Data-Driven and Knowledge-Based Multiscale Modeling of Viral Dynamics

by

ODAKA Mitsuhiro

Dissertation

Doctor of Philosophy

Department of Informatics
School of Multidisciplinary Sciences

The Graduate University for Advanced Studies, SOKENDAI

MaSTIC (Mathematics and Digital, Information and Communication Sciences and Technologies)
Specialty: Informatics

Centrale Nantes

March 2024

ACKNOWLEDGEMENT

I am indebted to the Japanese and French scientists involved in my doctoral research. First, I express my sincere gratitude to my co-supervisors, Dr. Katsumi Inoue (Professor of National Institute of Informatics [NII], The Graduate University for Advanced Studies, SOKENDAI [SOKENDAI], and École Centrale de Nantes [ECN]) and Dr. Morgan Magnin (Professor of ECN and NII), for their proper supervision of work and life alongside my journey, including a one-year stay in France. They advised me on the direction or ideas of the doctoral research that I had planned and drafted and offered suggestions for improvement on the analytical results and the papers. For example, Inoue's cutting-edge thinking on artificial intelligence and specific suggestions from the perspective of knowledge representation and reasoning led to ideas for utilizing data and knowledge. Magnin provided more general ideas from the perspective of formal methods as a principal investigator of the MeForBio team in Le Laboratoire des Sciences du Numérique de Nantes (LS2N) while providing me with a wide range of support from the life and psychological aspects of being a doctoral student. Their advice kept me motivated and led to the promotion of my doctoral research. In addition, since I was the first applicant to the Dual Degree Program (DDP) between SOKENDAI and ECN, my two supervisors, staff of SOKENDAI, NII, ECN, and LS2N, made efforts and worked hard out of my sight to realize the DDP. It was fortunate that I was able to take advantage of the opportunities they had prepared in advance. Second, I would also like to thank the researchers in the jury and thesis monitoring committee (comité de suivi de thèse; CSI): Dr. Takeaki Uno, Dr. Mahito Sugiyama, Dr. Loïc Paulevé, Dr. Sylvain Soliman, Dr. Sophie Limou, Dr. Bertrand Michel, Dr. François Fages, and Dr. Maxime Folschette for their constructive suggestion from the early to the last stage of my research. For example, Uno provided remarks from a general scientific research perspective, Sugiyama provided comments on data mining, Paulevé and Soliman provided evaluations and feedback as external reviewers, Limou provided insights from a computational biology perspective, Michel provided opinions on data analysis and statistical methods, and Fages and Folschette offered suggestions and advice on network inference and automation. Third, I am grateful to the secretaries and staff who helped me with paperwork and budgetary control: Aya Gotoda, Asaki Kamada, Junko Inose, Emily Thureau, Edith Dauve, and Linda Serandour. Their help allowed me to go to France and the U.S. and manage the budget. Fourth, I would like to

thank my Inoue Lab and MeForBio colleagues for providing me with seminal ideas supporting my research. In particular, I ensured my psychological safety in conversations with younger colleagues. Fifth, my heartfelt appreciation goes to my parents, Dr. Hiroyuki Odaka and Mutsumi Odaka, and my wife, Kimino Odaka, who emotionally encouraged and sometimes financially supported my pursuit of a Ph.D. Finally, my doctoral research was supported by JSPS KAKENHI Grant Numbers JP22KJ1417 and JP21H04905, JST CREST Grant Number JPMJCR22D3, and ROIS Strategic Research Project Grant Number 2022-SRP-09. Without the above persons and support, I would not have been able to stand where I am.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Brief summary of this chapter

— COVID-19, an infectious disease that emerged as a global issue in 2019, still leaves humankind with various concerns.

— This manuscript is our response to social demands for epidemic situations like COVID-19.

— We clarify how to understand the biological system from two viewpoints: viral dynamics and multiscale modeling.

— Looking ahead to uncovering the mechanism, we discover a hypothesis lacking in existing COVID-19 knowledge repositories about pathways as the substance of the mechanism.

— Two types of studies about viral dynamics are positioned in the context of the scientific discovery loop to find a hypothesis. These two studies are linked through multiscale modeling, in which pathways at the microscopic level verify a hypothesis about viral dynamics at the macroscopic level.

1.1 Social background

In December 2019, the human species experienced an unknown global issue, coronavirus disease 2019 (COVID-19). COVID-19 is an emerging infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [1]. SARS-CoV-2 infects bronchial epithelial cells, pneumocytes, or alveolar macrophages and causes severe symptoms such as acute respiratory distress syndrome (ARDS) due to excessive production of inflammatory cytokines known as a cytokine storm [2]. The first death was confirmed in Wuhan City in January 2020 [3]. By the end of 2020, COVID-19 became the third leading cause of death in the United

States [4]. Under this situation, concerted attempts were made, including the development of drugs, vaccines, or related treatment guidelines. The notable one was the administration of the first doses of Pfizer–BioNTech COVID-19 ribonucleic acid (RNA) vaccines in December 2020 [5]. The vaccination strategy starting from this initial success resulted in population immunity, thereby reducing the number of deaths worldwide dramatically [6]. In May 2023, COVID-19 turned to be no longer defined as the Public Health Emergency of International Concern (PHEIC) [7].

The global situation stated above does not mean the optimistic end of COVID-19. Namely, there remain various concerns. For example, the risk of new or de-escalated variants needs to be monitored [8]. Additionally, the global burden of COVID-19 to health or economy is ambiguous and requires continued measurement of disability-adjusted life years (DALYs), the sum of the life years lost due to disability or premature death [9]. Moreover, understanding longer-term health consequences (Long COVID) has been poor [10]. Alongside these concerns, a strategic shift occurs. The social demands have changed from infection control for discovering leverage points to prevention. For instance, Preparedness 2.0, the 5-year action plan to strengthen health emergency preparedness, response, and resilience in the WHO European Region, is planned from 2024 to 2029 [11].

Motivated by the circumstances above, we respond to the social demands from computational aspects through the following two ways:

1. elucidating the mechanisms of SARS-CoV-2 infection system and finding new knowledge about them;
2. proposing a framework applicable to pandemics, not limited to COVID-19, for reducing health and economic loss at the early decision-making phase.

1.2 Understanding biological mechanisms as systems of pathways

To elucidate the "mechanism" of biological systems in COVID-19 requires fully understanding the COVID-19-specific interactome, i.e., an entire body of interaction networks among different components, such as signaling pathways, protein-protein interaction (PPI) networks, or gene regulatory networks. For example, the COVID-19 Disease Map is a graphical knowledge repository about signaling pathways involved in COVID-19 based on pathway enrichment analysis or manual curation from external knowledge bases, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways [12, 13]. The COVID-19 Disease Map enlarges the pathway volume; however, it is indicated that the map lacks genes without pathway annotations [14]. Moreover, the initial version of the COVID-19 Disease Map was built on the fly [15], so it has inherently been incomplete and a work in progress. COVID-19 Signaling Network Open Resources (SIGNOR) is another knowledge repository as a part of the COVID-19 Disease Map [16]. Given that these knowledge repositories are imperfect, discovering missing pathways would contribute to compensating the knowledge repositories and understanding mechanisms.

To discover missing pathways, we refer to a closed loop of scientific discovery through triadic reasoning proposed by Charles Sanders Peirce [17]. Figure 1.1 shows an outline of scientific discovery empowered by different reasoning processes: induction, abduction, and deduction. Induction is generalizing multiple examples to find common patterns or rules (surprising facts). The collection of similar cases strengthens the soundness of inferences. Abduction is discovering knowledge (hypotheses) not yet explored by humankind. Deduction is deriving verifiable predictions from hypotheses, where conclusions are included in the premise and thus do not lead to new findings. Based on the prediction, one can plan to record or collect case studies to obtain further observable data. Considering that these three types of reasoning processes realize scientific discovery, we can roughly place our purpose on finding and verifying a "hypothesis" that is unknown despite of its potential significance. We achieve our purpose from the two viewpoints: *viral dynamics* and *multiscale modeling*.

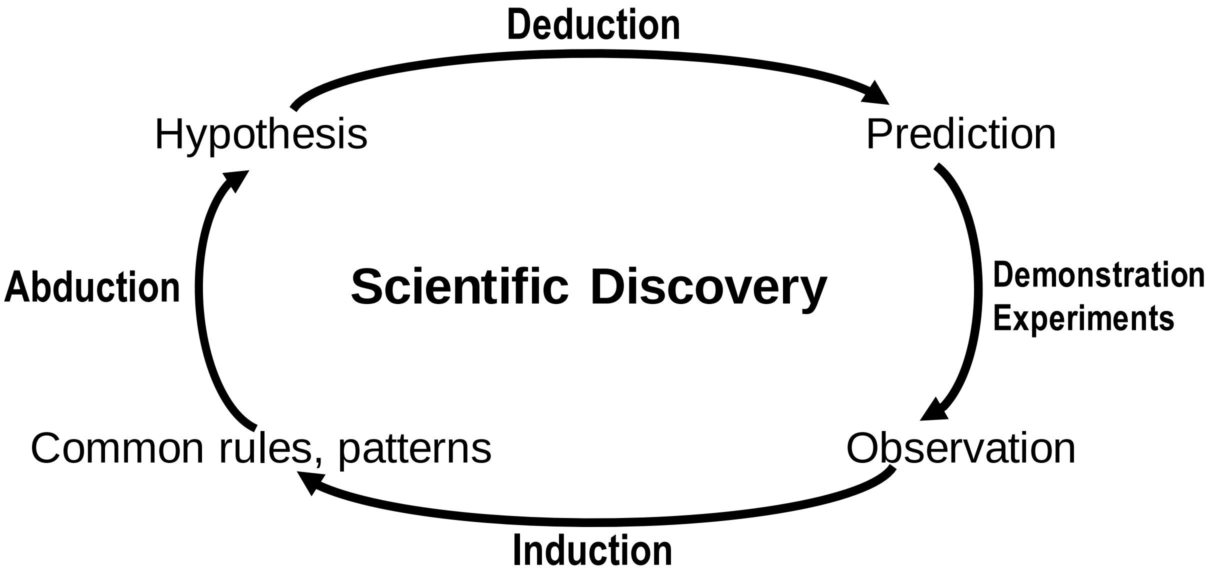


Figure 1.1. Closed loop for scientific discovery

1.3 Viral dynamics

Modeling and simulation studies of transmission dynamics *between individuals* for any pathogen have originated mathematical epidemiology more than a century before [18]. For example, Hamer built a transmission dynamics model of measles in 1906 [19], and Ross presented a model of malaria in 1911 [20]. Kermack and McKendrick *et al.* established mathematical theory for epidemics around the 1930s [21, 22].

On the other hand, the endeavors to quantify *in vivo* temporal change of cellular population or virulence within individuals have originated the isolation of the Human Immunodeficiency Virus (HIV) in 1983 [23]. Modeling techniques concerning transmission dynamics between individuals met this HIV isolation, forming a significant starting point for advances in the studies on *viral dynamics*, transmission dynamics within individuals based on computational modeling and simulation with time-series clinical or experimental data [24].

Under the COVID-19 situation, the attempts to address the COVID-19 pandemic by computational modeling and simulation of transmission dynamics have been promoted for uncovering the principle of SARS-CoV-2 pathogenesis. The computational models have influenced the development of the various models describing the SARS-CoV-2 transmission dynamics between individuals [25, 26]. Nevertheless, the underlying mechanism of SARS-CoV-2 pathogenesis has not been understood because modeling of the SARS-CoV-2 transmission dynamics *within individuals* is not investigated enough to reproduce *in vivo* data on COVID-19. Namely, few studies exist on SARS-CoV-2 spreading within a host, called viral dynamics in terms of population dynamics, compared to the number of papers about its spreading between hosts. Additionally, plausible and straightforward models have been required for explaining the mechanism of SARS-CoV-2 pathogenesis. Therefore, the story hook we can conceive for finding a new hypothesis would be to narrow down our scope to viral transmission within a host. Exploring and comparing different SARS-CoV-2 dynamics models should provide a novel envision of the dynamical system's behavior within the COVID-19 patients.

Now that our scope is on finding the unknown hypothesis about viral dynamics of SARS-CoV-2, verifying the hypothesis on viral dynamics requires information about components on the same layer as pathways for understanding biological mechanisms as systems of pathways. Therefore, we set the second viewpoint, multiscale modeling.

1.4 Multiscale modeling

The origin of multiscale modeling can be seen in the hierarchy of life in molecular biology in the 20th century. Molecular biology typically adopts reductionism, in which a component can be explained as the sum of components on the lower layer [27]. In contrast, systems theory since the 1960s encountered the development in microscopic observation techniques such as deoxyribonucleic acid (DNA) microarray since the 1990s or high-throughput sequencing of omics data since the 2000s, forming systems biology from holism, a component is more than the sum of components on the lower layer. Afterward, integrating these two standpoints is also argued as relational biology from neo-reductionism [28]. Multiscale modeling is one topic in relational biology [29]. Multiscale modeling injects reductionism into systems biology, which enables macro-to-micro mapping, recognizing microstates are impossible to analyze [27] and verification of macro phenomena from microscale to predict overall dynamics [30]. Thus, we adopt multiscale modeling to verify the hypothesis on macroscopic SARS-CoV-2's viral dynamics from the microscopic system of pathways. Figure 1.2 shows a multiscale model consisting of multiple scales: genes, transcripts, proteins, metabolites, and cells. The components

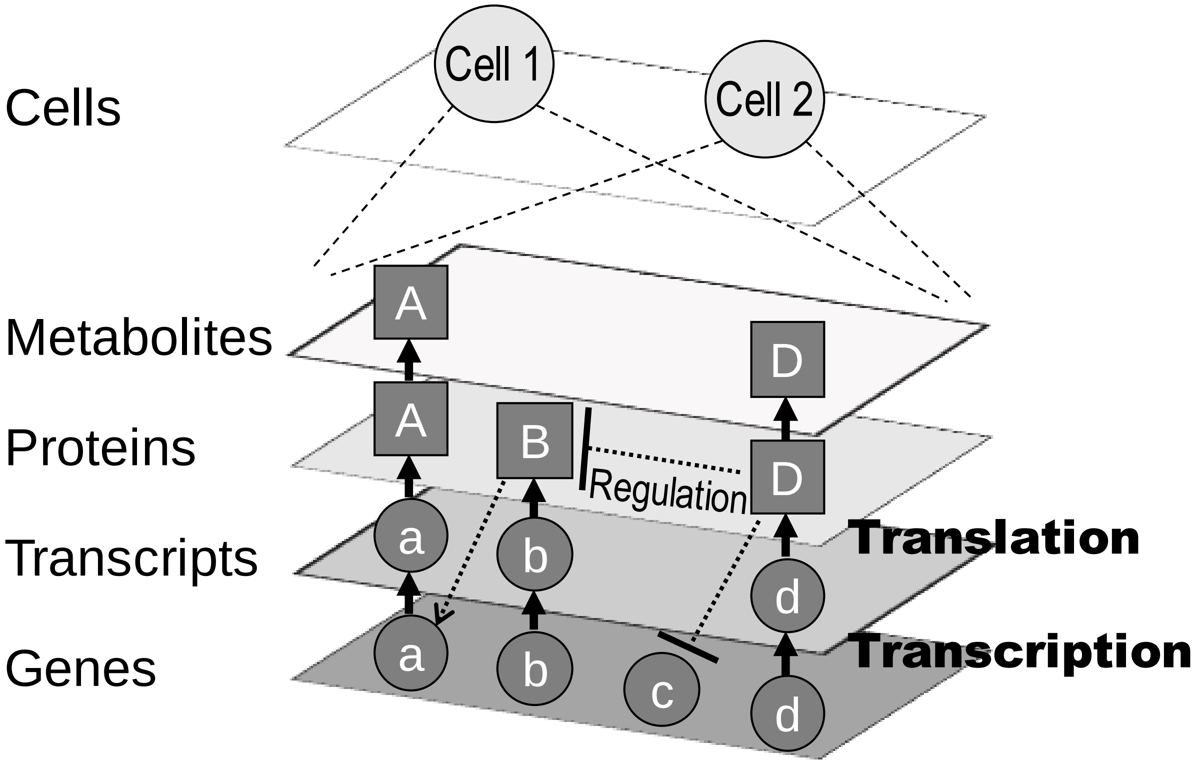


Figure 1.2. Multiscale model

on each scale interact with components on the same or different scales. For example, proteins are related via PPI on the same protein scale. In the case of genes, they receive genetic control of transcription from transcripts and proteins at higher scales. Thus, the components of a multiscale model form a network structure through interactions on the same scale or across different scales. We adopt multiscale modeling to verify the hypothesis on macroscopic SARS-CoV-2's viral dynamics from the microscopic system of pathways.

1.5 Two studies for scientific discovery toward uncovering COVID-19 mechanism

Projecting the above viewpoints onto a closed loop of scientific discovery allows us to carry out two studies. One is the attempt to find a new hypothesis about within-host SARS-CoV-2 dynamics as a population dynamics from a macroscopic view (Study 1). The other verifies this hypothesis as a result of a system of pathways from a microscopic view (Study 2). Specifically, we start by creating hypothesized viral dynamics models for SARS-CoV-2 (Family: Coronaviridae) based on parsing tree structure of viral dynamics models for human T-lymphocytropic virus or HIV-1 (Family: Retroviridae) and literature about the structural similarity of viral surface protein between these retroviruses and SARS-CoV-2. Simulation of hypothesized viral dynamics models, while fitting models to data quantifying the magnitude of viral infection, results in prediction on the specific viral dynamics. Subsequently, microscale gene expression data that would include information on viral dynamics are analyzed, and network structure underlying data is extracted. Here in the figure, a grey-colored path from prediction to observation means that we do not conduct animal or infection experiments or demonstrations. In practice, we utilize observed data provided by other experimental studies. Finally, we unify this data-driven network structure and background knowledge to yield hypotheses about unknown signal pathways, thereby fulfilling our contribution to elucidating the COVID-19 mechanism.

This manuscript is a compilation of the results of our studies of COVID-19, taking into account its social impact and significance in terms of social demands. The contributions for each aspect of computer science and biology are summarized as below.

1. Our contributions to computer science
   - Creating a pipeline for acquiring optimum models and parameters from time-series data and literature on state variables
     Novelty:

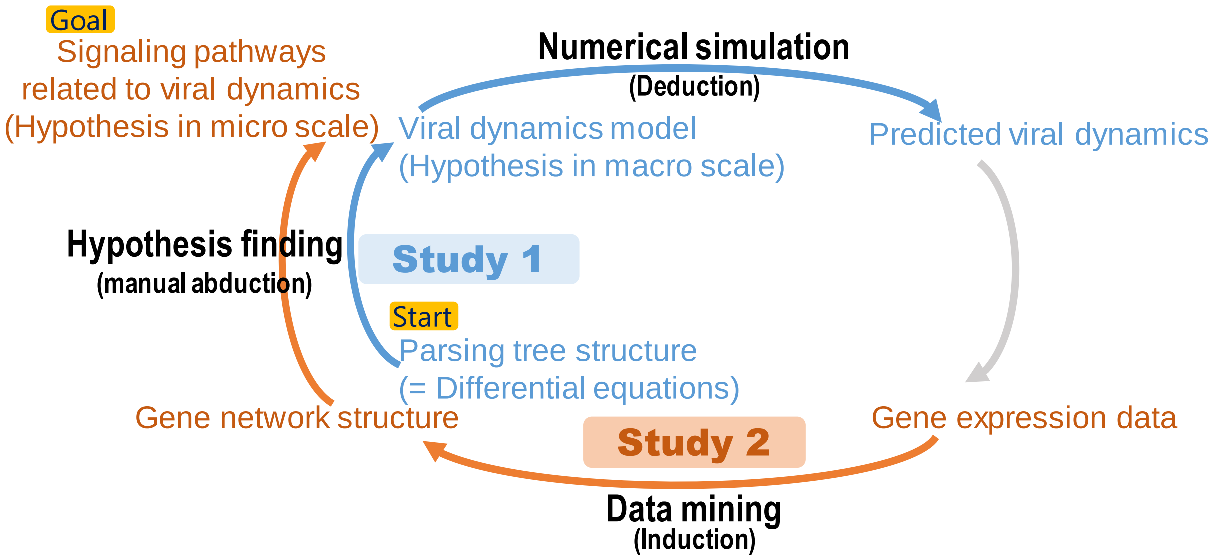


Figure 1.3. Relationship between Chapters 3 and 4

(a) Existing model diversion and original model construction (The comparison itself would be the first roll-out. Building an original model that outperforms the existing models in predictive accuracy is not the purpose of the study.)

(b) Application of four viral dynamics models for SARS-CoV-2 to empirical data

— Developing and proposing an original framework for automatically inferring systems by combining graphical modeling from the large-scale sparse matrix of multivariate (time-series) data and model validation with multiple knowledge bases and successfully forming the basis for further frameworks
Novelty:

(a) Proposal of a novel framework that combines existing data mining and database integration methods (Exception: Two-step clustering through gene-wise and cell-wise single-cell omics data analyses is an original technique.)

(b) Automation of hypothesis discovery from data and knowledge

2. Our contributions to biology

— Comparing estimated parameters of mathematical models of within-host viral dynamics to the COVID-19 dataset, thereby finding SARS-CoV-2 cell-to-cell transmission hypothesis based on modeling and simulation

— Demonstrating the framework applicability to the COVID-19 gene expression data and biological background knowledge, thereby reproducing existing pathways, discovering novel signaling pathways that might be related to viral dynamics, especially cell-to-cell transmission, and analyzing their spatiotemporal variation for different genes of interest not previously available in the knowledge repositories

The manuscript is organized as follows. Chapter 2 is dedicated to give preliminaries for the techniques used in the following studies. Chapters 3 and 4 constitute a main body part interconnected with the relationship described in Chapter 2. Following them, we review computational approaches related to our approach in Chapter 5. Finally, we conclude this manuscript with several remarks, such as a total summary, limitations, and future perspectives.

To the next Chapter

We provide the necessary preparation for the two studies: a more detailed background on viral dynamics and multiscale modeling.

CHAPTER 2

PRELIMINARIES

Brief summary of this chapter

- Two types of modeling styles: equation-based and agent-based. The viral dynamics model in our study is equation-based.
- Basic viral dynamics model and its practical use
- How to quantitatively measure the state of viral infection
- Integrating domain knowledge into data-driven results for interpretability
- The difference between model verification and model validation

2.1 Equation-based Modeling vs. Agent-based Modeling

There are various extensions and improvements related to model-based research considering two computational modeling types: Equation-based modeling (EBM) and Agent-based modeling (ABM) [31]. These two modeling approaches differ in heterogeneity and homogeneity, social behavior, and schematic representation [32].

ABM is characterized by heterogeneity, i.e., different characteristics at the individual level; state, location coordinates in space, age, gender, speed, degree of interaction [33]. Each individual is assumed to be a social, intelligent agent that constantly modifies its own behavioral rules through feedbacks called micro-macro loops [34, 35]. The schematic representation indicates the state transition diagram, which shows the principles of individual agents' behavioral rules. ABM is often employed to formulate the cellular population dynamics [33]. Some studies use ABM for describing the interaction of the immune system with intestines and lymph nodes, or between tuberculosis and cancer [36]. Despite being oriented more towards heterogeneity, some work exists that efficiently models large homogeneous populations using ABM [32].

In contrast, EBM is not subject to heterogeneity but rather heterogeneous, i.e., the stratification is more straightforward than that of ABM, and the individual characteristics differ depending on the categorized units such as age groups, rather than on the personal level [37]. EBM often assumes neither social individual nor behavioral change; every individual is habituated as an identical particle [38]. For example, assuming that viruses and cells lack sociality is regarded as reasonable. Accordingly, EBM represents a sum of individual state transitions as a stock-flow diagram or a compartment model.

The choice of which modeling approach to employ requires being consistent with the goals of the modeling. In this thesis, we opt for EBM instead of ABM for two reasons: first, we have the purpose of modeling the dynamics of a homogeneous group of cells and molecules; second, the spatial topography information required for ABM is difficult to procure. As one solution, bioimage informatics methods can provide image frames of cell movements, as some studies have tracked the dynamics of cellular reaction-diffusion in slime mold [39]. However, this would only reveal the cell population dynamics, which is not our goal. Besides, other factors need to be taken into account, such as temporal changes in the host immune response, rather than assuming a uniform probability of infection. Continuous ordinary differential equations (ODEs) are only useful if the number of molecules in the reaction volume is sufficiently large. Otherwise, we cannot ignore the molecules' discrete nature. In that case, stochastic or discrete stochastic models can be more appropriate [40].

2.2 Equation-based model of viral dynamics

The first viral dynamics model with ODEs was the HIV-1 dynamics model introduced by Perelson in 1996 [41]. This Perelson's model has described experimental or clinical data on HIV-1 or hepatitis B virus and quantified the virulence at the cellular scale, including viral burst size, basic reproduction number, and viral particle copies' or cells' mean lifetime [42, 43]. Based on the Perelson's model, varied viral dynamics models have been constructed, including a model of macrophage with immune cell influx in inflammation [44], a model of neural progenitor cells' dynamics in neurogenesis [45], or a model with mixed infection [46]. Moreover, the viral dynamics models have succeeded in predicting intervention outcomes or planning practical experiments [47].

Perelson's viral dynamics model has described the time course of three time-dependent state variables called T, I, and V. These state variables $(T, I, V)^{\mathrm{T}} \in \mathbb{R}^3$ correspond to the host's target cell density (susceptible cell count), the host's infectious cell density, and viral quantification measure density, respectively. Here, the dynamical system is assumed as a homogeneous wellstirred reaction system, independent of spatial distribution within each compartment. The state transition diagrams of a single target cell, a single infectious cell, and a viral particle per unit time are illustrated in Figure 2.1. A single target cell becomes infectious proportionally to viral particles density V. Let β be the proportionality constant involved in this infection establishment (virus infection rate). A single target cell turns into an infectious cell at a rate of βV. A single target cell also dies at a rate of μ_1 (target cell mortality). A single infectious cell is removed at a rate of μ_2 (infectious cell mortality) due to activated cell death or cell degeneration associated with virus replication or cytotoxicity during the immune response. A single viral particle is removed at a rate of μ_3 (virus mortality) by the culture medium exchange or physiological reaction or antibody neutralization reaction. Summing up the population of target cells, infected cells, and viral particles whose state transitions are described above for any individual, viral dynamics is regarded as population dynamics. Additionally, infectious cells replicate, release and replenish new viral particles to V in proportion to I. Let k be this proportionality constant (viral shedding rate). Therefore, the ODEs of the baseline viral dynamics model are as follows:

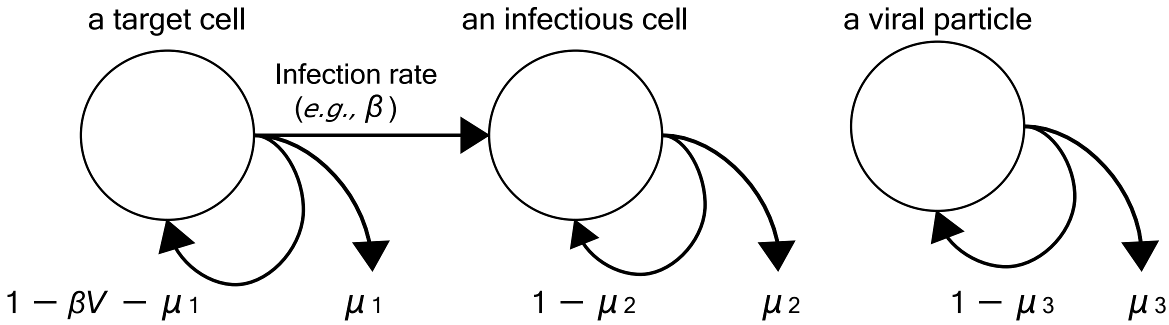


Figure 2.1. State transition diagrams of viral dynamics model. The diagrams illustrate three states, including a target (susceptible) cell, an infectious cell, and a viral particle, and their transitions. A single target cell turns into an infectious cell at an infection rate β proportional to viral particles density V. These are dead or killed at each mortality rate μ_1, μ_2, and μ_3.

$$\begin{aligned}
\frac{dT}{dt} &= -\beta TV - \mu_1 T, \\
\frac{dI}{dt} &= \beta TV - \mu_2 I, \\
\frac{dV}{dt} &= -\mu_3 V + kI.
\end{aligned}$$

2.3 Benefit of modeling viral dynamics

Given elucidation of the *in vivo* infection system of viral dynamics, it is necessary to obtain the time-series changes of viral indicators and explain why the time-series changes occur. For the time-series data, it is possible to use the observed data in clinical tests and experiments. These include genetic tests (quantification of viral RNA levels by real-time PCR), antibody tests (Enzyme-Linked Immuno Sorbent Assay [ELISA] and the faster and simpler immunochromatographic method), routined blood tests, single-cell data, and Electronic Health Record (EHR), and so on [48]. The visualization and statistical processing of such observed data can help us to understand phenomenological aspects of what is happening [49].

However, the above-mentioned techniques are difficult to explain the contents of the black box of etiology, pathogenesis, and why it is happening and provide a biological interpretation [50]. In this thesis, we adopt model-based approach, combining computational models that describe the phenomena' causality and mechanisms to look into the viral dynamics. The model-based approach emphasizes the premise that computational models are necessary before data and involves repeated model fitting with actual observation data to search for plausible models that explain the phenomena well and make future predictions [51]. This type of model-based forecasting is a useful tool for making policy advocacy. In fact, in the 1970s and 1980s, the United Kingdom suppressed Congenital Rubella Syndrome (CRS) through immunizations based on computational models [52]. These success stories illustrate the effectiveness of model-based studies. Akin to CRS, for COVID-19, individuals' population dynamics at the macroscopic scale have been explored by computational models such as Susceptible-Infectious-Removed (SIR) models [25, 53].

In this manuscript, viral dynamics model of HIV serves as the primary computational model because SARS-CoV-2 has several similarities with HIV-1, including the fact that it belongs to positive-sense single-stranded RNA viruses and the similarity of SARS-CoV-2 main protease (Mpro) and HIV-1 protease structure [54, 55].

2.4 Quantification of viral infection

To date, we accumulate knowledge such as the correlation between the severity of the disease and various factors (*e.g.,* angiotensin-converting enzyme 2 [ACE2], transmembrane protease serine 2 [TMPRSS2], and Furin) in the symptomatic period [56]. However, one does not fully understand the *in vivo* transmission system dynamics of SARS-CoV-2 to explain the pathogenesis. For example, little is known on the temporal changes in infectious burden during the not symptomatic phase, the switching mechanism between asymptomatic recovery and symptomatic disease. Also, there exist cases of *reactivation* (exacerbation) after hospital discharge, i.e., patients who do not recover, meet the discharge criteria, and become symptomatic again. In other words, there is a lack of understanding of viral dynamics involving host cell density, dynamics of virus quantification (*e.g.*, viral load, plaque forming units [PFU], and fifty-percent tissue culture infective dose [TCID50] [57]), and temporal changes in the expression of genes that mediate immune responsive signaling.

The notable examples of reactivation are some false-negative cases [58], which met the standard discharge criteria, including polymerase chain reaction (PCR) testing [59]. Such reactivation cases may mean that there are problems with the accuracy and interpretation of the real-time PCR in some cases [60].

2.5 Data-Driven and Knowledge-Based (DD-KB) approach

As mentioned in the introduction, informatics research related to infectious disease control includes data-driven research such as the visualization of the cumulative number of infection indicators [61] or the prediction and classification using machine learning [62]. However, results from data alone are insufficient to explain or interpret the identified data. For example, when considering a system as a pathway, simply mapping a model, such as a network obtained from data, to a knowledge repository calls into question the model's validity. Therefore, knowledge-based research should be conducted in conjunction with data-driven research. Here, data are given as observed values of state variables. Knowledge representation encompasses logical formulas (logical models), a set of causal relation triplets (head entity, relation, tail entity), graphical representations of correlations and causal relationships among components (graphical models or knowledge graphs), representations of equations in reverse Polish notation, and syntax tree structures. Let us designate such integration of data and knowledge as a Data-Driven and Knowledge-Based (DD-KB) approach. DD-KB approach can be seen in other studies. Mattioli et al. propose a unique pipeline of data-driven and connectionist artificial intelligence (AI), knowledge-based AI, and hybrid AI in the context of safety, i.e., freedom from intolerable risk [63]. Vafaee et al. applied their DD-KB approach to new microRNA biomarker discovery of colorectal cancer prognosis [64].

2.6 Model Verification and Validation (V & V)

Model Verification and Validation (V & V) is a methodology for the development of computational models that can be used to quantify confidence and build credibility in modeling [65]. The V & V definitions in this manuscript is adopted from the 1998 AIAA Guide [66]:

— Model verification

The process of determining that a model implementation accurately represents the developer's conceptual description of the model and the solution to the model

— Model validation

The process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model

Model V & V cannot prove that a model is correct and accurate for all conditions, but it can provide evidence that a model is sufficiently accurate [65]. In order to expect outcome that holds an agreement between experimental data and model prediction, we check whether a model is plausible compared to background knowledge.

To the next Chapter

Study 1 uses modeling and simulation of macroscale viral dynamics to find a hypothesis. Among viral dynamics, we mainly focus on two types of viral transmission from one cell to another.

Chapter 3

Modeling Viral Dynamics in SARS-CoV-2 Infection Based on Differential Equations and Numerical Analysis

Brief summary of this chapter

- — Exploring four viral dynamics models based on differential equations
- — Sensitivity analysis and stability analysis
- — Fitting models to data across mild and severe COVID-19 patients
- — Comparison of optimized key parameters for finding hypothesis on SARS-CoV-2 cell-to-cell transmission, a direct viral transfer from one cell to another cell

Published content and contributions:

1. Journal paper (IF=3.8, h-index=69, Q1)
   Odaka M, Inoue K. 2021. Modeling viral dynamics in SARS-CoV-2 infection based on differential equations and numerical analysis. *Heliyon* vol. 7, 10 (2021): e08207. DOI: 10.1016/j.heliyon.2021.e08207
2. Proceeding paper
   Odaka M, Inoue K. 2020. Computational Modeling and Simulation of Viral Load Kinetics in SARS-CoV-2 Replication. *In Proceedings of the Eleventh International Conference on Computational Systems-Biology and Bioinformatics (CSBio2020).* Association for Computing Machinery, New York, NY, USA, 75–82 DOI: 10.1145/3429210.3429214
3. (+α) Oral presentation in France
   Odaka M, Magnin M, Inoue K. 2022. Exploring Differential Equations for Modeling SARS-CoV-2 Dynamics with Sensitivity and Stability Analysis. *Statistical Methods for Post Genomic Data (SMPGD)*. Nantes, France.

Recap on what we do in Study 1 (Figure 3.1):

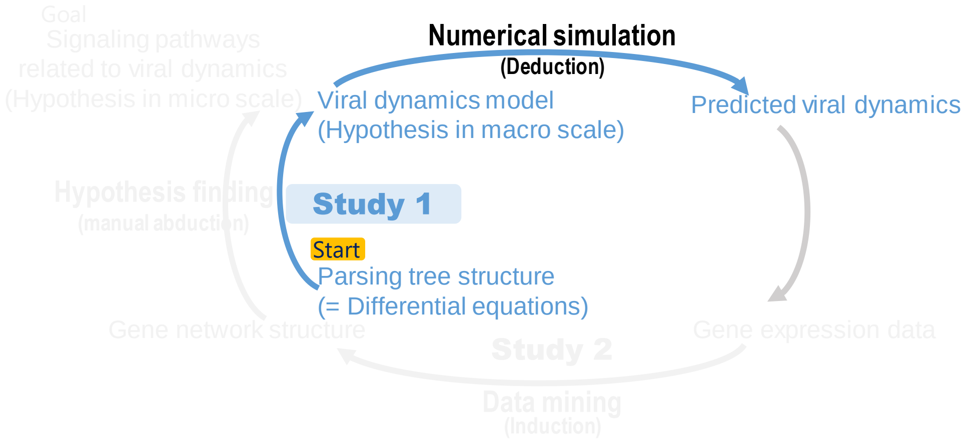


Figure 3.1. Study 1 overview. Finding new hypothesis about within-host SARS-CoV-2 dynamics as population dynamics (macroscopic view)

3.1 Introduction

This study is positioned as a pilot study for the future establishment of a pipeline that outputs an optimal model and parameters as long as the strings of the state variables are available. If such a pipeline can be established, it will be possible to combine it with natural language processing to automatically obtain models from information on the state variables of the user's interest more efficiently.

The purpose of this study towards establishing the pipeline includes the following two things: first, to build and compare multiple SARS-CoV-2 dynamics models based on ODEs; second, to fit the models to two cases of the observed COVID-19 experimental data. Compared to the existing research, the foci of this chapter are on constructing different SARS-CoV-2 dynamics models by abstracting *in vivo* SARS-CoV-2 pathogenesis as dynamical systems and distilling beneficial models that describe the population dynamics of host cells and viral particles.

On building the SARS-CoV-2 dynamics models, numerical analysis enhances the quality of modeling and simulation. In particular, pruning the fixable parameters based on sensitivity indices simplifies the redundant models, thereby balancing the model complexity and simplicity. Calculating the eigenvalues of the simplified models guarantees the solutions' orbital stability.

Further, the calibration experiments fit the simulated data generated from the models to two cases of actual observed data. Here, the comparison of the parameter values estimated from the viral load data sequence in mild patients and those in severe cases clarifies the relationship between the key parameters and the COVID-19 severity.

From another perspective, the content of Study 1 is an extension of the previous work, which has contributed to computational modeling and computer simulation of SARS-CoV-2 viral load kinetics and suggested a qualitative relationship between the asymptomatic carriers' reactivation risk and the COVID-19 severity [67]. As an improvement of the previous work, this chapter extends the scope of sensitivity analysis from one model to four models to simplify the models, evaluates the equilibrium solution's stability to ensure stable calibration, and conducts the calibration experiments to avoid local minima.

The rest of the chapter is organized as follows: Section 3.2 introduces four viral dynamics models. Section 3.3 explains the methods for data preparation, sensitivity analysis, stability analysis, and calibration experiments. Section 3.4 shows the results and expands the discussion. Section 3.7 is devoted to providing related work. Section 3.8 concludes with a summary of contributions, limitations, and future work.

3.2 Proposed models

This section formalizes the SARS-CoV-2 dynamical system consisting of host cells and viral particles with four computational models. One is Perelson's viral dynamics model as a baseline model described in Chapter 2. Other two models are the models derived from literature as the extensions of the Perelson's model. These three of four have successfully explained viral dynamics on other viruses, whereas the appropriate investigation of these models for SARS-CoV-2 dynamics has not been conducted. The last one is a newly constructed model.

3.2.1 Huang's model (functional response)

While Perelson's baseline model has demonstrated virus replication or host-pathogen interactions well, some experts have regarded it as a too simple model due to its linear infection rate β. Huang *et al.* expressed a more realistic infection rate bound to overhead by introducing a nonlinear term (*functional response*) [68]. By introducing the functional response, the shape of a rectangular hyperbola indicates the actual incidence rate well. This nonlinear term is $\beta TV/(1 + aT + bV)$, where a and b are constant values greater than or equal to zero. The term is similar to the Holling type II incidence functional response. Still, the additional term bV representing a mutual interference between viruses makes it different from Holling type II [69, 70].

3.2.2 Pearce-Pratt-Phillips model (viral synapse)

While the above models have taken a single transmission chance into account, in 1994, Pearce-Pratt and Phillips *et al.* explicitly presented a scheme of HIV transmission via two routes: cell-free transmission and cell-to-cell transmission [71]. Specifically, the structure mediating the cell-to-cell transmission as a counterpart of the cell-free transmission is called *viral synapse* [72]. Given that both SARS-CoV-2 and HIV have the spike glycoprotein on the surface of the viral envelope [73] and that it has a similar function such as viral entry, receptor recognition, cell attachment, and fusion [74], the viral synapse is presumably in the SARS-CoV-2 life cycle as well. Figure 3.2 shows a schematic representation of the SARS-CoV-2 life cycle to explain the differences in two types of transmissions and wherein the viral shedding constant k is also relevant.

A free viral particle attaches to a target cell binding to ACE2 receptor on the cell membrane supported by spike protein degradation by TMPRSS2 [56, 75]. Without elaborating on the

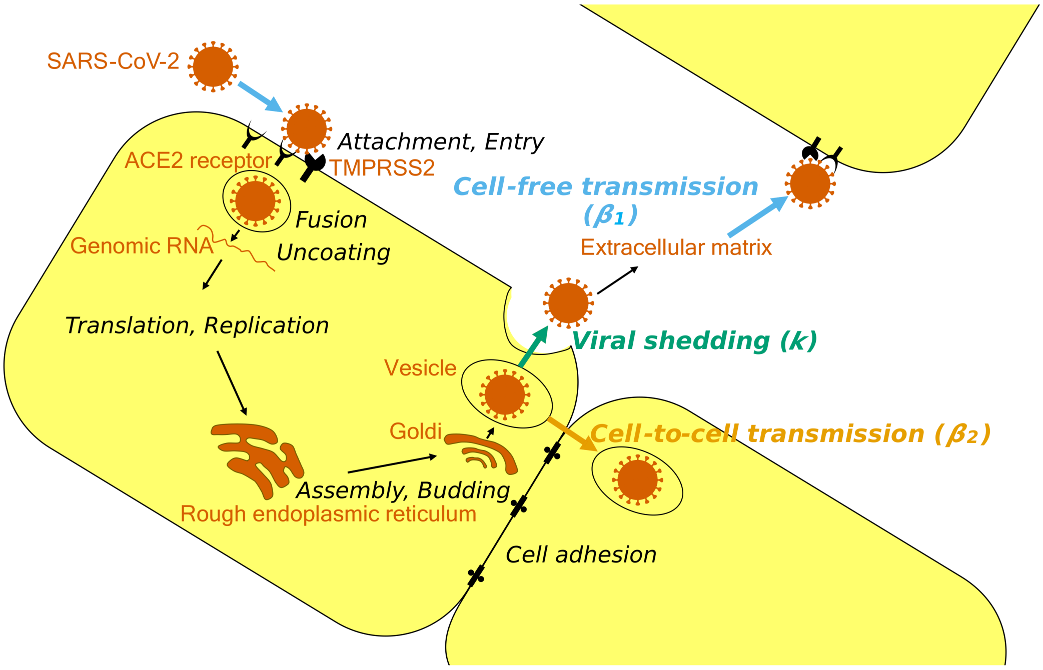


Figure 3.2. Schematic representation of SARS-CoV-2 life cycle. Here we assume the unknown direct viral transfer, cell-to-cell transmission (highlighted dense yellow), to be included in the SARS-CoV-2 life cycle, contrasting indirect viral transfer, cell-free transmission (highlighted skyblue). A particle of SARS-CoV-2 infects a host cell at a cell-free transmission rate β_1, binding to ACE2 receptor helped by TMPRSS2. The virus undergoes the subsequent typical processes, finally being released (highlighted green) at a viral shedding rate k. Here, the virus in the host cell infects another cell at a cell-to-cell transmission rate β_2. The rest of colors are as follows: black-colored texts and arrows are life-cycle processes; black-colored borders are cellular/vesicular membranes and membrane proteins; orange-colored texts and objectives are viruses, organelles, and extracellular matrix.

detailed translation process to replication, the copied viral particles are released at the magnitude of k. The cell-free transmission involves these multiplied viral particles' attachment to other cells after shedding to the extracellular matrix [76]. Consequently, the degree of cell-free transmission is proportional to the viral particle density. β_1 denotes this proportionality constant.

During the cell-to-cell transmission, viral particles directly enter neighboring cells through viral synapse mediated by cellular adhesion molecules [77]. Thus, the level of this direct entry is supposed to be proportional to the infectious cell density. β_2 is set as this proportionality constant. Reflecting the two transmission types, one obtains a term for infection rate as $\beta_1 TV + \beta_2 TI$.

3.2.3 New model (functional response and viral synapse)

The models reviewed above could have caused one to have a bias in exploring models due to one's subjective point of view [78]. Procedurally generating a model outside the scope of subjective bias compensated for the above models [79]. For simplicity, M_1, M_2, M_3 denote the above models in short. The machinery manipulation of subtree mutation of M_2 and M_3 generated a new model M_4. Figure 3.3 shows the parse trees reflecting the infection rate terms of M_1, M_2, M_3, and M_4. Substituting the dashed subtree of M_3 with the dashed subtree of M_2 generated the parse tree for the infection rate term of M_4.

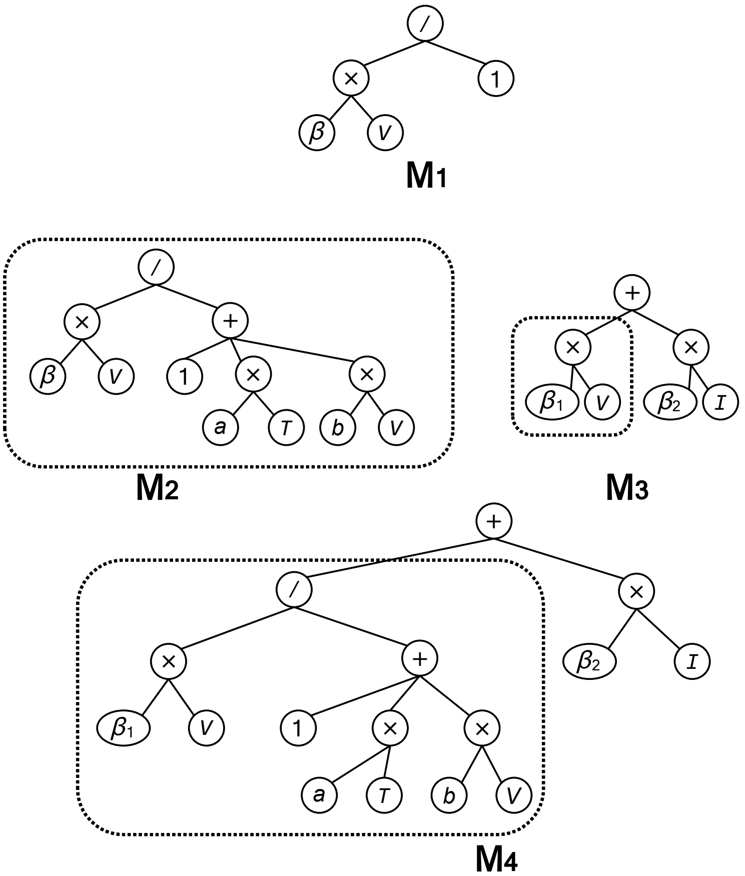


Figure 3.3. Infection rate terms of viral dynamics models. The parse trees stand for the infection rate terms of different viral dynamics models; M_1, M_2, M_3, and M_4. The trees consist of arithmetic operators and the variables and constants in Table 3.2. The trees of Huang's model M_2 and Pearce-Pratt-Phillips model M_3 originate from that of the baseline model M_1. The dashed subtree mutation between M_2 and M_3 generates the tree of an original model M_4.

This section has prepared the four models with different terms for the infection rate. Table 3.1 is a summary of the difference among the models. Table 3.2 is a summary of the symbols, definitions, and ranges of the variables and constants of the ODEs. As a pilot study, the new model $M4$ was introduced only procedurally and without any particular motivation, nor did we assume any situation in which there was a motivation to focus on the specific models prior to the analysis.

Table 3.1. Summary of four viral dynamics models and their corresponding ordinary differential equations (ODEs).

Models	ODEs
M_1: Perelson's model (baseline) [41]	$dT/dt = -\beta TV - \mu_1 T$
	$dI/dt = \beta TV - \mu_2 I$
	$dV/dt = -\mu_3 V + kI$
M_2: Huang's model (functional response) [68]	$dT/dt = -\beta TV/(1 + aT + bV) - \mu_1 T$
	$dI/dt = \beta TV/(1 + aT + bV) - \mu_2 I$
	$dV/dt = -\mu_3 V + kI$
M_3: Pearce-Pratt-Phillips model (viral synapse) [71]	$dT/dt = -\beta_1 TV - \beta_2 TI - \mu_1 T$
	$dI/dt = \beta_1 TV + \beta_2 TI - \mu_2 I$
	$dV/dt = -\mu_3 V + kI$
M_4: New model (functional response and viral synapse)	$dT/dt = -\beta_1 TV/(1 + aT + bV) - \beta_2 TI - \mu_1 T$
	$dI/dt = \beta_1 TV/(1 + aT + bV) + \beta_2 TI - \mu_2 I$
	$dV/dt = -\mu_3 V + kI$

Table 3.2. Summary of variables and constants and their corresponding symbols, definitions, and ranges.

Symbol	Definition	Range
t	unit time (*e.g.,* day) since symptom onset or the start of the experiment	$t \in [0\ \infty)$
T, I, V	target cell density, infectious cell density, virus density	$(T, I, V) : T \geq 0,$ $I \geq 0, V \geq 0$
β	virus infection rate	$\beta \in (0, 1)$
k	viral shedding rate	$k \in (0, 1)$
μ_1	target cell mortality	$\mu_1 \in (0, 1)$
μ_2	infectious cell mortality	$\mu_2 \in (0, 1)$
μ_3	virus mortality	$\mu_3 \in (0, 1)$
a	proportional constant	$a \in (0, 1)$
b	proportional constant	$b \in (0, 1)$
β_1	cell-free transmission rate	$\beta_1 \in (0, 1)$
β_2	cell-to-cell transmission rate	$\beta_2 \in (0, 1)$

3.3 Proposed methods

This section covers data and the remaining steps; numerical analysis and calibration experiments. Figure 3.4 shows a pipeline of the research methods.

This pipeline of methods (Figure 3.4) explicitly sees input as observable state variable(s) and output as models and optimum conditions. The intermediate computation process is a workflow

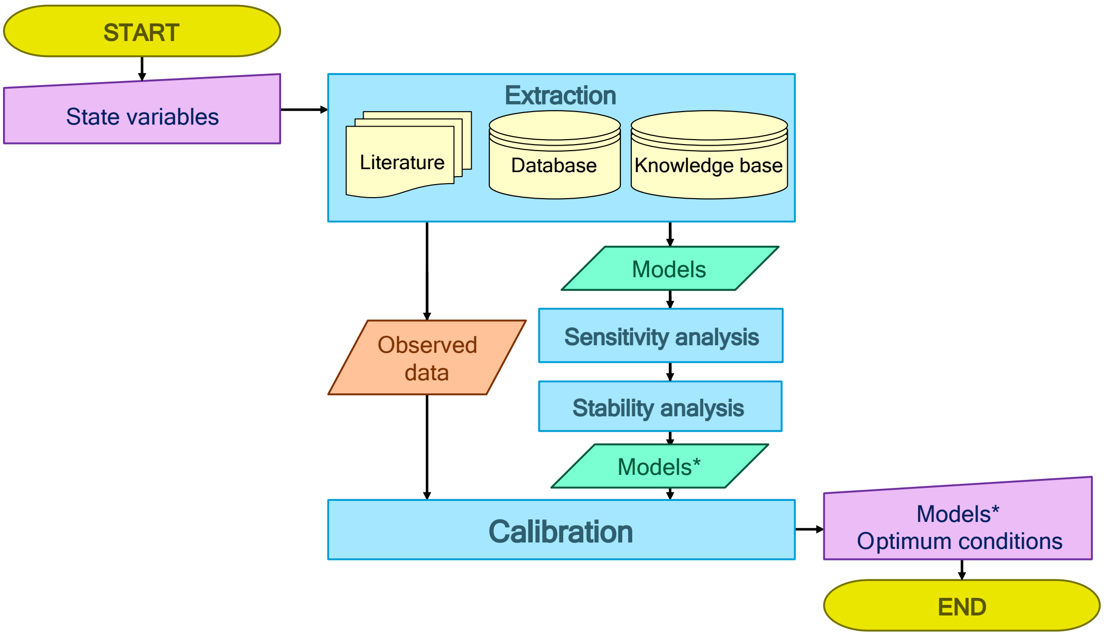


Figure 3.4. Pipeline of research methods, explicitly seeing input as observable state variable(s) and output as models and optimum conditions.

of three tasks: the extraction of data and models as to the inputted state variables, the numerical analysis simplifying the extracted models with sensitivity and stability, and the calibration between data and the simplified models. Here, the system of interest is assumed to be closed and determined only by the state variables of the extracted models.

3.3.1 Observed SARS-CoV-2 data

The literature, knowledge bases, and databases were searched to extract actual time-series data and the models with the state variable. Here, the state variable must be an observable viral quantification in clinical tests or experiments. The viral quantification includes viral load, which one can estimate from total viral particle copies by quantitative reverse transcription-polymerase chain reaction (qRT-PCR) of the specimen such as mucus in nasopharyngeal swab collection [48].

As a case study, viral load data was used. The data was from an image of time-series data sequences of median viral load in the mild and severe patient populations (anonymized) published in the previous literature [80]. The primary source originated 96 patients with SARS-CoV-2 infection (22 mild patients and 74 severe patients) collected by a COVID-19 designated hospital in Zhejiang Province, China, from January 19, 2020, to March 20, 2020. The severity diagnosis was according to the Chinese guideline for diagnosis and treatment of SARS-CoV-2 by National Health and Family Planning Commission of the People's Republic of China. This source has been licensed under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license, which has permitted to edit, process, and use it as a secondary source, subject to the author's acknowledgment [81]. The image processing via an open software *WebPlotDigitizer* version 4.3 transformed the viral load data points into coordinate values [82, 83].

Figure 3.5 illustrates daily viral load sequences since symptom onset in mild and severe cases.

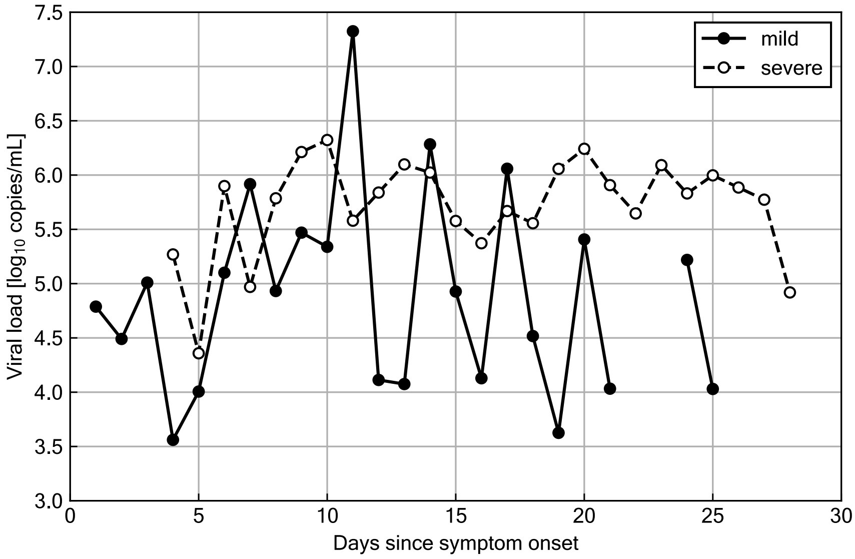


Figure 3.5. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) viral load data across the days since symptom onset in mild cases (solid line with black markers) and severe cases (dashed line with white markers). Each of these data sequences is a derivative of original figure by Zheng *et al.*, licensed under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license.

Viral load fluctuates and attenuates in both cases over time, often higher in severe cases except on days 7, 11, 14, and 17. The missing values in the original sequences have undergone an imputation by linear interpolations.

3.3.2 Sensitivity analysis

Subsequently, the sensitivity analysis was devised as the model's complexity reduction process. Sensitivity analysis identifies the parameters with little effect on the output even if fixed within the boundary conditions, and thereby one is reduced to calibrating simplified models only [84, 85]. The simple sensitivity analysis method is One-Factor-at-a-Time (OFAT), which examines a linear relationship between inputs and outputs. However, using OFAT is limited to exploratory modeling because it does not consider the combinatorial parameter variability [86]. Given a nonlinear system of viral dynamics and interactivity among multiple parameters, global sensitivity analysis (GSA) was employed in our study. GSA based on simultaneous perturbations of the entire model parameter space can investigate the parameter effects on the model's output individually and in combinations [87].

Suppose d-th dimensional parameter set $(p_1, p_2, \cdots, p_d)$, where each p_i is standardized as $p_i \in [0, 1]$. i and X_i denote parameter index running over natural numbers $\{1, \ldots, d\}$ and parameter set samples with only p_i fixed respectively. The contribution of p_i to output Y variance with all parameters varied is given by:

$$S_{T_i} = 1 - \frac{\mathrm{Var}_{X_{\sim i}}(\mathrm{E}(Y \mid X_{\sim i}))}{\mathrm{Var}(Y)}$$

Where:

Var : variance

E : expected value

The Quasi-Monte Carlo sampling method generated parameter value sets (Sobol sequences) with lower discrepancy than random value sets, and thereby yielding p_i with small S_{T_i} [88, 89].

3.3.3 Stability analysis

Additionally, the stability analysis examined the dynamic behavior of the solution trajectory in the neighborhood of the fixed point in phase space. The purpose of stability analysis is to guarantee that any solution is stable [90]. In other words, this process can imply the necessity of other separate simulations or detailed analysis near the bifurcation parameter conditions whenever the equilibrium solution bifurcates [91]. To perform a stability analysis of stationary equilibrium solutions, one can ground the Routh-Hurwitz theorem wherein the behavior of the system near the steady-state is related to the eigenvalues of the Jacobian matrix [92, 93].

Theorem (Routh-Hurwitz theorem)**.** If all the eigenvalues of the Jacobian matrix have negative real parts, the stationary solution is asymptotically stable. If any eigenvalue has a positive real part, the solution is unstable; if the maximum real part of the eigenvalues equals zero, the Jacobian matrix cannot characterize the stability.

Consequently, the eigenvalues of the Jacobian matrices of the two equilibrium solutions were calculated: the disease-free equilibrium (DFE) point, where the disease dies out, and the endemic equilibrium (EE) point, where the disease remains persistent [94]. For example, the Jacobian matrix of the M_1's DFE point $E_1 = (T_0, 0, 0)$ was

$$J(E_1) = \begin{bmatrix} -\mu_1 & 0 & -\beta T_0 \\ 0 & -\mu_2 & \beta T_0 \\ 0 & k & -\mu_3 \end{bmatrix}.$$

The Jacobian matrix of the M_1's EE point $E_1^* = (T^*, I^*, V^*)$ was

$$J(E_1^*) = \begin{bmatrix} -\mu_1 - \beta V^* & 0 & -\beta T^* \\ \beta V^* & -\mu_2 & \beta T^* \\ 0 & k & -\mu_3 \end{bmatrix}.$$

Likewise, the Jacobian matrix of the M_3's DFE point $E_3 = (T_0, 0, 0)$ was

$$J(E_3) = \begin{bmatrix} -\mu_1 & -\beta_2 T_0 & -\beta_1 T_0 \\ 0 & -\mu_2 + \beta_2 T_0 & \beta_1 T_0 \\ 0 & k & -\mu_3 \end{bmatrix}.$$

The Jacobian matrix of the M_3's EE point $E_3^* = (T^*, I^*, V^*)$ was

$$J(E_3^*) = \begin{bmatrix} -\mu_1 - \beta_1 V^* - \beta_2 I^* & -\beta_2 T^* & -\beta_1 T^* \\ \beta_1 V^* + \beta_2 I^* & -\mu_2 + \beta_2 T^* & \beta_1 T^* \\ 0 & k & -\mu_3 \end{bmatrix}.$$

The eigenvalues were calculated from these Jacobian matrices of M_1 and M_3 by SymPy 1.6.2. Finally, the artificially generated data by quadrature of the models' ODEs got calibrated to the observed data. In the calibration experiments, dynamic time warping (DTW) provided a similarity measure between the artificial time series of viral particles from the models and the actual time series of viral load [95]. Here, DTW computes the shortest path two time-series data by finding the absolute error value per point across them, which enables one to obtain the similarity even if their lengths and periods are different [96]. Global optimization of DTW distance as a cost function avoided dropping into local minima by Algorithm 1. Given that the well-posed inverse problems require that any solution is identifiable [97, 98], the calibration experiments estimated the parameter values with the finite prediction bands allowed.

Algorithm 1

```
Input: ODEs, Sobol sequences (n = 1000), observed data (mild or severe)
Output: estimated parameter value sets (n = 1000)
  $Param \leftarrow$ Sobol sequences
  for int $i = 1, i \leq 50, ++i$ do
    initialize DataFrame ($DF$) to empty
    for int $j = 1, j \leq 1000, ++j$ do
      for int $days = 0, days \leq 200, ++days$ do
        $SimData[j] \leftarrow$ ODEs integration with $Param[j]$
      end for
      $DTWdist[j] \leftarrow$ DTW distance between $SimData[j]$ and observed data
      stack ($Param[j], DTWdist[j]$) to $DF$
      sort $DF$ (in descending order by $DTWdist$)
      initialize $Param^*$ to top 250 sets of $Param$
      for int $l = 1, l \leq 3, ++l$ do
        for int $r = 1, r \leq 250, ++r$ do
          add random float value $\in [-0.01, 0.01]$ to one element of $Param[r]$ of $DF[r]$ and
          stack the new parameter value set to $Param^*$
        end for
      end for
    end for
    $Param \leftarrow Param^*$
  end for
```

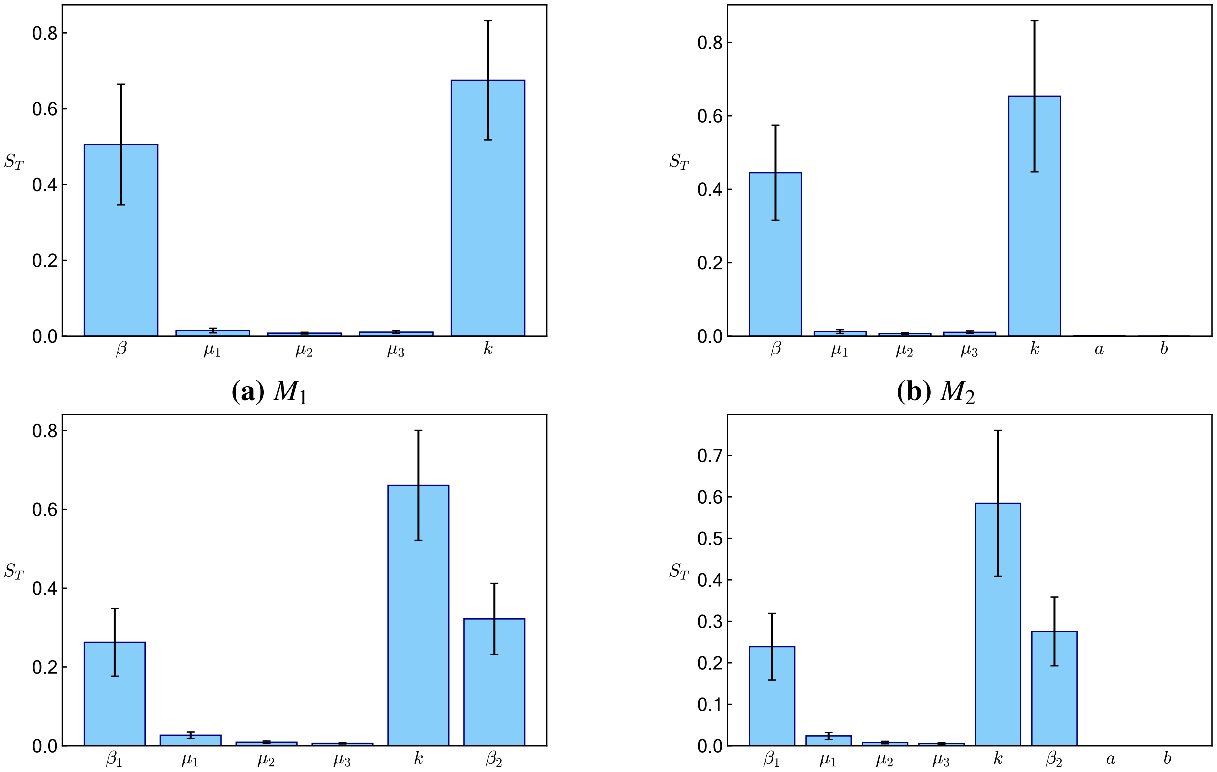
These methods resulted in the optimum set of models with parameter estimates. The experimental configurations were as follows: Intel(R) Core(TM) i7-7500U CPU@2.70GHz, 2904Mhz, 16GB of memory, and Microsoft(R) Windows(R) 10 Operating System.

3.4 Results and discussion

This section shows the results and expands the discussion.

3.4.1 Sensitivity analysis

As the results of GSA, the sensitivity indices with error bars are illustrated in Figure 3.6.



(a) M_1

(b) M_2

(c) M_3

(d) M_4

Figure 3.6. Sensitivity analysis results. Each figure includes the bars reflecting sensitivities to the parameters about four models; (a) Perelson's model M_1 (b) Huang's model M_2 (c) Pearce-Pratt-Phillips model M_3 (d) original model M_4. Each error bar is a 95% confidence interval.

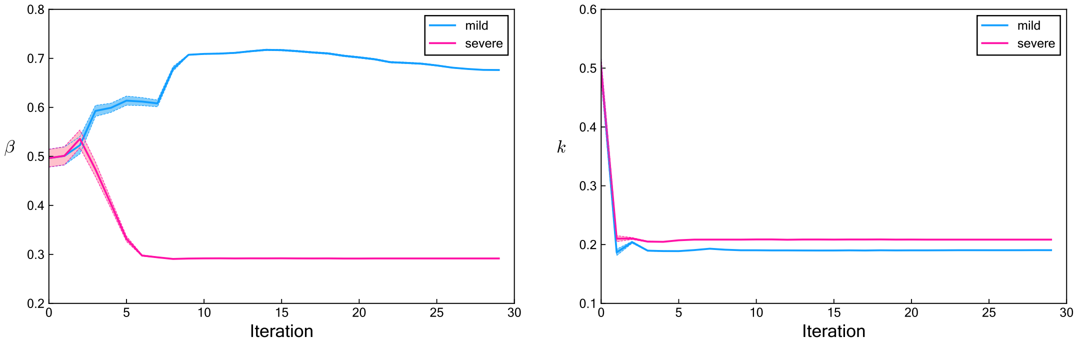
For all models, μ had low sensitivity. a and b had almost zero sensitivities. In contrast, β and k had high sensitivities, where β became distributed between β_1 and β_2 in the models considering viral synapse. Considering that the parameters with small sensitivities can be fixable [99] and that the sensitivities for parameters other than the fixable parameters were similar, the parameter values μ_1, μ_2, μ_3, a, and b were set to zero. This parameter pruning simplified the four models into two models. In particular, M_2 was reduced into M_1 and M_4 merged into M_3. The above model simplification implied that it would be sufficient to perform stability analysis and calibration experiments only for M_1 and M_3.

3.4.2 Stability analysis

Next, according to the stability analysis results, all the eigenvalues of $J(E_0)$, $J(E_0^*)$, $J(E_3)$, and $J(E_3^*)$ had negative real parts. These eigenvalues guaranteed the solution's orbital stability based on the Routh-Hurwitz theorem. Namely, it could be postulated that the two equilibrium solutions, DFE point and EE point, would remain asymptotically stable, which meant no requirement of specific constraints on parameter conditions in the calibration experiments. However, it would be curious that there existed no chaos or bifurcation, and the models' stability did not correspond to the fluctuation in the observed data sequences. Therefore, further searching experimental data sequences without fluctuation over a longer period would deal with this inconsistency.

3.5 Calibration experiment

Hereinafter, the calibration results of M_1 and M_3 are shown. Figure 3.7 shows the calibration results of M_1.

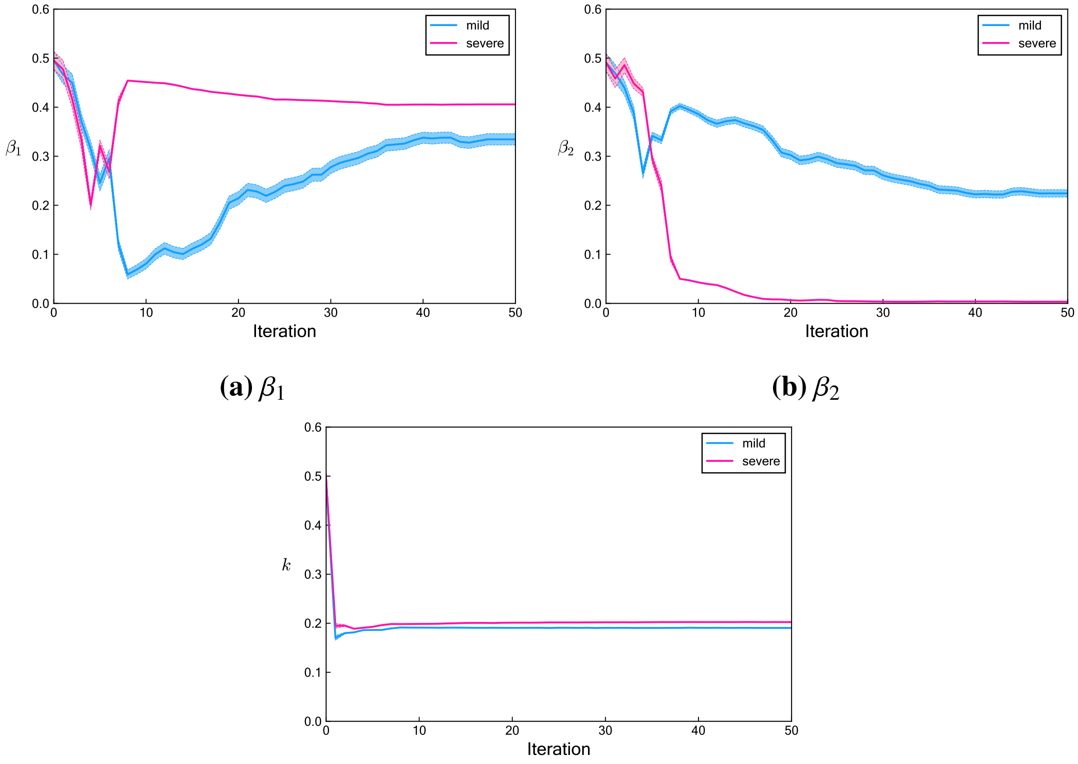


(a) β **(b)** k

Figure 3.7. Calibration results (Perelson's model M_1). The curves (blue: mild) (red: severe) are plotting the mean of estimated parameter value sets of (a) virus infection rate β (b) viral shedding rate k corresponding to the iteration number with prediction bands allowed. The dashed lines and the filled areas are the margins of errors and the prediction bands ±2SE (standard error of the mean), respectively.

The horizontal and vertical axes correspond to the iteration number and the estimated parameter value sets. The blue curve stands for mild cases and the red one for severe cases. The solid lines are not regression curves but the plots of the mean of estimated values. The dashed lines are margins of errors, and the filled areas are prediction bands ±2SE (standard error of the mean). The narrower prediction band reflects the higher prediction accuracy of the mean parameter value. Considering β converged to $(0.70, 0.30)$, whereas k to $(0.19, 0.21)$ for mild and severe cases, M_1 would be an identifiable model. If the relationship between the COVID-19 severity and infection rate were not subject to other factors, it could be speculated that smaller β would have reproduced severe cases. As for k, there was little difference in the estimates between mild and severe as for viral shedding, making it difficult to give a biologically meaningful interpretation.

Figure 3.8 shows the calibration results of M_3.



(a) β_1

(b) β_2

(c) k

Figure 3.8. Calibration results (Pearce-Pratt-Phillips model M_3). The curves (blue: mild) (red: severe) are plotting the mean of estimated parameter value sets of (a) cell-free transmission rate β_1 (b) cell-to-cell transmission rate β_2 (c) viral shedding rate k corresponding to the iteration number with prediction bands allowed. The dashed lines and the filled areas are the margins of errors and the prediction bands ±2SE (standard error of the mean), respectively.

β_1 converged to $(0.32, 0.42)$, β_2 to $(0.25, 0.0050)$, and k to $(0.195, 0.200)$ for mild and severe cases. Regarding viral shedding term k, the same discussion as above for the M_1 results holds. The calibration experiments could not determine the true values of β_1 and β_2 accompanied with the prediction bands. Although M_3 is more complicated than M_1, M_3 would be a partially identifiable model. This difference in the prediction bands would reflect that the model complexity could be a trade-off with the identifiability in the simple model. As for the comparison between β_1 and β_2, β_2 was eightieth of β_1 in severe cases. Suppose it is true that the smaller β in Figure 3.7a results in the more severe COVID-19 symptoms. Then, β_2, which is smaller in severe cases in Figure 3.8b, would be related to the severity rather than β_1. In other words, the cell-to-cell transmission would be essential for severe COVID-19 than cell-free transmission. The recent papers have reported the association between the COVID-19 severity and the expression level of the specific genes related to the cell-to-cell transmission on other viruses [100]. Therefore, one ideal interpretation from the calibration results would be the association between the cell-to-cell transmission and the COVID-19 severity. If accurate, it would lead to claiming the efficacy of drug intervention for β_2 such as a cell-to-cell transmission blocking. However, it has been still unclear whether the genes are involved in the cell-to-cell transmission in COVID-19. Hence, it is necessary to carefully validate the relationship between the cell-to-cell transmission and the COVID-19 severity. Table 3.3 shows the summary of the above calibration results.

Table 3.3. Summary of models and their corresponding converged values of estimated parameters. The values of (β, k) in Perelson's model M_1 and (β_1, β_2, k) in Pearce-Pratt-Phillips model M_3 are shown in mild cases and severe cases.

Model	Parameter	Mild	Severe	Description
M_1	β	0.70	0.30	virus infection rate
	k	0.19	0.21	viral shedding rate
M_3	β_1	0.32	0.42	cell-free transmission rate
	β_2	0.25	0.0050	cell-to-cell transmission rate
	k	0.195	0.200	viral shedding rate

3.6 Interpretation of estimated parameters

The above results show that the smaller β becomes, the more severe the disease is. This relationship may be contrary to our intuition. The comparison of the dynamical behavior of the viral dynamics model M_3 with estimated parameters between mild cases and severe cases (Figure 3.9) also indicates that the smaller β yields the bigger viral load, which is the observed data property in Figure 3.5. Therefore, parameters are successfully estimated from the data. One interpretation is that T has a larger influence in the infection rate term $T(\beta_1 V + \beta_2 I)$. For T on the supply side and I on the demand side, the smaller β_2 means the lower the decrease in T on the supply side, and thus the larger $T(\beta_1 V + \beta_2 I)$ is transferred. Similarly, the smaller β_1 means the larger infection rate term. Figure 3.10 is a quick check of infection rate term in mild cases and

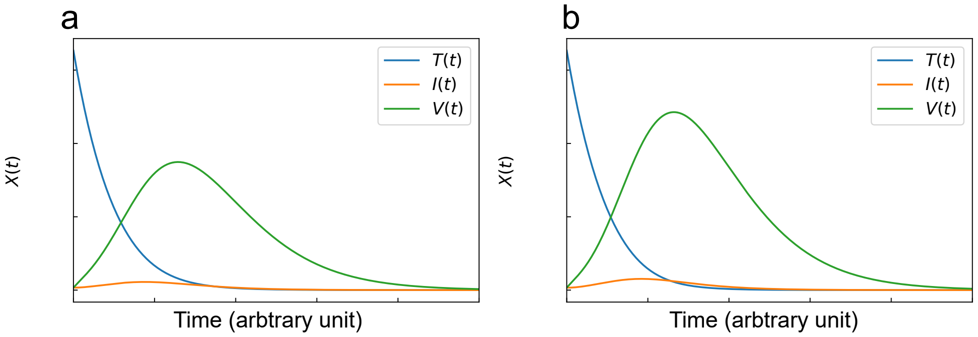


Figure 3.9. Dynamical behavior simulated with estimated parameters. **(a)** Simulation with estimated parameters in mild cases. $\beta_1 = 0.32, \beta_2 = 0.25, k = 0.195, \mu_1 = 0.01, \mu_2 = 0.01, \mu_3 = 0.01$. **(b)** Simulation with estimated parameters in severe cases. $\beta_1 = 0.42, \beta_2 = 0.0050, k = 0.20, \mu_1 = 0.01, \mu_2 = 0.01, \mu_3 = 0.01$.

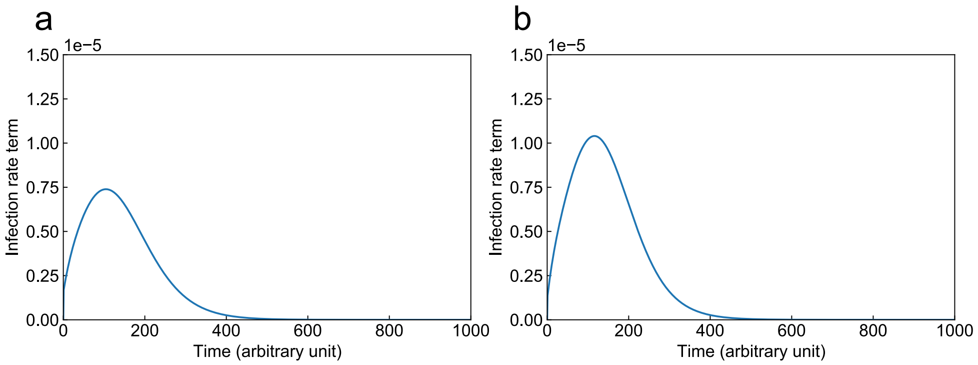


Figure 3.10. Temporal change in infection rate term $T(\beta_1 V + \beta_2 I)$ in mild and severe cases. **(a)** Mild cases **(b)** Severe cases. Parameter values are set as well as Figure 3.9.

severe cases, in which we can confirm that the smaller β_2 yields the higher infection rate term. On the other hand, we also interpret the result where the difference in estimated parameters is more prominent for β_2 than for β_1. This difference can be attributed to the fact that β_1 is multiplied by V and β_2 by I. In viral dynamics models, V is only affected by I, and the effect from T is indirect, whereas I is directly affected by both T and V. Therefore, the difference might become more pronounced in β_2, which favors the direct influence of T.

3.7 Related work

The choice of which modeling approach, ABM or EBM, has required consistency with the modeling goals, such as immunization policymaking [52]. This chapter has opted for EBM in place of ABM for two reasons: first, the homogeneous group of cells and viral particles; second, the difficulty in procuring the spatial information required for ABM.

The model extensions on EBM include the models taking into account the discrete nature of the molecules and temporal changes in the host immune response, rather than assuming a uniform probability of infection. If the number of molecules in the reaction volume is sufficiently large, the continuous ODEs, including stochastic or discrete stochastic models as more appropriate models, are sometimes helpful [101]. Compared with the asymptotically stable models, fractional models with a non-integer order derivative can reproduce more complex behavior [102]. For example, the fractional model in the Caputo-Fabrizio derivative with a nonsingular kernel has successfully described the dynamics of hepatitis B virus or tuberculosis [103, 104]. Otherwise, the fractional model in the Atangana-Baleanu derivative with nonsingular and nonlocal kernels for the crossover behavior in the model has described the complexity of dynamics [105].

3.8 Conclusion

Study 1 investigated the different SARS-CoV-2 dynamics models with numerical analysis based on ODEs. GSA simplified the models, and stability analysis revealed that the models satisfied the stability criterion. The subsequent calibration experiments fitted the models to the observed viral load data across two types of hospitalized COVID-19 patients. The comparison of optimum parameter conditions in mild cases and severe cases indicated that cell-to-cell transmission would significantly correlate to the COVID-19 severity.

As a limitation, our interpretations of estimated parameters do not mean that the parameter estimation results against our intuition escape from controversy. We can only claim association because the argument is valid but unsound only from Study 1's interpretations. Given that the experimental data fluctuating in mild cases is inconsistent with the model's solution curve, which is stable in the equilibrium state, we need to improve the data fidelity.

Further investigation to surmount the limitation would include three things. First, data fidelity can be improved by finding fine-grained SARS-CoV-2 data in a longer duration. Otherwise, systematic review and meta-synthesis on the open data platform [106] could also ensure the integrated data with higher fidelity. Second, more original population dynamics models representing fluctuated data can be generated by equation discovery with a genetic algorithm or inductive bias in syntax tree mutation. Third, the relationship between cell-to-cell transmission and COVID-19 severity can be validated, and the essential factors in severe cases need to be identified for infection control or prevention. Overall, future work remained, including data integration and the above relationship's validation. Still, the experiments for modeling and simulation in this chapter would have contributed to exploring the plausible SARS-CoV-2 dynamics models based on numerical analysis and differential equations.

To the next Chapter

— Verifying SARS-CoV-2 cell-to-cell transmission hypothesis by using larger data at different scale, focusing on molecules specific to cell-to-cell transmission

Chapter 4

Gene Network Inference from Single-Cell Omics Data and Domain Knowledge for Constructing COVID-19-Specific *ICAM1*-Associated Pathways

Brief summary of this chapter

- Scientific discovery of *ICAM1*-associated pathways (putative) involved in cell-to-cell transmission currently absent from COVID-19 Disease Map
- Verifying SARS-CoV-2 cell-to-cell transmission hypothesis from microscopic scale
- DD-KB gene network inference framework integrating single-cell omics data analysis and model validation using multiple knowledge bases

Published content and contributions:

1. Journal paper (IF=4.8, h-index=107, Q2)
   Odaka M, Magnin M, Inoue K. 2023. Gene network inference from single-cell omics data and domain knowledge for constructing COVID-19-specific *ICAM1*-associated pathways. Frontiers in Genetics vol. 14 (2023). DOI: 10.3389/fgene.2023.1250545
   HAL: hal-04195846v1

2. (+α) Oral presentation in USA
   Odaka M, Magnin M, Inoue K. 2022. A Data-Driven and Knowledge-Based Approach to Inferring Temporal Gene Networks for COVID-19. *Critical Assessment of Massive Data Analysis (CAMDA)*. Madison, Wisconsin, United States.

Recap on what we do in Study 2 (Figure 4.1):

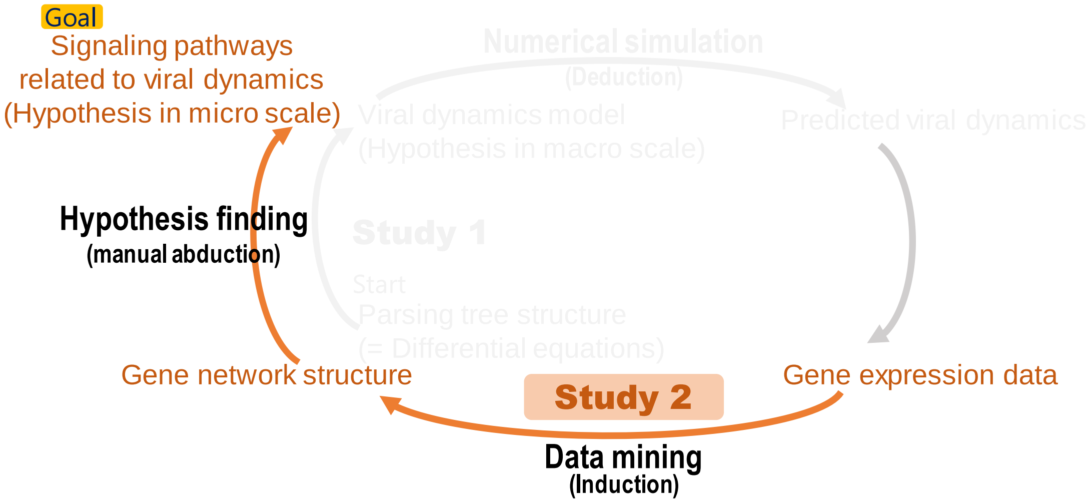


Figure 4.1. Study 2 overview. Verifying the hypothesis as a result of a system of pathways (microscopic view)

4.1 Introduction

The previous model-driven study suggested that compared to the other viral transfer manner called cell-free transmission, cell-to-cell transmission would be more associated with COVID-19 severity based on simulation of hypothesized ODEs with cell-to-cell transmission effect [107].

From this SARS-CoV-2 cell-to-cell transmission hypothesis, this chapter focuses on a noteworthy molecule responsible for the interactions between cells, such as cell-to-cell transmission, called intercellular adhesion molecule 1 (ICAM-1; also known as CD54), encoded by *ICAM1*. ICAM-1 is a transmembrane glycoprotein expressed on leukocytes, vascular endothelial cells, and respiratory epithelial cells. Its differential expression is critical for proinflammatory immune responses and viral infection. Additionally, ICAM-1 enables interactions between cells by controlling leukocyte migration, homing, and adhesion from outside to inside the cell (outside-in) and regulation from inside to outside the cell (inside-out) [108]. These functionalities make ICAM-1 an attractive drug target and a clinically essential molecule [109]. Another premise regarding ICAM-1 as an essential molecule in this study is grounded by several facts on cell-to-cell transmission. For example, cell-to-cell transmission is observed in other retroviruses, such as HIV-1 and human T cell leukemia virus type 1 (HTLV-1), whose functionalization is similar to that of SARS-CoV-2 [74]. Specifically, both these retroviruses and SARS-CoV-2 have a structurally homologous spike glycoprotein on the surface of the viral envelope that binds to a surface protein on the recipient cell during cell adhesion [73]. In HIV-1 or HTLV-1, the cell-to-cell transmission occurs after ICAM-1 triggers the peculiar pathways for cell adhesion [110] and induces the formation of the microtubule-organizing center (MTOC) and virological synapse (VS) [111]. The above arguments provide a rationale for focusing on ICAM-1 in this study and for hypothesizing the *in vivo* existence of ICAM-1 and the interactions between cells featured with ICAM-1 involved in cell-to-cell transmission in COVID-19.

In fact, there have been different *in vitro* experimental results on the expression level of ICAM-1 in SARS-CoV-2-infected cells. One study shows the time-dependent ICAM-1 expression level changes in COVID-19 patients [112]. Another study also shows that the ICAM-1 level increases in the severe phase and decreases in the convalescent phase of COVID-19 [113]. Another study shows the opposite result on the decrease of ICAM-1 after the immune cell infiltration in COVID-19 while leaving room for controversy regarding the reasons for downregulation [114]. Morevoer, in December 2021 (after publication of the Study 1), cell-to-cell transmission was observed in *in vitro* experiments of COVID-19's pathogen, SARS-CoV-2 [115].

Nevertheless, the interactions arising from ICAM-1 are not explicitly recognized as indispensable in the case of COVID-19. In particular, there is little insight into the signaling pathways surrounding ICAM-1, that is, the upstream and downstream signal cascades that occur upon the functional activation of ICAM-1 and its specific signaling molecules interacting with ICAM-1 (for simplicity, *ICAM1-associated pathways* for short). Consequently, it is significant to uncover the *ICAM1*-associated pathways to understand better the interactions between cells in the context of COVID-19.

Another substantial consequence of revealing *ICAM1*-associated pathways contributes to completing the COVID-19 Disease Map. As for ICAM-1, the pathways and even ICAM-1 are absent in the current COVID-19 Disease Map [13]. Thus, this study regards it challenging to find unknown *ICAM1*-associated pathways, expecting these pathways to include the molecules driving the cell-to-cell transmission.

Given the above, this study constructs the *ICAM1*-associated pathways based on gene networks. For inferring gene networks, we harness data and domain knowledge by extracting relationships between gene pairs from data while rectifying them with multiple knowledge bases. Such integration of data-driven and knowledge-based approaches allows us to avoid biologically meaningless interpretations based only on data characteristics. Identifying the unknown pathways with biologically meaningful interpretation will lead to a deeper understanding of the mechanisms of COVID-19.

4.2 Materials and Methods

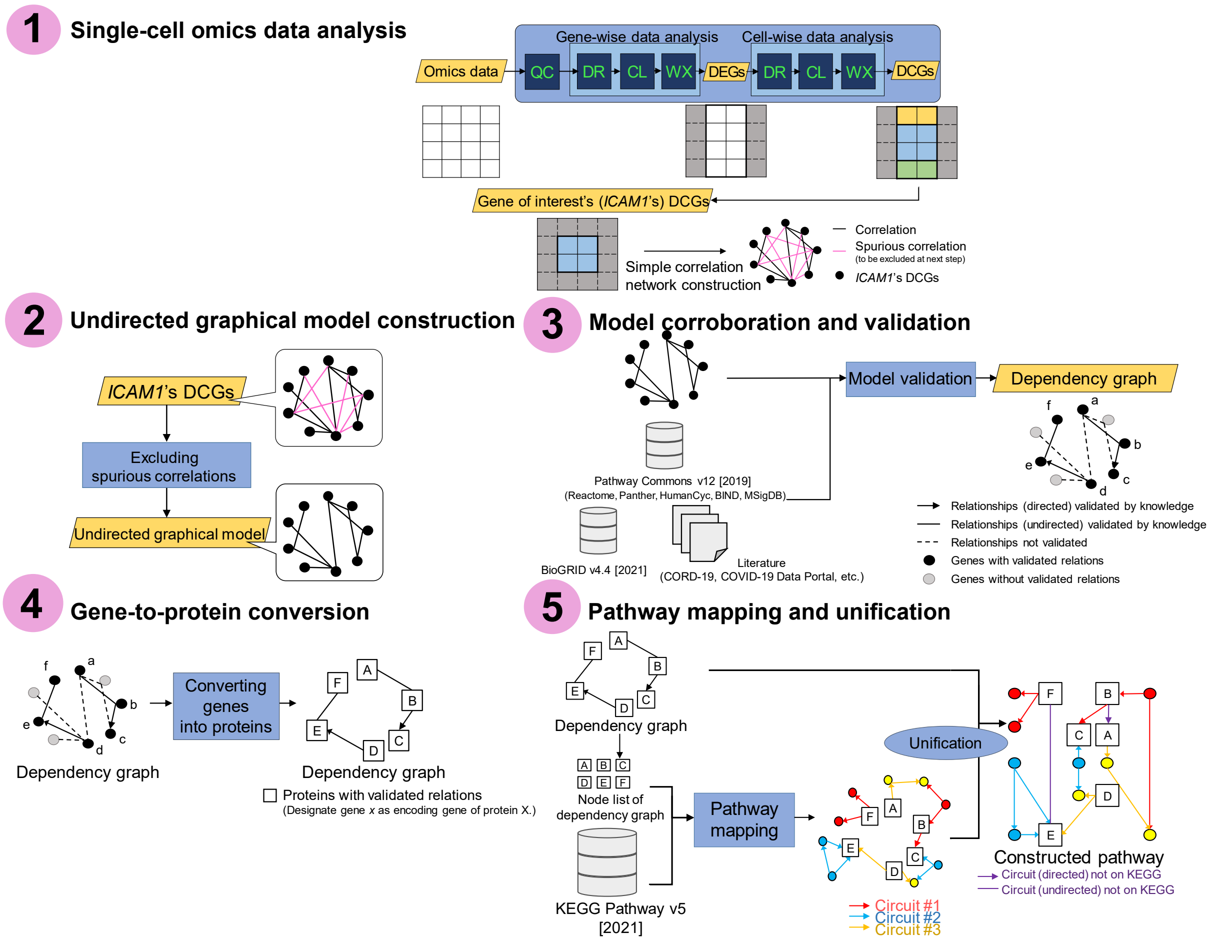
4.2.1 Overview

Figure 4.2 illustrates the framework of this study. This framework constructs the disease-specific pathways from single-cell omics data and domain knowledge via gene network inference. The framework consists of the following five steps.

1. Single-cell omics data analysis
2. Undirected graphical model construction
3. Model corroboration and validation
4. Gene-to-protein conversion
5. Pathway mapping and unification

Steps 1 & 2 are dedicated to gene network inference purely from data, and **Step 3** validates the data-driven gene network with domain knowledge. In this study, we call the rectification of data-driven objects with knowledge a DD-KB approach. In **Step 1**, we obtain the COVID-19-specific differentially expressed genes (DEGs) and a network of differentially coexpressed genes (DCGs) via single-cell omics data analysis. Here, DEGs are the genes whose expression levels differ significantly in COVID-19-positive patients and negative controls, and DCGs are coexpressed DEGs. **Step 2** removes spurious edges from the correlation networks, thereby building *de novo* undirected graphical models. In corroboration (**Step 3**), undirected graphical models are edited as dependency graphs with validated relationships.

Steps 4 & 5 are pathway construction steps. In **Step 4**, a functional annotation tool converts genes into encoded proteins. Pathway mapping and unification (**Step 5**) refine the results as the final outputs, the *ICAM1*-associated pathways. Through the framework, single-cell omics data and multiple knowledge bases are integrated, which allows the inference of gene networks containing the components absent from the current COVID-19 Disease Map.



1 Single-cell omics data analysis
Gene-wise data analysis
Cell-wise data analysis
Omics data
QC
DR
CL
WX
DEGs
DR
CL
WX
DCGs
Gene of interest's (ICAM1's) DCGs
Simple correlation network construction
Correlation
Spurious correlation (to be excluded at next step)
ICAM1's DCGs
2 Undirected graphical model construction
ICAM1's DCGs
Excluding spurious correlations
Undirected graphical model
3 Model corroboration and validation
Pathway Commons v12 [2019] (Reactome, Panther, HumanCyc, BIND, MSigDB)
BioGRID v4.4 [2021]
Literature (CORD-19, COVID-19 Data Portal, etc.)
Model validation
Dependency graph
a
b
c
d
e
f
Relationships (directed) validated by knowledge
Relationships (undirected) validated by knowledge
Relationships not validated
Genes with validated relations
Genes without validated relations
4 Gene-to-protein conversion
a
b
c
d
e
f
Dependency graph
Converting genes into proteins
A
B
C
D
E
F
Dependency graph
Proteins with validated relations (Designate gene x as encoding gene of protein X.)
5 Pathway mapping and unification
A
B
C
D
E
F
Dependency graph
A B C D E F
Node list of dependency graph
KEGG Pathway v5 [2021]
Pathway mapping
A
B
C
D
E
F
Circuit #1
Circuit #2
Circuit #3
Unification
F
B
C
A
D
E
Constructed pathway
Circuit (directed) not on KEGG
Circuit (undirected) not on KEGG

Figure 4.2. Schematic representation of the framework. Step 1: Single-cell omics data analysis. **Step 2**: Undirected graphical model construction. **Step 3**: Model corroboration and validation. **Step 4**: Gene-to-Protein conversion. **Step 5**: Pathway mapping and unification. The *circuits* are subpathways transmitting a signal from input receptor nodes to output effector nodes, where the nodes mostly represent proteins such as metabolic enzymes. **QC**: Quality Control; **DR**: Dimensionality Reduction; **CL**: Clustering; **WX**: Wilcoxon rank-sum test; **DEGs**: Differentially Expressed Genes; **DCGs**: Differentially Coexpressed Genes. See also DOI: 10.6084/m9.figshare.18095717.

4.2.2 Gene Network Inference and Pathway Construction

In this subsection, explanations for each step of the framework are provided.

Step 1: Single-Cell Omics Data Analysis

Single-cell omics data analysis adopts a combination of the standard methods defined as three subroutines, including dimensionality reduction, clustering, and Wilcoxon rank-sum test, for each gene pair and each cell pair [116]. This step can extract COVID-19-specific DEGs and *ICAM1*-associated DCGs from the omics data.

Standard protocol for extracting differentially expressed genes In single-cell omics data analysis, there exists a "standard" protocol that is used in multiple tutorials of single-cell gene expression data tools, such as R's package Seurat [117], Python's library Scanpy [118], squidpy [119], or MUON [120]. Specifically, a typical workflow consists of Quality Control (QC) to select cells for further analysis, dimensionality reduction, embedding and clustering the neighborhood graph, and finding differentially expressed features (cluster biomarkers that can assign cell type specific to clusters).

Dimensionality Reduction Dimensionality reduction is executed after the imputation of zeros representing either technically missing data or biologically absent genes within a matrix of single-cell omics data [121]. To reduce dimensionality, we employ two methods: principal component analysis (PCA) and uniform manifold approximation and projection (UMAP). These methods detect possible batch effects and embed the matrix in the latent space. By computing 50 PCA coordinates on the sparse matrix for mean centering [122], eigenvalues, and eigenvectors with the singular value decomposition solver ARPACK (ARnoldi PACKage) [123], PCA reduces the dimension to 100 by a Gaussian kernel. Given the 50 decomposed coordinates, the connectivities (weighted adjacency matrix) of the k-nearest neighborhood graph are computed and thresholded at the closest neighbors defined for data points of the manifold in Euclidean space. Following PCA, UMAP [124] projects the data points onto the two-dimensional latent space.

Clustering Afterward, clustering is enforced to classify data points in the latent space into subgroups by similarity measurements and filtering out the genes unassociated with the gene of interest. The Louvain algorithm, a greedy optimization of local modularity to detect the groups [125], is applied for clustering. Clustering allows to obtain the data points of subgroups with similar gene expression profiles.

Biclustering is a clustering method that clusters rows and columns to find correlated feature subsets. We conduct biclustering of omics data referring to the above-mentioned standard protocol and Single Cell Representation Learning (SCRL), which combines a bipartite cell-context gene network and a bipartite gene-context gene network to learn the low-dimensional vector representations for cells, genes and context-genes [116].

Wilcoxon Rank-Sum Test The Wilcoxon rank-sum test is conducted to sort the data points and pick up the top 200 data points within a cluster. This test compares the signal values between each subgroup and the union of the other subgroups with the Benjamini–Hochberg method for adjusting the false discovery rate and correcting the *p-value* [126]. The comparison allows us to detect significant differences in expression levels between COVID-19-positive and COVID-19-negative patients and rank the genes characteristic of each subgroup.

The above analysis, including dimensionality reduction, clustering, and Wilcoxon rank-sum test, is conducted for each gene pair and each cell pair. Genewise analysis filters the DEGs to distinguish those whose gene expression levels are correlated. Here, given that functionally related genes are coexpressed in the same clusters, the identified gene clusters can be considered to include the genes with significant differences in expression levels from the negative control, and the genes within the same cluster share a common differential expression pattern [127]. Likewise, cellwise analysis filters the DCGs to classify all the cells into cell clusters based on the correlation coefficients as similarity measurements for embedding, which means that the genes within the same cell cluster are more strongly correlated with each other than with the genes in other clusters. Constraining the DCGs with the gene of interest, *ICAM1*, provides a subset of DCGs correlated with *ICAM1*.

Step 2: Undirected Graphical Model Construction

Next, we infer gene networks from *ICAM1*'s DCGs obtained by the single-cell omics data analysis. The gene networks are inferred as undirected graphical models with a partial correlation method, displaying *de novo*-produced direct linear associations [128]. Considering that correlated gene pairs are coexpressed with similar functions, designating any gene pair as nodes and the correlation coefficient of gene expression levels as edges forms the simple correlation networks of *ICAM1*'s DCGs. Calculating the second-order partial correlation coefficients between all gene pairs and removing the edges of the gene pairs with almost zero partial correlation coefficients for any combination yield undirected graphical models without spurious correlations [129].

The equations for the zero-order, first-order, and second-order partial correlations are shown in Eqs. (4.1), (4.2), and (4.3).

$$\text{Zero-order correlation}: \quad r_{xy} = \frac{\operatorname{cov}(xy)}{\sqrt{\operatorname{var}(x)\operatorname{var}(y)}} \tag{4.1}$$

$$\text{First-order partial correlation}: \quad r_{xy,z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{\left(1 - r_{xz}^2\right)\left(1 - r_{yz}^2\right)}} \tag{4.2}$$

$$\text{Second-order partial correlation}: \quad r_{xy,zq} = \frac{r_{xy,z} - r_{xq,z}r_{yq,z}}{\sqrt{\left(1 - r_{xq,z}^2\right)\left(1 - r_{yq,z}^2\right)}} \tag{4.3}$$

The random variables denoted by x, y, z, and q represent the gene names. r_{xy} is Pearson's correlation coefficient between the gene expression-level vector running over all cells of any gene x and that of any gene y. The simple correlation network starts by connecting x and y if and only if $r_{xy} \neq 0$. Undirected graphical modeling removes the linear effect of all second-order partial correlation coefficients $r_{xy,zq}$ between two variables (x, y) conditional on all other variables.

The edge is weighted as $(0.5 + 0.5 \cdot r_{xy,zq})^{12}$ to follow the scale-free law, which typically used in Weighted gene correlation (coexpression) network analysis (WGCNA) [130]. WGCNA is a widely used data mining technique for inferring biological networks based on correlations between gene pairs [131]. In WGCNA, the absolute value of the correlation is regarded as a coexpression similarity measure, which is defined as the following expression:

$$a_{i,j} = \{0.5 + 0.5PartialCorr(x_i, x_j)\}^{\beta}$$

, where the power β is a soft thresholding parameter [132]. While unweighted coexpression network set the hard thresholding value to determine binary connection between gene pairs, WGCNA enables to make more realistic biological networks, leaving continuous-valued edges [132]. Additionally, community detection of such networks can be an alternative method to clustering in DEG analysis [133]. However, the major drawback of WGCNA is that once two objects are clustered together, it cannot be reversed [134]. We adopt our two-step biclustering, i.e., extracting disease-specific DEGs and then DCGs from DEGs, to avoid extracting DCGs from DEGs unrelated to COVID-19.

Step 3: Model Corroboration and Validation

Until the previous data analysis, gene networks consisting of the *ICAM1* gene of interest are inferred without guaranteeing the validity of each edge. Namely, possible errors within data, such as noise, could result in nodes or edges with no biological meaning. Hence, the models require corroboration and validation with heuristics based on domain knowledge. To corroborate and validate each relationship of gene networks, we query multiple knowledge bases, including Pathway Commons Web Service 12 [135], BioGRID REST Web Service [136], and STRING version 11.5 [137]. Pathway Commons' application programming interface (API) provides access to the significant pathway databases Reactome, Panther, HumanCyc, BIND, and MSigDB. BioGRID is used as a complementary source of the latest knowledge since Pathway Commons is not up-to-date [138]. HumanCyc is used because it has richer information on biochemical reactions and regulatory relationships than the KEGG pathways alone [139] and enables the obtained model to include more information than a subset of the KEGG pathways. STRING is used for annotations of functional or physical interactions between the queried proteins. Fetching relations between gene pairs in the simple interaction format (SIF) through these knowledge bases enables us to convert a subset of undirected edges to directed edges, thereby editing undirected graphical models as dependency graphs.

The subsequent two steps are dedicated to the pathway construction by overlaying the inferred DD-KB gene networks onto the KEGG pathways.

Step 4: Gene-to-Protein Conversion

There exists a gap between the gene network and the KEGG pathways because the nodes of the DD-KB gene networks are DCGs (genes), while the nodes of the KEGG pathways are primarily proteins. Therefore, this gap needs to be filled before overlaying the DD-KB gene networks onto the KEGG pathways. The DAVID functional annotation tools 6.8 [140, 141] allows us to fill the gap by converting gene symbols into Entrez IDs. We apply the DAVID tools to the node lists of the DD-KB gene networks to give the corresponding protein attributes for each DCG.

Step 5: Pathway Mapping and Unification

In order to examine what types of pathways are activated, we conduct pathway enrichment analysis by mapping the protein node lists and edge lists of the DD-KB gene networks onto the KEGG pathways. In particular, the DD-KB gene networks and the KEGG pathways in KEGG Markup Language (KGML) format are unified by Cytoscape 3.9.0 [142], resulting in the final COVID-19-specific *ICAM1*-associated pathways, visualized in yFiles Hierarchical Layout.

4.2.3 Application

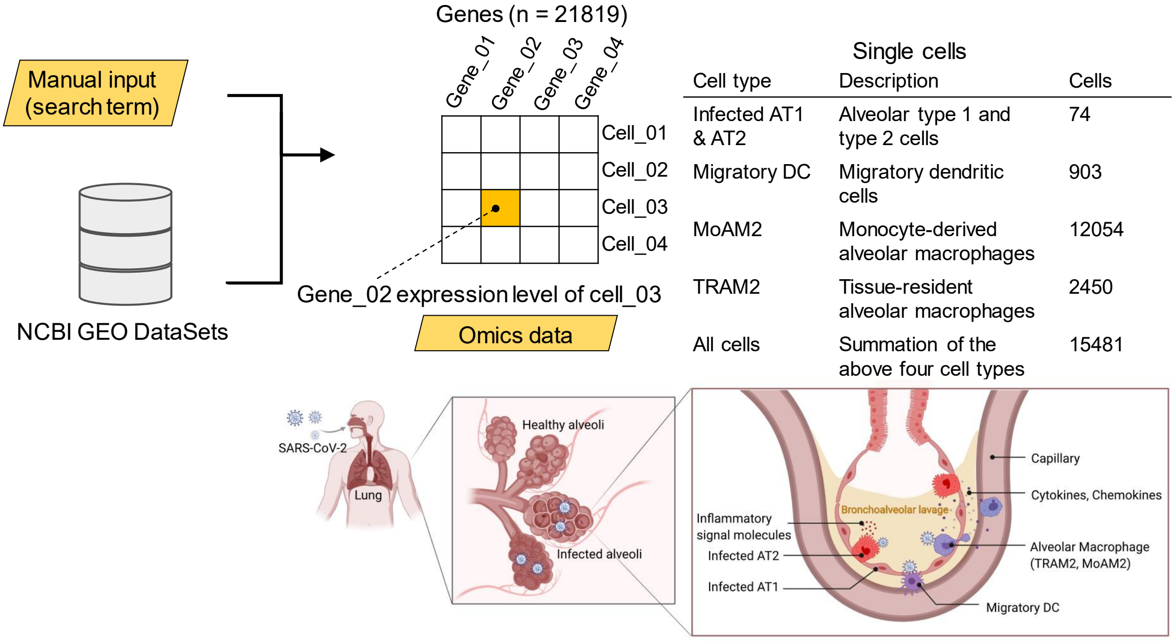
We applied the above framework to the two COVID-19 datasets for comparing the *ICAM1*-associated pathways between different locations where *ICAM1* is expressed (case study 1) and between different time points starting from hospitalization (case study 2). The machine configuration was as follows: Python 3.7, GPU Tesla V100-SXM2-16GB, and 51.01 GB of RAM.

Case Study 1: Comparison of *ICAM1*-Associated Pathways Between Different Cell Types

Inputting the search term `((COVID-19 OR SARS-CoV-2) AND gse[entry type]) AND "Homo sapiens" AND h5ad` to the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) DataSet [143] provided the data. In case study 1, the data included the gene expression profiles of bronchoalveolar lavage fluid samples isolated from 10 patients with severe COVID-19 and two negative controls via high-throughput single-cell RNA sequencing [144]. Especially, we used the data of messenger ribonucleic acid (mRNA) expression levels from four antigen-presenting cell types in which virus particles were detectable. The single-cell omics data analysis in the original paper had already annotated the cell types, cell subpopulation partitioning the cell heterogeneity into nonoverlapping classes, for clusters according to the reference biomarkers present in the cluster [145]. The cell types included infected alveolar type 1 and 2 cells (infected AT1 & AT2), migratory dendritic cells (migratory DCs), tissue-resident alveolar macrophages type 2 (TRAM2), and monocyte-derived alveolar macrophages type 2 (MoAM2), as well as the summation of these four cell types at the level of full single-cell resolution (Figure 4.3).

Case Study 2: Comparison of *ICAM1*-Associated Pathways Between Different Time Points

Likewise, another transcriptome omics data were GSE180578 [146] fetched from NCBI GEO DataSet. The omics data were 86 samples obtained by single-cell RNA sequencing, including peripheral blood from COVID-19 patients or negative control at the intensive care unit (ICU) of the University of Pittsburgh Medical Center. These samples included three-time points (day 1, day 5, and day 10 post-enrollment in the ICU). The cell counts and gene counts were (34970, 2000), (23616, 2000), and (32105, 2000), respectively. The cell types annotated with biomarker genes are identified by single-cell omics data analysis in the original paper. The cell types include canonical immune lineages, such as B cells, CD1c+ DCs, CD34+ cells, CD4+ T cells, CD8+ T cells, NK cells, Monocytes, pDCs, and plasmablasts [146].



Single cells

Cell type	Description	Cells
Infected AT1 & AT2	Alveolar type 1 and type 2 cells	74
Migratory DC	Migratory dendritic cells	903
MoAM2	Monocyte-derived alveolar macrophages	12054
TRAM2	Tissue-resident alveolar macrophages	2450
All cells	Summation of the above four cell types	15481

Figure 4.3. The cell types for which data were collected. Pulmonary tissue illustrations: Created with BioRender.com. See also DOI: 10.6084/m9.figshare.18095714.

Additional Case Studies

Not limited to *ICAM1*, our framework would be re-usable in another context for mixing quantitative data and domain knowledge into building models capturing pathways. To allege the generality of the framework, in case study 2, we also applied the framework to other genes related to the interaction between cells, including *ACTB* and *C15orf48*. *ACTB* encodes β-actin, a non-muscle cytoskeletal filament implicated in cell motility, structure and integrity. *C15orf48* encodes modulator of cytochrome C oxidase during inflammation (MOCCI), which inhibits inflammatory response, as indicated by the downregulation of proinflammatory biomarkers, such as NF-κB, ICAM-1 and VCAM-1 [147]. MOCCI is also likely to cooperate with ICAM-1 and β-actin in COVID-19 for intercellular adhesion.

To grasp the time-dependent change of activated pathways, we drew the parallel coordinates for each matched KEGG and Reactome pathway, ranking in descending order of the gene counts in the pathways. Here, the Reactome pathway is an alternative pathway database that includes information more about pathological names than about specific disease names that the KEGG pathway tends to label.

4.2.4 Quality Control

Before succeeding to the further steps, the single-cell omics data underwent quality control. Quality control included filtering, scaling, and normalization by Scanpy version 1.8.2 [118]. Given cell quality, we regarded the cells with overexpressed mitochondrial RNA per data count tagged by a unique molecular identifier (UMI) [148] as dead or broken cells. Similarly, cells with many genes per data count tagged by UMI were identified as doublets. Subsequently, the genes detected in fewer than three cells were filtered out to ensure gene quality. The count data were scaled with regression on total UMI counts and normalization per feature based on standard deviation. Normalization of the gene expression data adjusted for the RNA composition bias and allowed a comparison of the values among the cells. Finally, log-transformation prepared the data for calculating the log-fold changes reflecting the gene expression difference.

4.2.5 Data Availability Statement

We display the results in case study 1 only. But any reader who wants to check the figures or datasets in case study 2 in Study 2 can refer to DOI: `10.6084/m9.figshare.23590755`. The panel labels (a) - (e) in the figures stand for infected alveolar type 1 and 2 cells, migratory dendritic cells, tissue-resident alveolar macrophages, monocyte-derived alveolar macrophages, and a summation of all cells. Some readers may find it difficult to read some figures because of the small font size due to the paper size limitation. In that case, they can access the *figshare* link in the caption of each figure to see it in a larger view.

The single-cell omics dataset analyzed for this study can be found in the NCBI GEO DataSet `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE155249` (case study 1) and `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE180578` (case study 2). The source codes to reproduce the results in this study are available in *figshare* repository `10.6084/m9.figshare.23590713`.

4.3 Results

4.3.1 Case Study 1: *ICAM1*-Associated Pathways at Different Locations (Cell Types)

Quality Control

The data initially contained 77,650 cells × 24,714 genes. Removing the cells with a high proportion of mitochondrial RNA resulted in 15,220 cells. After filtering, the whole dataset contained 68,734 cells × 24,001 genes. Among this dataset, we used the data of four cell types, 21,819 genes in the SARS-CoV-2-infected 15,481 cells (cells with SARS-CoV-2 transcripts detected). The doublet discrimination provided 14,723 cells. The filtering processes excluded 8,916 cells with more than 5,000 expressed genes, 700 genes detected in fewer than three cells, mitochondrially encoded genes, and cells with a low percentage ($< 10\%$) of mitochondrial genes, leaving 17,644 genes. Cells with less than one gene count were filtered out, leaving 9,050 cells. The quality control ultimately yielded log-transformed normalized gene expression data for 9,050 cells × 17,644 genes.

Single-Cell Omics Data Analysis

The genewise analysis extracted the 18 gene clusters with differential expression patterns specific to COVID-19. Excluding the duplicated genes extracted 1,434 DEGs in 9,050 single cells. The results of genewise clustering, heatmap of DEGs, and rank-sum test are shown in Figures 4.4, 4.5, and 4.6, respectively.

The cellwise analysis yielded 11 clusters based on the correlations between gene pairs in the embedded space and distinguished the DEGs whose gene expression levels were correlated. One of the 11 clusters included *ICAM1*, and this cluster was made of 178 *ICAM1*'s DCGs. The results of cellwise clustering, heatmap of DCGs, and rank-sum test are shown in Figures 4.7, 4.8, and 4.9, respectively.

Genewise analysis extracted the gene clusters with differential expression patterns specific to COVID-19. Each cluster includes genes with significant differences in expression from the negative control. Genes within the same cluster share a common differential expression pattern. Signal magnitudes are logarithms of fold changes with cluster identifiers. Supplementary Table 1 is a hash table of DEGs, including the cluster number, gene name string, log fold change, and p-values. See also doi: 10.6084/m9.figshare.17273156.

Cellwise analysis filtered the DCGs via three subroutines to classify all the cells into cell

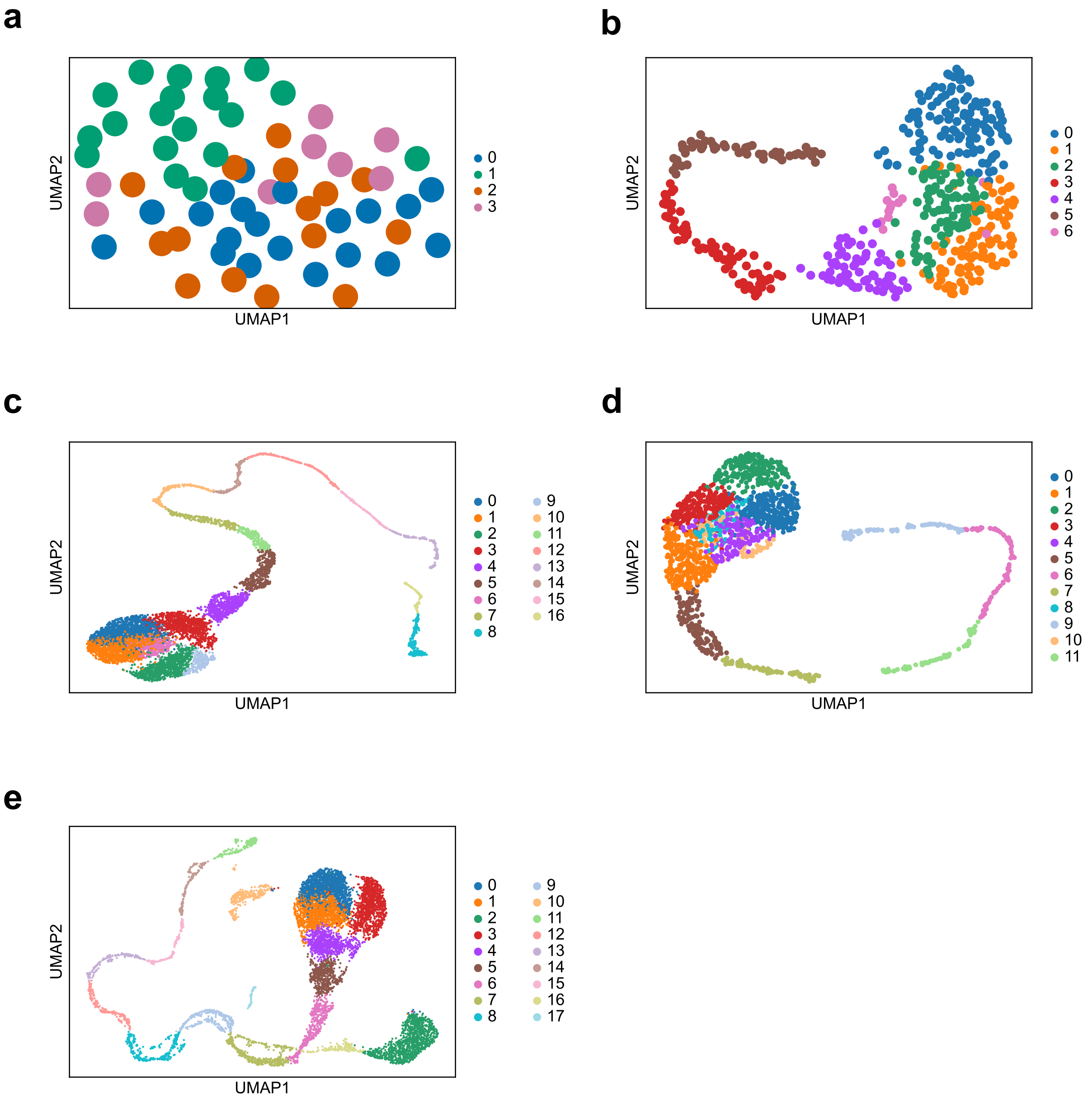


Figure 4.4. Genewise clustering result. Genewise clustered data points within an embedded latent space. This procedure extracts genes within a specific cluster, including a gene of interest. Euclidean distance measures the distance between clusters. Calculating the log fold change (the magnitude of differential expression) yielded a list of the differential expression genes (DEGs), whose expression levels significantly increased or decreased. Each cluster is assigned a unique cluster number with a different color. See also doi: 10.6084/m9.figshare.17263814.

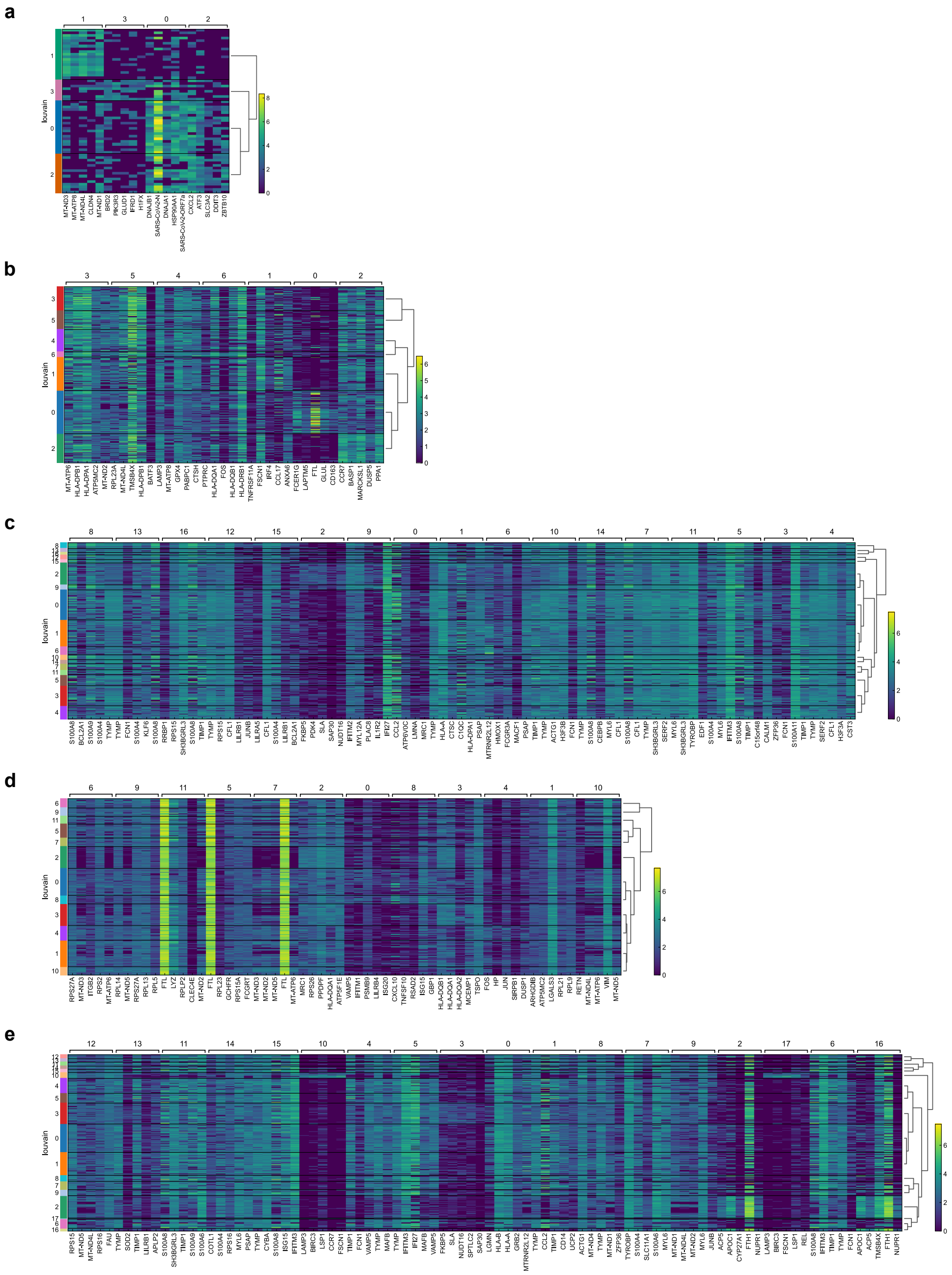


Figure 4.5. The heatmap of differentially expressed genes. The top five differentially expressed genes (DEGs) for each cluster. Each cluster with a number given by the Louvain algorithm enumerates the corresponding gene names. A brighter color means a higher tendency for differential expression. See also doi: 10.6084/m9.figshare.17263841.

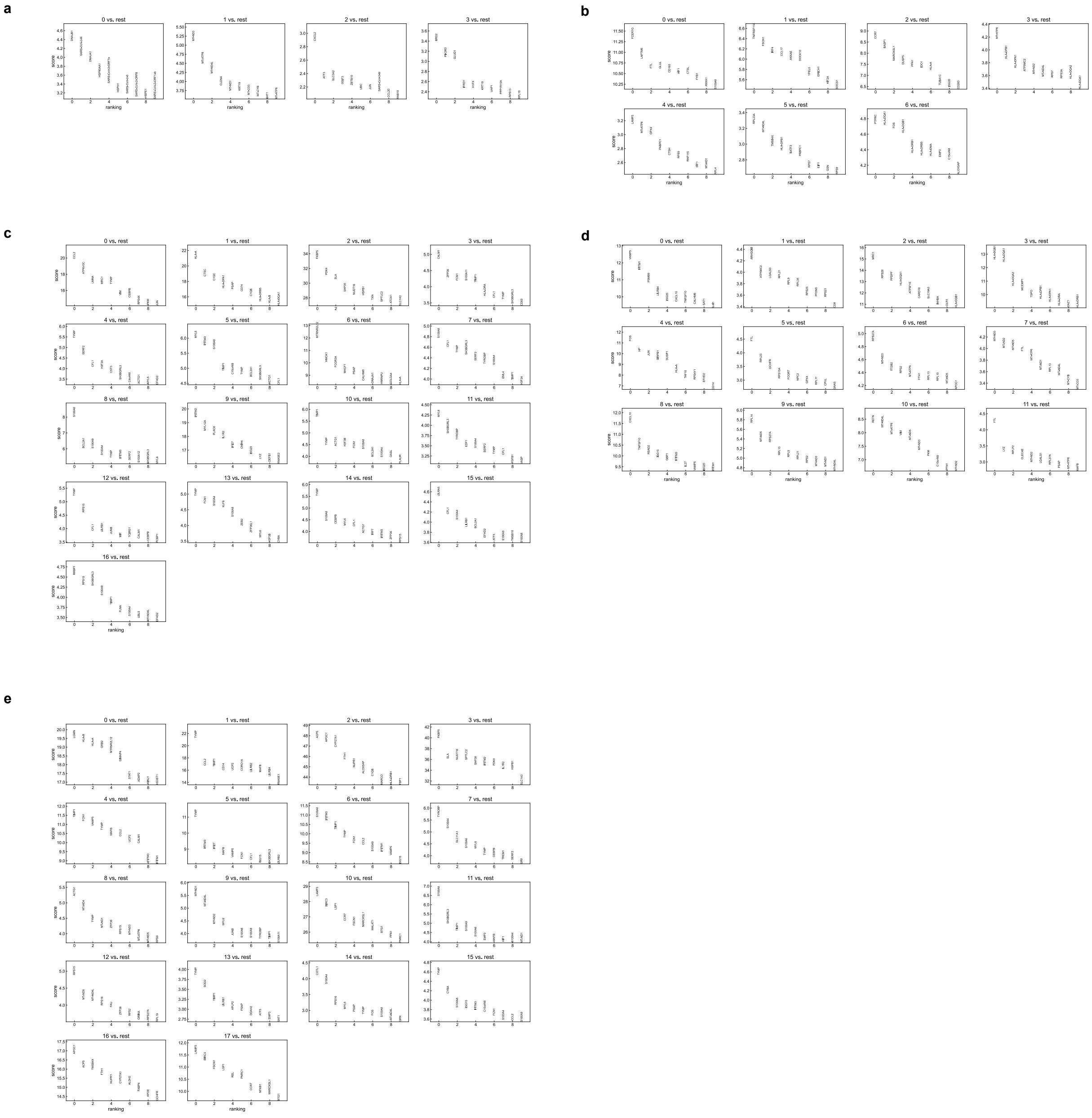


Figure 4.6. Rank-sum test result for differentially expressed genes. The Wilcoxon rank-sum test determined the clusters' rank-ordered differentially expressed genes (DEGs). Each cluster of each subfigure contains the top 10 genes' names in descending order from left to right, with rank-sum scores assigned to the vertical axes. See also doi: 10.6084/m9.figshare.17263877.

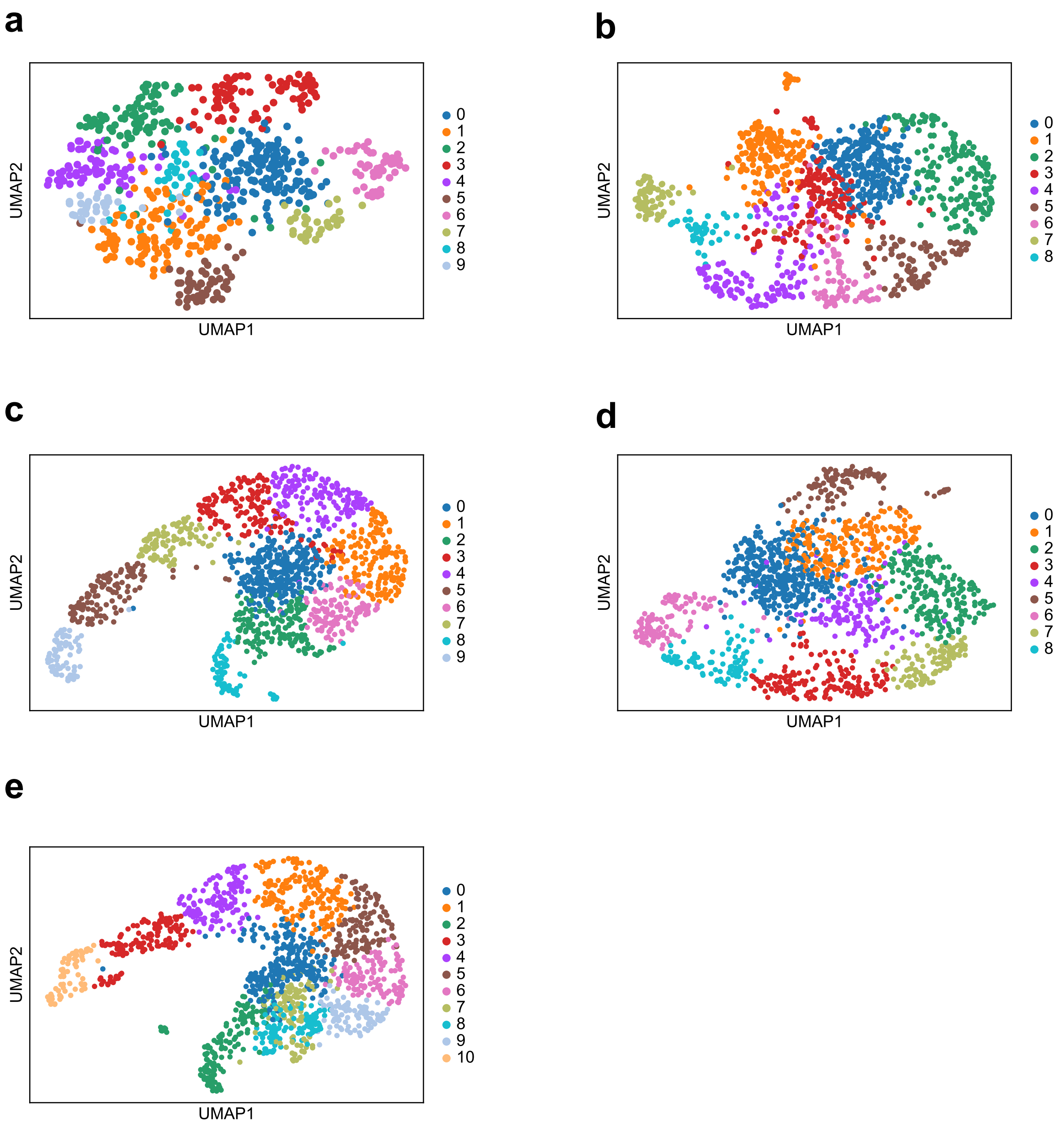


Figure 4.7. Cellwise clustering result. Cellwise clustered data points within an embedded latent space. Pearson's correlation coefficient measures the distance between clusters. Calculating the log fold change (the magnitude of differential expression) filters a list of the differential coexpression genes (DCGs) from DEGs. Each cluster is assigned a unique cluster number with a different color. See also doi: 10.6084/m9.figshare.17263889.

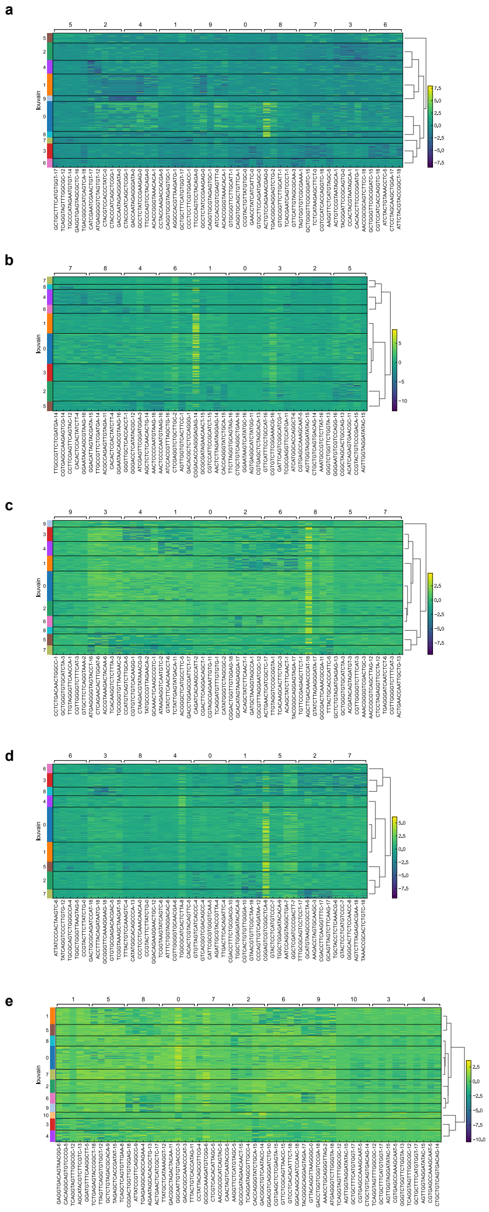


Figure 4.8. The heatmap of differentially coexpressed genes. Top five cell IDs for each cell cluster with a similar gene coexpression pattern. Each cluster with a number given by the Louvain algorithm enumerates the corresponding cell IDs. A brighter color means a higher tendency for differential coexpression. See also doi: 10.6084/m9.figshare.17263892.

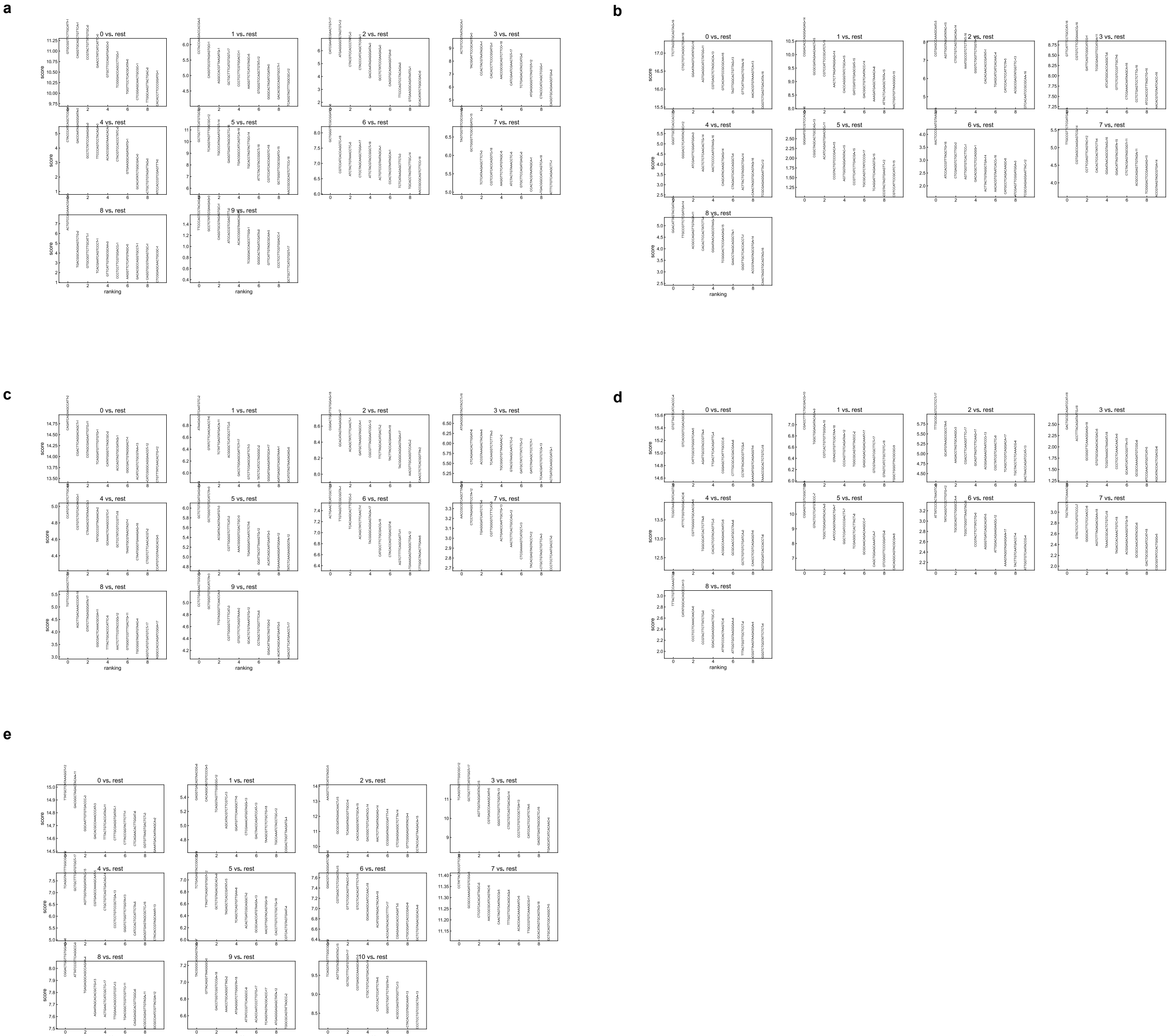


Figure 4.9. Rank-sum test result for differentially coexpressed genes. Differentially coexpressed genes (DCGs) for each cluster. Each cluster of each subfigure contains the top 10 cells' IDs in descending order from left to right, with rank-sum scores assigned to the vertical axes. See also doi: 10.6084/m9.figshare.17263898.

clusters based on the correlation coefficients as similarity measurements. Genes within the same cell cluster are more strongly correlated than those in other clusters. Constraining the DCGs with the gene of interest, *ICAM1*, provided a subset of DCGs correlated with *ICAM1*. Supplementary Table 2 is a hash table of DCGs, including gene expression levels for each single cell. The *p*-values of *ICAM1* expression variation and the computation times (sec.) for each cell type were as follows: p = 0.250, time = 51.6 (Infected AT1 & AT2), p = 2.50E-4, time = 26.1 (Migratory DC), p = 2.57E-12, time = 46.6 (TRAM2), p = 7.55E-2, time = 109.0 (MoAM2), and p = 0.241, time = 178.9 (Summation). See also doi: 10.6084/m9.figshare.17273177.

Undirected Graphical Model Construction

Removal of spurious correlations yielded undirected graphical models (Table 4.1). See also the finally obtained undirected graphical models in Figure 4.10.

Table 4.1. Spurious correlation removal in case study 1

Cell types	Nodes	Edges (full)	Edges (excluded)	Edges (output)
Infected AT1 & AT2	116	6670 (100 %)	6296 (94%)	374 (6%)
Migratory DCs	248	30628 (100 %)	28695 (94%)	1933 (6%)
TRAM2	150	11175 (100 %)	8690 (78%)	2485 (22%)
MoAM2	152	11476 (100 %)	7819 (68%)	3657 (32%)
Summation	179	15931 (100 %)	12529 (79%)	3402 (21%)

The table depicts the number of nodes, the number of edges in the simple correlation network (full model), the number of spurious edges removed by calculating the second-order partial correlation coefficients, and the number of ultimately left edges.

Model Corroboration and Validation

Dependency graphs are shown in Figure 4.11.

The entire list of relationships of dependency graphs with knowledge bases used for model validation can be found in Supplementary Table 3. Edge weights include unweighted, weighted unsigned, and weighted signed correlation coefficients. Gray-filled cells in the table are gene pairs that could not be validated based on background knowledge. Therefore, the relations and data sources are missing. The appended tabs marked with (2) contain the tables that exclude unvalidated gene pairs and combine gene pairs with overlapping data sources, i.e., gene pairs contained in two or more different background knowledge. See also doi: 10.6084/m9.figshare.17273120.

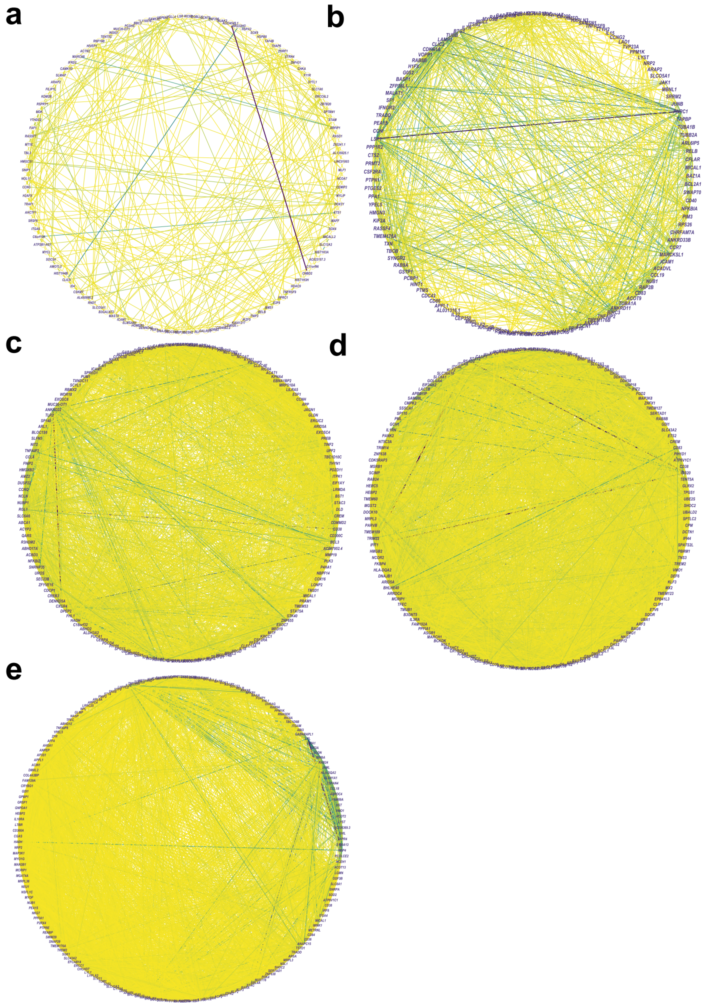


Figure 4.10. Undirected graphical models. (a) Infected alveolar type 1 and 2 cells; (b) Migratory dendritic cells; (c) Tissue-resident alveolar macrophages; (d) Monocyte-derived alveolar macrophages; (e) All cells. Nodes, edges, and weights are DCGs, relationships between DCGs, and second-order partial correlation coefficients. See also doi: 10.6084/m9.figshare.17261825.

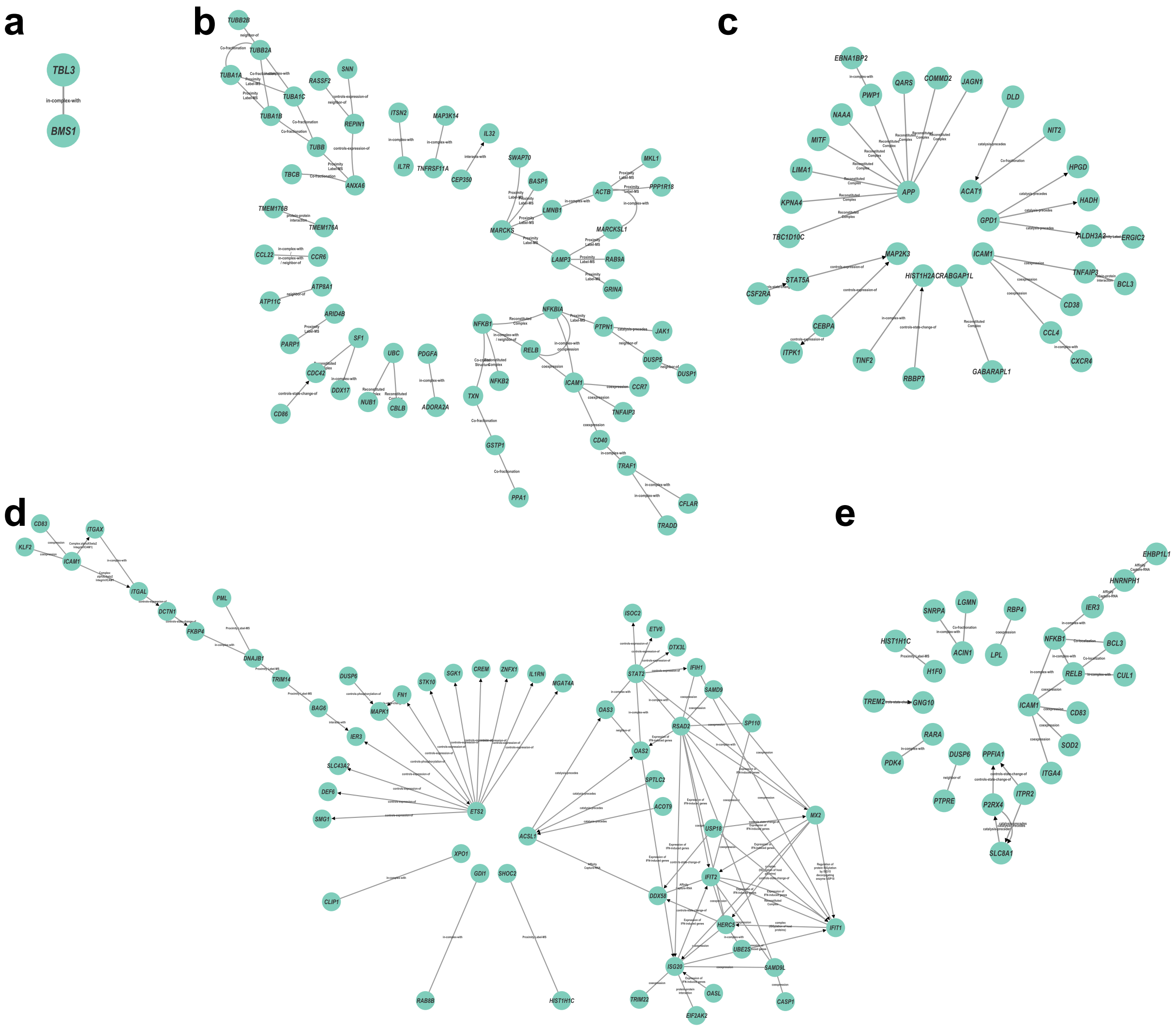


Figure 4.11. Dependency graphs. (a) Infected alveolar type 1 and 2 cells; (b) Migratory dendritic cells; (c) Tissue-resident alveolar macrophages; (d) Monocyte-derived alveolar macrophages; (e) All cells. Nodes are DCGs, and edges are relationships between DCGs with annotated function names. These function names are validated from knowledge bases. The directed edge is given when the edge is a regulatory relationship, such as activation or inhibition. Undirected edges represent coexpression or other functions without direction. See also doi: 10.6084/m9.figshare.17261780.

Gene-to-Protein Conversion

The nodes in the dependency graphs were annotated with protein names, which helped us map the nodes onto the KEGG pathways in the next step.

Pathway Mapping and Unification

Pathway mapping discovered which subpathways within existing signaling pathways reflect the activity of a group of genes varying and coexpressed in a disease-specific manner in the observed gene expression data. Table 4.2 shows the typical pathways selected from the mapping results. See also a complete list of mapping results (Figure 4.12).

Table 4.2. Mapping results of the dependency graphs for each cell type onto the KEGG pathways

Cell types	Scores	Matched KEGG pathways
Infected AT1 & AT2	0 genes (no match)	
Migratory DCs	40 genes (63.5% match)	NF-κB signaling pathway (hsa04064) (12), HTLV-1 infection (hsa05166) (9)
TRAM2	23 genes (65.7% match)	Influenza A (hsa05164) (13), HTLV-1 infection (hsa05166) (5)
MoAM2	31 genes (54.4% match)	NF-κB signaling pathway (hsa04064) (3)
Summation	18 genes (64.3% match)	TNF signaling pathway (hsa04668) (3), NF-κB signaling pathway (hsa04064) (3)

Scores count the “matched” genes on the dependency graphs, whose encoding proteins are on any of the KEGG pathways and their proportion to total gene counts. Matched KEGG pathways exemplify how many matched genes are included in a specific pathway. For example, if gene *x*’s encoded protein *X* is on KEGG pathways A and B, one is added to the score, and both A and B are represented.

The final COVID-19-specific *ICAM1*-associated pathways for each cell type are shown in Figure 4.13. Although pathway mapping was performed by converting gene symbols to protein IDs before mapping, the nodes of pathways in the figure are assigned only gene symbol names for space limitation. For the *ICAM1*-associated pathway for infected AT1 & AT2 (Fig 4.13a), there are no hits among the KEGG pathways, which is attributed to only one pair of validated *ICAM1*-associated DCGs remained.

The characteristics common to the obtained pathways and the characteristics of those path-

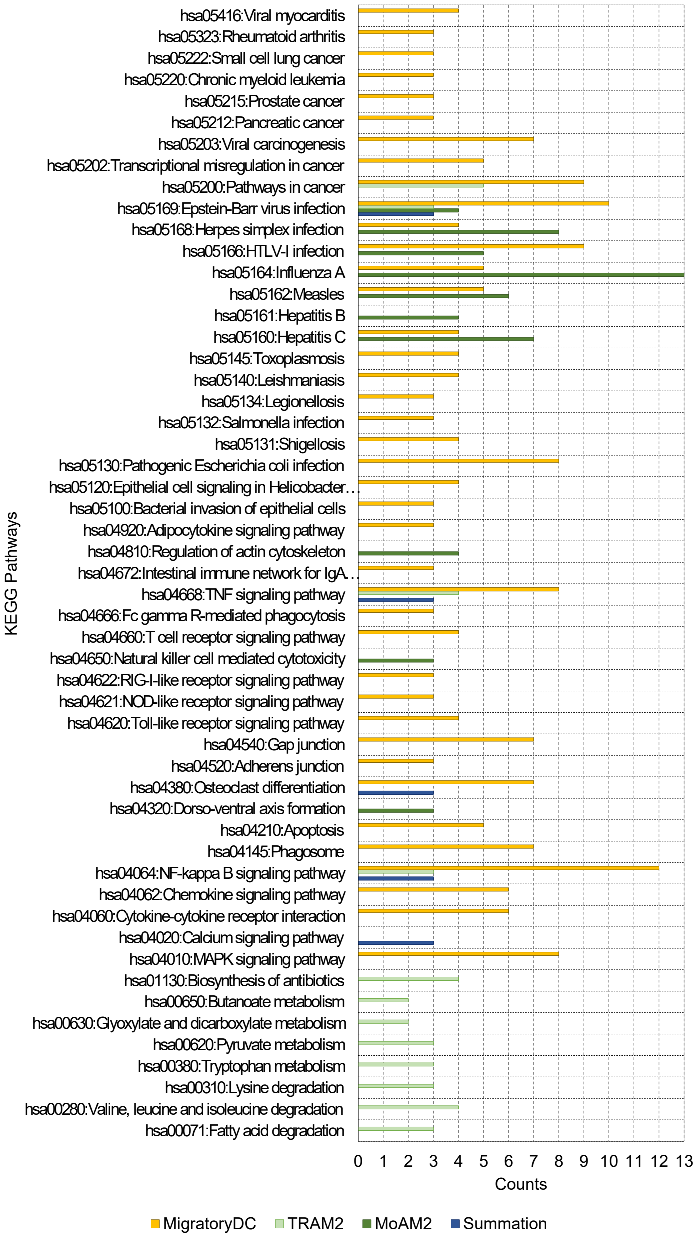


Figure 4.12. Pathway mapping result. Querying the gene lists of dependency graphs yielded the activated Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways for each cell type. The bars colored yellow, lime green, olive, and navy blue stand for the count of genes that appeared on the KEGG pathways for migratory dendritic cells (DCs), tissue-resident alveolar macrophages (TRAM2), monocyte-derived alveolar macrophages (MoAM2), and all cells. See also doi: 10.6084/m9.figshare.17263907.

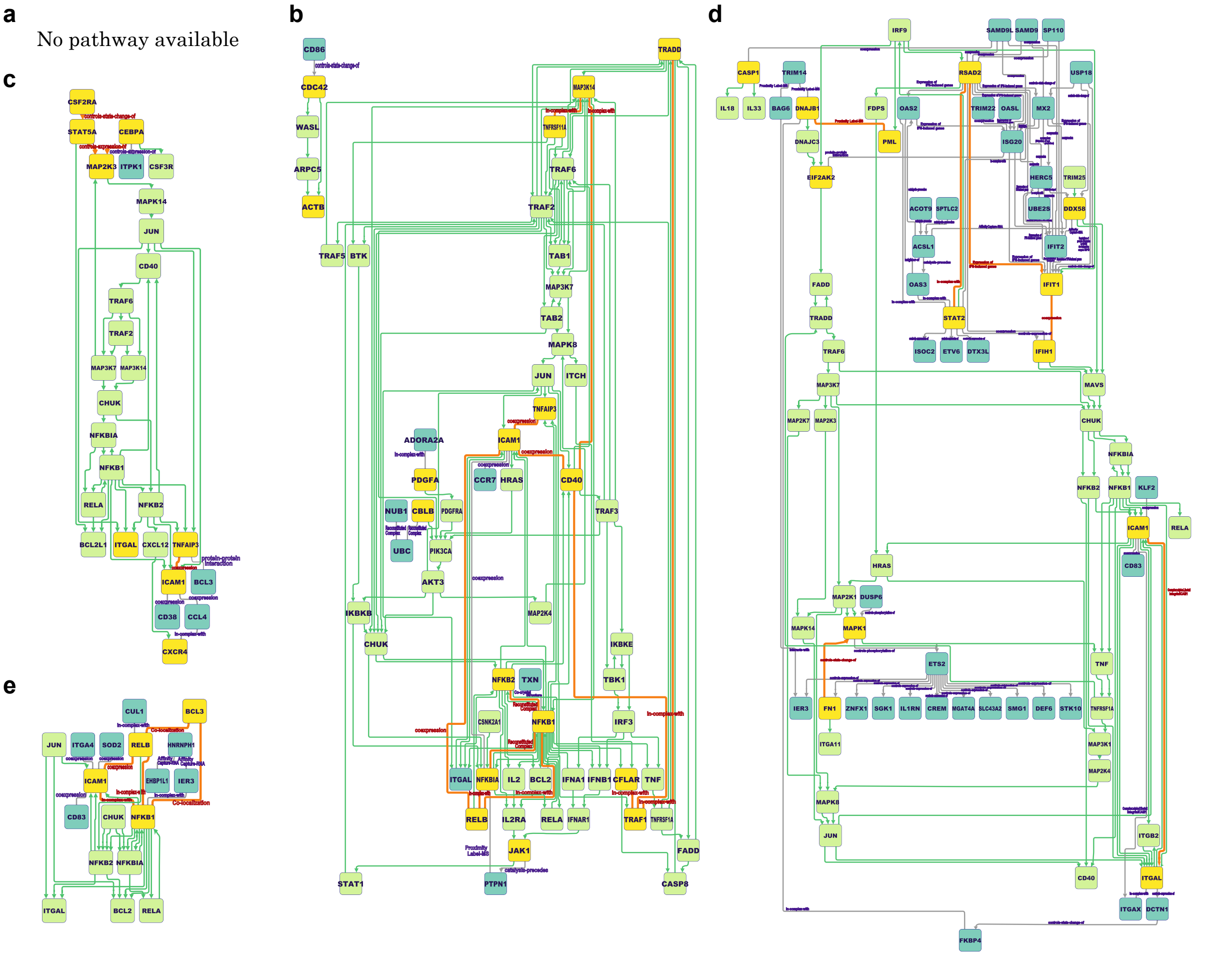


Figure 4.13. *ICAM1*-associated pathways at different locations (cell types). a: No pathway available (Infected alveolar type 1 and 2 cells); **b**: NF-κB/non-canonical NF-κB/Integrin pathway putative (Migratory dendritic cells); **c**: NF-κB/Integrin pathway putative (Tissue-resident alveolar macrophages); **d**: NF-κB/Integrin pathway putative (Monocyte-derived alveolar macrophages); **e**: TNF/NF-κB/non-canonical NF-κB/Integrin pathway putative (Summation). The rectangular nodes colored blue, yellow, and lime green reflect the proteins only on the dependency graphs, the proteins common to both the dependency graphs and the KEGG pathways, and the proteins only on the KEGG pathways, respectively. Gray lines are the directed or undirected edges only on the dependency graphs. Orange lines represent the directed or undirected edges between yellow nodes on the dependency graphs. Green lines are the directed edges only on the KEGG pathways. Orange edges do not have direction if the KEGG pathways indirectly connect its yellow node pair. See also DOI: 10.6084/m9.figshare.17261540.

ways for each cell type are as follows. One common feature of the pathways for all other cell types except Fig 4.13a is the presence of some integrins, such as *ITGAL* (gene encoding CD11a; also known as LFA1A) (Fig 4.13b – 4.13e), *ITGAX* (gene encoding CD11c) (Fig 4.13d), *ITGB2* (gene encoding CD18) (Fig 4.13d), and *ITGA4* (gene encoding CD49d) (Fig 4.13e). Some integrins are downstream, such as *ACTB* (gene encoding β-actin) in the pathway for migratory DCs (Fig 4.13b) and *DCTN1* (gene encoding Dynactin subunit 1) in the pathway for MoAM2 (Fig 4.13d). Integrins are molecules interacting with ICAM-1 to stabilize cell adhesion. Especially, Dynactin recruits and tethers dynein to microtubules.

Another common feature of the pathways for all other cell types except Fig 4.13a is the presence of the molecules responsible for NF-κB pathways, such as *NFKB1* (gene encoding NF-κB p105 subunit 1), *NFKB2* (gene encoding NF-κB p105 subunit 2), *RELA* (gene encoding NF-κB p65 subunit), *JUN* (Jun proto-oncogene; also known as AP-1 transcription factor subunit), *CHUK* (gene encoding inhibitor of nuclear factor κ-B kinase subunit α; also known as IKK-α) (Fig 4.13b - 4.13e). These molecules are not DCGs but nodes in the KEGG pathway, but they are located upstream of *ICAM1* and flanked by DCGs.

As the specific features of some pathways, the pathway for migratory DCs (Fig 4.13b) and summation (Fig 4.13e) include *RELB* (gene encoding transcription factor RelB). In general, *NFKB2* and *RELB* lie in the noncanonical NF-κB pathway, which is an upstream pathway of *ICAM1* [149]. The pathway for migratory DCs (Fig 4.13b) also includes *TNFRSF11A* (gene encoding receptor activator of NF-κB; also known as RANK) and *MAP3K14* (gene encoding NF-κB-inducing kinase; also known as NIK). These molecules are known as triggers of noncanonical NF-κB pathway [150]. While previous analysis or curation work found the canonical NF-κB pathway [151], the noncanonical pathways were not known to be involved in the COVID-19 Disease Map.

The pathway for summation (Fig 4.13e) has some commonalities with other pathways, such as integrins and molecules related to the NF-κB pathway, but it also has some differences. *SOD2* (gene encoding superoxide dismutase 2), for example, is a gene that is not found in the other pathways and could not be found without taking summation. *SOD2* is known as a gene whose expression variation has been confirmed accompanied with *ICAM1* in COVID-19 [152].

4.3.2 Case Study 2: *ICAM1*-Associated Pathways at Different Time Points

Like case study 1, the original dataset underwent single-cell omics data analysis, and *ICAM1* and its DCGs were extracted. The p-values of *ICAM1* expression variation and the computation times (sec.) for each time point were as follows: p = 1.50E-05, time = 77.9 (day 1), p = 3.09E-2, time = 55.4 (day 5), and p = 3.04E-06, time = 56.3 (day 10). This manuscript does not explain the other detailed results of the single-cell omics data analysis because the procedures were the same as in case study 1. These other results can be found at `10.6084/m9.figshare.23590755`. As for the rest steps, we explain the results of spurious correlation removal and pathway construction.

Undirected Graphical Model Construction

Table 4.3 depicts how spurious correlated edges were removed for each of the three-time points. Of the number of edges in the simple correlation network (full model), more than 84% of the edges were deleted by calculating the second-order partial correlation coefficients.

Table 4.3. Spurious correlation removal in case study 2

Day	Nodes	Edges (full)	Edges (excluded)	Edges (output)
1	121	7,503 (100%)	6,309 (84%)	1,194 (16%)
5	198	20,706 (100%)	18,914 (91%)	1,792 (9%)
10	126	8,001 (100%)	6,748 (84%)	1,253 (16%)

Pathway Mapping and Unification

The pathways resulting from combining the KEGG pathways with the partial correlation networks of *ICAM1*-associated DCGs extracted from the omics data are shown in Figure 4.14.

The common characteristics and the unique attributes of the found pathways for each time point are as follows. One common feature is the presence of molecules responsible for the immune response included across the three time points. The immunoreactive molecules include chemokine ligand (*CXCL1, 2, 3*) induced by interleukin-1 (*IL1B*) or TNF-α-induced protein 6 (*TNFAIP6*). These molecules are along the green-colored molecules, such as NF-κB p105 subunit (*NFKB1*), NLRP3 inflammasome (*NLRP3*, *PYCARD*, and *CASP1*), which are related to pro-inflammatory effects and activation of NF-κB pathway or the MAPK pathway [153]. Other common molecules include modulator of cytochrome C oxidase during inflammation (*C15orf48*), chemokine ligands acting as macrophage inflammatory protein (*CCL3* and *CCL3L1*), transmembrane protein (*TMEM176A/B*), nuclear factor erythroid (*NFE2*), Peptidyl Arginine Deiminase 4 (*PADI4*), RAS Oncogene (*RAB20*), and proto-oncogene (*ETS2*). These are also related to inflammation, immune response, or membrane fusion [147, 154–156].

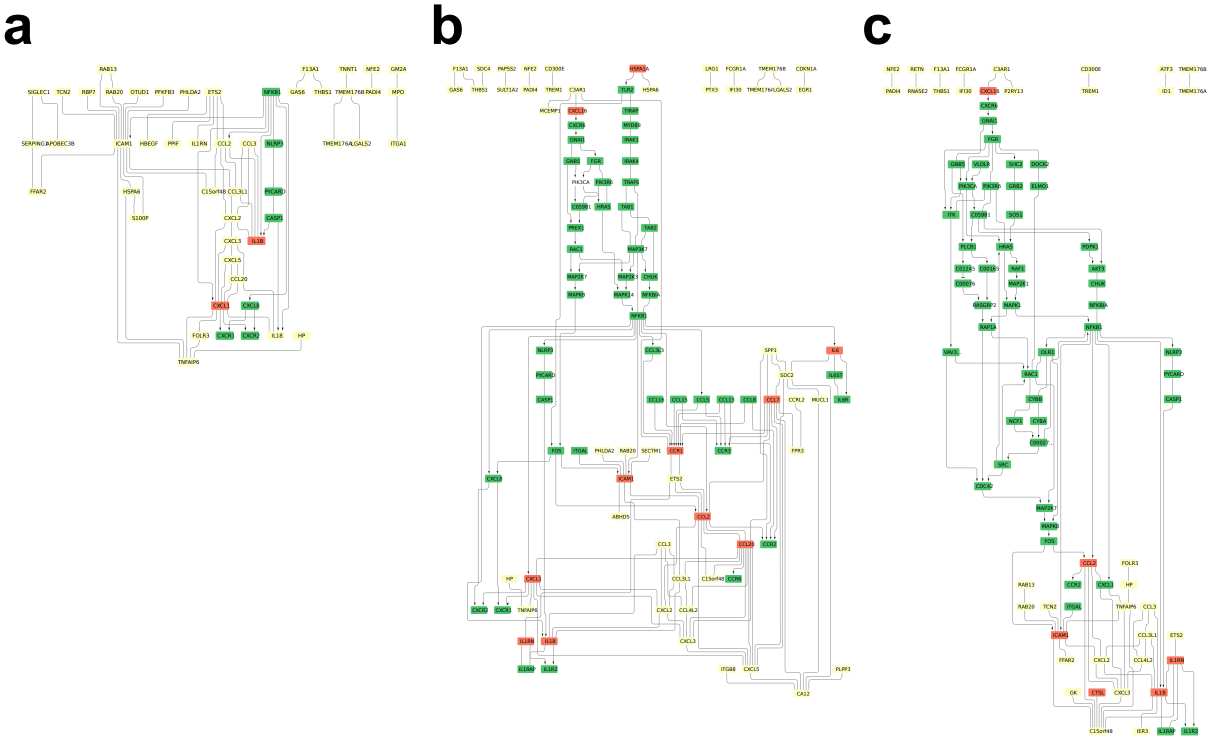


Figure 4.14. *ICAM1*-associated pathways at different time points. a: NF-κB/MAPK pathway putative (day 1) **b:** NF-κB/MAPK pathway putative (day 5) **c:** NF-κB/MAPK pathway putative (day 10). The light yellow, green, and orange nodes represent data-driven DCGs, the genes listed only in the KEGG pathways, and the genes derived from both data and the KEGG pathways. The directed edges are the edges whose directions are given in the KEGG pathways. See this figure for a larger view in *figshare*: 10.6084/m9.figshare.23576226.

As the specific features of two pathways, the pathways at days 1 and 5 include Pleckstrin homology-like domain family A member 2 (*PHLDA2*) and chemokine ligand (*CXCL5*) (Fig 4.14a–b). The pathways at days 5 and 10 include chemokine ligand acting as macrophage inflammatory protein (*CCL4L2*) (Fig 4.14b–c). The pathways at days 1 and 10 include folate receptor (*FOLR3*), Haptoglobin (*HP*), and chemokine ligand (*CXCL16*) (Fig 4.14a–c). These are related to acute inflammatory response, immunity, or membrane attachment [157–159].

The obtained pathways do not contain the molecules in the NF-κB pathway (*NFKB1*, *NFKBIA*, and *CHUK*) or the MAPK pathway (*HRAS*, *RAF1*, and *MAP2K7*). The absence of these molecules may be due to insufficient gene coverage in the original omics data.

4.3.3 Additional Case Studies

ACTB-associated pathways

At all three time points, molecules for immune responses such as inflammation or chemotaxis were abundant. For example, Ficolin-1 (*FCN1*) or leukocyte immunoglobulin like receptor (*LILRA5*, *LILRB1*, *LILRB*) were all involved in innate immune responses. As a unique feature of the *ACTB*-associated pathway, the network motif for cell-membrane fusion was conserved. This motif was consisting of the genes encoding major histocompatibility complex (MHC) class II regulating membrane fusion, including *HLA-DPA1*, *HLA-DPB1*, *HLA-DRA*, *HLA-DRB1*, *HLA-DRB5*, and *HLA-DQB1*. Relevant to MHC class II, HLA class II histocompatibility antigen γ chain (*CD74*) involved in the formation of MHC class II peptide complexes for CD4+ T cell responses or Cathepsin S (*CTSS*) related to antigen presentation with MHC class II were present. Moreover, Vimentin (*VIM*) for cytoskeleton formation or actin-binding protein allograft inflammatory factor 1 (*AIF1*) were identified.

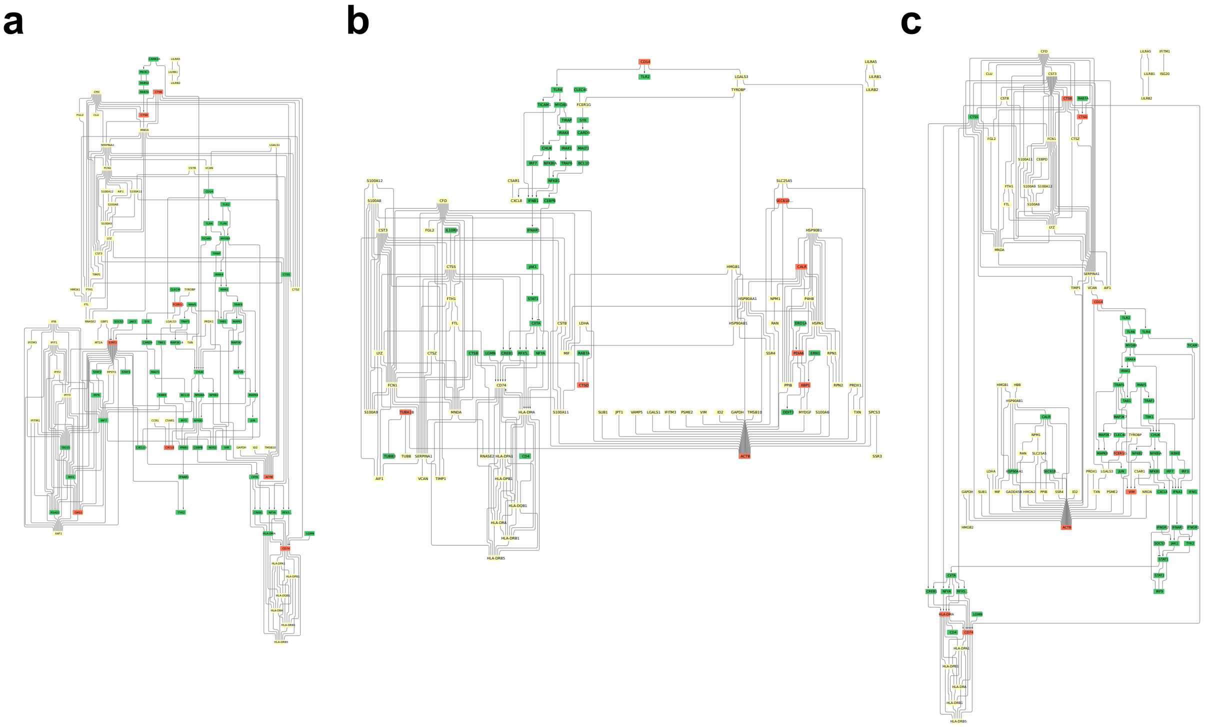


Figure 4.15. *ACTB*-associated pathways at different time points **(a)** Day1, **(b)** Day 5, and **(c)** Day 10. The figures and datasets regarding *ACTB*-associated pathways can be referred in *figshare*: 10.6084/m9.figshare.23591490.

C15orf48-associated pathways

Throughout the three-time points, molecules for immune response or macrophage polarization were apparent. There were also similar features with the *ICAM1*-associated pathways. For instance, both *ICAM1* and *C15orf48* pathways contained C-X-C Motif Chemokine Ligand (*CXCL*), interleukin 1 β (*IL1B*), transmembrane protein 176A/B (*TMEM176A*, *TMEM176B*), TNF-α-induced protein 6 (*TNFAIP6*), interleukin 1 receptor antagonist (*IL1RN*).

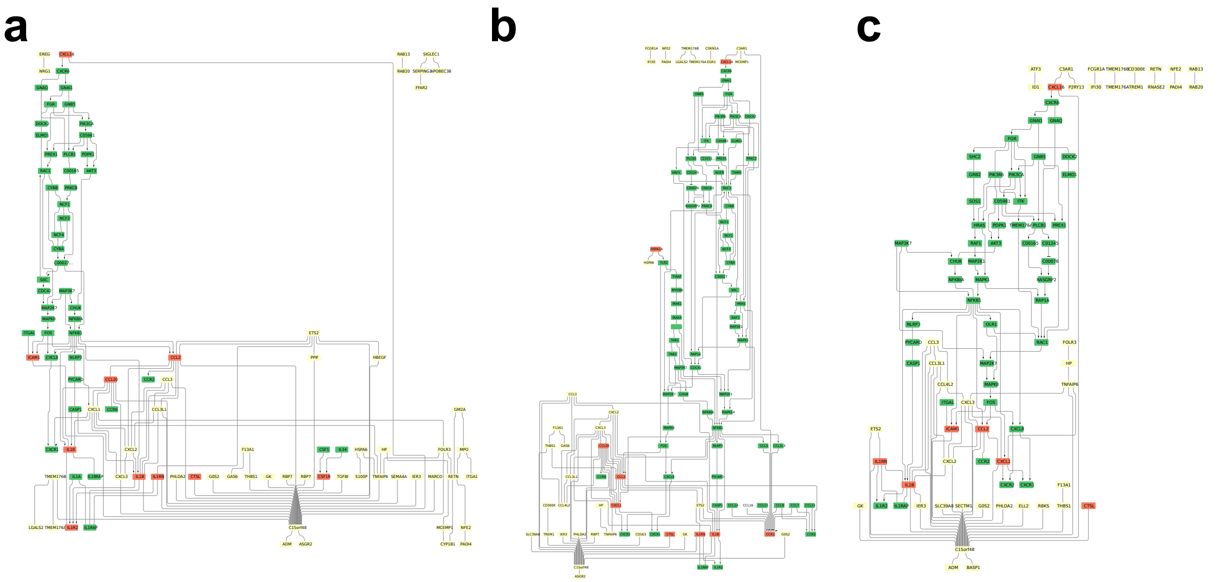


Figure 4.16. *C15orf48*-associated pathways at different time points **(a)** Day1, **(b)** Day 5, and **(c)** Day 10. The figures and datasets regarding *C15orf48*-associated pathways can be referred in *figshare*: 10.6084/m9.figshare.23591505.

Time-dependent change of the three pathways

In the figure, the different line colors indicate the variation of temporal pattern over the three-time points. Purple, green, yellow, and dark blue colored lines reflect an increase followed by a decrease in the number of hit genes, a decrease followed by an increase, a monotonic decrease, and a monotonic increase.

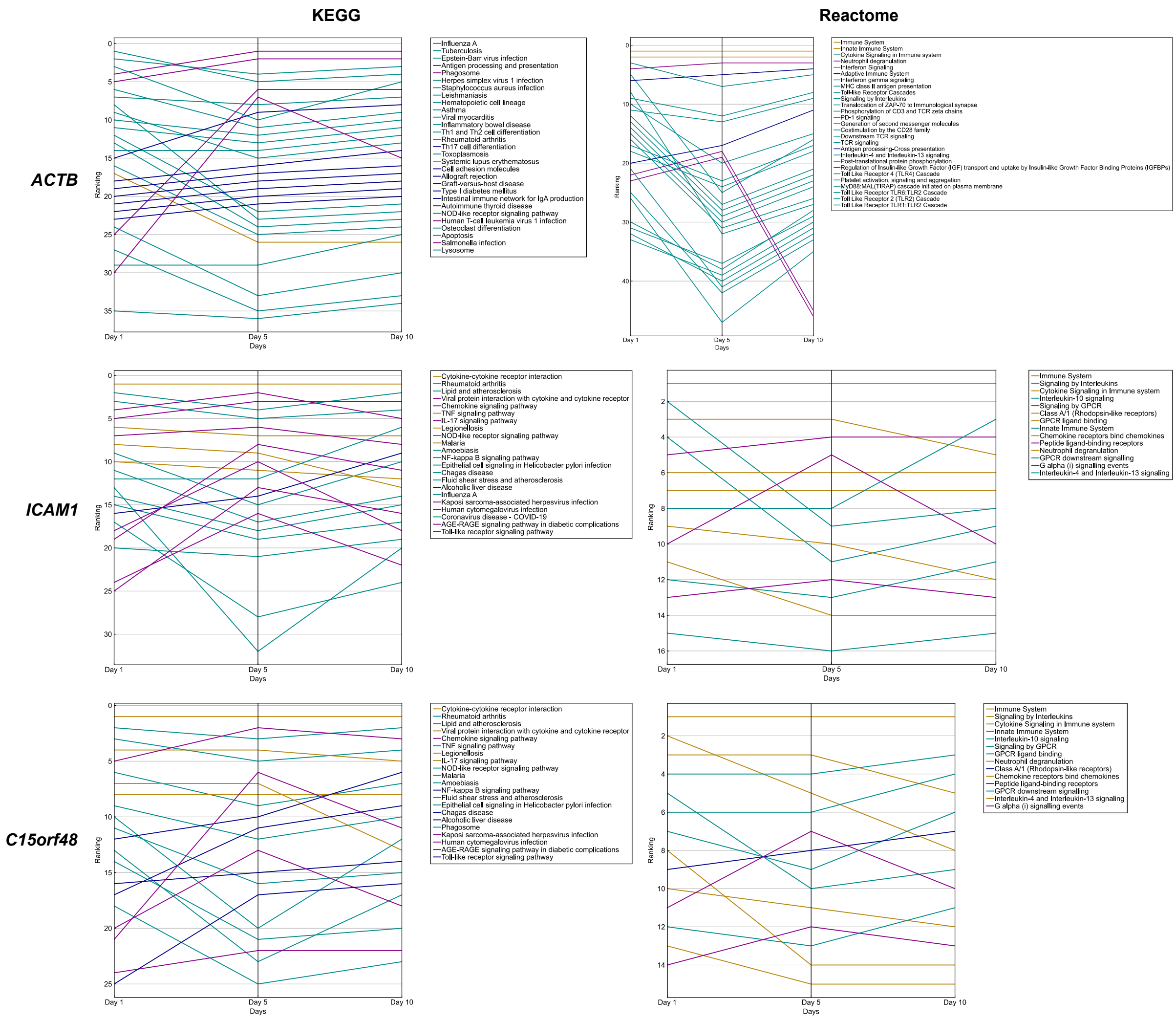


Figure 4.17. Parallel coordinate ranking plots of the matched KEGG and Reactome pathways. See this figure for a larger view in *figshare*: 10.6084/m9.figshare.23576238.

4.4 Discussion

4.4.1 *ICAM1*-Associated Pathways from Case Studies

Comparison between the obtained *ICAM1*-associated pathways in the case study 1 with the COVID-19 Disease Map reveals existing and unknown nodes. For example, *MAP2K3*, *MAPK14*, *JUN*, *FOS*, *ITGA2*, *ITGB1*, *RSAD2*, *OAS*, and *STAT2* have already been mapped onto the COVID-19 Disease Map, while *RELB*, *ITGAL*, *CDC42*, *ACTB*, *CD40*, *DCTN1*, *BCL3*, and *CD83* in the obtained pathways are still absent in the current COVID-19 Disease Map.

Likewise, we can identify the difference between the obtained *ICAM1*-associated pathways and the current COVID-19 Disease Map from the results of case study 2. For instance, *IL1B*, *NFKB1*, *NLRP3*, *PYCARD*, and *CASP1* are listed in the COVID-19 Disease Map, while there are molecules absent from it, including *CCL3, 3L1, 4L2*, *CXCL1, 2, 3, 5, 16*, *TNFAIP6*, *C15orf48*, *TMEM176A/B*, *NFE2*, *PADI4*, *RAB20*, *ETS2*, *PHLDA2*, *FOLR3*, and *HP*.

The results from both case studies indicated that the NF-κB pathway would likely be activated, which reflects that our framework can reproduce the already-known fact that the NF-κB pathway is activated in COVID-19, as seen in the KEGG's COVID-19 pathway (hsa05171). As new insight into the unknown pathways missing from the current COVID-19 Disease Map, the results imply that the mechanism that might cause cell-to-cell transmission involves the following two up/downstream pathways.

— Upstream pathway with proteins on the noncanonical NF-κB pathway

— Downstream pathway with integrins and cytoskeletal elements associated with actin and the motor protein dynein for cell transformation

The noncanonical NF-κB pathway is reasonable because it is relevant to the proinflammatory response in viral infections such as COVID-19. It is also creditable that *TNFRSF11A* is found only in the pathway of dendritic cells (Fig 4.13b) since *TNFRSF11A* is known to be expressed on dendritic cells and T cells to facilitate their interaction with each other [160]. The involvement of downstream pathways leading to the cytoskeleton (the internal filaments of eukaryotic cells), including actin filaments and microtubules, in COVID-19 is also plausible. After the interaction between ICAM-1 and integrin regulates cell adhesion, the motor protein myosin would move on actin filaments, inducing cell transformation and movement. The motor protein dynein would move on microtubules transporting molecules in the cytoplasm to the MTOC. Given the argument mentioned in the Introduction that MTOC or VS spawned by ICAM-1 invoke cell-to-cell transmission in HIV-1 or HTLV-1, the existence of these downstream molecules of *ICAM1*-associated pathways raises the possibility of pathways involved in the formation of MTOC or VS in SARS-CoV-2. In this study, the Ras-Raf-MEK-ERK pathway for MTOC or VS in HTLV-1 was inactive. Meanwhile, *RAC1* and *CDC42* were conserved. Ras-related C3 botulinus toxin substrate 1 (Rac1, encoded by *RAC1*) and cell division control protein 42 homolog (Cdc42, encoded by *CDC42*) are essential for VS formation in HTLV-1 cells [161]. Although it is unclear whether SARS-CoV-2 has a VS formation mechanism analogous to that of HIV-1 or HTLV-1, we cannot rule out the possibility that MTOC formation and VS formation never occur. To verify these inferred phenomena, observing MTOC and VS formation through infection experiments or molecular dynamics tracking using high-end live-cell imaging techniques [162] would be desirable.

4.4.2 Time-dependent change of the three pathways

Comparison between the three ranking plots tells us that the inferred pathways are plausible in terms of the COVID-19 typical clinical time courses such as innate immunity or neutrophil degranulation. In addition to that, both in *ICAM1*-associated pathways and *C15orf48*-associated pathways, signaling by G-Protein Coupled Receptor (GPCR) and Class A/1 (Rhodopsin-like receptors) are found. Considering that GPCR largely includes the Rhodopsin-like family proteins and they regulates microtubule stabilization [163], we cannot deny that MTOC formation and viral cell-to-cell transmission could be observed in COVID-19 as well. Verifying the MTOC formation or cell-to-cell transmission in COVID-19 would require further *in vitro* infection experiments with microscopy.

4.4.3 Related Work

The need to identify unknown pathways has accelerated the work related to gene network inference in COVID-19. For example, Hasankhani *et al.* obtained signaling pathways associated with the main hallmarks of COVID-19 by differential coexpression network analysis [164]. Tanaka *et al.* revealed host cellular gene networks by Bayesian network [165]. Generally, several methods for gene network inference from single-cell omics data exist, which can be classified into data-driven methods and knowledge-based methods. Data-driven gene network inference methods include statistical approaches such as regression, mutual information, correlation, and a combination of different techniques [166–168]. Among those techniques, correlation analysis is the typically first choice to gain insights into systems for further investigation (56% of papers in 2023 on the preprint server bioRxiv contain the word “correlation”) [169]. Especially, partial correlation is used for inference of features regulating coexpression or other features’ activities within the network by estimating conditional dependencies [170]. Alternatively, knowledge-based gene network inference uses prior knowledge for information retrieval or logic programming. Fabris *et al.* quantified the influence by creating interpretable KEGG feature types for the hierarchical classification of aging-related protein functions [171]. Chen *et al.* provided the biological relevance by analyzing the gene ontology terms and KEGG pathways of each drug category enriched in the literature and clinical trials for predicting drug-target interaction [172]. There also exist hybrid methods incorporating data-driven and knowledge-based methods. Soh *et al.* enumerated the minimal network components by adopting a Boolean satisfiability problem (SAT) solver for KEGG pathways [173]. Zuo *et al.* integrated information at gene expression and network topology levels by differentially weighted graphical LASSO [174]. However, full-scale integration of data-driven and knowledge-based methods is still under development for gene network inference. Our method favors this development by extending the correlation network by integrating data and knowledge. Especially, two-step extraction of DCGs in Step 1, narrowing down DCGs after filtering DEGs, is a mixture of detecting the significant differences in the gene expression levels and checking the pairwise correlation between gene pairs. This extraction is substitutive to other methods for extracting DCGs, such as WGCNA or gene sets net correlations analysis (GSNCA) [175].

4.4.4 Conclusion

As a summary of contributions, this study proposed a DD-KB framework for automatically inferring systems by graphical modeling from large-scale sparse matrix and model validation with multiple knowledge bases. Using the framework, we demonstrated its applicability to empirical COVID-19 data and three types of genes. We realized reproducing existing pathways, discovering novel pathways currently absent from the COVID-19 Disease Map, and analyzing their spatiotemporal variation. The discovered pathways suggested the existence of unknown pathways in the map, an upstream noncanonical NF-κB pathway, and a downstream pathway that may lead to MTOC formation subject to observation.

In addition to the scientific findings, our framework, which integrates existing data mining and database integration methods and automates hypothesis discovery from single-cell omics data and multiple knowledge bases, is also original and versatile. Single-cell omics data analysis in Step 1 and model validation by multiple knowledge bases in Step 3 realized constructing pathways in different cases. For these reasons, our work would contribute to a remarkable development in the DD-KB gene network inference methods.

The existence of undirected edges within the final pathways would be a limitation of our framework. These edges without direction arise from correlation networks that find direct and indirect relationships but do not distinguish between causality and correlation [176]. Our methodology requires its extension to infer causal directions of the edges.

Consequently, future work will include the following three tasks. First, we will infer causal networks based on data and knowledge via Bayesian networks or other observational causal discovery techniques [177]. Second, we will compare our framework with network inference methods. For example, its generalization performance can be improved by replacing biclustering with other techniques, such as WGCNA. Third, we will analyze the obtained pathways for verifying or modifying them in terms of dynamics. For example, modeling and simulation of differential equations based on state transitions would help us comprehend the dynamics [67].

Otherwise, the perturbation experiments can simulate the intervention effects on dynamics by explicitly using direct transcription factor knockout or overexpression [178]. Indeed, such a study has significantly improved prediction accuracy for downstream targets [179].

Overall, the *ICAM1*-associated pathways constructed from the data and knowledge in this study will expedite the repair and completion of the COVID-19 Disease Map for a deeper understanding of SARS-CoV-2 pathogenesis.

To the next Chapter

— Reviewing related work on multiscale modeling
— Especially related to Study 1: Benefit of modeling viral dynamics, possible extensions, the detail of viral quantification, different approaches to building viral dynamics models, reactivation risk
— Especially related to Study 2: How to infer causal directions in biological network

Chapter 5

Related work

Brief summary of this chapter

- — Scientific discovery
- — Causal discovery
- — Other possible conditions for viral dynamics modeling
- — Multiomics data

Although the two studies in this manuscript differ in scale and methodology, they commonly attempt to discover unknown hypotheses or interpretable models using structural information (syntax tree structures or network structures derived from knowledge) on actual observational data. The discipline of such discovery is called discovery science and constitutes a field of study. This chapter reviews discovery science as a related study of the two studies. In particular, given that the edges in Study 2 could not be directed as causal relationships based on data alone, we review causal discovery methods separately. In addition, we will review other possible conditions for viral dynamics models to refine Study 1 and other data that could be selected to further refine the multiscale model beyond the omics data of gene expression levels as used in Study 2.

5.1 Scientific discovery

Scientific discovery is the culmination of humankind's creative thinking [180]. This endeavor has evolved alongside the computational perspective, forming a field of research that formulates laws or models from data or knowledge and finding new knowledge, which domain is called *computational scientific discovery* [181]. The focus of computational scientific discovery is not to find black box models, such as the traditional Gaussian process or neural network, but to find laws or models with symbolic structures for model interpretability [182]. Here, symbolic structure refers to the relationships between entities or inferred components represented by logical formula, process, or tree structure seen in decision tree or numerical equation as syntax tree [183].

Research on symbolic structures has been conducted in this area for more than 20 years. Development of deep learning with high-end GPU computing around 2010 or the groundbreaking work on symbolic regression with genetic algorithm [184] have led to research dealing with both symbolic structures and continuous distributed systems. Such efforts can be found in several cutting-edge studies, exemplified by finding physical concepts based on linkage inside neural networks [185] or learning Boolean functions as matricized logic programs in vector spaces [186]. Additionally, computing symbolic structure in continuous space can be found in several work on equation discovery. For example, the recent state-of-the-art studies are found in identification of systems of partial differential equations by sparse regression from data and its applications to gene network estimation [187], physics-informed learning from small data [188], learning ODEs with data symmetry and separability [189], discovery of physical principles via symbolic regression of the model learned by Graph Neural Network (GNN) [190], parametric latent space dynamics identification [191], and discovery of closed-form ODEs from observed trajectories [192]. These attempts successfully go beyond heuristic search through discrete space to symbolic structure search in continuous parameter space, thereby representing dynamics in parametric form with interpretability.

On the other hand, it remains challenging to realize a *parsimonious* model preserving interpretability while balancing accuracy, model complexity, computational scalability, and generalizability [193]. In particular, discovered models often fall into redundant models due to overfitting to models with unnecessary terms and a lack of model validation, which makes them unrealistic. The previous methods also address these computational problems, but there have not been adequate criteria for filtering variables into minimum ones. To reduce dull terms and improve low fidelity in the model, there are some attempts to discover a parsimonious model with a symbolic structure that background knowledge can corroborate or rectify. For example, data and knowledge integration has been accomplished in several work, such as process-based modeling from observed data and background knowledge on observed dynamical behavior for induction of hypothesized processes about entities' interactions as syntax tree [194], grammar-based equation discovery [195], and symbolic regression and reasoning by incorporating axiom or physics theory [196].

5.2 Causal discovery

5.2.1 Traditional methods

Traditional causal discovery methods are constraint-based method and score-based method. Constraint-based methods, such as Peter-Clark (PC) algorithm and Fast Causal Inference (FCI) algorithm, perform conditional independence tests to construct Directed Acyclic Graphs (DAGs) [197]. Score-based methods, such as Greedy Equivalent Search (GES), optimize score function for structure learning [198]. Max-Min Hill Climbing algorithm (MMHC) is a hybrid of these two methods which integrates conditional independence tests to construct constraint-based causal networks and score optimization for structure learning. MMHC outperforms PC or GES [199], while the performance depends on statistical evaluation criteria [200]. Constraint-based, score-based, and their hybrid methods rely on the following two assumptions [201].

Assumption 5.2.1 (Causal Markov Assumption)**.** Given a DAG G over variable set X and probability distribution Pr over X, G and Pr satisfy the Causal Markov Assumption (CMA) iff. $\forall x_i \in X$ is conditionally independent of non-descendants (nodes without direct causes) $X \setminus descendants(X) \cup parents(X)$ given $parents(X)$, where $Pr(x_i, \cdots, x_d) = \Pi Pr(x_i | parent(x_i))$.

Assumption 5.2.2 (Faithfulness)**.** Given a DAG G over variable set X and probability distribution Pr over X, G is faithful to Pr iff. every conditional independence relation true in Pr is entailed by CMA (**Assumption 5.2.1**) applied to G.

These **Assumptions 5.2.1** and **5.2.2** allows one to regard an outputted causal network as a separated DAG consisting of chain, fork, and collider directions, i.e. a DAG obeying d-separation rule [202].

Additionally, PC requires the following assumption whereas FCI does not.

Assumption 5.2.3 (No latent confounder)**.** There is no unobserved common cause that directly affects two or more observed variables.

These methods are successful in learning causal network even in high-dimensional settings [203]. However, they do not distinguish graphs that entail the same d-separation properties, that is, the Markov equivalence class of conditional independence [201].

5.2.2 Methods based on causal graphs

In contrast, the recent spurred attention has been paid to methods based on functional causal models (FCMs) because they can uniquely determine the true model by distinguishing from different DAGs in the same class [204]. FCM is a directed causal graph with variables determined by structural equations and assumptions on joint distribution [205]. This type of method includes addictive noise model (ANM) [206], linear non-Gaussian acyclic model (LiNGAM) [207], non-Gaussian structural vector auto-regressive model (VAR-LiNGAM) [208], regression error based causal inference (RECI) [209], repetitive causal discovery (RCD) algorithm [210]. Their property of *identifiability* is owed to the benefit of the following additional assumptions on the data generation process.

Assumption 5.2.4 (Exogenous variables' i.i.d.)**.** Exogenous variables are independent and identically distributed. In particular, probabilistic distribution of exogenous variables follow non-Gaussian distribution. Under this assumption, the two models with different parameters never produce identical distributions.

In addition, the following assumption is necessary in case of causality between time series.

Assumption 5.2.5 (Stationary process)**.** The data are generated from a stationary process, a property in which the mean and variance of the data do not change over time.

In essence, VAR-LiNGAM constructs linear structural equations with time lag commonly relying on the **Assumptions 5.2.1–5.2.5** so that contemporaneous and past causal effects can be considered.

LiNGAM LiNGAM discovers linear structural equations without time lag representing causal relationships between state variables from inputted data based on **Assumptions 5.2.1–5.2.5**. Its linear function f is expressed as below.

$$X = BX + e \tag{5.1}$$

, where $B^{d \times d}$ is an adjacency matrix of its causal graph. Regarding one variable x_i, the equation is written by

$$x_i = \Sigma_{j \in pa_i} b_{j,i} x_j + e_i \tag{5.2}$$

, where $k(j) < k(i)(i \neq j)$ holds. e_i and e_j are mutually i.i.d. (**Assumption 5.2.4**), and $b_{j,i}$ is a real-valued constant representing causal effect from x_j to x_i. The adjacency matrix B is permuted for the right permutation of its rows (corresponding to causal order $k(i)$) to be a lower-triangular matrix producing a DAG and neglecting self loops.

VAR-LiNGAM VAR-LiNGAM is the LiNGAM's extension that considers contemporaneous and past causal effects by introducing structural VAR model. VAR model, not a causal discovery method but a stochastic process for multivariate time series, assumes past effect only. VAR is expressed by

$$x(t) = \nu + \Sigma_{\tau=1}^{\delta} B_\tau x(t-\tau) + e(t) \tag{5.3}$$

, where ν, τ, and δ are a dth-order constant vector acting as intercept, a time lag ($\tau \in 0, 1, \cdots, \delta$), and an autoregressive coefficient ($\delta \in \mathbb{N}$). Structural VAR model is the extended VAR model that allows contemporaneous time point, expressed by

$$x(t) = \nu + \Sigma_{\tau=0}^{\delta} B_\tau x(t-\tau) + e(t). \tag{5.4}$$

Both models have a dth-order white noise vector as error term $e(t)$ which does not always satisfy **Assumption 5.2.4** (*e.g.,* additive white Gaussian noise).

Combining different estimation steps as structural VAR model and as LiNGAM, VAR-LiNGAM estimates B_τ. VAR-LiNGAM is expressed as follows:

$$x(t) = \Sigma_{\tau=0}^{\delta} B_\tau x(t-\tau) + e(t) \tag{5.5}$$

, where $e(t)$ satisfies **Assumption 5.2.4**. The adjacency matrix B_0 is permuted to be lower triangular.

5.2.3 Methods based on symbolic reasoning or deep learning

We can also refer to several studies on causal learning through different approaches. Learning from interpretation transition (LFIT) is a method for learning rules about state transitions by discretizing time-series data of signal values or changes and constructing a Boolean network [211]. Logical formulas can be constructed from time-series data by applying LFIT for Boolean network inference. Additionally, causal discovery has been realized with deep learning, such as DAG learning by GNN [212] or reinforcement learning [213]. Among the existing deep learning for causal discovery, we focus on Structural Agnostic Modeling (SAM) [214]. SAM is a penalized adversarial learning method that receives noise vector and real data as the generator's input and incorporates a binary adjacency matrix without time lag, named as *structural gate*, in the generator to predict this matrix as causal structure.

As more recent deep learning technique, BaCaDi is a differentiable Bayesian method for identifying causal structure from partial observations, even when data are scarce [215]. Inspired by the previous causal inference techniques, such as the Peter-Clark algorithm (JCI-PC) [216], constraint-based inference (UT-IGSP) [217], and score-based inference (DCDI) [218], BaCaDi can estimate Causal Bayesian networks in a gradient-based manner by observing differences in posterior distributions after intervention. In practice, it estimates a gene regulatory network from synthetic single-cell data, showing its outperformance over existing methods in several indices. Hereinafter, we review the causal discovery methods, focusing more on the biological context.

5.2.4 Causal discovery in the biological context

Various methods have also been proposed for data-driven network inference in biological context. Marbach *et al.* classified the data-driven gene network inference into the three groups: statistical approach, probabilistic approach, and dynamical models [166]. Statistical approach includes correlation networks (WGCNA [219]), information theoretic scores (ARACNE; an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context [220], CLR; context likelihood of relatedness algorithm [221], MRNET; minimum redundancy networks [222]), and regression-based methods (TIGRESS; trustful inference of gene regulation with stability selection [223], GENIE3; gene network inference with ensemble of trees [224]). Probabilistic approach includes Gaussian graphical models [225] or Bayesian signaling network [226]. Dynamical models includes dynamic Bayesian networks [227] or differential equation models [228].

In contrast to data-driven gene network inference, gene network inference can be done by knowledge-based approach. Knowledge-based gene network inference includes hierarchical classification of functions [171] or combining gene ontology and the KEGG pathways [172]. Hybrid of the data-driven and knowledge-based methods also exist, exemplified by Boolean SAT solver for KEGG pathways [173] or differentially weighted graphical least absolute shrinkage and selection operator (LASSO) [174].

As for metabolic network construction from metabolite data, contrary to network inference from lower scales, such as genes and transcripts, inference from upper scales, such as metabolites, can be considered and be a subject for comparing the performance of our framework in Study 2. Reactomine uses a combination of greedy methods, heuristics, and unsupervised learning to estimate chemical reaction networks from time-series data. It requires neither domain knowledge nor supervised data and has the strength of excellent interpretability, but has the weakness of not being scalable [229]. Structural learning estimates Bayesian networks based on Bayes' theorem and conditional independence. In light of our study, we can apply structural learning, such as MMHC, to undirected edges and convert them to the Systems Biology Markup Language

(SBML) format to make the model easier to analyze. Process Based Modeling Tool (ProBMoT) explores chemical reaction networks by adding edges using probabilistic context-free grammars and evolving syntax trees of algebraic equations [230].

5.3 Other possible conditions in modeling

5.3.1 System's boundary and metabolic flux balance analysis

Although no boundaries were defined for the infection system in this research, metabolic flux equilibrium could serve to define the system boundaries. Since the signal value of a metabolite evolves with time, Michaelis-Menten's law and the law of mass action, a type of reaction kinetics, can be applied to metabolites. Metabolic flux equilibrium analysis is an analytical method for solving constrained optimization problems for chemical reaction network models by linear programming, assuming that the entire model is in equilibrium based on these laws applicable at the metabolic scale. This analysis can be performed using tools such as the COnstraint-Based Reconstruction and Analysis (COBRA) tool [231] to find the flux vector that maximizes the metabolic activity and brings the system to equilibrium.

5.3.2 Asymptomatic SARS-CoV-2 careers

Although we ignored it in this manuscript, we might consider the existence of asymptomatic careers, which is one of characteristics specific to COVID-19. The same betacoronaviruses as SARS-CoV-2 have caused outbreaks before: SARS caused by SARS-CoV in 2002-2004 and Middle East respiratory syndrome (MERS) caused by MERS-CoV in 2012 [232]. Considering that the cumulative number of deaths in both SARS and MERS was less than 1,000, the absolute scale of COVID-19 is dramatically different [233]. Since most of those infected with SARS were symptomatic, the epidemic was limited to about two years by isolation policies [234].

In contrast, COVID-19 is highly infectious but often asymptomatic, making the traceability of infected individuals difficult. About half of those infected with COVID-19 are asymptomatic carriers, and 45% of secondary infections occur before the onset of the symptoms [235]. The infection fatality rate (IFR) of COVID-19, the total death ratio of infectious persons with or without a confirmed diagnosis, is estimated to be 0.3% to 0.6% as of February 2020 [236]. This fact indicates that many potentially infected people exist without being counted in the reported cases. Therefore, there is a need for biological indicators that can track potentially infectious populations.

5.3.3 Reactivation

Specific related work investigating the phenomena of viral reactivation with the above-mentioned other techniques includes the following. Miura *et al.* indicate the rapid reactivation of HTLV-1 by cellular stress response like a transmission into a new host using computer simulation and observed transcripts data [237]. Garnett and Grenfell described the relationship between the reactivation of varicella-zoster virus (VZV) and the host's age through mathematical models and observed epidemiological data [238]. A stochastic model distinct from the deterministic model in this study also provides new insights as exemplified by the successful description of the HIV-1 replication dynamics model proposed by Yuan *et al.* [101]. Furthermore, as seen in the remarks by Eissing *et al.* [239], we would use multiscale models in which cell scales connect higher hierarchical levels with omics data [240].

5.4 Multiomics data

While Study 2 constructed models from gene expression data pertaining to the gene and transcript scales, there are attempts to fit models to data at all scales simultaneously. Such data, including protein and metabolite data, are called multiomics data, and studies using this multiomics data have been attracting attention in recent years. For example, attempts to learn and infer multiscale models based on multiomics data are seen in Bayesian relational learning from multiomics data [241] or metabolic activity prediction for each cell by incorporating the expression levels of metabolite-related genes as a penalty matrix in metabolic flux equilibrium analysis [242]. Considering these attempts are published in top journal or conference papers, linking results at the gene, transcript, and protein scales to results at the metabolite scale for multiscale systems would be expected to have a large impact worldwide.

To the next Chapter

The next chapter concludes this manuscript with a total summary through the manuscript, limitations, and future work for overcoming these limitations.

Chapter 6

Concluding remarks

Brief summary of this chapter

- — Motivated by social demands to global issue COVID-19
- — Elucidating the mechanism of viral dynamics (SARS-CoV-2 cell-to-cell transmission)
- — Verifying hypothesis from a system of pathways while proposing DD-KB Framework
- — Discovering missing knowledge through multiscale modeling
- — Future work: Framework extension, development, and expansion

In summary, motivated by social demands to the global issue of COVID-19, we conducted two studies towards elucidating the mechanism of viral dynamics for SARS-CoV-2, that is, cell-to-cell transmission. SARS-CoV-2 cell-to-cell transmission was verified from a system of pathways while proposing the DD-KB framework, discovering missing knowledge through multiscale modeling. Study 2 has conducted gene network inference considering multiscale properties and led to knowledge discovery in pathway identification. Beyond knowledge discovery, we have developed an original framework that utilizes coexpressed genes of specific genes such as *ICAM1* and graphical modeling of these genes, establishing a foundation for future research.

As other achievements than publishing journal papers, Study 2 preprint paper was registered on the FAIRDOMHub and linked to a unique ID (`https://fairdomhub.org/publications/641`). This platform integrates the interactions of proteins, such as signaling molecules and metabolic activators, from pathological conditions based on the sources, including the COVID-19 Disease Map, and provides them as a graphical representation, like knowledge graphs. In addition, to ensure transparency of the research content and to promote public disclosure, the output data obtained from the research were uploaded to the online repository *figshare*. The data obtained from NCBI GEO has already been processed for anonymization and de-identification, and we also specified this point in our study.

As a limitation throughout this research, our multiscale modeling attempts to link different scale studies in the same scientific discovery loop as a concept. However, linking such components in one closed multiphysics system is ideal. Compared to a linkage of macroscale tumor models governed by physical laws of metastasis with microscale models of gene regulatory networks [243], the perspective of elucidating infectious systems and using multiscale modeling has yet to be explored. Therefore, modeling microscale metabolites whose dynamics can be represented by ODEs, such as the genome-scale metabolic model (GEM) [244], would help our attempt to connect genes and population dynamics.

Future perspectives include our DD-KB framework extension, development, and expansion. The framework extension includes identifying nonlinear systems and uncertainty in chemical reaction networks and modifying models by embedding discretized network structure in symbolic space onto parameters in continuous algebraic space. In addition to causal discovery methods mentioned in Chapter 5, we can also refer to conversion from causal network into ODEs based on ODEs-to-FCM conversion via equilibrium equation with intervention [245] or FCM building asymptotic behavior of ODEs under intervention [246] would be able to pave the way for verifying dynamics inherent in the causality by continuous ODEs. As another verification method, we can see whether the dynamic behavior of the model satisfies the specification. For example, a model described by Computation Tree Logic (CTL) can be verified by a symbolic model checking tool, such as NuSMV, to analyze the reachable states [247].

The development of the framework would include outreach activities, such as opening a platform for propagating the extensions and releasing application tools. Here, we assume reaching a wide range of users by delivering Graphical User Interface (GUI)-based applications rather than Python libraries. One feasible way is to use MATLAB. MATLAB language is compatible with Python and provides tools for converting programs into applications, such as user interface development tools and visually integrated development environments. With these tools, MATLAB algorithms can be integrated into existing C, C++, and Java applications, and the developed framework can be distributed. If the framework being developed can be coupled with MATLAB, models in SBML format can be more easily analyzed using toolboxes, such as the PottersWheel Toolbox [248], for model reduction and identifiability analysis. In addition, the framework can be integrated with existing knowledge bases such as Pathway Commons, BioGRID, and Signor via APIs and then released on the website as an original knowledge base to which the framework can be applied. In the GUI-based platform, users can apply the framework using data on hand and update them daily. Furthermore, if personal data are accepted after de-identification and anonymization processing and providing users with estimated results, the framework realized online could be operated as a Software-as-a-Service (SaaS) type of AI that performs personalized medicine (patient-tailored therapeutics). Accepting registration of data and knowledge graphs from users as a hub will also expand the knowledge base according to access, automating knowledge acquisition and promoting lifelong learning.

As another aspect, the framework would be expanded to application to other global issues. This manuscript dealt with macroscopic viral dynamics as a within-host viral infection. However, we can work on a more macroscopic Earth and planetary system as an application of the DD-KB approach framework. This topic may seem far from the viruses, but it can be very close. For example, both viral infection systems and Earth and planetary systems are complex systems with multiscale and multiphysics properties. They share the same formal expression in that they are based on a network structure composed of time-evolving entities. They also have in common that they should be compatible with the DD-KB approach. There are two representative models related to the Earth and planetary systems: the Earth System Model (ESM) and the Integrated Assessment Model (IAM). The ESM is closer to natural science and data-driven (downscaling by deep learning to achieve large scale and high accuracy) model consisting of material cycles such as biogeochemical cycles or the Atlantic Meridional Overturning Circulation (AMOC). At the same time, the IAM has a higher affinity to knowledge, such as socio-economic processes, behavior change of human groups, and policy-making (emphasizing causality and interpretability and aiming for simplicity). More recently, coupling these two models into a global-scale model has also been studied [249, 250]. The DD-KB approach would enable ESM-IAM coupled modeling for evidence-based policy making (EBPM), which is based on big data but can also explain the rationale behind the results. Because of the real-time properties of Earth's satellite data, the model may be called the "Earth Digital Twin," built on the metaverse rather than a simulation model. Furthermore, given that marine virome forms a vast biomass and significantly impacts marine ecosystems, omics data would be combined with the Earth and planetary systems. To realize the above, cooperation with various stakeholders will be more critical than this doctoral research, with the significant goals of further understanding the mechanisms of the Earth and planetary systems and contributing to evidence-based policy-making and advocacy to realize a safe and secure society and sustainable development in the Anthropocene.

BIBLIOGRAPHY

1. Zhou, P., Yang, X. L., Wang, X. G., *et al.,* A pneumonia outbreak associated with a new coronavirus of probable bat origin, *Nature* **579,** 270–273 (2020).
2. Remuzzi, A. & Remuzzi, G., COVID-19 and Italy: what next?, *Lancet* **395,** 1225–1228 (2020).
3. Dong, E., Du, H. & Gardner, L., An interactive web-based dashboard to track COVID-19 in real time, *Lancet Infect. Dis.* **20,** 533–534 (2020).
4. Woolf, S. H., Chapman, D. A. & Lee, J. H., COVID-19 as the Leading Cause of Death in the United States. *JAMA* **325,** 123–124 (2021).
5. *Pfizer Press release: Pfizer and BioNTech Celebrate Historic First Authorization in the U.S. of Vaccine to Prevent COVID-19* https://www.pfizer.com/news/press-release/press-release-detail/pfizer-and-biontech-celebrate-historic-first-authorization, December, 2020.
6. *World Health Organization: WHO Coronavirus (COVID-19) Dashboard* https://covid19.who.int/, Retrieved on November 5, 2023.
7. *Centers for Disease Control and Prevention: End of Public Health Emergency* https://www.cdc.gov/coronavirus/2019-ncov/your-health/end-of-phe.html, May, 2023.
8. *European Centre for Disease Prevention and Control: SARS-CoV-2 variants of concern as of 20 October 2023* https://www.ecdc.europa.eu/en/covid-19/variants-concern (Retrieved on November 5, 2023).
9. Gebeyehu, D. T., East, L., Wark, S. & Islam, M. S., Disability-adjusted life years (DALYs) based COVID-19 health impact assessment: a systematic review. *BMC Public Health* **23** (2023).
10. *Centers for Disease Control and Prevention: Long COVID* https://www.cdc.gov/coronavirus/2019-ncov/long-term-effects/index.html, July, 2023.

11. *World Health Organization: Development of a strategy and action plan on health emergency preparedness, response, and resilience in the WHO European Region (Preparedness 2.0)* `https://www.who.int/europe/teams/who-health-emergencies-programme-(whe)/preparedness-2.0`, Retrieved on November 5, 2023.

12. Ostaszewski, M. *et al.,* Community-driven roadmap for integrated disease maps, *Brief. Bioinform.* **20,** 659–670 (2019).

13. Ostaszewski, M. *et al.,* COVID19 Disease Map, a computational knowledge repository of virus–host interaction mechanisms, *Mol. Syst. Biol.* **17,** e10387 (2021).

14. Reimand, J. *et al.,* Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap, *Nat. Protoc.* **14,** 482–517 (2019).

15. Ostaszewski, M. *et al.,* Author Correction: COVID-19 Disease Map, building a computational repository of SARS-CoV-2 virus-host interaction mechanisms, *Sci. Data* **7** (2020).

16. Surdo, L. P. *et al.,* SIGNOR 3.0, the SIGnaling network open resource 3.0: 2022 update. *Nucleic Acids Res.* **51,** D631–D637 (2022).

17. Peirce, C. S., *Collected Papers of Charles Sanders Peirce* (Harvard University Press, Cambridge, 1958).

18. Smith, D. L., Battle, K. E., Hay, S. I., *et al.,* Ross, Macdonald, and a Theory for the Dynamics and Control of Mosquito-Transmitted Pathogens, *PLOS Pathog.* **8,** 1–13 (2012).

19. Hamer, W. H., The Milroy Lectures on Epidemic Disease in England—The Evidence of Variability and Persistence of Type, *Lancet* **167,** 569–574 (1906).

20. Ross, R., Some quantitative studies in epidemiology, *Nature* **87,** 466–467 (1911).

21. Kermack, W. O., McKendrick, A. G. & Walker, G. T., A contribution to the mathematical theory of epidemics, *Proc. R. Soc. A* **115,** 700–721 (1927).

22. Kermack, W. O. & McKendrick, A. G., Contributions to the mathematical theory of epidemics. III.—Further studies of the problem of endemicity, *Proc. R. Soc. A* **141,** 94–122 (1933).

23. Murray, A. & Jackson, G., Viral dynamics: a model of the effects of size shape, motion and abundance of single-celled planktonic organisms and other particles, *Mar. Ecol. Prog. Ser.* **89,** 103–116 (1992).

24. Barré-Sinoussi, F., Chermann, J. C., Rey, F., *et al.,* Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS), *Science* **220,** 868–871, ISSN: 0036-8075 (1983).

25. Giordano, G., Blanchini, F., Bruno, R., *et al.,* Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy, *Nat. Med.* **26,** 855–860 (2020).

26. He, S., Peng, Y. & Sun, K., SEIR modeling of the COVID-19 and its dynamics, *Nonlinear Dyn.* **101,** 1667–1680 (2020).

27. Rosenberg, A., Reductionism (and antireductionism) in biology, *The Cambridge Companion to the Philosophy of Biology. Cambridge University Press,* 120–138 (2007).

28. Gatherer, D., So what do we really mean when we say that systems biology is holistic?, *BMC Syst. Biol.* **4** (2010).

29. Sorger, P. K., A reductionist's systems biology: Opinion, *Curr. Opin. Cell Biol.* **17,** 9–11 (2005).

30. Alber, M., Buganza, T. A., Cannon, W., *et al.,* Integrating machine learning and multiscale modeling-perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences, *npj Digit. Med.* **2** (2019).

31. Parunak, H. V. D., Savit, R. & Riolo, R. L., Agent-Based Modeling vs. Equation-Based Modeling: A Case Study and Users' Guide, *Multi-Agent Systems and Agent-Based Simulation,* 10–25 (1998).

32. Holcombe, M., X-machines as a basis for dynamic system specification, *Softw. Eng. J.* **3,** 69–76 (1988).

33. Vodovotz, Y. & An, G., Agent-based models of inflammation in translational systems biology: A decade later, *Wiley Interdiscip. Rev. Syst. Biol. Med.* **11,** e1460 (2019).

34. Andrighetto, G., Conte, R., Turrini, P., *et al.,* Emergence In the Loop: Simulating the two way dynamics of norm innovation, *Normative Multi-agent Systems* (2007).

35. Nguyen, T. N. A., Zucker, J.-D., Nguyen, D. H., *et al.,* A Hybrid Macro-Micro Pedestrians Evacuation Model to Speed Up Simulation in Road Networks, *Advanced Agent Technology: AAMAS Workshops 2011,* 371–383 (2011).

36. Shinde, S. B. & Kurhekar, M. P., Review of the systems biology of the immune system using agent-based models, *IET Syst. Biol.* **12,** 83–92 (2018).

37. Bobashev, G., Goedecke, D., Yu, F., *et al.,* A Hybrid Epidemic Model: Combining The Advantages Of Agent-Based And Equation-Based Approaches, *In Proceedings of the Winter Simulation Conference,* 1532–1537 (2007).

38. Mach, R. & Schweitzer, F., Modeling Vortex Swarming In Daphnia, *Bull. Math. Biol.* **69,** 539–562 (2007).

39. Xu, R. & Wunsch, D., Survey of Clustering Algorithms, *IEEE Transactions on Neural Networks and Learning Systems* **16,** 645–678 (2005).

40. Rao, C. V., Wolf, D. M. & Arkin, A. P., Control, exploitation and tolerance of intracellular noise, *Nature* **420,** 231–237, ISSN: 1476-4687 (2002).

41. Perelson, A. S., Neumann, A. U., Markowitz, M., *et al.,* HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time, *Science* **271,** 1582–1586 (1996).

42. Nowak, M. A. & Bangham, C. R., Population dynamics of immune responses to persistent viruses, *Science* **272,** 74–79 (1996).

43. Perelson, A. S., Modelling viral and immune system dynamics, *Nat. Rev. Immunol.* **2,** 28–36 (2002).

44. Torres, M., Wang, J., Yannie, P. J., *et al.,* Identifying important parameters in the inflammatory process with a mathematical model of immune cell influx and macrophage polarization, *PLOS Comput. Biol.* **15,** 1–27 (2019).

45. Postel, M., Karam, A., Pézeron, G., *et al.,* A multiscale mathematical model of cell dynamics during neurogenesis in the mouse cerebral cortex, *BMC Bioinform.* **20** (2019).

46. Guo, T., Qiu, Z., Kitagawa, K., *et al.,* Modeling HIV multiple infection, *J. Theor. Biol.* **509,** 110502 (2021).

47. Neumann, A., Lam, N., Dahari, H., *et al.,* Hepatitis C Viral Dynamics in Vivo and the Antiviral Efficacy of Interferon-α Therapy, *Science* **282,** 103–107 (1998).

48. He, X., Lau, E. H. Y., Wu, P., *et al.,* Author Correction: Temporal dynamics in viral shedding and transmissibility of COVID-19, *Nat. Med.* **26,** 1491–1493 (2020).

49. Barabási, A. L. & Albert, R., Emergence of Scaling in Random Networks, *Science* **286,** 509–512 (1999).

50. Hanahan, D. & Weinberg, R. A., Hallmarks of Cancer: The Next Generation, *Cell* **144,** 646–674 (2011).

51. Luo, J. D., Liu, J., Yang, K. & Fu, X., Big data research guided by sociological theory: a triadic dialogue among big data analysis, theory, and predictive models, *J. Chin. Sociol.* **6,** 11, ISSN: 21982635 (Dec. 2019).

52. Anderson, R. M. & Grenfell, B. T., Quantitative investigations of different vaccination policies for the control of congenital rubella syndrome (CRS) in the United Kingdom, *Epidemiol. Infect.* **96,** 305–333 (1986).

53. Bertozzi, A. L., Franco, E., Mohler, G., Short, M. B. & Sledge, D., The challenges of modeling and forecasting the spread of COVID-19, *arXiv preprint arXiv:2004.04741* (2020).

54. Schlesinger, S., Terminology for model credibility, *Simulation* **32,** 103–104 (1979).

55. Hattori, S.-I. *et al.,* A small molecule compound with an indole moiety inhibits the main protease of SARS-CoV-2 and blocks virus replication, en, *Nat. Commun.* **12,** 668 (2021).

56. V'kovski, P., Kratzel, A., Steiner, S., *et al.,* Coronavirus biology and replication: implications for SARS-CoV-2, *Nat. Rev. Microbiol.* **19,** 155–170 (2021).

57. Gordon, D. E. *et al.,* A SARS-CoV-2 protein interaction map reveals targets for drug repurposing, en, *Nature* **583,** 459–468 (July 2020).

58. Sethuraman, N., Jeremiah, S. S. & Ryo, A., Interpreting Diagnostic Tests for SARS-CoV-2, *JAMA* **323,** 2249–2251 (2020).

59. Huang, C. *et al.,* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China, en, *Lancet* **395,** 497–506 (Feb. 2020).

60. Tillett, R. L. *et al.,* Genomic evidence for reinfection with SARS-CoV-2: a case study, *Lancet Infect. Dis.* **21,** 52–58 (2021).

61. *Worldometer: World Coronavirus Pandemic* `https://www.worldometers.info/coronavirus/`, 2020.

62. Wang, S., Kang, B., Ma, J., *et al.,* A deep learning algorithm using CT images to screen for Corona virus disease (COVID-19), *Eur. Radiol.* **31,** 6096–6104 (2021).

63. Mattioli, J., Pedroza, G., Khalfaoui, S. & Leroy, B., Combining Data-Driven and Knowledge-Based AI Paradigms for Engineering AI-Based Safety-Critical Systems, *CEUR-WS* (2022).

64. Vafaee, F., Diakos, C., Kirschner, M. B., *et al.,* A data-driven, knowledge-based approach to biomarker discovery: application to circulating microRNA markers of colorectal cancer prognosis, *npj Syst Biol Appl* **4** (2018).

65. Thacker, B. H. *et al., Concepts of Model Verification and Validation* (Los Alamos National Laboratory, 2004).

66. *American Institute of Aeronautics and Astronautics: Guide for the Verification and Validation of Computational Fluid Dynamics Simulations* AIAA-G-077-1998, Reston, VA, 1998.

67. Odaka, M. & Inoue, K., Computational Modeling and Simulation of Viral Load Kinetics in SARS-CoV-2 Replication, *In Proceedings of the International Conference on Computational Systems-Biology and Bioinformatics,* 75–82 (2020).

68. Huang, G., Ma, W. & Takeuchi, Y., Global properties for virus dynamics model with Beddington-DeAngelis functional response, *Appl. Math. Lett.* **22,** 1690–1693 (2009).

69. Beddington, J. R., Mutual interference between parasites or predators and its effect on searching efficiency, *J. Anim. Ecol.* **44,** 331–340 (1975).

70. DeAngelis, D. L., Goldstein, R. A. & O Neill, R. V., A model for tropic interaction, *Ecology* **56,** 881–892 (1975).

71. Pearce-Pratt, R., Malamud, D. & Phillips, D. M., Role of the cytoskeleton in cell-to-cell transmission of human immunodeficiency virus, *J. Virol.* **68,** 2898–2905 (1994).

72. Igakura, T., Stinchcombe, J. C., Goon, P. K. C., *et al.,* Spread of HTLV-I between lymphocytes by virus-induced polarization of the cytoskeleton, *Science* **299,** 1713–1716 (2003).

73. Weissenhorn, W., Dessen, A., Calder, L. J., *et al.,* Structural basis for membrane fusion by enveloped viruses, *Mol. Membr. Biol.* **16,** 3–9 (1999).

74. Walls, A. C., Park, Y., Tortorici, M. A., *et al.,* Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein, *Cell* **181,** 281–292.e6 (2020).

75. Hoffmann, M., Kleine-Weber, H., Simon, S., *et al.,* SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor, *Cell* **181,** 271–280.e8 (2020).

76. Jones, K. S., Petrow-Sadowski, C., Huang, Y. K., *et al.,* Cell-free HTLV-1 infects dendritic cells leading to transmission and transformation of CD4+ T cells, *Nat. Med.* **14,** 429–436 (2008).

77. Pais-Correia, A., Sachse, M., Guadagnini, S., *et al.,* Biofilm-like extracellular viral assemblies mediate HTLV-1 cell-to-cell transmission at virological synapses, *Nat. Med.* **16,** 83–89 (2010).

78. Koza, J. R., Human-competitive results produced by genetic programming, *Genet. Program. Evolvable Mach.* **11,** 251–284 (2010).

79. Page, J., Poli, R. & Langdon, W. B., Smooth Uniform Crossover with Smooth Point Mutation in Genetic Programming: A Preliminary Study, *In Proceedings of the European Conference on Genetic Programming,* 39–48 (1999).

80. Zheng, S., Fan, J., Yu, F., *et al.,* Viral load dynamics and disease severity in patients infected with SARS-CoV-2 in Zhejiang province, China, January-March 2020: retrospective cohort study, *BMJ* **369:m1443** (2020).

81. *Creative Commons Attribution-NonCommercial 4.0 International* Last accessed on September 4, 2021, `https://creativecommons.org/licenses/by-nc/4.0/`.

82. Drevon, D., Fursa, S. R. & Malcolm, A. L., Intercoder Reliability and Validity of WebPlotDigitizer in Extracting Graphed Data, *Behav. Modif.* **41,** 323–339 (2017).

83. Rohatgi, A., *Webplotdigitizer: Version 4.3* Last accessed on September 4, 2021, 2020, `https://automeris.io/WebPlotDigitizer`.

84. Kent, E., Neumann, S., Kummer, U., *et al.,* What Can We Learn from Global Sensitivity Analysis of Biochemical Systems?, *PLOS ONE* **8,** 1–13 (Nov. 2013).

85. Anderson, B., Borgonovo, E., Galeotti, M., *et al.,* Uncertainty in climate change modeling: can global sensitivity analysis be of help?, *Risk Anal.* **34,** 271–293 (2014).

86. Ligmann-Zielinska, A., Kramer, D. B., Spence-Cheruvelil, K. & Soranno, P. A., Using Uncertainty and Sensitivity Analyses in Socioecological Agent-Based Models to Improve Their Analytical Performance and Policy Relevance, *PLOS ONE* **9** (2014).

87. Lilburne, L. & Tarantola, S., Sensitivity analysis of spatial models, *Int. J. Geogr. Inf. Sci.* **23,** 151–168 (2009).

88. Sobol, I. M., Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates, *Math. Comput. Simul.* **55,** 271–280 (2001).

89. Saltelli, A., Making best use of model evaluations to compute sensitivity indices, *Comput. Phys. Commun.* **145,** 280–297 (2002).

90. Kuznetsov, Y. A., *Elements of Applied Bifurcation Theory* (Springer-Verlag, 1998).

91. Dai, L., D., V., S., K. K., *et al.,* Generic Indicators for Loss of Resilience Before a Tipping Point Leading to Population Collapse, *Science* **336,** 1175–1177 (2012).

92. Routh, E. J., *A Treatise on the Stability of a Given State of Motion, Particularly Steady Motion* (Macmillan and Company, 1877).

93. Hurwitz, A., Ueber die Bedingungen, unter welchen eine Gleichung nur Wurzeln mit negativen reellen Theilen besitzt, *Math. Ann.* **46,** 273–284 (1895).

94. Sun, R., Global stability of the endemic equilibrium of multigroup SIR models with nonlinear incidence, *Comput. Math. with Appl.* **60,** 2286–2291 (2010).

95. Berndt, D. J. & Clifford, J., Using Dynamic Time Warping to Find Patterns in Time Series, *In Proceedings of the International Conference on Knowledge Discovery and Data Mining,* 359–370 (1994).

96. Petitjean, F., Ketterlin, A. & Gançarski, P., A global averaging method for dynamic time warping, with applications to clustering, *Pattern Recognit.* **44,** 678–693 (2011).

97. Tarantola, A., Popper, Bayes and the inverse problem, *Nat. Phys.* **2,** 492–494 (2006).

98. Yu, G., Sapiro, G. & Mallat, S., Solving Inverse Problems With Piecewise Linear Estimators: From Gaussian Mixture Models to Structured Sparsity, *IEEE Trans. Image Process.* **21,** 2481–2499 (2012).

99. Saltelli, A., Annoni, P., Azzini, I., *et al.,* Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index, *Comput. Phys. Commun.* **181,** 259–270 (2010).

100. Tong, M., Jiang, Y., Xia, D., *et al.,* Elevated Expression of Serum Endothelial Cell Adhesion Molecules in COVID-19 Patients, *J. Infect. Dis.* **222,** 894–898 (2020).

101. Yuan, Y. & Allen, L. J. S., Stochastic models for virus and immune system dynamics, *Math. Biosci.* **234,** 84–94 (2011).

102. Khan, M. A. & Atangana, A., Modeling the dynamics of novel coronavirus (2019-nCov) with fractional derivative, *Alex. Eng. J.* **59,** 2379–2389 (2020).

103. Ullah, S., Khan, M. A. & Farooq, M., A new fractional model for the dynamics of the hepatitis B virus using the Caputo-Fabrizio derivative, *Eur. Phys. J. Plus* **133,** 1–14 (2018).

104. Ullah, S., Khan, M. A. & Farooq, M., A fractional model for the dynamics of TB virus, *Chaos Solit. Fractals* **116,** 63–71 (2018).

105. Khan, M. A., Ullah, S. & Farooq, M., A new fractional model for tuberculosis with relapse via Atangana-Baleanu derivative, *Chaos Solitons Fractals* **116,** 227–238 (2018).

106. Bwire, G. M., Majigo, M. V., Njiro, B. J., *et al.,* Detection profile of SARS-CoV-2 using RT-PCR in different types of clinical specimens: A systematic review and meta-analysis, *J. Med. Virol.* **93,** 719–725 (2021).

107. Odaka, M. & Inoue, K., Modeling viral dynamics in SARS-CoV-2 infection based on differential equations and numerical analysis, *Heliyon* **7,** e08207 (2021).

108. Luo, B. H., Carman, C. V. & Springer, T. A., Structural basis of integrin regulation and signaling, *Annu. Rev. Immunol.* **25,** 619–647 (2007).

109. Wilson, R. W. *et al.,* Gene targeting yields a CD18-mutant mouse for study of inflammation, *J. Immunol.* **151,** 1571–8 (1993).

110. Bracq, L., Xie, M., Benichou, S. & Bouchet, J., Mechanisms for Cell-to-Cell Transmission of HIV-1, *Front. Immunol.* **9** (2018).

111. Nejmeddine, M. *et al.,* HTLV-1-Tax and ICAM-1 act on T-cell signal pathways to polarize the microtubule-organizing center at the virological synapse, *Blood* **114,** 1016–1025 (2009).

112. Smith-Norowitz, T. A., Loeffler, J., Norowitz, Y. M. & Kohlhoff, S., Intracellular Adhesion Molecule-1 (ICAM-1) Levels in Convalescent COVID-19 Serum: A Case Report, *Ann. Clin. Lab. Sci.* **51,** 730–734 (2021).

113. Tong, M. *et al.,* Elevated Expression of Serum Endothelial Cell Adhesion Molecules in COVID-19 Patients, *J. Infect Dis.* **222,** 894–898 (2020).

114. Won, T. *et al.,* Endothelial thrombomodulin downregulation caused by hypoxia contributes to severe infiltration and coagulopathy in COVID-19 patient lungs, *eBioMedicine* **75,** 103812 (2022).

115. Zeng, C. *et al.,* SARS-CoV-2 spreads through cell-to-cell transmission, *Proc. Natl. Acad. Sci. USA* **119,** e2111400119 (2021).

116. Li, X. *et al.,* Network embedding-based representation learning for single cell RNA-seq data, *Nucleic Acids Res.* **45,** e166–e166 (2017).

117. Hao, Y. *et al.,* Integrated analysis of multimodal single-cell data, *Cell* **184,** 3573–3587 (2021).

118. Wolf, A. A., Angerer, P. & Theis, F. J., SCANPY: large-scale single-cell gene expression data analysis, *Genome Biol.* **19** (2018).

119. Palla, G. *et al.,* Squidpy: a scalable framework for spatial omics analysis, *Nat. Methods* **19,** 171–178 (2022).

120. Bredikhin, D., Kats, I. & Stegle, O., MUON: multimodal omics analysis framework, *Genome Biol* **42** (2022).

121. Hicks, S. C., Townes, F. W., Teng, M. & Irizarry, R. A., Missing data and technical variability in single-cell RNA-sequencing experiments, *Biostatistics* **19,** 562–578 (2018).

122. Pedregosa, F. *et al.,* Scikit-learn: Machine Learning in Python, *J. Mach. Learn. Res.* **12,** 2825–2830 (2011).

123. Lehoucq, R. B., Sorensen, D. C. & Yang, C., ARPACK users'guide — solution of large-scale eigenvalue problems with implicitly restarted Arnoldi methods, *In: Software, Environments, Tools (SIAM)* **6** (1998).

124. McInnes, L. & Healy, J., UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, *ArXiv e-prints,* abs/1802.03426 (2018).

125. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E., Fast unfolding of communities in large networks, *J. Stat. Mech.* **2008,** P10008 (2008).

126. Benjamini, Y. & Hochberg, Y., Controlling the false discovery rate: a practical and powerful approach to multiple testing, *J. R. Stat. Soc. Series B* **57,** 289–300 (1995).

127. Eisen, M. B., Spellman, R. T., Brown, P. O. & Botstein, D., Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. USA* **96,** 10943–10943 (1999).

128. De la Fuente, A., Bing, N., Hoeschele, I. & Mendes, P., Discovery of meaningful associations in genomic data using partial correlation coefficients, *Bioinform.* **20,** 3565–3574 (2004).

129. Zuo, Y., Yu, G., Tadesse, M. G. & Ressom, H. W., Biological network inference using low order partial correlation, *Methods* **69,** 266–273 (2014).

130. Langfelder, P. & Horvath, S., WGCNA: an R package for weighted correlation network analysis, *BMC Bioinform.* **9** (2008).

131. Horvath, S., *Weighted Network Analysis: Application in Genomics and Systems Biology* (Springer, 2011).

132. Zhang, B. & Horvath, S., A general framework for weighted gene co-expression network analysis, *Stat. Appl. Genet. Mol. Biol.* **4** (2005).

133. Hiraoka, S., Hirai, M., Matsui, Y., *et al.,* Microbial community and geochemical analyses of trans-trench sediments for understanding the roles of hadal environments, *ISME J.* **14,** 740–756 (2020).

134. Hou, J., Ye, X., Li, C. & Wang, Y., K-Module Algorithm: An Additional Step to Improve the Clustering Results of WGCNA Co-Expression Networks, *Genes* **12,** 87 (2021).

135. Rodchenkov, I. *et al.,* Pathway Commons 2019 Update: integration, analysis and exploration of pathway data, *Nucleic Acids Res.* **48,** D489–D497 (2020).

136. Stark, C. *et al.,* BioGRID: a general repository for interaction datasets, *Nucleic Acids Res.* **34,** D535–D539 (2006).

137. Szklarczyk, D. *et al.,* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets, *Nucleic Acids Res.* **47,** D607–D613 (2019).

138. Wadi, L., Meyer, M., Weiser, J., Stein, L. D. & Reimand, J., Impact of outdated gene annotations on pathway enrichment analysis, *Nat. Methods* **13,** 705–706 (2016).

139. Altman, T., Travers, M., Kothari, A., Caspi, R. & Karp, P. D., A systematic comparison of the MetaCyc and KEGG pathway databases, *BMC Bioinform.* **14** (2013).

140. Huang, D. W., Sherman, B. T. & Lempicki, R. A., Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, *Nat. Protoc.* **4,** 44–57 (2009).

141. Huang, D. W., Sherman, B. T. & Lempicki, R. A., Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists, *Nucleic Acids Res.* **37,** 1–13 (2009).

142. Shannon, P. *et al.,* Cytoscape: a software environment for integrated models of biomolecular interaction networks, *Genome Res.* **13,** 2498–2504 (2003).

143. Barrett, T. *et al.,* NCBI GEO: archive for functional genomics data sets–update, *Nucleic Acids Res.* **41,** D991–D995 (2012).

144. Grant, R. A. *et al.,* Circuits between infected macrophages and T cells in SARS-CoV-2 pneumonia, *Nature* **590,** 635–641 (2021).

145. Duò, A., Robinson, M. D. & Soneson, C., A systematic performance evaluation of clustering methods for single-cell RNA-seq data, *F1000Res.* **7** (2018).

146. Cillo, A. R. *et al.,* People critically ill with COVID-19 exhibit peripheral immune profiles predictive of mortality and reflective of SARS-CoV-2 lung viral burden, *Cell. Rep Med.* **2,** 100476 (2021).

147. Lee, C. Q. E. *et al.,* Coding and non-coding roles of MOCCI (C15ORF48) coordinate to regulate host inflammation and immunity, *Nat. Commun.* **12,** 2130 (2021).

148. Kivioja, T. *et al.,* Counting absolute numbers of molecules using unique molecular identifiers, *Nat. Methods* **9,** 72–74 (2012).

149. Liu, Z. *et al.,* A NIK–SIX signalling axis controls inflammation by targeted silencing of non-canonical NF-κB, *Nature* **568,** 249–253 (2019).

150. Malinin, N. L., Boldin, M. P., Kovalenko, A. V. & Wallach, D., MAP3K-related kinase involved in NF-kappaB induction by TNF, CD95 and IL-1, *Nature* **385,** 540–544 (1997).

151. Fujisawa, K., Shimo, M., Taguchi, Y., Ikematsu, S. & Miyata, R., PCA-based unsupervised feature extraction for gene expression analysis of COVID-19 patients, *Sci. Rep.* **11** (2021).

152. Zheng, W. *et al.,* Glycyrrhizic Acid for COVID-19: Findings of Targeting Pivotal Inflammatory Pathways Triggered by SARS-CoV-2, *Front. Pharmacol.* **12,** 631206 (2021).

153. Bouwmeester, T. *et al.,* A physical and functional map of the human TNF-α/NF-κB signal transduction pathway, *Nat. Cell Biol.* **6,** 97–105 (2004).

154. Gold, R., Kappos, L., Arnold, D. L., Bar-Or, A., Giovannoni, G., *et al.,* Placebo-Controlled Phase 3 Study of Oral BG-12 for Relapsing Multiple Sclerosis, *N. Engl. J. Med.* **367,** 1098–1107 (2012).

155. Tanikawa, C. *et al.,* Regulation of histone modification and chromatin structure by the p53-PADI4 pathway, *Nat. Commun.* **3** (2012).

156. Pylayeva-Gupta, Y., Grabocka, E. & Bar-Sagi, D., RAS oncogenes: weaving a tumorigenic web, *Nat. Rev. Cancer* **11,** 761–774 (2011).

157. Duan, Y., Du, Y., Gu, Z., Zheng, X. & Wang, C., Prognostic Value, Immune Signature, and Molecular Mechanisms of the PHLDA Family in Pancreatic Adenocarcinoma, *Int. J. Mol. Sci.* **23,** 10316 (2022).

158. Palomino, D. & Marti, L., Chemokines and immunity, *Einstein (Sao Paulo)* **13,** 469–473 (2015).

159. Cohen, D., Pilozzi, A. & Huang, X., Network Medicine Approach for Analysis of Alzheimer's Disease Gene Expression Data, *Int. J. Mol. Sci.* **21,** 332 (2020).

160. Anderson, D. M. *et al.,* A homologue of the TNF receptor and its ligand enhance T-cell growth and dendritic-cell function, *Nature* **390,** 175–179 (1997).

161. Gross, C. & Thoma-Kress, A. K., Molecular Mechanisms of HTLV-1 Cell-to-Cell Transmission, *Viruses* **8,** 74 (2016).

162. Real, F., Sennepin, A., Ganor, Y., Schmitt, A. & Bomsel, M., Live Imaging of HIV-1 Transfer across T Cell Virological Synapse to Epithelial Cells that Promotes Stromal Macrophage Infection, *Cell Rep.* **23,** 1794–1805 (2018).

163. Palazzo, A. F., Joseph, H. L., Chen, Y. J., *et al.,* Cdc42, dynein, and dynactin regulate MTOC reorientation independent of Rho-regulated microtubule stabilization, *Curr. Biol.* **11,** 1536–1541 (2001).

164. Hasankhani, A. *et al.,* Differential Co-Expression Network Analysis Reveals Key Hub-High Traffic Genes as Potential Therapeutic Targets for COVID-19 Pandemic, *Front. Immunol.* **12,** 5371 (2021).

165. Tanaka, Y. *et al.,* Dynamic changes in gene-to-gene regulatory networks in response to SARS-CoV-2 infection, *Sci. Rep.* **11,** 11241 (2021).

166. Marbach, D. *et al.,* Wisdom of crowds for robust gene network inference, *Nat. Methods.* **9,** 796–804 (2012).

167. Mochida, K., Koda, S., Inoue, K. & Nishii, R., Statistical and Machine Learning Approaches to Predict Gene Regulatory Networks From Transcriptome Datasets, *Front. Plant Sci.* **9,** 1770 (2018).

168. Agamah, F. E., Bayjanov, J. R. & Niehues, A., Computational approaches for network-based integrative multi-omics analysis, *Front. Mol. Biosci.* **9** (2022).

169. Becker, M., Nassar, H., Espinosa, C., *et al.,* Large-scale correlation network construction for unraveling the coordination of complex biological systems, *Nat. Comput. Sci.* **3,** 346–359 (2023).

170. Hawe, J. S., Theis, F. J. & Heinig, M., Inferring interaction networks from multi-omics data, *Front. Genet.* **10** (2019).

171. Fabris, F. & Freitas, A. A., New KEGG pathway-based interpretable features for classifying ageing-related mouse proteins, *Bioinform.* **32,** 2988–2995 (2016).

172. Chen, L. *et al.,* Gene Ontology and KEGG Pathway Enrichment Analysis of a Drug Target-Based Classification System, *PLOS ONE* **10** (2015).

173. Soh, T., Inoue, K., Baba, T., Takada, T. & Shiroishi, T., Evaluation of the Prediction of Gene Knockout Effects by Minimal Pathway Enumeration, *Adv. Life Sci.* **4,** 154–165 (2012).

174. Zuo, Y., Cui, Y., Yu, G., Li, R. & Ressom, H. W., Incorporating prior biological knowledge for network-based differential gene expression analysis using differentially weighted graphical LASSO, *BMC Bioinform.* **18** (2017).

175. Rahmatallah, Y., Emmert-Streib, F. & Glazko, G. V., Gene Sets Net Correlations Analysis (GSNCA): a multivariate differential coexpression test for gene sets, *Bioinform.* **30,** 360–368 (2014).

176. Opgen-Rhein, R. & Strimmer, K., From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data, *BMC Syst. Biol.* **1,** 37 (2007).

177. Pearl, J., Causality: Models, Reasoning and Inference, `https://doi.org/10.5860/choice.47-3771` (2000).

178. Loucera, C. *et al.,* Drug repurposing for COVID-19 using machine learning and mechanistic models of signal transduction circuits related to SARS-CoV-2 infection, *Sig. Transduct. Target Ther.* **5** (2020).

179. Noh, H., Shoemaker, J. E. & Gunawan, R., Network perturbation analysis of gene transcriptional profiles reveals protein targets and mechanism of action of drugs and influenza A viral infection, *Nucleic Acids Res.* **46,** e34 (2018).

180. Langley, P., *Scientific discovery: Computational explorations of the creative processes* (MIT press, 1987).

181. Langley, P., The computational support of scientific discovery, *Int. J. Hum. Comput.* **53,** 393–410 (2000).

182. Sun, S., Ouyang, R., Zhang, B. & Zhang, T. Y., Data-driven discovery of formulas by symbolic regression, *MRS Bull.* **44,** 559–564 (2019).

183. Makke, N. & Chawla, S., Interpretable Scientific Discovery with Symbolic Regression: A Review, *arXiv preprint* **abs/2211.10873** (2022).

184. Schmidt, M. & Lipson, H., Distilling Free-Form Natural Laws from Experimental Data, *Science* **324,** 81–85 (2009).

185. Iten, R., Metger, T., Wilming, H., Rio, L. & Renner, R., Discovering Physical Concepts with Neural Networks, *Phys. Rev. Lett.* **124,** 010508 (2020).

186. Sato, K. & Inoue, K., Diferentiable learning of matricized DNFs and its application to Boolean networks, *Mach. Learn.* **112,** 2821–2843 (2023).

187. Brunton, S., Proctor, J. & Kutz, J. N., Discovering governing equations from data by sparse identification of nonlinear dynamical systems, *Proc. Natl. Acad. Sci. USA* (2016).

188. Raissi, M., Perdikaris, P. & Karniadakis, G. E., 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 (2019).

189. Udrescu, S. M. & Tegmark, M., AI Feynman: A physics-inspired method for symbolic regression, *Sci. Adv.* **6,** eaay2631 (2020).

190. Cranmer, M. *et al.,* Discovering Symbolic Models from Deep Learning with Inductive Biases, *In Proceedings of the Conference on Neural Information Processing Systems* (2020).

191. Fries, W. D., He, X. & Choi, Y., Lasdi: Parametric latent space dynamics identification, *Comput. Methods Appl. Mech. Eng.* **399,** 115436 (2022).

192. Qian, Z., Kacprzyk, K. & Schaar, M., D-CODE: Discovering Closed-form ODEs from Observed Trajectories, *In Proceedings of the International Conference on Learning Representations* (2022).

193. Kutz, J. N. & Brunton, S. L., Parsimony as the ultimate regularizer for physics-informed machine learning, *Nonlinear Dyn.* **107,** 1801–1817 (2022).

194. Bridewell, W., Langley, P. & Todorovski, L., Inductive process modeling, *Mach. Learn.* **71,** 1–32 (2008).

195. Brence, J., Todorovski, L. & Džeroski, S., Probabilistic grammars for equation discovery, *Knowl. Based Syst.* **224,** 107077 (2021).

196. Cornelio, C. *et al.,* Combining data and theory for derivable scientific discovery with AI-Descartes, *Nat. Commun.* **14,** 1777 (2023).

197. Spirtes, P. *et al.,* Constructing Bayesian network models of gene expression networks from microarray data, *In Proceedings of the Atlantic Symposium on Computational Biology* (2000).

198. Chickering, D. M., Optimal structure identification with greedy search, *J. Mach. Learn. Res.* **3,** 507–554 (2003).

199. Tsamardinos, I., Brown, L. E. & Aliferis, C. F., The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm, *Mach. Learn.* **65,** 31–78 (2006).

200. Scutari, M., Graafland, C. E. & Gutiérrez, J. M., Who Learns Better Bayesian Network Structures: Constraint-Based, Score-based or Hybrid Algorithms, *In Proceedings of the International Conference on Probabilistic Graphical Models. PMLR* **72,** 416–427 (2018).

201. Spirtes, P. *et al., Causation, Prediction, and Search* (MIT press, 2000).

202. Pearl, J., *Causality: Models, Reasoning, and Inference* (Cambridge university press, 2009).

203. Colombo, D. & Maathuis, M. H., Order-Independent Constraint-Based Causal Structure Learning, *J. Mach. Learn. Res.* **15,** 3921–3962 (2014).

204. Glymour, C., Zhang, K. & Spirtes, P., Review of Causal Discovery Methods Based on Graphical Models, *Front. Genet.* **10** (2019).

205. Schölkopf, B. *et al.,* On causal and anticausal learning, *arXiv preprint,* arXiv:1206.6471 (2012).

206. Kano, Y. & Shimizu, S., Causal inference using nonnormality, *In Proceedings of the International Symposium on Science of Modeling, the 30th Anniversary of the Information Criterion,* 261–270 (2003).

207. Shimizu, S., Hoyer, P. O., Hyvärinen, A. & Kerminen, A. J., A linear non-gaussian acyclic model for causal discovery, *J. Mach. Learn. Res.* **7,** 2003–2030 (2006).

208. Hyvärinen, A., Zhang, K., Shimizu, S. & Hoyer, P. O., Estimation of a Structural Vector Autoregression Model Using Non-Gaussianity, *J. Mach. Learn. Res.* **11,** 1709–1731 (2010).

209. Bloebaum, P., Janzing, D., Washio, T., Shimizu, S. & Schoelkopf, B., Cause-Effect Inference by Comparing Regression Errors, *In Proceedings of the International Conference on Artificial Intelligence and Statistics. PMLR* **84,** 900–909 (2018).

210. Maeda, T. N. & Shimizu, S., RCD: Repetitive causal discovery of linear non-Gaussian acyclic models with latent confounders, *In Proceedings of the International Conference on Artificial Intelligence and Statistics, PMLR* **108,** 735–745 (2020).

211. Inoue, K., Ribeiro, T. & Sakama, C., Learning from interpretation transition, *Mach. Learn.* **94** (2014).

212. Yu, Y., Chen, J., Gao, T. & Yu, M., DAG-GNN: DAG Structure Learning with Graph Neural Networks, *In Proceedings of the International Conference on Machine Learning* (2019).

213. Zhu, S., Ng, I. & Chen, Z., Causal Discovery with Reinforcement Learning, *In Proceedings of the International Conference on Learning Representations* (2020).

214. Kalainathan, D., Goudet, O., Guyon, I., Lopez-Paz, D. & Sebag, M., Structural Agnostic Modeling: Adversarial Learning of Causal Graphs, *J. Mach. Learn. Res.* **23,** 1–62 (2022).

215. Hägele, A. *et al.,* BaCaDI: Bayesian Causal Discovery with Unknown Interventions, *In Proceedings of the International Conference on Artificial Intelligence and Statistics, PMLR* **206,** 1411–1436 (2023).

216. Mooji, J. M., Magliacane, S. & Claassen, T., Joint Causal Inference from Multiple Contexts, *J. Mach. Learn. Res.* **21,** 1–108 (2020).

217. Squires, C., Wang, Y. & Uhler, C., Permutation-Based Causal Structure Learning with Unknown Intervention Targets, *In Proceedings of the Conference on Uncertainty in Artificial Intelligence, PMLR* **124** (2020).

218. Brouillard, P., Lachapelle, S., Lacoste, A., Lacoste-Julien, S. & Drouin, A., Differentiable Causal Discovery from Interventional Data, *In Proceedings of the Conference on Neural Information Processing Systems* (2020).

219. Zhang, B. & Horvath, S., A general framework for weighted gene coexpression network analysis, *Stat. Appl. Genet. Mol. Biol.* **4** (2005).

220. Margolin, A. A. *et al.,* ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context, *BMC Bioinform.* **7,** S7 (2006).

221. Faith, J. J. *et al.,* Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles, *PLOS Biol.* **5,** e8 (2007).

222. Meyer, P. E., Kontos, K., Lafitte, F. & Bontempi, G., Information-theoretic inference of large transcriptional regulatory networks, *EURASIP J Bioinform. Sys. Biol.* **1,** 879 (2007).

223. Haury, A. C., Mordelet, F., Vert, J. P. & Vera-Licona, P., TIGRESS: trustful inference of gene regulation using stability selection, *BMC Sys. Biol.* **6,** 145 (2012).

224. Huynh-Thu, V. A., Irrthum, A., Wehenkel, L. & Geurts, P., Inferring regulatory networks from expression data using tree-based methods, *PLOS ONE* **5,** e12, 776 (2010).

225. Schafer, J. & Strimmer, K., An empirical bayes approach to inferring large scale gene association networks, *Bioinform.* **21,** 754–764 (2004).

226. Hill, S. M. *et al.,* Bayesian inference of signaling network topology in a cancer cell line, *Bioinform.* **28,** 2804–2810 (2012).

227. Thorne, T. & Stumpf, M. P., Inference of temporally varying Bayesian Networks, *Bioinform.* **28,** 3298–3305 (2012).

228. Wang, Y., Joshi, T., Zhang, X. S., Xu, D. & Chen, L., Inferring gene regulatory networks from multiple microarray datasets, *Bioinform.* **22,** 2413–2420 (2006).

229. Martinelli, J., Grignard, J., Soliman, S. & Fages, F., A Statistical Unsupervised Learning Algorithm for Inferring Reaction Networks from Time Series Data, *ICML Workshop on Computational Biology,* `https://hal.archives-ouvertes.fr/hal-02163862` (2019).

230. Bridewell, W., Langley, P., Todorovski, L. & Džeroski, S., Inductive process modeling, *Mach. Learn.* **71,** 1–32 (2008).

231. Heirendt, L., Arreckx, S., Pfau, T., *et al.,* Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0, *Nat. Protoc.* **14,** 639–702 (2019).

232. Lavine, J. S., Bjornstad, O. N. & Antia, R., Immunological characteristics govern the transition of COVID-19 to endemicity, *Science* **371,** 741–745, ISSN: 0036-8075 (2021).

233. Petersen, E. *et al.,* Comparing SARS-CoV-2 with SARS-CoV and influenza pandemics, *Lancet Infect. Dis.* **20,** e238–e244, ISSN: 1473-3099 (2020).

234. Letko, M., Marzi, A. & Munster, V., Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses, *Nat. Microbiol.* **5,** 562–569 (2020).

235. Guan, W. *et al.,* Clinical characteristics of coronavirus disease 2019 in China, *N. Engl. J. Med.* **382,** 1708–1720 (2020).

236. Nishiura, H. *et al.,* The Rate of Underascertainment of Novel Coronavirus (2019-nCoV) Infection: Estimation Using Japanese Passengers Data on Evacuation Flights, *J. Clin. Med.* **9,** ISSN: 2077-0383 (2020).

237. Miura, M. *et al.,* Kinetics of HTLV-1 reactivation from latency quantified by single-molecule RNA FISH and stochastic modelling, *PLOS Pathog.* **15,** e1008164 (2019).

238. Garnett, G. & Grenfell, B., The epidemiology of varicella–zoster virus infections: A mathematical model, *Epidemiol. Infect.* **108,** 495–511 (1992).

239. Eissing, T. *et al.,* A computational systems biology software platform for multiscale modeling and simulation: integrating whole-body physiology, disease biology, and molecular reaction networks, *Front. Physiol.* **2,** 4 (2011).

240. Hoops, S., Sahle, *et al.,* COPASI–a COmplex PAthway SImulator, *Bioinform.* **22,** 3067–3074 (2006).

241. Hajiramezanali, E., Hasanzadeh, A., Duffield, N., Narayanan, K. & Qian, X., BayReL: Bayesian Relational Learning for Multi-omics Data Integration, *In Proceedings of the Conference on Neural Information Processing Systems* **33,** 19251–19263 (2020).

242. Wagner, A., Wang, C., Fessler, J., *et al.,* Metabolic modeling of single Th17 cells reveals regulators of autoimmunity, *Cell* **184,** 4168–4185.e21 (2021).

243. Deisboeck, T. S., Wang, Z., Macklin, P. & Cristini, V., Multiscale cancer modeling, *Annu. Rev. Biomed. Eng.* **13,** 127–155 (2011).

244. Klamt, S., Saez-Rodriguez, J. & Gilles, E. D., Structural and functional analysis of cellular networks with CellNetAnalyzer, *BMC Syst. Biol.* **1** (2007).

245. Mooij, J. M., Janzing, D. & Schölkopf, B., From Ordinary Differential Equations to Structural Causal Models: The Deterministic Case, *In Proceedings of the Conference on Uncertainty in Artificial Intelligence* **9,** 440–448 (2013).

246. Rubenstein, P. K., Bongers, S., Schölkopf, B. & Mooij, J. M., From Deterministic ODEs to Dynamic Structural Causal Models, *In Proceedings of the Conference on Uncertainty in Artificial Intelligence,* 114–123 (2018).

247. Clarke, E., McMillan, K., Campos, S. & Hartonas-Garmhausen, V., Symbolic model checking, *In Proceedings of the International Conference on Computer Aided Verification,* 419–422 (1996).

248. Maiwald, T. & Timmer, J., Dynamical Modeling and Multi-Experiment Fitting with PottersWheel, *Bioinform.* **24,** 2037–2043 (2008).

249. Vuuren, D. P. v. *et al.,* A comprehensive view on climate change: coupling of earth system and integrated assessment models, *Environ. Res. Lett.* **7,** 024012 (2012).

250. Bond-Lamberty, B. *et al.,* Coupling earth system and integrated assessment models: the problem of steady state, *Geosci. Model Dev. Discuss.* **7,** 1499–1524 (2014).

Titre : Modélisation multi-échelle de la dynamique virale basée sur l'analyse des données et des connaissances

Mots clés : Modélisation et simulation, Analyse numérique, Transmission de cellule à cellule, Inférence de réseaux de gènes, Construction de voies de signalisation, COVID-19

Résumé : Le système intra-hôte qui est à la base de l'infection est un système complexe composé d'éléments interconnectés à la même échelle de la hiérarchie biologique et à différentes échelles de la hiérarchie. Ces liens forment des structures de réseau sous forme de corrélations ou de relations causales entre les composants, qui peuvent être calculées en intégrant des données d'observation et des connaissances de base. Dans cette recherche, compte tenu de la demande sociale d'acquisition de connaissances sur le mécanisme et les stratégies de contrôle de l'infection concernant le problème global du COVID-19, nous visons à découvrir de nouvelles connaissances sur son virus pathogène, le SARS-CoV-2, par la modélisation et l'exploration de données des systèmes d'infection virale aux échelles macroscopique et microscopique. Plus spécifiquement, nous construisons d'abord une hypothèse basée sur la simulation de la dynamique de transmission de cellule à cellule du virus en modélisant la dynamique de la population virale avec des équations différentielles. Pour vérifier cette hypothèse à l'échelle microscopique, nous estimons ensuite le réseau de gènes à partir de données omiques unicellulaires et de multiples graphes de connaissances, en nous concentrant sur une molécule d'adhésion intercellulaire. En conséquence, nous découvrons des voies de signalisation inconnues, absentes de la base de connaissances existante sur COVID-19. Enfin, cette recherche contribue à la découverte scientifique en exploitant les données et les connaissances dans la modélisation multi-échelle et l'exploration des données de la dynamique virale de la transmission de cellule à cellule du SARS-CoV-2.

Title: Data-Driven and Knowledge-Based Multiscale Modeling of Viral Dynamics

Keywords: Modeling and simulation, Numerical analysis, Cell-to-cell transmission, Gene network inference, Signaling pathway construction, COVID-19

Abstract: The within-host system underlying infection is a complex system consisting of components interconnected at the same scale of the biological hierarchy and different scales across the hierarchy. Such linkages form network structures as correlations or causal relationships among components, which can be computed by integrating observational data and background knowledge. In this study, considering the social demand for knowledge acquisition on the mechanism and infection control strategies regarding the global issue of COVID-19, we aim to discover novel knowledge about its pathogenic virus, SARS-CoV-2, through modeling and data mining of viral infection systems at both macroscopic and microscopic scales. Specifically, we first construct a simulation-based hypothesis on viral cell-to-cell transmission dynamics by modeling viral population dynamics with differential equations. To verify this hypothesis at the microscopic scale, we subsequently estimate the gene network from single-cell omics data and multiple knowledge graphs, focusing on an intercellular adhesion molecule. As a result, we discover unknown signaling pathways missing from the existing knowledge base on COVID-19. Overall, this study contributes to scientific discovery by harnessing data and knowledge in multiscale modeling and data mining of the viral dynamics of SARS-CoV-2 cell-to-cell transmission.