Leveraging Structured Biological Knowledge for Counterfactual Inference: A Case Study of Viral Pathogenesis

Jeremy Zucker, Kaushal Paneri, Sara Mohammad-Taheri, Somya Bhargava, Pallavi Kolambkar, Craig Bakker, Jeremy Teuton, Charles Tapley Hoyt, Kristie Oxford, Robert Ness, and Olga Vitek

Abstract—Counterfactual inference is a useful tool for comparing outcomes of interventions on complex systems. It requires us to represent the system in form of a structural causal model, complete with a causal diagram, probabilistic assumptions on exogenous variables, and functional assignments. Specifying such models can be extremely difficult in practice. The process requires substantial domain expertise, and does not scale easily to large systems, multiple systems, or novel system modifications. At the same time, many application domains, such as molecular biology, are rich in structured causal knowledge that is qualitative in nature. This article proposes a general approach for querying a causal biological knowledge graph, and converting the qualitative result into a quantitative structural causal model that can learn from data to answer the question. We demonstrate the feasibility, accuracy and versatility of this approach using two case studies in systems biology. The first demonstrates the appropriateness of the underlying assumptions and the accuracy of the results. The second demonstrates the versatility of the approach by querying a knowledge base for the molecular determinants of a severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)-induced cytokine storm, and performing counterfactual inference to estimate the causal effect of medical countermeasures for severely ill patients.

Index Terms—Biological expression language, structural causal model, counterfactual inference, causal biological knowledge graph, systems biology, SARS-CoV-2

1 Introduction

Each time a cell senses changes in its environment, it marshals a complex choreography of molecular interactions to initiate an appropriate response. When a virus infects the cell, this delicate balance is disrupted and can result in a cascade of systemic failures leading to disease. In particular, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the novel pathogen responsible for the COVID-19 pandemic, has a complex etiology that differs in subtle and substantial ways from previously studied viruses. To make informed decisions about the risk that a new pathogen presents, it is imperative to rapidly predict the determinants of pathogenesis and identify potential targets for medical countermeasures. Current solutions for this task include systems biology data-driven models, which correlate biomolecular expression to pathogenicity, but cannot go beyond associations in the data to reason about causes of the disease [1], [2]. Alternatively, hypothesis-driven mathematical models capture causal relations, but are hampered by limited parameter identifiability and predictive power [3], [4].

We argue that counterfactual inference [5] helps bridge the gap between data-driven and hypothesis-driven approaches. It enables questions of the form: "Had we known the eventual outcome of a patient, what would we have done differently?" At the heart of counterfactual inference is a formalism known as a structural causal model [5], [6]. It represents prior domain knowledge in terms of causal diagrams, assumes a probability distribution on exogenous variables, and assigns a deterministic function to endogenous variables. SCM are particularly attractive in systems biology, where structured domain knowledge is extracted from the biomedical literature and is readily available through advances in natural language processing [7], [8], [9], large-scale automated assembly systems [10], and semi-automated curation workflows [11]. This knowledge is curated by multiple organizations [12], [13], [14], [15], [16] and stored in structured knowledge bases [17], [18], [19], [20]. It can be brought to bear for answering causal questions regarding SARS-CoV-2.

This manuscript contributes a three-part algorithm that leverages existing structured biological knowledge to answer counterfactual questions about viral pathogenesis. Algorithm 1 formalizes biologically relevant questions as

- *Jeremy Zucker, Craig Bakker, Jeremy Teuton, Kristie Oxford are with the Pacific Northwest National Laboratory, Richland, WA 99354 USA. E-mail: {jeremy.zucker, craig.bakker, Jeremy.Teuton, kristie.oxford}@pnnl.gov.*
- *Kaushal Paneri is with the Microsoft, Redmond, WA 98052 USA. E-mail: kaushalpaneri@gmail.com.*
- *Sara Mohammad-Taheri, Somya Bhargava, Pallavi Kolambkar, and Olga Vitek are with the Northeastern University, Boston, MA 02115 USA. E-mail: {mohammadtaheri.s, o.vitek}@northeastern.edu, bhargavasomyav2@gmail.com, kolambkar.p@husky.neu.edu.*
- *Charles Tapley Hoyt is with the Enveda Biosciences, 53225 Bonn, Germany. E-mail: charles.hoyt@envedatx.com.*
- *Robert Ness is with the Altdeep, Boston, MA 02115 USA. E-mail: robertness@gmail.com.*

Manuscript received 17 July 2020; revised 11 Nov. 2020; accepted 14 Dec. 2020. Date of publication 18 Jan. 2021; date of current version 1 Mar. 2021. (Corresponding author: Olga Vitek.)
Recommended for acceptance by the Guest Editors of the Special Issue On AI for COVID-19.
Digital Object Identifier no. 10.1109/TBDATA.2021.3050680

queries to an existing causal knowledge graph. Algorithm 2 converts the query result into a structural causal model. Algorithm 3 operationalizes the counterfactual inference by interrogating the model with the observed data to estimate a causal effect.

We illustrate the benefits of this approach using two case studies. Case study 1 illustrates the increased precision of counterfactual estimates, as compared to the ODE- and SDE-based forward simulation, in a situation with known ground truth mechanisms of data generation. Case study 2 demonstrates the automated construction of an SCM and the value of counterfactual reasoning in novel situations with limited treatment options (as is the case for SARS-CoV-2). It shows that counterfactual inference enables more precise predictions regarding who would be likely to survive without receiving treatment, who would be likely to die even if they did receive treatment, and who would likely survive only if they received treatment.

2 Background

Biological Signaling Pathways. Signaling pathways are composed of entities that engage in activities [21]. Examples of entities are proteins and metabolites, but also higher level biological processes such as an immune response. Activities are the producers of change. Examples include catalytic activity, kinase activity, or transcriptional activity.

The basic unit of causality in signaling pathways is a directed molecular interaction, where the activity of an upstream molecule increases or decreases the activity of a downstream molecule. For example, the mitogen-activated protein kinase (MAPK) intracellular signaling pathway is a causal chain of directed molecular interactions shown in eq. (1)

$$a(S_1) \rightarrow kin(p(Raf)) \rightarrow kin(p(Mek)) \rightarrow kin(p(Erk)). \quad (1)$$

The interactions transmit information about a stimulus at the cell surface to the nucleus, where proteins called transcription factors activate an appropriate biological process [22]. A causal diagram of MAPK consists of a signaling molecule S_1 and three proteins Raf, Mek, and Erk, each of which engage in kinase activity. We represent signaling molecule abundance with $a()$, protein abundance with $p()$ and the kinase activity of a protein with $kin()$. In the case of MAPK, the abundance or activity of an upstream entity causes the abundance or activity of a downstream entity to increase, and is represented with a sharp edge $\rightarrow$. The diagram is a abstraction showing that the abundance of the signaling molecule S_1 increases the kinase activity of Raf, which increases the kinase activity of Mek, which increases the kinase activity of Erk. In other cases, if the abundance or activity of an upstream entity causes the abundance or activity of a downstream entity to decrease, we represent this with a blunt edge.

Viral Dysregulation. Viral disruptions of a signaling pathway take form of overactivation or repression of its activities. For example, by amplifying the release of intercellular signaling molecules that overstimulate the immune response, known as Cytokine Release Syndrome (cytokine storm), a virus can cause severe system-level cellular damage.

Quantitative Modeling of Biological Processes With ODE/SDE. Temporal dynamics of biological processes can be expressed quantitatively using ordinary (or stochastic) differential equations. A small number of high quality, validated models have been published in the literature and stored in a computable form in repositories such as Biomodels [23], [24]. For example, the MAPK signaling pathway in eq. (1) is well characterized. We denote $R(t)$, $M(t)$, and $E(t)$ as the respective amounts of active Raf, Mek, and Erk at time t; We denote T_R, T_M, and T_E as their total amounts, which we assume do not change during the considered timeframe; v_R^{act}, v_R^{inh}, v_M^{act}, v_M^{inh}, v_E^{act}, and v_E^{inh} are experimentally derived activation or inhibition kinetic rate constants; and S_1 is the amount of the input signal. The system of ordinary differential equations (ODEs) is specified as follows [25], [26]:

$$\begin{aligned}
\frac{\mathrm{d}R}{\mathrm{d}t} &= v_R^{\text{act}} S_1 (T_R - R(t)) - v_R^{\text{inh}} R(t) \\
\frac{\mathrm{d}M}{\mathrm{d}t} &= \frac{(v_M^{\text{act}})^2}{v_M^{\text{inh}}} R(t)^2 (T_M - M(t)) - v_M^{\text{act}} R(t) M(t) - v_M^{\text{inh}} M(t) \\
\frac{\mathrm{d}E}{\mathrm{d}t} &= \frac{(v_E^{\text{act}})^2}{v_E^{\text{inh}}} M(t)^2 (T_E - E(t)) - v_E^{\text{act}} M(t) E(t) - v_E^{\text{inh}} E(t).
\end{aligned} \quad (2)$$

Given initial conditions, forward simulations from the ODEs can be used to generate the temporal trajectories of the amounts of activated proteins , such as $R(t)$, $M(t)$, and $E(t)$ in the MAPK example. In this manuscript we refer to such simulated data as *observational data*. We define an *ideal intervention* as an event that fixes the amount of an activated protein. For example, if we fix the kinase acivity of Mek at $M(t) = m$, the second equality $\frac{\mathrm{d}M}{\mathrm{d}t}$ in eq. (2) becomes zero. We can simulate data from eq. (2) with $\frac{\mathrm{d}M}{\mathrm{d}t} = 0$, and refer to these as *interventional data*. Contrasting observational and interventional data helps evaluate the outcome of the intervention [27].

The deterministic ODE ignore the fact that at low concentration, stochasticity becomes a significant factor in determining the reaction [28]. As the collisions between molecules participating in biochemical process become stochastic, a stochastic model is required. In contrast to ODE, a stochastic differential equation model or stochastic differential equation (SDE) specifies biological process as a random process. For example, in the case of MAPK, the random process of the reaction $Mek \rightarrow Erk$ is specified with

$$\frac{dP_E(t)}{dt} = g_E(t, v_E^{act}, v_E^{inh}, M(t)), \; E(0) = e_0 \quad (3)$$

where $P_E(t)$ is marginal probability density of $E(t)$, function g_E determines the probability of a state change between $E(t)$ and $E(s), s > t$, e_0 is initial condition, and $M(t)$ is the value of its parent Mek at t. Once stochastic differential equation are fully specified, one can use, e.g., Gillespie's stochastic simulation algorithm [29] to simulate observational and interventional data, and evaluate the outcomes of interventions.

Unfortunately, even simple ODEs such as the one in the MAPK example are difficult to build *de novo*. This is nearly impossible for novel and poorly studied systems that lack

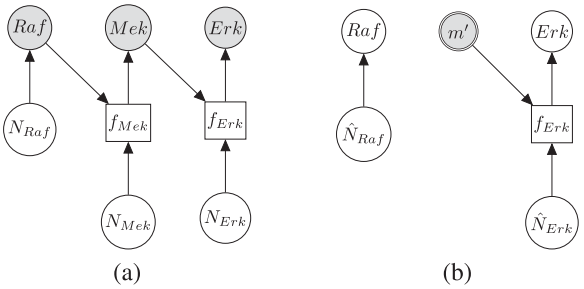

Fig. 1. *Causal modeling of MAPK signaling pathway.* Circles are variables, double circles are variables intervened upon, squares are deterministic functional assignments, gray nodes are observed variables, and white nodes are hidden variables. (a) Structural causal model. N_{Raf}, N_{Mek} and N_{Erk} are statistically independent noise variables. Root node Raf is only dependent on noise variable N_{Raf}. Non-root nodes Mek and Erk are dependent on their parent and on the associated noise variable. (b) Counterfactual model. The intervention fixes the count of phosphorylated Mek to m', such that Mek is no longer dependent on Raf and N_{Mek}. Given an observed data point, counterfactual inference infers the noise variables $\hat{N}_{Raf}$, and $\hat{N}_{Erk}$.

the existence or findability of experimental information describing the structure or boundaries of the process, kinetic equations governing their dynamics [30], rate constants for these equations, or rules governing each agents' states and functions.

Equilibrium Enzyme Kinetics. Simpler and more general quantitative models can be specified when a reaction reaches the state of chemical equilibrium [31]. One commonly used such model is *Hill function* in the form of

$$X = \beta \frac{\mathbf{PA}_X^n}{K^n + \mathbf{PA}_X^n}, \tag{4}$$

where X is the abundance of a protein in a causal diagram (such as Erk in eq. (1)), $\mathbf{PA}_X$ is the set of its parents, n is a parameter interpreted as the number of ligand binding sites of the protein, and β is the total number of molecules of the protein. A special and frequently used case of the Hill function, called *Michaelis-Menten* function, occurs when $n = 1$. Although simple to use, these models are deterministic, and do not describe the stochasticity that is a distinctive property of biological systems at low concentrations.

Modeling Biological Processes With Structural Causal Models. The stochastic nature of biological processes at steady-state can be represented by an SCM such as in Fig. 1a [27], [32]. SCMs represent the dependencies between a child node X and its parents $\mathbf{PA}_X$ in terms of a deterministic function $X = f_X(\mathbf{PA}_X, N_X)$ called *structural assignment*, and a noise variable N_X. In Fig. 1a, f_{Mek} and f_{Erk} are linear or non-linear structural assignments, and N_{Raf}, N_{Mek}, and N_{Erk} are statistically independent noise variables with defined probability distributions

$$\begin{aligned} &Raf = N_{Raf};\ \ Mek = f_{Mek}(Raf; N_{Mek}) \\ &Erk = f_{Erk}(Mek, N_{Erk}), \end{aligned} \tag{5}$$

An ideal intervention in an SCM is performed on a functional assignment. For example, an ideal intervention on Mek sets $Mek = m'$, defining a new SCM

$$Raf = N_{Raf}; Mek = m'; Erk = f_{Erk}(Mek, N_{Erk}). \tag{6}$$

An ideal intervention can also be thought of as a process of mutilating the causal graph. For example, intervening on Mek eliminates its dependence upon Raf, and therefore the edge from Raf to Mek is removed as shown in Fig. 1b.

Counterfactual Inference With SCM. Beyond direct model-based predictions, SCMs enable *counterfactual inference*, i.e., the process of inferring the unseen outcomes of a hypothetical intervention given data observed in absence of the intervention [5]. In the context of SCM, counterfactuals are defined as operations

$$Y_{do(T=t')}(u) \triangleq Y_{M_{do(T=t')}}(u), \tag{7}$$

In other words, the outcome Y that individual u would have had she received treatment t' is defined as the value that Y would have in a structural causal model M mutilated to replace $T = f_T(\cdot)$ with $T = t'$.

For example, in the MAPK signaling pathway, we may be interested in the counterfactual question: *Having observed the kinase activities of $Raf = r$, $Mek = m$, $Erk = e$, what would be the kinase activity of Erk in a hypothetical experiment where the kinase activity of Mek was fixed to m'?* This counterfactual query is more formally translated into

$$P(Erk_{do(Mek=m')} | Raf = r, Mek = m, Erk = e). \tag{8}$$

The probability distribution in eq. (8) is estimated with the following steps:

1) *Abduction:* Given observational data, estimate the posterior distribution of the noise variables. In the MAPK example, we estimate the posterior distribution of the noise variables:

$$\hat{N}_{Raf} = \{N_{Raf} | Raf = r, Mek = m, Erk = r\}$$
$$\hat{N}_{Erk} = \{N_{Erk} | Raf = r, Mek = m, Erk = r\}$$

   Several inference algorithms are available for this task, e.g., Markov Chain Monte Carlo [33], Gibbs sampling [34], or no-u-turn Hamiltoninan Monte Carlo (HMC) [35]. In recent years, gradient-based inference algorithms such as stochastic variational inference [36] have become popular, because they can scale to larger models by converting an inference problem into an optimization problem.
2) *Intervention:* Apply the intervention to the SCM to generate a mutilated SCM as in Fig. 1b. In the MAPK SCM, $Mek = f_{Mek}(Raf, N_{Mek})$ is replaced with $Mek = m'$ as shown in Fig. 1b.
3) *Prediction:* Generate samples from the mutilated SCM using the estimated posterior distribution over the exogenous variables $\hat{N}_{Raf}$ and $\hat{N}_{Erk}$ to obtain the counterfactual distribution, as shown in Fig. 1b.

Causal Effects. We distinguish between two causal effects. The first is the average treatment effect (ATE), defined as the difference between the outcome of a hypothetical intervention and the observed outcome in the entire population. In the MAPK example, the ATE of Erk upon an intervention

fixing $Raf = r'$ is:

$$\{Erk_{do(Raf=r')} - Erk\}. \tag{9}$$

This requires no observational data, and therefore the ATE can be inferred with forward simulation.

On the other hand, the individual treatment effect (ITE) is defined as the difference between the outcome of a hypothetical intervention and the observed outcome for a specific individual or context. In the MAPK example, the individual treatment effect of Erk upon an intervention fixing $Raf = r'$ in a context where $Raf = r$, $Mek = m$, $Erk = e$ is:

$$\{Erk_{do(Raf=r')} - Erk\}|Raf = r, Mek = m, Erk = e \tag{10}$$

The ITE shares stochastic components of the noise variables between observational and interventional data, and is therefore often more precise than a comparison based on a direct simulation [27].

In cases where domain knowledge is available to describe the systems dynamics in the form of an SDE, the system at equilibrium can be translated into an SCM to enable counterfactual reasoning and estimation of the individual treatment effect [27], [37]. Unfortunately, this process is challenging in novel and poorly studied systems, due to our limited ability to establish the structure of the causal graph.

Structured Knowledge Graphs. Although there exist a multitude of biological knowledge bases that are manually curated from the literature [12], [13], [14], [15], [16], the systems biology community has coalesced around a small number of structured knowledge representations that differ mainly in their intended purpose. For example, the Biological Pathway Exchange Language (BioPAX) was designed for pathway database integration [17], and the Systems Biology Graphical Notation (SBGN) was designed for graphical layout [19].

In contrast, the Biological Expression Language (BEL) was specifically designed for manual extraction and automated integration of author statements about causal relationships among biological entities, biological processes, and cellular-level observable phenomena [11]. The syntax of a BEL statement is comprised of a triple in the form of {*subject*, *predicate*, *object*}. Each subject and object represents an activity or abundance whose entities are grounded using terms from standard namespaces. If the subject directly increases the abundance or the activity of the object, we represent this with `=>`, and for directly decreasing relationships, we use `=|`. BEL statements can be chained together from the object of the first statement to the subject of the next statement, as shown in Fig. 2 for the case of the MAPK pathway.

BEL provides a number of valuable features for causal modeling. First, the restriction of BEL edges to causal relations implies the topology of the BEL graph can be reflected in the topology of the causal model. Second, the language is expressive enough for humans to manually curate a wide range of biological concepts, but formal enough to serve as a training corpus for corpus for natural language processing of biomedical literature competitions [38]. Third, the BEL ecosystem is sufficiently mature that causal knowledge represented in other languages can be readily converted to BEL [39], [40].

```
kin(p(fplx:RAF)) => kin(p(fplx:MEK))
kin(p(fplx:MEK)) => kin(p(fplx:ERK))
```

Fig. 2. *Example BEL statement.* The statement details the processes in the MAPK signaling pathway in eq. (1). The first line states that the kinase activity of RAF directly increases the kinase activity of MEK. The second line states that kinase activity of MEK directly increases the kinase activity of ERK.

3 Methods

3.1 Notation, Definitions, and Assumptions

Let $\mathbf{X} = \{X_i\}$ be a set of variables, such as molecular activities in a signaling pathway. Let $\mathbf{P} = \{P_j\}$ be a set of causal predicates that link these variables, such as increases, or regulates. Using this notation, we define a knowledge graph $\mathbb{K}$ as a set of k triples

$$\mathbb{K} = \{X_i, P_j, X_{i'} \,|X_i \in \mathbf{X}, P_j \in \mathbf{P}, X_{i'} \in \{\mathbf{X} \setminus X_i\}\}_{j=1}^{k}. \tag{11}$$

We define a causal query $\mathbb{Q}$ as a set $\{\mathbf{X}^{\mathrm{c}},\ \mathbf{X}^{\mathrm{e}},\ \mathbf{X}^{\mathrm{z}}\}$ of variables that are potential causes, effects and covariates of interest for the biological investigation, where

$$\mathbf{X}^{\mathrm{c}} \subset \mathbf{X},\ \ \mathbf{X}^{\mathrm{e}} \subset \mathbf{X}\backslash\mathbf{X}^{\mathrm{c}},\ \text{and}\ \mathbf{X}^{\mathrm{z}} \subset \mathbf{X}\backslash\mathbf{X}^{\mathrm{c}}\backslash\mathbf{X}^{\mathrm{e}}.$$

A pathway $\mathbb{P}(X_1, X_{k'+1})$, $k \leq k'$ is a sequence of a subset of triples from $\mathbb{K}$, where the object of the previous triple is subject of the next triple

$$\{(X_1, P_1, X_2), (X_2, P_2, X_3), \ldots, (X_{k'}, P_{k'}, X_{k'+1})\}. \tag{12}$$

Our goal is to query the knowlege graph to generate a qualitative causal model $\mathbb{B}$ that links the causes, the effects and the covariates of interest. Importantly, the query result $\mathbb{B}$ induces a directed acyclic graph G with p variables from $\mathbf{X}$ as nodes, and causal relations from $\mathbf{P}$ as edges.

We assume that every variable in $\mathbb{B}$ is continuous. We denote $\mathbb{D} = \{X_{1j}, X_{2j}, \ldots, X_{pj}\}_{j=1}^{m}$ the observational data of m samples from the joint distribution $\mathcal{P}(\mathbf{X}; \theta)$. The distribution is specified in terms of parameters θ. We denote $\mathbf{R} \subset \mathbf{X}$ a set of nodes in G without parents.

3.2 Querying a Knowledge Graph to Obtain a Qualitative Causal Model

Given a biological knowledge graph $\mathbb{K}$ and a causal query of interest $\mathbb{Q}$, our first objective is to generate a qualitative causal model $\mathbb{B}$ capable of answering the query. To this end, we need to explore all potential directed acyclic paths in $\mathbb{K}$ from the cause to the effect in $\mathbb{Q}$, and then consider all covariates that may act as confounders of the causal question. This is done with the steps in Algorithm 1. The algorithm can be implemented on any knowledge graph that represents causal relationships as directed edges, such as BEL or the Systems Biology Graphical Notation Activity Flow (SBGN-AF) language [41].

In the case of MAPK, the qualitative causal model that is capable of answering the counterfactual question in eq. (8) corresponds to the result of this query: $\mathbb{Q} = \{\mathbf{X}^{\mathrm{c}} = kin(p(\mathrm{MEK})), \mathbf{X}^{\mathrm{e}} = kin(p(\mathrm{ERK})), \mathbf{X}^{\mathrm{z}} = kin(p(\mathrm{RAF}))\}$.

Algorithm 1. *Causal query to Biological Expression Language* (QUERY2BEL) *algorithm*

Inputs: knowledge graph $\mathbb{K}$
causal query $\mathbb{Q} = \{\mathbf{X}^{\mathrm{c}},\ \mathbf{X}^{\mathrm{e}},\ \mathbf{X}^{\mathrm{z}}\}$
Outputs: $\mathbb{B}$

1: **procedure** QUERY2BEL($\mathbf{X}^{\mathrm{c}}, \mathbf{X}^{\mathrm{e}}, \mathbf{X}^{\mathrm{z}}, \mathbb{K}$)
2: ▶ *Get all pathways from cause to effect*
3: **for each** cause $X_i^{\mathrm{c}} \in \mathbf{X}^{\mathrm{c}}$ and for each effect $X_j^{\mathrm{e}} \in \mathbf{X}^{\mathrm{e}}$ **do**
4: find all pathways $\{\mathbb{P}(X_i^{\mathrm{c}}, X_j^{\mathrm{e}})\}$
5: ▶ *Get all pathways from covariates to causes*
6: **for each** covariate $X_i^{\mathrm{z}} \in \mathbf{X}^{\mathrm{z}}$ and for each cause $X_j^{\mathrm{c}} \in \mathbf{X}^{\mathrm{c}}$ **do**
7: find all pathways $\{\mathbb{P}(X_i^{\mathrm{z}}, X_j^{\mathrm{c}})\}$
8: ▶ *Get all pathways from covariates to effects*
9: **for each** covariate $X_i^{\mathrm{z}} \in \mathbf{X}^{\mathrm{z}}$ and for each effect $X_j^{\mathrm{e}} \in \mathbf{X}^{\mathrm{e}}$ **do**
10: find all pathways $\{\mathbb{P}(X_i^{\mathrm{z}}, X_j^{\mathrm{e}})\}$
11: $\mathbb{B} = \{\mathbb{P}(X_i^{\mathrm{c}}, X_j^{\mathrm{e}})\} \cup \{\mathbb{P}(X_i^{\mathrm{z}}, X_j^{\mathrm{c}})\} \cup \{\mathbb{P}(X_i^{\mathrm{z}}, X_j^{\mathrm{e}})\}$
12: **return** $\mathbb{B}$

We execute Algorithm 1 step 2 to obtain all pathways from the cause to the effect:

$$kin(p(\mathrm{MEK})) \rightarrow kin(p(\mathrm{ERK})).$$

We execute Algorithm 1 step 6 to obtain all pathways from the covariate to the cause:

$$kin(p(\mathrm{RAF})) \rightarrow kin(p(\mathrm{MEK})).$$

We execute Algorithm 1 step 10, but since there are no new pathways from the covariate $kin(p(\mathrm{RAF}))$ to the effect $kin(p(ERK))$, we obtain the empty set. The final returned model is:

$$kin(p(\mathrm{RAF})) \rightarrow kin(p(\mathrm{MEK})) \rightarrow kin(p(\mathrm{ERK})).$$

3.3 Compiling a Qualitative Causal Model to a Quantitative Structural Causal Model

Our second objective is to express the qualitative causal structure in $\mathbb{B}$ into a quantitative SCM, and estimate the parameters of the SCM from experimental data. These steps are described in Algorithm 2.

Input. The algorithm takes as input a BEL causal query result $\mathbb{B}$ and observed measurements on its variables $\mathbb{D}$.

Get Network Structure G From $\mathbb{B}$ *(Algorithm 2 Line 3).* Since a set of BEL statements identifies parents and children, it induces a causal network structure. We determine this structure by traversing BEL statements with the breadth first search approach, starting with root variables (such as Raf in Fig. 2). For all the non-root variables, the algorithm waits until all the parents are traversed.

For Each Root Node R, Use $\mathbb{D}$ *to Estimate Parameters* θ *of* $\mathcal{P}(R;\theta)$*(Algorithm 2 Line 5).* In order to specify the SCM, we need to define the type and parameters of the marginal probability distributions of the root variables $\mathcal{P}(R;\theta)$. The BEL statements provide prior knowledge about the distribution in a parametric form. Therefore, this step involves techniques such as maximum likelihood to estimate the parameters of this distribution.

Algorithm 2. *Biological Expression Language to Structural Causal Models* (BEL2SCM) *algorithm*

Inputs: BEL statements $\mathbb{B}$
$\mathbb{D} \sim P(X_1, \ldots, X_p)$
Outputs: *SCM* $\mathbb{M} = \{f_i(\mathbf{PA}_i, N_i)\}_{i=1}^{p}$

1: **procedure** BEL2SCM($\mathbb{B}$, $\mathbb{D}$)
2: $\mathbb{M} = \{\}$
3: Get network structure G from $\mathbb{B}$.
4: **for each** $R \in \mathbf{R}$ in G **do**
5: ▶ *Use* $\mathbb{D}$ *to estimate parameters* θ *of* $\mathcal{P}(R;\theta)$
6: $\theta = arg\,max_\theta \mathcal{P}(R;\theta \mid \mathbb{D})$
7: ▶ *Reparameterize* $\mathcal{P}(R;\theta)$ *in terms of* f_R *and* N_R
8: $N_R \sim \mathcal{N}(0,1)$
9: $f_R(N_R) = F^{-1}_{\mathcal{P}(R;\theta)}(N_R)$
10: $\mathbb{M}$.Add($f_R(N_R)$)
11: **for each** $X \in \{\mathbf{X} \setminus \mathbf{R}\}$ in G **do**
12: ▶ *Estimate parameters* $\mathbf{w}$ *and* b *of sigmoid function*
13: $\log(\frac{X}{\beta_X - X}) = \mathbf{w}'\mathbf{PA}_X + b$
14: ▶ *Define distribution of* N_X *from model residuals.*
15: $residual = X - \frac{\beta_X}{1+\exp(-\mathbf{w}'\mathbf{PA}_X - b)}$
16: $N_X \sim \mathcal{N}(0, MSE(residual))$
17: ▶ *Get* $f_X(\mathbf{PA}_X, N_X)$ *with additive* N_X.
18: $f_X(\mathbf{PA}_X, N_X) = \frac{\beta_X}{1+exp(-\mathbf{w}'_X\mathbf{PA}_X - b_X)} + N_X$
19: $\mathbb{M}$.Add($f_X(\mathbf{PA}_X, N_X)$).
20: **return** $\mathbb{M}$

For example, in a stochastic MAPK system at equilibrium the root variable the number of active Raf in a cell follows a Binomial distribution. When the maximum number of active or inactive particles in the system is large, the Binomial distribution can be approximated with a Normal distribution with $\theta_{Raf} = (\mu_{Raf}, \sigma^2_{Raf})$. We then estimate θ_{Raf} using maximum likelihood from the observed Raf in $\mathbb{D}$.

For Each Root Node R, Reparameterize $\mathcal{P}(R;\theta)$ *in Terms of* f_R *and* N_R *(Algorithm 2 Line 7).* The specification of an SCM requires us to separate the deterministic and the stochastic components of variation of each variable as shown in Fig. 1. We accomplish this using a reparameterization technique popularized by variational autoencoders [42], which was shown to make counterfactual inference consistent with core biological assumptions [43]. In the case of root nodes, we reparameterize $\mathcal{P}(R;\theta)$ with Uniform(0,1), and then pass it to the inverse CDF of $\mathcal{P}(R;\theta)$, as follows

$$\begin{aligned} \text{Original :} \quad & R \sim \mathcal{P}(R;\theta) \\ \text{Reparametrized :} & N_R \sim \text{Uniform}(0,1) \\ & f_R(N_R) = F^{-1}_{\mathcal{P}(R;\theta)}(N_R), \end{aligned} \tag{13}$$

where $F^{-1}_{\mathcal{P}(R;\theta)}(N_R)$ is the inverse cumulative distribution function of $\mathcal{P}(R;\theta)$. In the case of MAPK, since Raf follows a Normal distribution with parameters θ_{Raf}, the reparameterization simplifies even further to

$$\begin{aligned} \text{Original :} \quad & Raf \sim \mathcal{N}(\mu_{Raf}, \sigma^2_{Raf}) \\ \text{Reparametrized :} \quad & N_{Raf} \sim \mathcal{N}(0,1) \\ & f_{Raf}(N_{Raf}) = \sigma_{Raf} N_{Raf} + \mu_{Raf}. \end{aligned} \tag{14}$$

Add R to $\mathbb{M}$ *(Algorithm 2 line 10)* For each root node, we add the corresponding function $f_R(N_R)$ and its noise

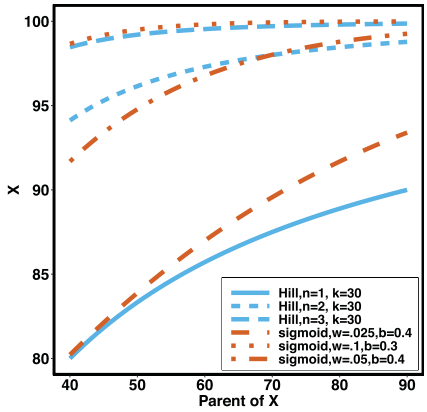

Fig. 3. *Examples of hill function and sigmoid function for two variables.* X is a single node that has a single parent $\mathbf{PA}_X$. We use the Hill function ($X = \beta \frac{\mathbf{PA}_X^n}{K^n + \mathbf{PA}_X^n}$) and sigmoid function as in eq. (15) to predict the value of X given its parent value. In the Hill function, K is the activation rate, n defines the steepness of function and β is fixed at 100. Blue lines correspond to Hill equation with $K = 30$ and $n \in \{1, 2, 3\}$. Brown lines correspond to sigmoid function where $b \in \{0.4, 0.3, 0.4\}$ and $w \in \{0.025, 0.1, 0.5\}$.

variable N_R to $\mathbb{M}$. For example, since MAPK has only one root node Raf, the Algorithm adds $f_{Raf}(N_{Raf})$ to $\mathbb{M}$.

For Each $X \in \{\mathbf{X} \setminus \mathbf{R}\}$, Estimate Parameters $\mathbf{w}$ and b of Sigmoid Function (Algorithm 2 Line 12). In order to specify the SCM for non-root nodes, we need to define the form (polynomial, linear, non-linear, sigmoid, etc.) of functional assignments linking the measurements on the parent nodes to the measurements on the child. We chose the functional assignment in the form of a sigmoid function

$$\log\left(\frac{X}{\beta_X - X}\right) = \mathbf{w}'\mathbf{PA}_X + b, \tag{15}$$

where β_X is the maximum number of activated protein molecules. For a node X with q parents, $\mathbf{PA}_X$ is a $q \times 1$ vector of measurements on the parent nodes, $\mathbf{w}$ is a $1 \times q$ vector of weights, $\mathbf{w}'$ is the transpose of $\mathbf{w}$, and b is a scalar bias. Parameters $\mathbf{w}$ and b of the sigmoid function are estimated from the data, e.g., using smooth L_1 loss function.

In the example of the MAPK pathway, f_{Mek} has only one parent. Therefore f_{Mek} has the form

$$f_{Mek}(Raf, N_{Mek}) = \frac{\beta_{Mek}}{1 + exp(-w_{Mek}Raf - b)} + N_{Mek}. \tag{16}$$

We use the sigmoid function in eq. (15) as a special case of the Hill equation. The full parametric description of the Hill equation has a nuanced precise biochemical interpretation. For example, the parameter n represents the number of times a protein must be phosphorylated before it becomes active and can therefore be obtained from domain knowledge. However, it is difficult to estimate this parameter from data. The sigmoid function maintains the Hill equation's functions, but with a reduced set of parameters that are easier to estimate. Fig. 3 shows that the approximation is reasonable for a range of parameter values.

Define Distribution of N_X From Model Residuals (Algorithm 2 Line 14). Similarly to the root variables, for non-root variables we assume that the noise variables follow Normal distribution with 0 mean. The variance of this distribution is estimated from the residuals of the model fit in the previous step. For example, in the MAPK pathway, f_{Mek} has only one parent Raf. Therefore, the residuals of the sigmoid curve fit for Mek are defined as

$$residual_{Mek} = Mek - \frac{\beta_{Mek}}{1 + \exp(-w_{Mek}Raf - b)}, \tag{17}$$

and the distribution of the noise variable is defined as $N_{Mek} \sim \mathcal{N}(0, MSE(residual_{Mek}))$

Get $f_X(\mathbf{PA}_X, N_X)$ With Additive N_X (Algorithm 2 Line 17). The step combines the sigmoid functional assignment and the independent noise variable. In the example of Mek in the MAPK pathway, the step outputs

$$f_{Mek}(Raf, N_{Mek}) = \frac{\beta_{Mek}}{1 + exp(-w_{Mek}Raf - b)} + N_{Mek} \tag{18}$$

Add $f_X(\mathbf{PA}_X, N_X)$ to SCM (Algorithm 2 Line 19). The step iteratively adds (f_X, N_X) for all $X \in \mathbf{X}$.

Output (Algorithm 2 Line 20). The algorithm returns a generative structural causal model $\mathbb{M} = \{f_i(\mathbf{PA}_i, N_i)\}_{i=1}^p$ where $\mathbf{PA}_i \subset \mathbf{X}$. For example, in the case of the MAPK model, it returns $[N_{Raf}, N_{Mek}, N_{Erk}, f_{Raf}(N_{Raf}), f_{Mek}(Raf, N_{Mek}), f_{Erk}(Mek, N_{Erk})]$.

3.4 Counterfactual Inference Procedure

The generated SCM enables counterfactual inference using a standard procedure [5]. Given a new observation $\mathbb{D}^{new}$,

1) *Abduction:* Update the probability $P(N_X)$ to obtain $P(N_X|\mathbb{D}^{new})$.
2) *Action:* Replace the equations determining the variables in set $\mathbf{X}^c$ by $\mathbf{X}^c = \mathbf{x}^{c\prime}$.
3) *Prediction:* Sample from the modified model to generate the target distribution $\mathbf{X}^e_{do(\mathbf{X}^c = \mathbf{x}^{c\prime})}$.

After generating the target distribution of the intervention model, we estimate causal effects. Algorithm 3 describes the detailed steps of both counterfactual inference (with $\mathbb{D}^{new}$) and forward simulation (if $\mathbb{D}^{new}$ is empty)

3.5 Implementation

QUERY2BEL was implemented manually using a publicly available instance of BioDati Studio, then validated using Integrated Dynamical Reasoner and Assembler (INDRA)'s [10] interactive dialogue system Bob with BioAgents [10]. Parameter estimation in BEL2SCM was implemented in PyTorch. Let C be the number of nodes in causal graph G with parents. Let k be the number of iterations for gradient descent, let N be the number of samples in data, and let d be the maximum number of parents in graph G. Computational complexity of parameter estimation step is given by $O(CkNd)$.

SCM-based counterfactual inference was performed with Pyro [44], due to its ability to perform interventions on probabilistic models and scalability to larger models, as described in Algorithm 3. Specifically, the implementation

relies on the following functionalities in Pyro. The `pyro.do` method is an implementation of Pearl's do-operator used for causal inference. The `pyro.infer.SVI` method performs abduction using stochastic variational inference with ELBO loss. The `pyro.infer.Importance` method performs posterior inference by importance sampling. The `pyro.infer.EmpiricalMarginal` method performs empirical marginal distribution from the trace posterior's model.

Algorithm 3. *Estimate causal effect on X^E upon intervening on X^C*

Inputs: New data point $\mathbb{D}^{new}$
effect node X^E
observational data for effect node $\mathbb{D}^E \in \mathbb{D}^{new}$
intervention value c
node to intervene upon X^C
number of iteration I
network structure G
SCM $\mathbb{M}$
Outputs: Causal Effect CE

procedure GETCAUSALEFFECT($\mathbb{D}^{new}, E, \mathbb{D}^E, X^C, c, I, G, \mathbb{M}$)
  $\hat{N} = \{\}$
  ▶ *Interventional data for effect node X^E*
  $\mathbb{ID}^E = \{\}$
  for I **do**
    for each $X \in \{\mathbf{X} \setminus X^C\}$ in G **do**
      ▶ ***Abduction:*** *Apply stochastic variational inference*
      $\hat{N}_X = SVI(\mathbb{D}^{new})$
      $\hat{N}$.Add($\hat{N}_X$)
    ▶ ***Action:*** *Apply intervention on X^C*
    $CM = pyro.do(\mathbb{M}, X^C = c)$
    ▶ *Get posterior of CM with importance sampling*
    $CMP = pyro.infer.Importance(CM, \hat{N})$
    ▶ ***Prediction:*** *Get EmpiricalMarginal (EM) for X^E*
    $CMM = pyro.infer.EM(CMP, X^E)$
    $\mathbb{ID}^E$.Add(CMM)
  $CE = \mathbb{ID}^E - \mathbb{D}^E$
  return CE

Experiments in this manuscript took between 13 to 82 seconds depending on the graph size on a system with Intel Core i7 8th Gen CPU, 16 GB RAM and Ubuntu 18.04 Operating System. The code is available at https://github.com/bel2scm.

4 Case Studies

Below we introduce two biological case studies investigated using the approach proposed in this manuscript. The first case study allows us to evaluate the accuracy of the results based on known ground truth. The second uses counterfactual reasoning to pinpoint the mechanism by which SARS-CoV-2 infection can lead to a cytokine storm in severely ill coronavirus disease 2019 (COVID- 19) patients. The details of the case studies, parameter values of the simulations, and of the results are at https://github.com/bel2scm.

4.1 Case Study 1: The IGF Signaling System

The System. The IGF signaling pathway (Fig. 4) regulates growth and energy metabolism of a cell. The IGF system

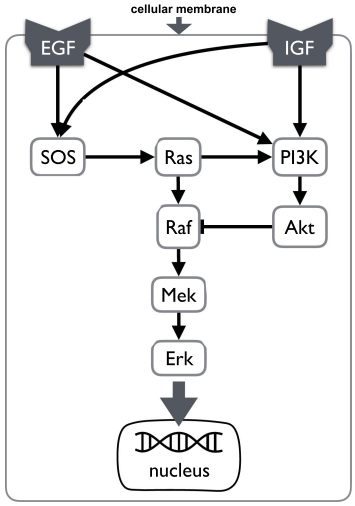

Fig. 4. *Case Study 1: the IGF signaling system.* The insulin-like growth factor (IGF) and epidermal growth factor (EGF) are receptors of external stimuli, triggering downstream signaling pathways that include the MAPK pathway. All the relationships between abundances of activated proteins in this network are of the type *increase*, except for the relationship between Akt and Raf which is of the type *decrease*.

has been extensively investigated, and its dynamics are well characterized in form of ODE and SDE models [25]. Activated by external stimuli, insulin-like growth factor (IGF) or epidermal growth factor (EGF) triggers a signaling event, which includes the MAPK signaling pathway in eq. (1). Similarly to eq. (1), nodes in the system are kinase activities, and edges represent whether the kinase activity of the upstream protein directly increases or decreases the kinase activity of the downstream protein. However, the system is larger and more complex. It includes two different paths from Ras to Erk, one direct and the other through $PI3K$ and Akt. This challenges estimates of outcomes of interventions. In this case study, we assume that the IGF system has no unobserved confounders.

Intervention. We considered two interventions. The first fixes the kinase activity of Mek to 40. The second fixes the kinase activity of Ras to 30.

Causal Effects of Interest. We are interested in two causal questions. First, *what would have been the kinase activity of Erk had we intervened to fix the kinase activity of Mek to 40?* The second query is as above, but with the intervention fixing the kinase activity of Ras to 30. More formally, we are interested in the average treatment effect

$$\{Erk_{do(Mek=40)} - Erk\} \tag{19}$$

$$\{Erk_{do(Ras=30)} - Erk\}. \tag{20}$$

Next, we introduce a new piece of information about a specific data point generated from the ODE-based simulation. We wish to estimate the causal effect of intervention for this specific data point. More formally, we are interested in the individual treatment effect

$$\{Erk_{do(Mek=40)} - Erk\}|\mathbb{D}^{new} \tag{21}$$

$$\{Erk_{do(Ras=30)} - Erk\}|\mathbb{D}^{new}, \tag{22}$$

where $\mathbb{D}^{new}$ is a new data point. We note that this counterfactual inference can only be performed with an SCM. We wish to compare these estimates of causal effects, in order to characterize the ability of counterfactual inference via D^{new} to improve the precision of the estimates.

Evaluation. The kinetic equations described by the ODE and SDE represent the true underlying dynamics of the IGF signaling pathway. Since the ODE and the SDE can estimate the causal effects by forward simulation, we view the estimates as the ground truth. We then wish to compare the estimates from the SCM against the ground-truth estimates from the ODE and the SDE. Since an SCM represents causal relationships at steady state, we train the parameters of the SCM using data generated from the ground-truth SDE after it has reached steady state.

We consider two types of evaluations. First, we compare the estimates of the forward simulation of the ODE and SDE with the forward simulation of the SCM. This allows us to characterize the impact of SCM specification and estimates of weights on the accuracy of causal effects. We do not expect to see a substantial difference between these two approaches for a correctly specified SCM. We then compare the SCM-based counterfactual inference of causal effects with the estimates based on forward simulation. We expect that the counterfactual inference will provide more precise estimates, illustrating the statistical efficiency of counterfactual inference as compared to the forward simulation.

4.2 Case Study 2: Host Response to Viral Infection

The System. Retrospective studies have indicated that high levels of pro-inflammatory cytokine Interleukin 6 (IL6) are strongly associated with severely ill COVID-19 patients [45]. One recently proposed explanation for this is the viral induction of a positive feedback loop, known as Interleukin 6 Amplifier (IL6-AMP) [46]. IL6-AMP is stimulated by simultaneous activation of nuclear factor kappa-light-chain-enhancer of activated B cell (NF-κB) and Signal Transducer and Activator of Transcription 3 (STAT3) [47]. This in turn induces various pro-inflammatory cytokines and chemokines, including Interleukin 6, which recruit activated T cells and macrophages. This strengthens the Interleukin 6 Amplifier into a positive feedback loop leading to a cytokine storm [48], which is believed to be responsible for the tissue damage observed in patients with acute respiratory distress syndrome (ARDS) [46].

Intervention. Originally developed to treat autoimmune disorders such as rheumatoid arthritis [49], Tocilizumab (Toci) is an immunosuppressive drug consisting of a recombinant monoclonal antibody that targets the soluble Interleukin 6 receptor and can effectively block the IL6 signal transduction pathway [50]. Tocilizumab has emerged as a promising drug repurposing candidate to reduce mortality in severely ill COVID-19 patients [51], [52].

Causal Effect of Interest. We define a severely ill COVID-19 patient as someone with CytokineStorm > 65. We are interested in the individual treatment effect (ITE)

$$\Big\{\text{CytokineStorm}_{do(Toci=0)} - \text{CytokineStorm}\Big\}|\mathbb{D}^{new}, \tag{23}$$

where $\mathbb{D}^{new}$ is an observed patient who received Tocilizumab treatment and became severely ill. We wish to characterize the severity of cytokine storm which would have occurred had she not received the treatment. We further wish to compare the ITE with the ATE

$$\Big\{\text{CytokineStorm}_{do(Toci=0)} - \text{CytokineStorm}\Big\}. \tag{24}$$

Evaluation. Tocilizumab is known to have a strong inhibitory effect on soluble Interleukin 6 receptor. We therefore expect that the severity of the cytokine storm would have been worse had the patient not received treatment. Unfortunately, at the time of writing, there were no ODE or SDE-based models of the pathway, nor were there publicly available COVID-19 datasets quantifying the kinase activity of the Interleukin 6 Amplifier pathway at the single-cell level. Therefore, we simulated data from a "ground-truth" sigmoidal structural causal model, where the topology reflects the causal structure of the pathway, and the numeric values of the parameters were fixed to reflect our prior qualitative knowledge of the IL6-AMP pathway.

We evaluate the ITE the proposed approach in two ways. First, we train the parameters of the SCM using the simulated data, and compare the counterfactual inference of the ITE obtained from the "trained" SCM to the counterfactual inference of the ITE from the "ground-truth" SCM. This comparison allows us to characterize the impact of weight estimation on the accuracy of causal effects. We expect that the need to estimate the weights will inflate the variance of the estimates. Second, we compare the estimates of ITE to the estimates of the ATE using the trained SCM. This comparison allows us to characterize the statistical efficiency of counterfactual inference when estimating causal effects. We expect that the ITE will provide much more precise estimates.

5 Results

5.1 Case Study 1: The IGF Signaling System

Generating BEL Causal Model. The BEL representation of the IGF system was manually curated using PyBEL [40], to match the existing ODE and SDE. The BEL representation of the IGF system specified all the node types as in category *abundance*. All the relationships between parents and children nodes were of type *increase*, except for the parent node Akt, where the relationship was of type *decrease*.

Observational Data. We mimicked the process of collecting observational data by simulating kinase activity from the corresponding ODE and SDE. The initial number of particles for the receptor was 37 for EGF and 5 for IGF. The deterministic simulation numerically solved the ODE using the *deSolve* [53] R package. The stochastic simulation used the Gillespie algorithm [29] from the *smfsb* [54] R package.

Appropriateness of Model Assumptions. SCM-based estimates of functional assignments with sigmoid approximations were well within the range of the SDE-based data (as shown for Raf and Mek in Fig. 5). Similar results were obtained for estimates of Ras, $PI3K$, AKT, Raf, and Erk. The fitted functional assignment had little curvature. This indicates that a more complicated function with more parameters, such as Hill equation, was unnecessary in this case.

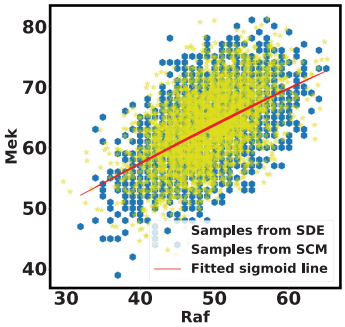

Fig. 5. *Case Study 1: IGF Model* Scatter Plot of Mek Versus Raf. Blue points are the data points generated by SDE. Yellow points are the estimates from SCM. The red line is the fitted sigmoid curve in Algorithm 2 line 12.

To further evaluate the plausibility of the assumptions, Fig. 6 shows the histograms of the SDE-generated abundances of root nodes, which were not affected by functional assignments in SCM. The shape of the histograms indicate that the assumption of Normal distribution was plausible.

Accuracy of Causal Effects. Figs. 7c and 7d show that the average treatment effects (ATEs) on Erk of fixing Mek and Raf, based on forward simulation of ODE, SDE and SCM, were consistent. Figs. 7a and 7b show that the based on counterfactual inference has a smaller variance than the ATE. Since counterfactual inference reduces nuisance variation by sharing stochastic components in contexts with and without intervention, it increases the statistical efficiency of the estimation.

The individual treatment effect on Erk by fixing Mek was much stronger than the ITE on Erk by fixing Ras for the following reason. While Mek directly influences Erk (i.e., there is a single path from Mek to Erk), Ras has two pathways to Erk. The path through AKT has an inhibiting (deactivation) effect on Raf, and estimated negative weights in the sigmoid function in eq. (15). The alternative path, a cascade from Ras to Erk, has the opposite (activating) effect on Erk. The two paths mitigate the overall causal effect of Ras on Erk.

5.2 Case Study 2: Host Response to Viral Infection

Generating BEL Causal Model. The steps of the proposed Algorithm 1 produced the qualitative causal model in Fig. 8, and the corresponding BEL causal model $\mathbb{B}$, as follows. In accordance with the inputs to Algorithm 1, we defined the knowledge base $\mathbb{K}$ as the Covid-19 knowledge network automatically assembled from the Covid-19

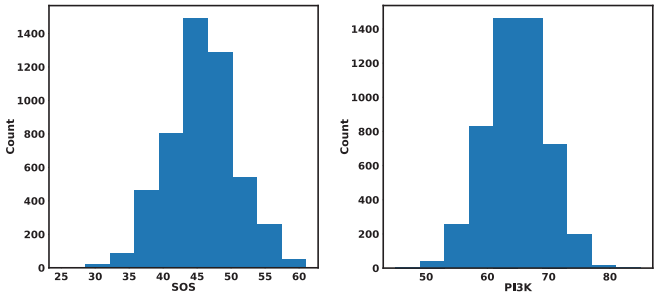

Fig. 6. *Case Study 1: Probability distributions of the root nodes of IGF Model* (a) Histogram of SOS generated from SDE simulation (b) As in (a), for $PI3K$.

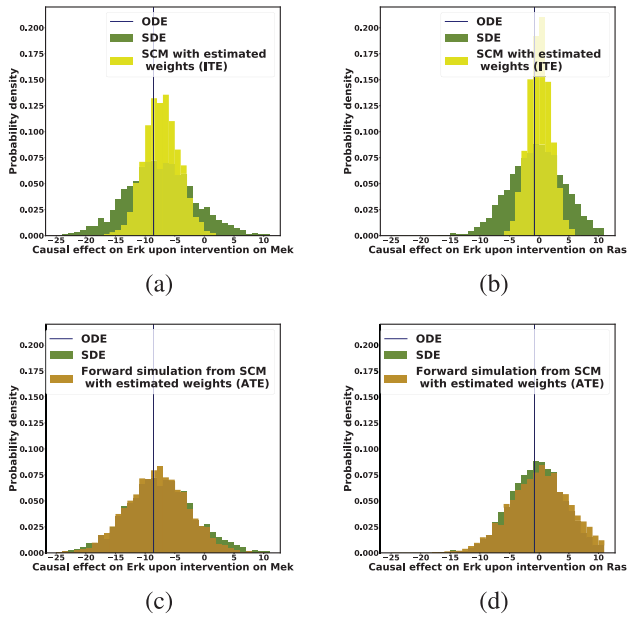

Fig. 7. *Case Study 1: Estimated causal effects of the IGF signaling pathway using algorithm 3.* The ODE and SDE represent the true underlying dynamics of the IGF signaling pathway. The ODE and SDE-based forward simulation can only estimate the average treatment effect. These estimates are viewed as ground truth. In contrast, an SCM can estimate both the average treatment effect (ATE) and the individual treatment effect (ITE). (a) Comparison of ITE vs ATE for Erk when Mek is fixed. (b) Comparison of ITE vs ATE for Erk when Ras is fixed. (c) Comparison of SCM, SDE and ODE estimates of the ATE for Erk when Mek is fixed. (d) Comparison of SCM, SDE and ODE estimates of the ATE on Erk when Ras is fixed.

document corpus using the INDRA workflow. We defined the cause $\mathbf{X}^c$ as sIL6Rα, the effect $\mathbf{X}^e$ as cytokine storm, and the covariates $\mathbf{X}^z$ as SARS-CoV- 2 and Toci.Therefore the causal query of interest was defined as $\mathbb{Q}$ = sIL6R α, CytokineStorm, SARS-CoV-2,Toci}}.

Algorithm 1 line 2 generated all pathways from Interleukin 6 to Cytokine Release Syndrome, resulting in $kin(p(\text{sIL6R}\ \alpha))) \rightarrow kin(p(\text{IL6-STAT3}))\ ) \rightarrow bp(\text{IL6} - \text{AMP})$(CytokineStorm), where bp() is a biological process. Next, line 5 generated all pathways from Tocilizumab to Interleukin 6: $a(\text{Toci})kin\ (p((\text{sIL6R}\alpha)))$, where $a()$ is the dosage level of Tocilizumab. We then generated all pathways from severe acute respiratory syndrome coronavirus 2 to Interleukin 6 receptor: $pop(\text{SARS-CoV-2})cat(\text{ACE2})a(\text{Angiotensin II}) \rightarrow kin(\text{p(AGTR1)}) \rightarrow kin(p\ (\text{ADAM17})) \rightarrow kin(p(\text{sIL6R}\alpha))$, where $pop()$ is the viral load of SARS-CoV-2 and $cat()$ is the normal catalytic activity of Angiotensin Converting Enzyme 2.

Line 8 found no new branches from Tocilizumab to Cytokine Release Syndrome. Finally, we generated all pathways from severe acute respiratory syndrome coronavirus 2 to Cytokine Release Syndrome, which resulted in three new branches $pop(\text{SARS-CoV-2}) \rightarrow kin(\text{p(PRR)}) \rightarrow kin(\text{p(NF-}\kappa\text{B)}) \rightarrow bp(\text{(IL6-AMP)}), kin(\text{p(ADAM17)}) \rightarrow p(\text{(EGF)}) \rightarrow kin(\text{p}\ (\text{EGFR})) \rightarrow kin(\text{p(NF-}\kappa\text{B)}), and kin(\text{pEGFR}) \rightarrow kin(\text{p(TNF}\ \ \alpha)) \rightarrow kin(\text{p(NF-}\kappa\text{B)})$.

Observational Data. We simulated observational data from a "ground-truth" sigmoidal structural causal model, where the topology reflects the causal structure in Fig. 8, and the parameters reflect our prior qualitative knowledge of the IL6-AMP pathway. The root nodes SARS-CoV-2 and Tocilizumab were sampled from a Normal distribution

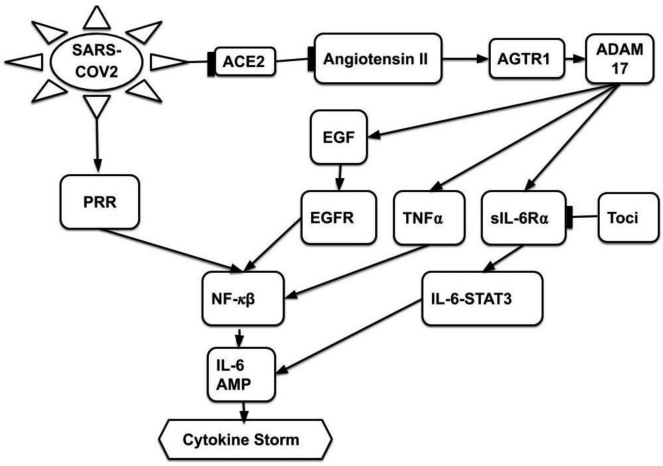

Fig. 8. Case Study 2: Host response to viral infection pointed edges represent relationships of type *increase*; flat-headed edges represent relationships of type *decrease*. Nodes SARS-COV2 and Toci are external stimuli.

with mean of 50 and standard deviation of 10. The non-root nodes were sampled from a sigmoid function as in eq. (15). Since we have prior qualitative knowledge that IL6-AMP is only activated due to simultaneous activation of NF-κB and IL6-STAT3, we set the threshold for activation above what could be achieved by NF-κB or IL6-STAT3 alone. Since we also know that Toci is a strong inhibitor of sIL6Rα, we set the inhibition coefficient to a large negative number. The parameters of the sigmoid function were chosen to ensure that the variables were in the desired range of 0–100. Finally, we randomly generated two new individuals $\mathbb{D}^{new}$ with Cytokine Release Syndrome > 65 to represent severely ill patients. The first patient had a higher viral load of SARS-CoV-2 and received a lower dose of Toci. The second patient had a lower viral load of and received a higher dose of Toci.

Estimation of Individual-Level Treatment Effect. Fig. 9 evaluates the SCM-based estimates of the individual treatment effect of withholding treatment from two COVID-19 patients who were severely ill. The distribution of the individual treatment effect obtained with the SCM trained using Algorithm 2 was consistent with, but had a slightly larger variance then, the distribution of ITE obtained with the "ground truth" SCM with known weights. Even though both patients had the same severity of illness prior to the intervention, patient B was estimated to have a more severe cytokine storm after Toci was withheld.

Fig. 10 further compared the individual treatment effect obtained with the SCM trained using Algorithm 2 with the average treatment effect estimated from the same model using forward simulation. The distribution of the individual treatment effect was patient-specific and had smaller variance, thus illustrating the statistical efficiency of counterfactual inference.

6 Discussion

We proposed a general approach that leverages structured qualitative prior knowledge, automatically generates a quantitative SCM, and enables answers to counterfactual research questions. In both case studies, the use of the

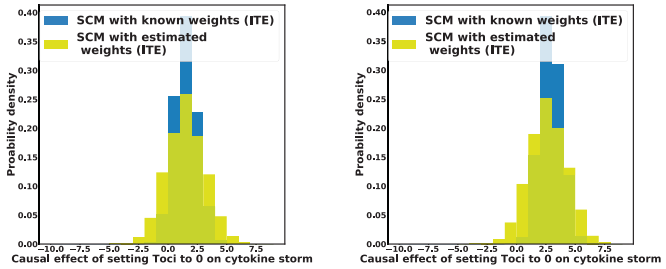

Fig. 9. *Case Study 2: SCM-Based estimates of the using algorithm 3.* Blue histogram: the ITE estimated from the ground-truth SCM using Algorithm 3. Yellow histogram: the ITE estimated from the Algorithm 3-trained SCM using Algorithm 3. (a) Patient has a high viral load and received a low dose of Tocilizumab. (b) Patient has a low viral load and received a high dose of Tocilizumab. Both patients were severely ill.

Biological Expression Language allowed us to leverage large repositories of structured biological knowledge to specify an SCM and perform counterfactual inference in an automated manner, which would otherwise require a substantial manual effort. The application to the IGF signaling system demonstrated the appropriateness of the underlying assumptions, and the accuracy of the results when compared to ODE- and SDE-based forward simulation. The application to a study of host response to SARS-CoV-2 infection demonstrated the feasibility, versatility and usefulness of this approach as applied to an urgent public health issue. In particular, the approach can help determine the amount of Tocilizumab (Toci) required to reduce the severity of each individual's cytokine storm. Furthermore, in situations where treatment options are limited (as is the case SARS-CoV-2), counterfactual estimates enable a more precise conclusion regarding who would likely live without receiving the treatment, who would likely die even if they did receive the treatment, and who would likely live only after receiving the treatment.

The approach opens multiple directions for future research. In particular, future work can extend the configurability of the BEL2SCM algorithm by incorporating the rich type information in BEL, mapping parent-child type signatures to functional forms such as post-nonlinear models, neural networks, mass action kinetics and Hill equations, and incorporating additional data types such as binary

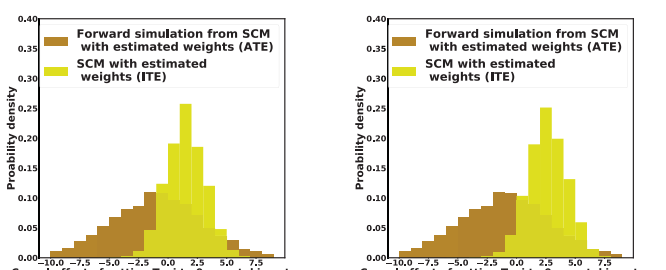

Fig. 10. *Case Study 2: SCM-Based estimates of the ATE and of the ITE using algorithm 3.* Yellow histogram: the ITE estimated using counterfactual inference. Brown histogram: the ATE estimated using forward simulation. (a) Patient has a high viral load and received a low dose of Tocilizumab. (b) Patient has a low viral load and a received a high dose of Tocilizumab. Both patients were severely ill.

variables, categorical variables, and continuous variables with constraints on their domains. In some cases, the variables in the model may not be directly observable, but may nonetheless be characterized by means of detectable molecular signatures. For example, even if interferon signaling may not be directly observable using transcriptomics measurements, it may still be possible to infer the activity of interferon signaling by an upregulation of interferon stimulated genes (ISG). Future work will focus on leveraging molecular signature databases to infer the activity of variables in the model, and on learning and/or evaluating the models using experimental data [55].

We also note that experimentalists typically formulate biological processes as linear pathways (e.g., from S_1 to Erk in the MAPK example) that can be effectively perturbed and measured in a laboratory setting. Yet such boundaries of biological processes are quite arbitrary, and are therefore highly susceptible to confounders. One way to address this issue is to search the knowledge graph for all common causes of variables in the causal model, use an identification algorithm [56] to find the minimal valid adjustment set of the augmented model, and then prune all common causes that do not contribute to that set. This approach will require us to tackle the issues of parameter and causal identifiability in the presence of confounders.

In addition to unobserved confounders, the validity of causal inferences can be threatened by feedback loops, model misspecification, missing data, and out-of-sample distributions. To address the possibility of feedback loops, we must consider the time scale at which these feedbacks reach steady-state: fast timescale feedback loops can be addressed with the chain graph interpretation of SCMs [57], [58]; intermediate timescale feedbacks can be addressed with non-recursive structural causal models [5]; slow timescale feedback loops can be handled by unrolling the structure of the SCM as is done with dynamic Bayesian networks [59], or simply by representing the entire feedback loop as a biological process, as we did with IL6-AMP. In the case of model misspecification, we will investigate the ability of counterfactual inference to improve the estimation [43]. For missing data, we can leverage causal inference recoverability algorithms that have been published recently [60], and for handling out-of-sample distributions, we can leverage recent results applying causal inference to the problem of external validity [61]. Future work will focus on addressing these threats to validity when applied to real biological data.

Acknowledgments

This work was supported by funds from the PNNL Mathematics and Artificial Reasoning Systems Laboratory Directed Research and Development Initiative. Knowledge curation environments were provided by BioDati.com and Causaly.com. The authors would also like to acknowledge Jessica Stothers and Rose Glavin at CoronaWhy.org and Marek Ostaszewski at the COVID-19 Disease Map Initiative for providing valuable feedback about the IL6-AMP model. Jeremy Zucker, Kaushal Paneri, Sara Mohammad-Taheri contributed equally to this work.

References

[1] A. Pezeshki, I. G. Ovsyannikova, B. A. McKinney, G. A. Poland, and R. B. Kennedy, "The role of systems biology approaches in determining molecular signatures for the development of more effective vaccines," *Expert Rev. Vaccines*, vol. 18, 2019, Art. no. 253.

[2] M. Pedragosa *et al.*, "Linking cell dynamics with gene coexpression networks to characterize key events in chronic virus infections," *Front. Immunol.*, vol. 10, 2019, Art. no. 1002.

[3] V. K. Nguyen, F. Klawonn, R. Mikolajczyk, and E. A. Hernandez-Vargas, "Analysis of practical identifiability of a viral infection model," *PloS One*, vol. 11, 2016, Art. no. e0167568.

[4] A. Arazi, W. F. Pendergraft, R. M. Ribeiro, A. S. Perelson, and N. Hacohen, "Human systems immunology: Hypothesis-based modeling and unbiased data-driven approaches," *Seminars Immunol.*, vol. 25, 2013, Art. no. 193.

[5] J. Pearl, *Causality: Models, Reasoning and Inference*. Cambridge, MA, USA: Cambridge Univ. Press, 2013.

[6] J. Peters, D. Janzing, and B. Schölkopf, *Elements of Causal Inference: Foundations and Learning Algorithms*. Cambridge, MA, USA: MIT press, 2017.

[7] J. F. Allen, M. Swift, and W. De Beaumont, "Deep semantic analysis of text," *Proc. Conf. Semantics Text Process.*, 2008, vol. 1, Art. no. 343.

[8] D. D. McDonald, "Issues in the Representation of Real Texts: The Design of KRISP," in *Proc. Natural Lang. Process. Knowl. Representation: Lang. Knowl. Knowl. Lang.*, 2000, pp. 77–110.

[9] M. A. Valenzuela-Escárcega *et al.*, "Large-scale automated machine reading discovers new cancer-driving mechanisms," *Database*, vol. 2018, 2018, Art. no. 1.

[10] B. M. Gyori, J. A. Bachman, K. Subramanian, J. L. Muhlich, L. Galescu, and P. K. Sorger, "From word models to executable models of signaling networks using automated assembly," *Mol. Syst. Biol.*, vol. 13, 2017, Art. no. 954.

[11] C. T. Hoyt *et al.*, "Re-curation and rational enrichment of knowledge graphs in biological expression language," *Database*, vol. 2019, 2019, Art. no. baz068.

[12] E. G. Cerami *et al.*, "Pathway Commons, a web resource for biological pathway data," *Nucl. Acids Res.*, vol. 39, pp. D685–D690, 2011.

[13] A. Fabregat *et al.*, "The Reactome pathway knowledgebase," *Nucl. Acids Res.*, vol. 46, pp. D649–D655, 2018.

[14] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, "KEGG: New perspectives on genomes, pathways, diseases and drugs," *Nucl. Acids Res.*, vol. 45, pp. D353–D361, 2017.

[15] L. Perfetto *et al.*, "SIGNOR: A database of causal relationships between biological entities," *Nucl. Acids Res.*, vol. 44, pp. D548–D554, 2016.

[16] D. N. Slenter *et al.*, "WikiPathways: A multifaceted pathway database bridging metabolomics to other omics research," *Nucl. Acids Res.*, vol. 46, pp. D661–D667, 2018.

[17] E. Demir *et al.*, "The BioPAX community standard for pathway data sharing," *Nat. Biotechnol.*, vol. 28, 2010, Art. no. 1308.

[18] M. Hucka *et al.*, "The Systems Biology Markup Language (SBML): Language specification for level 3 version 2 core," *J. Integrative Bioinf.*, 2018, Art. no. 20170081.

[19] N. Le Novere *et al.*, "The systems biology graphical notation," *Nat. Biotechnol.*, vol. 27, pp. 735–741, 2009.

[20] T. Slater, "Recent advances in modeling languages for pathway maps and computable biological networks," *Drug Discov. Today*, vol. 19, pp. 193–198, 2014.

[21] P. Machamer, L. Darden, and C. F. Craver, "Thinking about mechanisms," *Philosophy Sci.*, vol. 67, 2000, Art. no. 1.

[22] Y. Li, J. Roberts, Z. AkhavanAghdam, and N. Hao, "Mitogen-activated protein kinase (MAPK) dynamics determine cell fate in the yeast mating response," *The J. Biol. Chem.*, vol. 292, pp. 20354–20361, 2017.

[23] L. Chen, R. Wang, C. Li, and K. Aihara, *Modeling Biomolecular Networks in Cells: Structures and Dynamics*. Berlin, Germany: Springer, 2010.

[24] D. Gratie, B. Iancu, and I. Petre, "ODE analysis of biological systems," in *International School on Formal Methods for the Design of Computer, Communication and Software Systems*. Berlin, Germany: Springer 2013, Art. no. 29.

[25] F. Bianconi, E. Baldelli, V. Ludovini, L. Crino, A. Flacco, and P. Valigi, "Computational model of EGFR and IGF1R pathways in lung cancer: A systems biology approach for translational oncology," *Biotechnol. Adv.*, vol. 30, pp. 142–153, 2012.

[26] E. K. Kim and E.-J. Choi, "Pathological roles of MAPK signaling pathways in human diseases," *Biochimica et Biophysica Acta - Mol. Basis Disease*, vol. 1802, pp. 396–405, 2010.
[27] R. Ness, K. Paneri, and O. Vitek, "Integrating Markov processes with structural causal modeling enables counterfactual inference in complex systems," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, Art. no. 14211.
[28] K. Paneri, "Integrating Markov process and structural causal models enables counterfactual inference in complex systems," Northeastern Univ., 2019.
[29] D. T. Gillespie, "Exact stochastic simulation of coupled chemical reactions," *The J. Phys. Chem.*, vol. 81, pp. 2340–2361, 1977.
[30] S. K. Jha and C. J. Langmead, "Exploring behaviors of stochastic differential equation models of biological systems using change of measures," *BMC Bioinf.*, vol. 13, 2012, Art. no. S8.
[31] U. Alon, *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Boca Raton, FL, USA: CRC press, 2019.
[32] S. Bongers and J. M. Mooij, "From random differential equations to structural causal models: The stochastic case," *ArXiv*, vol. abs/1803.08784, 2018.
[33] M. Jerrum, A. Sinclair, and D. S. Hochbaum, "The markov chain monte carlo method: An approach to approximate counting and integration," *Approximation Algorithms NP-hard problems, PWS Publishing*, 1996.
[34] A. E. Gelfand, "Gibbs sampling," *J. Amer. Statist. Assoc.*, vol. 95, 2000, Art. no. 1300.
[35] M. D. Hoffman and A. Gelman, "The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo," *J. Mach. Learn. Res.*, vol. 15, pp. 1593–1623, 2014.
[36] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, "Stochastic variational inference," *The J. Mach. Learn. Res.*, vol. 14, pp. 1303–1347, 2013.
[37] T. Blom, S. Bongers, and J. M. Mooij, "Beyond structural causal models: Causal constraints models," in *Proc. 35th Conf. Uncertainty Artif. Intell.*, 2019, pp. 585–594.
[38] S. Madan *et al.*, "The extraction of complex relationships and their conversion to biological expression language (BEL) overview of the BioCreative VI (2017) BEL track," *Database, J. Biol. Databases Curation*, vol. 2019, 2019, Art. no. baz084.
[39] C. T. Hoyt *et al.*, "Integration of structured biological data sources using biological expression language," *BioRxiv*, Cold Spring Harbor Lab., pp. 631812, 2019.
[40] C. T. Hoyt, A. Konotopez, C. Ebeling, and J. Wren, "PyBEL: A computational framework for biological expression language," *Bioinformatics*, vol. 34, pp. 703/704, 2018.
[41] H. Mi *et al.*, "Systems biology graphical notation: Activity flow language level 1 version 1.2," *J. Integrative Bioinf.*, vol. 12, 2015, Art. no. 265.
[42] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and variational inference in deep latent gaussian models," in *Proc. Int. Conf. Mach. Learn.*, vol. 2, 2014.
[43] R. Ness, K. Paneri, and O. Vitek, "Integrating Markov processes with structural causal modeling enables counterfactual inference in complex systems," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, Art. no. 14234.
[44] E. Bingham *et al.*, "Pyro: Deep Universal Probabilistic Programming," *J. Mach. Learn. Res.*, vol. 20, pp. 1–6, 2018.
[45] Z. S. Ulhaq and G. V. Soraya, "Interleukin-6 as a potential biomarker of COVID-19 progression," *Med. Mal. Infect.*, vol. 50, pp. 382/383, 2020.
[46] T. Hirano and M. Murakami, "COVID-19: A new virus, but a familiar receptor and cytokine release syndrome," *Immunity*, vol. 52, pp. 731–733, 2020.
[47] M. Murakami and T. Hirano, "The pathological and physiological roles of IL-6 amplifier activation," *Int. J. Biol. Sci.*, vol. 8, pp. 1267–1280, 2012.
[48] H. Ogura *et al.*, "Interleukin-17 promotes autoimmunity by triggering a positive-feedback loop via interleukin-6 induction," *Immunity*, vol. 29, pp. 628–636, 2008.
[49] V. Oldfield, S. Dhillon, and G. L. Plosker, "Tocilizumab: A review of its use in the management of rheumatoid arthritis," *Drugs*, vol. 69, pp. 609–632, 2009.
[50] C. Zhang, Z. Wu, J.-W. Li, H. Zhao, and G.-Q. Wang, "Cytokine release syndrome in severe COVID-19: Interleukin-6 receptor antagonist Tocilizumab may be the key to reduce mortality," *Int. J. Antimicrob. Agents*, vol. 55, 2020, Art. no. 105954.
[51] E. A. Coomes and H. Haghbayan, "Interleukin-6 in COVID-19: A systematic review and meta-analysis," *MedRxiv*, Cold Spring Harbor Lab. Press, 2020.
[52] X. Xu *et al.*, "Effective Treatment of Severe COVID - 19 Patients with Tocilizumab," *Proc. Nat. Acad. Sci. USA*, vol. 117, pp. 10970–10975, 2020.
[53] K. E. R. Soetaert, T. Petzoldt, and R. W. Setzer, "Solving differential equations in R: Package deSolve," *J. Statist. Softw.*, vol. 33, pp. 77–83, 2010.
[54] D. Wilkinson, "Smfsb-stochastic modelling for systems biology," *R Package Version*, vol. 1, 2018.
[55] A. Liu, P. Trairatphisan, E. Gjerga, A. Didangelos, J. Barratt, and J. Saez-Rodriguez, "From expression footprints to causal pathways: Contextualizing large signaling networks with CARNIVAL," *Syst. Biol. Appl.*, vol. 5, 2019, Art. no. 40.
[56] S. Tikka and J. Karvanen, "Identifying causal effects with the R package causal effect," *J. Statist. Softw.*, vol. 76, 2017, Art. no. 1.
[57] S. L. Lauritzen and T. S. Richardson, "Chain graph models and their causal interpretations," *J. Roy. Statist. Soc.: Series B*, vol. 64, pp. 321–348, 2002.
[58] E. Sherman and I. Shpitser, "Identification and estimation of causal effects from dependent data," *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, vol. 2018, Art. no. 9446.
[59] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. Cambridge, MA, USA: MIT Press, 2009.
[60] R. Nabi, R. Bhattacharya, and I. Shpitser, "Full law identification in graphical models of missing data: Completeness results," 2020, *arXiv:2004.04872*.
[61] E. Bareinboim and J. Pearl, "Causal inference and the data-fusion problem," *Proc. Nat. Acad. Sci. USA*, vol. 113, 2016, Art. no. 7345.

Jeremy Zucker is currently the principal investigator for the MARS causal inference for viral pathogenesis project. He has more than 15 years of experience developing causal models to obtain actionable insights from systems biology data to advance knowledge in the study of metabolic engineering, circadian rhythms, evolution, human health and infectious disease.

Kaushal Paneri received the master's degree in data science from Northeastern university. He is a data scientist at Microsoft, currently working on counterfactual platform for Bing Ads Marketplace Optimization. His prominent research interests include causality, optimization and machine learning.

Sara Mohammad-Taheri received the bachelor's and master's degree in mathematics from the Sharif University of Technology. She is currently working toward the PhD degree in computers science with Northeastern University's Khoury College of Computer Sciences, advised by professor Olga Vitek. Her research interest includes causal inference techniques in computational biology and causal discovery of biomolecular data. She is also interested in developing statistical and computational methods and open source software for systems-wide molecular investigations of biological organisms including quantitative genomics, proteomics etc. She is a member of the statistical methods for studies of biomolecular systems group.

Somya Bhargava received the master's degree in data science from Northeastern University. She is currently working with Embedded Healthcare. She's been working in Healthcare industry and is experienced in using natural language processing, machine learning, statistical analysis and causal inference for researching for new products and enhancing existing ones.

Pallavi Kolambkar received the bachelor's degree in computer science, and the master's degree in computer applications, from India. She is majored in data science from Northeastern University and is currently working with Tesla. She has worked with companies from different domains to explore and visualize different dynamics of data.

Craig Bakker received a PhD degree in engineering from the University of Cambridge, where his research focused on optimization algorithms, differential geometry, and computational methods for model decomposition. Following this, he did postdoctoral research in climate change, food security, and economic modelling at Johns Hopkins University. He is currently a research scientist with the Pacific Northwest National Laboratory. He works in game theory, machine learning, and optimal control.

Jeremy Teuton received the PhD degree in cell and molecular biology (virology). He is an experienced interdisciplinary researcher and project leader. He is proficient in experimental design, trouble shooting, data analysis, and interdisciplinary application of scientific principles and approaches including cyber security and signal detection/classification. He excels in challenging environments, where problem-solving skills and experience in adapting technologies, systems and processes/approaches can be of most use.

Charles Tapley Hoyt received the PhD degree in computational life sciences from the University of Bonn. His research interests cover the interface of biocuration, knowledge graphs, and machine learning with systems biology, networks biology, and drug discovery. He is an advocate of open source software, reproducibility, and open science. His open source projects PyBEL and PyKEEN are used by several academic and industrial groups.

Kristie Oxford is a virologist, with expertise in host-pathogen interactions. Her research at Pacific Northwest National Laboratory (PNNL) primarily involves characterizing and interpreting host biomolecular responses to viral infection. She and her team analyze systems biology data from cells infected in vitro and in vivo with mammalian viruses representing many genera and families, in order to understand mechanisms of disease and to identify targets for medical countermeasures. The systems approach interrogates the host transcriptomic response to infection from microarray or RNA sequencing data as well as the proteomic, lipidomic, and metabolomic response from high resolution mass spectrometry analysis. She and her team have studied host-virus interactions from thousands of samples representing more than 12 human viruses, identifying gene, protein, and metabolite candidates for medical intervention and/or mechanistic studies.

Robert Ness received the PhD degree in mathematical statistics from Purdue University, and then he worked as a research engineer in various AI startups. He didn't start in machine learning. He started his career by becoming fluent in Mandarin Chinese and moving to Tibet to do developmental economics fieldwork. He later obtained a graduate degree from Johns Hopkins School of Advanced International Studies. After switching to the tech industry, his interests shifted to modeling data. He has published in journals and venues across these spaces, including *Research in Computational Molecular Biology* and NeurIPS, on topics including causal inference, probabilistic modeling, sequential decision processes, and dynamic models of complex systems. In addition to startup work, currently he is a machine learning professor with Northeastern University.

Olga Vitek received the PhD degree in statistics from Purdue University. She is currently a professor with the Khoury College of Computer Sciences at Northeastern University. Her research interests include statistical science, machine learning, mass spectrometry and systems biology. Statistical methods and open-source software MSstats and Cardinal developed in her lab are used in academia and industry, and were recently recognized with the Chan Zuckerberg Essential Open Source Software for Science Award. She is a senior member of the International Society for Computational Biology, and an elected member of the Council of HUPO and of the board of directors of USHUPO. She is a member of the editorial advisory board of *Molecular and Cellular Proteomics* and of *Journal of Proteome Research*.