ryu, H., McGregor, M., Babcock, B., Neidleman, J., Xie, G., et al. (2021). Protracted yet coordinated differentiation of long-lived SARS-CoV-2-specific CD8(+) T cells during convalescence. *J. Immunol.* 207, 1344–1356. doi:10.4049/jimmunol.2100465

Maeda, Y., Motooka, D., Kawasaki, T., Oki, H., Noda, Y., Adachi, Y., et al. (2022). Longitudinal alterations of the gut mycobiota and microbiota on COVID-19 severity. *BMC Infect. Dis.* 22, 572. doi:10.1186/s12879-022-07358-7

Martinez-Bartolome, S., Medina-Aunon, J. A., Lopez-Garcia, M. A., Gonzalez-Tejedo, C., Prieto, G., Navajas, R., et al. (2018). PACOM: a versatile tool for integrating, filtering, visualizing, and comparing multiple large mass spectrometry proteomics data sets. *J. Proteome Res.* 17, 1547–1558. doi:10.1021/acs.jproteome.7b00858

Menachery, V. D., Schafer, A., Burnum-Johnson, K. E., Mitchell, H. D., Eisfeld, A. J., Walters, K. B., et al. (2018). MERS-CoV and H5N1 influenza virus antagonize antigen presentation by altering the epigenetic landscape. *Proc. Natl. Acad. Sci. U. S. A.* 115, E1012-E1021–E1021. doi:10.1073/pnas.1706928115

Mercatelli, D., Balboni, N., Giorgio, F., Aleo, E., Garone, C., and Giorgi, F. M. (2021). The transcriptome of SH-SY5Y at single-cell resolution: a CITE-seq data analysis workflow. *Methods Protoc.* 4, 28. doi:10.3390/mps4020028

Metz, T. D., Clifton, R. G., Gallagher, R., Gross, R. S., Horwitz, L. I., Jacoby, V. L., et al. (2023). Researching COVID to enhance recovery (RECOVER) pregnancy study: rationale, objectives and design. *PLoS One* 18, e0285351. doi:10.1371/journal.pone.0285351

Michelhaugh, S. A., and Januzzi, J. L., Jr. (2023). Using artificial intelligence to better predict and develop biomarkers. *Clin. Lab. Med.* 43, 99–114. doi:10.1016/j.cll.2022.09.021

Mohan, D., Wansley, D. L., Sie, B. M., Noon, M. S., Baer, A. N., Laserson, U., et al. (2018). PhIP-Seq characterization of serum antibodies using oligonucleotide-encoded peptidomes. *Nat. Protoc.* 13, 1958–1978. doi:10.1038/s41596-018-0025-6

Moulton, V. R. (2018). Sex hormones in acquired immunity and autoimmune disease. *Front. Immunol.* 9, 2279. doi:10.3389/fimmu.2018.02279

Neidleman, J., Luo, X., Frouard, J., Xie, G., Gill, G., Stein, E. S., et al. (2020). SARS-CoV-2-Specific T cells exhibit phenotypic features of helper function, lack of terminal differentiation, and high proliferation potential. *Cell Rep. Med.* 1, 100081. doi:10.1016/j.xcrm.2020.100081

Neidleman, J., Luo, X., George, A. F., McGregor, M., Yang, J., Yun, C., et al. (2021b). Distinctive features of SARS-CoV-2-specific T cells predict recovery from severe COVID-19. *Cell Rep.* 36, 109414. doi:10.1016/j.celrep.2021.109414

Neidleman, J., Luo, X., McGregor, M., Xie, G., Murray, V., Greene, W. C., et al. (2021a). mRNA vaccine-induced T cells respond identically to SARS-CoV-2 variants of concern but differ in longevity and homing properties depending on prior infection status. *Elife* 10, e72619. doi:10.7554/eLife.72619

Niranjan, V., Uttarkar, A., Kaul, A., and Varghese, M. (2023). A machine learning-based approach using multi-omics data to predict metabolic pathways. *Methods Mol. Biol.* 2553, 441–452. doi:10.1007/978-1-0716-2617-7_19

O'Donnell, E., Floras, J. S., and Harvey, P. J. (2014). Estrogen status and the renin angiotensin aldosterone system. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* 307, R498–R500. doi:10.1152/ajpregu.00182.2014

Peluso, M. J., Deitchman, A. N., Torres, L., Iyer, N. S., Munter, S. E., Nixon, C. C., et al. (2021). Long-term SARS-CoV-2-specific immune and inflammatory responses in individuals recovering from COVID-19 with and without post-acute symptoms. *Cell Rep.* 36, 109518. doi:10.1016/j.celrep.2021.109518

Pervez, M. T., Hasnain, M. J. U., Abbas, S. H., Moustafa, M. F., Aslam, N., and Shah, S. S. M. (2022). A comprehensive review of performance of next-generation sequencing platforms. *Biomed. Res. Int.* 2022, 3457806. doi:10.1155/2022/3457806

Porritt, R. A., Binek, A., Paschold, L., Rivas, M. N., McArdle, A., Yonker, L. M., et al. (2021). The autoimmune signature of hyperinflammatory multisystem inflammatory syndrome in children. *J. Clin. Invest* 131, e151520. doi:10.1172/JCI151520

Proal, A. D., and VanElzakker, M. B. (2021). Long COVID or post-acute sequelae of COVID-19 (PASC): an overview of biological factors that may contribute to persistent symptoms. *Front. Microbiol.* 12, 698169. doi:10.3389/fmicb.2021.698169

Reel, P. S., Reel, S., Pearson, E., Trucco, E., and Jefferson, E. (2021). Using machine learning approaches for multi-omics data analysis: a review. *Biotechnol. Adv.* 49, 107739. doi:10.1016/j.biotechadv.2021.107739

Rodahl, I., Gotley, J., Andersen, S. B., Yu, M., Mehdi, A. M., Christ, A. N., et al. (2021). Acquisition of murine splenic myeloid cells for protein and gene expression profiling by advanced flow cytometry and CITE-seq. *Star. Protoc.* 2, 100842. doi:10.1016/j.xpro.2021.100842

Rojas, M., Rodriguez, Y., Acosta-Ampudia, Y., Monsalve, D. M., Zhu, C., Li, Q. Z., et al. (2022). Autoimmunity is a hallmark of post-COVID syndrome. *J. Transl. Med.* 20, 129. doi:10.1186/s12967-022-03328-4

Sacco, K., Castagnoli, R., Vakkilainen, S., Liu, C., Delmonte, O. M., Oguz, C., et al. (2022). Immunopathological signatures in multisystem inflammatory syndrome in children and pediatric COVID-19. *Nat. Med.* 28, 1050–1062. doi:10.1038/s41591-022-01724-3

Saigusa, R., and Ley, K. (2020). CITE-seq hits vascular medicine. *Clin. Chem.* 66, 751–753. doi:10.1093/clinchem/hvaa016

Schumacher, A., Costa, S. D., and Zenclussen, A. C. (2014). Endocrine factors modulating immune responses in pregnancy. *Front. Immunol.* 5, 196. doi:10.3389/fimmu.2014.00196

Sherif, Z. A., Gomez, C. R., Connors, T. J., Henrich, T. J., Reeves, W. B., and RECOVER Mechanistic Pathway Task Force (2023). Pathogenic mechanisms of post-acute sequelae of SARS-CoV-2 infection (PASC). *Elife* 12, e86002. doi:10.7554/eLife.86002

Shi, X., Baracho, G. V., Lomas, W. E., 3rd, Widmann, S. J., and Tyznik, A. J. (2021). Co-staining human PBMCs with fluorescent antibodies and antibody-oligonucleotide conjugates for cell sorting prior to single-cell CITE-Seq. *Star. Protoc.* 2, 100893. doi:10.1016/j.xpro.2021.100893

Solis, O., Beccari, A. R., Iaconis, D., Talarico, C., Ruiz-Bedoya, C. A., Nwachukwu, J. C., et al. (2022). The SARS-CoV-2 spike protein binds and modulates estrogen receptors. *Sci. Adv.* 8, eadd4150. doi:10.1126/sciadv.add4150

Sonawane, A. R., Aikawa, E., and Aikawa, M. (2022). Connections for matters of the heart: network medicine in cardiovascular diseases. *Front. Cardiovasc Med.* 9, 873582. doi:10.3389/fcvm.2022.873582

Sonawane, A. R., Tian, L., Chu, C. Y., Qiu, X., Wang, L., Holden-Wiltse, J., et al. (2019). Microbiome-transcriptome interactions related to severity of respiratory syncytial virus infection. *Sci. Rep.* 9, 13824. doi:10.1038/s41598-019-50217-w

Sposito, B., Broggi, A., Pandolfi, L., Crotta, S., Clementi, N., Ferrarese, R., et al. (2021). The interferon landscape along the respiratory tract impacts the severity of COVID-19. *Cell* 184, 4953–4968 e16. doi:10.1016/j.cell.2021.08.016

Sullivan, K. D., Galbraith, M. D., Kinning, K. T., Bartsch, K. W., Levinsky, N. C., Araya, P., et al. (2021). The COVIDome Explorer researcher portal. *Cell Rep.* 36, 109527. doi:10.1016/j.celrep.2021.109527

Sun, Y., Miao, N., and Sun, T. (2019). Detect accessible chromatin using ATAC-sequencing, from principle to applications. *Hereditas* 156, 29. doi:10.1186/s41065-019-0105-9

Szukiewicz, D., Wojdasiewicz, P., Watroba, M., and Szewczyk, G. (2022). Mast cell activation syndrome in COVID-19 and female reproductive function: theoretical background vs. Accumulating clinical evidence. *J. Immunol. Res.* 2022, 9534163. doi:10.1155/2022/9534163

Thaweethai, T., Jolley, S. E., Karlson, E. W., Levitan, E. B., Levy, B., McComsey, G. A., et al. (2023). Development of a definition of postacute sequelae of SARS-CoV-2 infection. *JAMA* 329, 1934–1946. doi:10.1001/jama.2023.8823

Thompson, R. C., Simons, N. W., Wilkins, L., Cheng, E., Del Valle, D. M., Hoffman, G. E., et al. (2023). Molecular states during acute COVID-19 reveal distinct etiologies of long-term sequelae. *Nat. Med.* 29, 236–246. doi:10.1038/s41591-022-02107-4

Valdés, A., Moreno, L. O., Rello, S. R., Orduña, A., Bernardo, D., and Cifuentes, A. (2022). Metabolomics study of COVID-19 patients in four different clinical stages. *Sci. Rep.* 12, 1650. doi:10.1038/s41598-022-05667-0

Wang, H., Sun, X., J, L. V., Kon, N. D., Ferrario, C. M., and Groban, L. (2021b). Estrogen receptors are linked to angiotensin-converting enzyme 2 (ACE2), ADAM metallopeptidase domain 17 (ADAM-17), and transmembrane protease serine 2 (TMPRSS2) expression in the human atrium: insights into COVID-19. *Hypertens. Res.* 44, 882–884. doi:10.1038/s41440-021-00626-0

Wang, K., Khoramjoo, M., Srinivasan, K., Gordon, P. M. K., Mandal, R., Jackson, D., et al. (2023). Sequential multi-omics analysis identifies clinical phenotypes and predictive biomarkers for long COVID. *Cell Rep. Med.* 4, 101254. doi:10.1016/j.xcrm.2023.101254

Wang, M., Li, A., Sekiya, M., Beckmann, N. D., Quan, X., Schrode, N., et al. (2021a). Transformative network modeling of multi-omics data reveals detailed circuits, key regulators, and potential therapeutics for alzheimer's disease. *Neuron* 109, 257–272 e14. doi:10.1016/j.neuron.2020.11.002

Xiong, R., Gunter, C., Fleming, E., Vernon, S. D., Bateman, L., Unutmaz, D., et al. (2023). Multi-'omics of gut microbiome-host interactions in short-and long-term myalgic encephalomyelitis/chronic fatigue syndrome patients. *Cell Host and Microbe* 31, 273–287. e5. doi:10.1016/j.chom.2023.01.001

Xu, Z., Heidrich-O'Hare, E., Chen, W., and Duerr, R. H. (2022). Comprehensive benchmarking of CITE-seq versus DOGMA-seq single cell multimodal omics. *Genome Biol.* 23, 135. doi:10.1186/s13059-022-02698-8

Yan, F., Powell, D. R., Curtis, D. J., and Wong, N. C. (2020). From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biol.* 21, 22. doi:10.1186/s13059-020-1929-3

Yang, Q., Song, W., Reheman, H., Wang, D., Qu, J., and Li, Y. (2024). PANoptosis, an indicator of COVID-19 severity and outcomes. *Brief. Bioinform* 25. doi:10.1093/bib/bbae124

Yin, K., Peluso, M. J., Luo, X., Thomas, R., Shin, M. G., Neidleman, J., et al. (2024). Long COVID manifests with T cell dysregulation, inflammation and an uncoordinated adaptive immune response to SARS-CoV-2. *Nat. Immunol.* 25 (2), 218–225. doi:10.1038/s41590-023-01724-6

Yin, K., Peluso, M. J., Thomas, R., Shin, M. G., Neidleman, J., Luo, X., et al. (2023). Long COVID manifests with T cell dysregulation, inflammation, and an uncoordinated adaptive immune response to SARS-CoV-2. *bioRxiv*.

Zhang, J., Garrett, S., and Sun, J. (2021). Gastrointestinal symptoms, pathophysiology, and treatment in COVID-19. *Genes Dis.* 8, 385–400. doi:10.1016/j.gendis.2020.08.013

Zhang, J., Zhang, Y., Xia, Y., and Sun, J. (2023). Microbiome and intestinal pathophysiology in post-acute sequelae of COVID-19. *Genes Dis.* 11, 100978. In press. doi:10.1016/j.gendis.2023.03.034

Zhang, X., Justice, A. C., Hu, Y., Wang, Z., Zhao, H., Wang, G., et al. (2016). Epigenome-wide differential DNA methylation between HIV-infected and uninfected individuals. *Epigenetics* 11, 750–760. doi:10.1080/15592294.2016.1221569

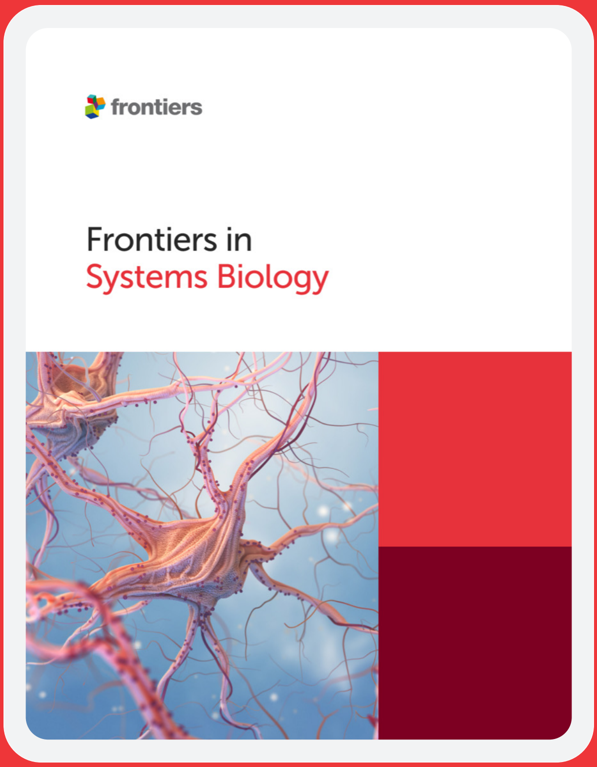
Zhou, S., Zhang, J., Xu, J., Zhang, F., Li, P., He, Y., et al. (2021). An epigenome-wide DNA methylation study of patients with COVID-19. *Ann. Hum. Genet.* 85, 221–234. doi:10.1111/ahg.12440

Zhu, T., Jin, J., Chen, M., and Chen, Y. (2022). The impact of infection with COVID-19 on the respiratory microbiome: a narrative review. *Virulence* 13, 1076–1087. doi:10.1080/21505594.2022.2090071

Ziegler, C. G. K., Miao, V. N., Owings, A. H., Navia, A. W., Tang, Y., Bromley, J. D., et al. (2021). Impaired local intrinsic immunity to SARS-CoV-2 infection in severe COVID-19. *Cell* 184, 4713–4733 e22. doi:10.1016/j.cell.2021.07.023

Zuo, T., Wu, X., Wen, W., and Lan, P. (2021). Gut microbiome alterations in COVID-19. *Genomics Proteomics Bioinforma.* 19, 679–688. doi:10.1016/j.gpb.2021.09.004

Zuo, T., Zhang, F., Lui, G. C. Y., Yeoh, Y. K., Li, A. Y. L., Zhan, H., et al. (2020). Alterations in gut microbiota of patients with COVID-19 during time of hospitalization. *Gastroenterology* 159, 944–955. doi:10.1053/j.gastro.2020.05.048



frontiers

Frontiers in
Systems Biology

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact