Interventionally Consistent Surrogates for Complex Simulation Models

Joel Dyer[*]
University of Oxford

Nicholas Bishop
University of Oxford

Yorgos Felekis
University of Warwick

Fabio Massimo Zennaro
University of Bergen

Anisoara Calinescu
University of Oxford

Theodoros Damoulas
University of Warwick

Michael Wooldridge
University of Oxford

Abstract

Large-scale simulation models of complex socio-technical systems provide decision-makers with high-fidelity testbeds in which policy interventions can be evaluated and *what-if* scenarios explored. Unfortunately, the high computational cost of such models inhibits their widespread use in policy-making settings. Surrogate models can address these computational limitations, but to do so they must behave consistently with the simulator under interventions of interest. In this paper, we build upon recent developments in causal abstractions to develop a framework for learning interventionally consistent surrogate models for large-scale, complex simulation models. We provide theoretical results showing that our proposed approach induces surrogates to behave consistently with high probability with respect to the simulator across interventions of interest, facilitating rapid experimentation with policy interventions in complex systems. We further demonstrate with empirical studies that conventionally trained surrogates can misjudge the effect of interventions and misguide decision-makers towards suboptimal interventions, while surrogates trained for *interventional* consistency with our method closely mimic the behaviour of the original simulator under interventions of interest.

1 Introduction

Large-scale, complex simulators are powerful tools for modelling distributed socio-technical systems and emergent phenomena across application domains, including the social sciences [Wiese et al., 2024], epidemiology [Kerr et al., 2021], and finance [Cont, 2007]. Many such systems consist of a multitude of autonomous, interacting, and decision-making agents, whose individual behaviours and interactions can be captured more readily and at a higher degree of fidelity in a computer program than through conventional modelling paradigms. This level of granularity can, in turn, allow for more effective control of the potentially deleterious effects that can arise from the endogenous dynamics of real-world systems by providing a testbed for experimentation with policy interventions. In economics, for example, such interventions may take the form of limits on loan-to-value ratios in housing markets to attenuate housing price cycles [Baptista et al., 2016], while in epidemiology they may be (non-)pharmaceutical interventions that aim to inhibit disease transmission [Kerr et al., 2021].

Whilst simulation modelling of this kind promises many benefits, the intricacy and multi-scale nature of the simulators that result from these modelling efforts can result in large computational costs even

[*]joel.dyer@cs.ox.ac.uk

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

for single forward simulations [Jagiella et al., 2017, Fadai et al., 2019, Wright and Davidson, 2020, Heppenstall et al., 2021]. Since extensive simulation studies are often required to aid decision-making with these models, such costs present a barrier to their use as synthetic test environments for potential policy interventions in practice. Moreover, the high-fidelity data generated by detailed simulation models can be difficult for decision-makers to interpret and relate to policy interventions that act system-wide [Haldane and Turrell, 2018]. This motivates the development of simpler *surrogate* models that model the underlying system at a higher level of abstraction. Such surrogates can also be used in place of the complex model for downstream tasks where computational resources are limited. In addition, surrogates may be viewed as interpretable explanations for the complex simulator, and they allow for rapid testing of population-wide interventions which may be difficult to implement or test within the orginal model.

However, for surrogates to be useful in downstream tasks involving experimentation with possible policy interventions, they must preserve the complex simulator's dynamics under the external interventions of interest. Without imposing this condition on the constructed surrogate, there is no guarantee that the surrogate will behave similarly under external policy interventions, which in turn may lead policy-makers away from effective policies and towards suboptimal interventions. Existing methods typically apply off-the-shelf machine learning methods to learn surrogates through observation [Lamperti et al., 2018, Platt, 2022], which fails to account for interventional consistency.

Our contribution. To address this, we build on recent developments in *causal abstraction* [Beckers and Halpern, 2019, Zennaro et al., 2023a]. We view the complex simulator and its surrogate as *structural causal models* [Pearl, 2009], and propose a framework for constructing and learning surrogate models for expensive simulators of complex socio-technical systems that are *interventionally consistent*, in the sense that they (approximately) preserve the behaviour of the simulator under equivalent policy interventions. This perspective enables *treating the surrogate model as a causal abstraction of the simulator*. We motivate our proposed methodology theoretically, and demonstrate with simulation studies that our method permits us to learn an abstracted surrogate model for an epidemiological agent-based model that behaves consistently in multiple interventional regimes.

Our approach establishes, for the first time, a connection between complex simulation models and causal abstraction, and a practical approach to learning interventionally consistent surrogates for complex simulators. Our work provides an avenue for researchers modelling complex socio-technical systems to draw on the rich literature in causality for integrating causal knowledge, evaluating *what-if* scenarios, and learning new abstracted models with guarantees about interventional consistency. Our contribution lays the groundwork for surrogate modelling methods that facilitate rapid experimentation with different scenarios and interventions, with assurances that the error introduced by experimenting at a higher level of abstraction is low. This line of work has the potential to enable decision- and policy-makers to use simulation models to quickly identify life-saving policy strategies during novel and rapidly unfolding emergencies, such as pandemics and economic crises. Indeed, a recent World Health Organisation report [World Health Organization et al., 2024] emphasises the importance of integrated modelling to concurrently address interdependent policy objectives, such as reducing disease transmission, mitigating hospital admissions overload, and minimising the economic costs of service closures on society during pandemics. It further discusses the intense time pressures involved in these efforts. Our work addresses these points by taking steps towards facilitating rapid experimentation with large and computationally expensive integrated simulation models.

2 Background

We first recall the key elements of causal inference, following Pearl [2009], and elucidate the connection between structural causal models (SCMs) and complex simulators. We also review the notion of exact transformations between SCMs, which theoretically motivates our framework.

2.1 Structural causal models

A SCM is a rigorous model describing a causal system:

Definition 1 (SCMs [Pearl, 2009])**.** *A structural causal model $\mathcal{M}$ is a tuple $\langle \mathbf{X}, \mathbf{U}, \mathcal{F}, \mathbb{P}(\mathbf{U}) \rangle$ where:*

- $\mathbf{X} = \{X_i\}_{i=1}^n$*, is a finite set of endogenous random variables X_i each with domain* $\mathrm{dom}[X_i]$*;*

- $\mathbf{U} = \{U_i\}_{i=1}^n$, *is a finite set of exogenous random variables, each with domain* $\mathrm{dom}[U_i]$ *and each associated with an endogenous variable;*
- $\mathcal{F} = \{f_i\}_{i=1}^n$, *is a finite set of measurable structural functions, one for each endogenous variable defined as* $f_i : \mathrm{dom}[PA(X_i)] \times \mathrm{dom}[U_i] \to \mathrm{dom}[X_i]$, *where* $PA(X_i) \subseteq \mathbf{X} \setminus X_i$.
- $\mathbb{P}_{\mathcal{M}}(\mathbf{U})$ *is a joint probability distribution over the exogenous variables factorizing as* $\prod_{i=1}^n \mathbb{P}_{\mathcal{M}}(U_i)$.

The model $\mathcal{M}$ *is associated with a Directed Acyclic Graph (*`DAG`*)* $\mathcal{G}_{\mathcal{M}} = \langle \mathcal{V}, \mathcal{E} \rangle$ *where the set* $\mathcal{V}$ *of vertices is given by* $\mathbf{X} \cup \mathbf{U}$ *and the set* $\mathcal{E}$ *of edges is given by* $\{(S_j, X_i) \mid S_j \in PA(X_i) \cup \{U_i\}\}_{i=1}^n$.

Definition 1 conforms to the standard definition of a *Markovian* `SCM` (see Appendix A for an explanation of the underlying assumptions). Thanks to the measurability of the structural functions in $\mathcal{F}$, the probability distribution $\mathbb{P}_{\mathcal{M}}(\mathbf{U})$ over the exogenous variables can be pushed forward over the endogenous variables, defining the probability distribution $\mathbb{P}_{\mathcal{M}}(\mathbf{X}) = \mathcal{F}_{\#}(\mathbb{P}_{\mathcal{M}}(\mathbf{U}))$. Joint distributions $\mathbb{P}_{\mathcal{M}}(\mathbf{S})$ can then be defined for any subset $\mathbf{S} \subseteq \mathbf{X}$.

External interventions on the system by an experimenter can be represented in an `SCM` through changes in the structural functions. Here, we restrict our attention to hard interventions, in which fixed values are assigned to subsets of endogenous variables:

Definition 2 (Interventions [Pearl, 2009])**.** *Given an* `SCM` $\mathcal{M}$, $\mathbf{S} \subseteq \mathbf{X}$, *and a set of values* $\mathbf{s}$ *realizing* $\mathbf{S}$, *an intervention* $\iota = \mathrm{do}(\mathbf{S} = \mathbf{s})$, *is an operator that replaces each function* f_i *associated with* S_i *with constant* s_i.

The intervention $\iota = \mathrm{do}(\mathbf{S} = \mathbf{s})$ induces a new *post-intervention* `SCM`, $\mathcal{M}_\iota = \langle \mathbf{X}, \mathbf{U}, \mathcal{F}_\iota, \mathbb{P}(\mathbf{U}) \rangle$, identical to the original one, except that in $\mathcal{F}_\iota$ the functions f_i are replaced with the constants s_i. The probability distribution of $\mathcal{M}_\iota$ is computed as $\mathbb{P}_{\mathcal{M}_\iota}(\mathbf{X} \setminus \mathbf{S})$. Graphically, the intervention ι transforms the `DAG` of $\mathcal{M}$ by removing incoming edges in each variable S_i.

We use $\mathcal{I}$ to denote a set of feasible interventions on the `SCM` $\mathcal{M}$ that are relevant to a policymaker. Intervention sets are equipped with a natural partial ordering: let $\iota_1 = (\mathbf{S} = \mathbf{s})$ and $\iota_2 = (\mathbf{T} = \mathbf{t})$; then $\iota_1 \preceq \iota_2$ iff (i) $\mathbf{S} \subseteq \mathbf{T}$, and (ii) for each $S_i = T_i$ it holds $s_i = t_i$; informally, ι_1 intervenes on a subset of the variables that ι_2 intervenes on, and it sets the same values as ι_2.

2.2 Complex simulators as structural causal models

Many simulation models of complex systems – such as, for example, agent-based models (`ABM`s) – can be modelled as a `SCM` by expressing its implicit underlying causal structure. Practically, this entails encoding quantities of interest as endogenous variables, deterministic dynamics into structural equations, and factoring sources of randomness into exogenous variables. The following example illustrates how a common `ABM` from epidemiology can be cast as a `SCM`.

Example 1 (Spatial SIRS `ABM`)**.** *We consider a susceptible-infected-recovered-susceptible (SIRS) epidemic model on an* $L \times L$ *lattice of cells, each of which represents one of* $N = L^2$ *agents. The state of each agent can be 0, 1, or 2, respectively, indicating that the agent is disease-free and susceptible to infection, infected, or is recovered from a recent infection. The infection status of all agents at discrete time step* $t \in [\![0, T]\!]$ *is written as* $\mathbf{x}_t \in \{0, 1, 2\}^N$, *where* T *is the total number of simulated time steps, and* $[\![l, m]\!] = \{l, l+1, \ldots, m-1, m\}$ *for integers* $l \leq m$. *The states* $\mathbf{x}_{t,n}$ *of each of the agents* $n \in [\![1, N]\!]$ *are updated synchronously as follows for* $t \in [\![0, T-1]\!]$*:*

(U1) *If* $\mathbf{x}_{t,n} = 0$, *then* $\mathbf{x}_{t+1,n} = 1$ *with probability*

$$p_{t,n}(\alpha_{t+1}) = 1 - (1 - \alpha_{t+1})^{\sum_{n' \in \mathcal{N}_n} \mathbb{I}[\mathbf{x}_{t,n'} = 1]} \tag{1}$$

where $\mathcal{N}_n$ *is the von Neumann neighbourhood for cell* n*; else remain susceptible.*

(U2) *If* $\mathbf{x}_{t,n} = 1$, *then* $\mathbf{x}_{t+1,n} = 2$ *with probability* β_{t+1}*; else remain infected.*

(U3) *If* $\mathbf{x}_{t,n} = 2$, *then* $\mathbf{x}_{t+1,n} = 0$ *with probability* γ_{t+1}*; else remain recovered.*

In the above, $\boldsymbol{\theta}_t = (\alpha_t, \beta_t, \gamma_t) \in [0, 1]^3$ *are the model parameters determining the transition probabilities between states. While these may vary over time, the simplest case consists of assigning all* $\boldsymbol{\theta}_t$ *the same vector,*

$$\boldsymbol{\theta}_t = \mathbf{v} \qquad \forall t \in [\![1, T]\!]. \tag{2}$$

The model is initialised by infecting each agent in the model at initial time $t = 0$ with probability $I_0 \in [0, 1]$. The value of I_0 for any forward simulation of the model can be chosen by drawing a random variable a from some distribution on $[0, 1]$ and setting

$$I_0 = a. \tag{3}$$

With this model in place, lockdowns over some time period $t_l : t_l + \Delta$ of length $\Delta \geq 0$ can be modelled (crudely) by setting $\boldsymbol{\theta}_{t_l:t_l+\Delta} = (0, \beta, \gamma)$ for $\beta, \gamma \in [0, 1]$. To express this ABM as an SCM, we define the following:

Endogenous variables *These consist of the variables of interest that may be set by the policymaker: I_0, $\{\mathbf{x}_t\}_{0 \leq t \leq T}$, and $\{\boldsymbol{\theta}_t\}_{1 \leq t \leq T}$.*

Exogenous variables *The model as described above is initialised randomly according to a, $\mathbf{v}$, and a collection $\mathbf{u}_0 = (\mathbf{u}_{0,n})_{1 \leq n \leq N}$ of N random variables distributed as $\mathcal{U}(0, 1)$, the nth of which helps determine whether agent n is infected at time $t = 0$. Similarly, further collections $\mathbf{u}_t, t \in [\![1, T]\!]$ of $\mathcal{U}(0, 1)$ random variables decide how each agent updates their state at each time step. Thus the exogenous variables for the model are a, $\mathbf{v}$, and the $\mathbf{u}_t$ for $t \in [\![0, T]\!]$.*

Structural equations *Equations 2 and 3, respectively, define the structural equations $f_{\boldsymbol{\theta}_t}$ and f_{I_0} for the endogenous variables $\boldsymbol{\theta}_t$ and I_0. The structural equation $f_{\mathbf{x}_{0,n}}$ for each $\mathbf{x}_{0,n}, n \in [\![1, N]\!]$ can furthermore be written as*

$$\mathbf{x}_{0,n} = f_{\mathbf{x}_{0,n}}(\mathbf{u}_{0,n}, I_0) = \mathbb{I}\left[\mathbf{u}_{0,n} < I_0\right]. \tag{4}$$

*Finally, update rules (**U1**)-(**U3**) can be written in the following way for $t \in [\![0, T-1]\!]$:*

$$\begin{aligned} \mathbf{x}_{t+1,n} &= f_{\mathbf{x}_{t+1,n}}(\boldsymbol{\theta}_{t+1}, \mathbf{u}_{t+1,n}, \mathbf{x}_{t,n}) \\ &= \mathbb{I}\left[\mathbf{x}_{t,n} = 0\right] \cdot \mathbb{I}\left[\mathbf{u}_{t+1,n} < p_{t,n}(\alpha_{t+1})\right] + \mathbb{I}\left[\mathbf{x}_{t,n} = 1\right] \cdot (1 + \mathbb{I}\left[\mathbf{u}_{t+1,n} < \beta_{t+1}\right]) \\ &\qquad + 2\mathbb{I}\left[\mathbf{x}_{t,n} = 2\right] \cdot (1 - \mathbb{I}\left[\mathbf{u}_{t+1,n} < \gamma_{t+1}\right]), \end{aligned} \tag{5}$$

Distribution over exogenous variables *The (random) behaviour of the exogenous variables is fully specified by the distribution over a and $\mathbf{v}$, along with $\mathcal{U}(0, 1)$ distributions over the $\mathbf{u}_{t,n}$.*

The underlying graph *The DAG corresponding to this SCM is shown in Figure 1 for $T = 3$.*

In this model, interventions in the form of, for example, lockdowns can be (crudely) modelled by intervening on one or more of the $\boldsymbol{\theta}_t$ as $\mathrm{do}(\boldsymbol{\theta}_t = (0, \beta, \gamma))$ for some $\beta, \gamma \in [0, 1]$, while in the observational regime the $\boldsymbol{\theta}_t$ will all be assigned the same value.

We emphasise that the above example is intended only to illustrate that complex simulators, such as ABMs, can be seen as SCMs; explicitly representing a given simulator as an SCM as in the example above is not required in the sequel.

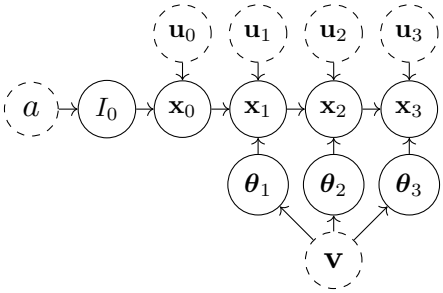

Figure 1: The directed acyclic graph induced by the structural causal model for the spatial SIRS agent-based model for $T = 3$ time steps.

2.3 Causal abstractions

Beside expressing interventions more rigorously, viewing complex simulation models as SCMs allows one to take advantage of the theory of *causal abstraction* to formalise the relationship between the simulator and its surrogate model. Indeed, causal abstraction provides a framework for relating SCMs representing an identical system at different levels of granularity. The notion of exact transformation formalizes this relation, providing a framework to relate complex models, such as ABMs, to simpler top-down models, while preserving causal structure.

Definition 3 (τ-ω Exact Transformation [Rubenstein et al., 2017])**.** *Given two SCMs, $\mathcal{M}$ and $\mathcal{M}'$, with respective intervention sets $\mathcal{I}$ and $\mathcal{I}'$, a τ-ω* transformation *is a pair (τ, ω) consisting of a*

map $\tau : \text{dom}[\mathbf{X}] \to \text{dom}[\mathbf{X}']$ *and a surjective, order-preserving map* $\omega : \mathcal{I} \to \mathcal{I}'$. *An* exact τ-ω *transformation is a* τ-ω *transformation such that*

$$\tau_{\#}(\mathbb{P}_{\mathcal{M}_\iota}) = \mathbb{P}_{\mathcal{M}'_{\omega(\iota)}}, \forall \iota \in \mathcal{I}. \tag{6}$$

An exact τ-ω transformation constitutes a form of abstraction between probabilistic causal models [Beckers et al., 2020] with the guarantee of commutativity between intervention and transformation as detailed in Figure 2: intervening via ι and then abstracting produces the same result as abstracting first and then intervening via $\omega(\iota)$. The map τ describes corresponding states in each of the models, while the map ω describes corresponding interventions in each model. Whenever the map τ is clear from context, we herein shorthand the pushforward measure $\tau_{\#}(\mathbb{P}_{\mathcal{M}_\iota})$ as $\mathbb{P}^{\#}_{\mathcal{M}_\iota}$.

An exact τ-ω transformation between the SCM $\mathcal{M}$ underlying a complex simulation model and the SCM $\mathcal{M}'$ underlying the candidate surrogate model would (a) certify that the surrogate preserves the causal behaviour of interest, guaranteeing interventional consistency when policymakers study interventions through the surrogate, and (b) allow to interpret the emergent causal structure of the simulator through $\mathcal{M}'$.

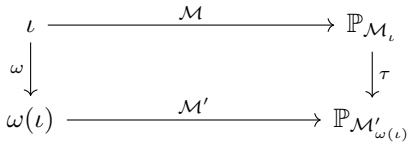

Figure 2: Computing $\tau_{\#}(\mathbb{P}_{\mathcal{M}_\iota})$ corresponds to moving right, then down, in the diagram. That is, running the intervention ι in a base model $\mathcal{M}$ such as an ABM. Computing $\mathbb{P}_{\mathcal{M}'_{\omega(\iota)}}$ corresponds to moving down, then right. That is, running the intervention $\omega(\iota)$ in an abstracted model $\mathcal{M}'$ such as a surrogate. If (τ, ω) is an exact transformation, then the diagram is commutative for all interventions. That is, $\tau_{\#}(\mathbb{P}_{\mathcal{M}_\iota}) = \mathbb{P}_{\mathcal{M}'_{\omega(\iota)}}$ for all $\iota \in \mathcal{I}$.

3 Abstraction error

It is often unrealistic to assume that an exact τ-ω transformation exists between a complex simulator and its surrogate A more pragmatic goal is to find an approximate abstraction [Beckers et al., 2020] from the simulator to the surrogate. We therefore define the *abstraction error*:

Definition 4 (Abstraction error)**.** *Let* (τ, ω) *be a* τ-ω *transformation between two* SCMs $\mathcal{M}$ *and* $\mathcal{M}'$ *with respective intervention sets* $\mathcal{I}$ *and* $\mathcal{I}'$. *Given a statistical divergence* d *between distributions, and a distribution* η *over the intervention set* $\mathcal{I}$, *we define the* abstraction error *as follows:*

$$d_{\tau,\omega}(\mathcal{M}, \mathcal{M}') = \mathbb{E}_{\iota \sim \eta}\left[d\left(\tau_{\#}(\mathbb{P}_{\mathcal{M}_\iota}), \mathbb{P}_{\mathcal{M}'_{\omega(\iota)}}\right)\right]. \tag{7}$$

A τ-ω *transformation is* α-approximate *for some* $\alpha \in \mathbb{R}_{\geq 0}$ *if* $d_{\tau,\omega}(\mathcal{M}, \mathcal{M}') \leq \alpha$.

A τ-ω abstraction with low abstraction error implies that $\tau_{\#}(\mathbb{P}_{\mathcal{M}_\iota})$ is close to $\mathbb{P}_{\mathcal{M}'_{\omega(\iota)}}$ in expectation with respect to the interventional distribution η. If the $d\left(\tau_{\#}(\mathbb{P}_{\mathcal{M}_\iota}), \mathbb{P}_{\mathcal{M}'_{\omega(\iota)}}\right)$ is zero for all interventions $\iota \in \mathcal{I}$, then (τ, ω) is an exact transformation (see Figure 3).

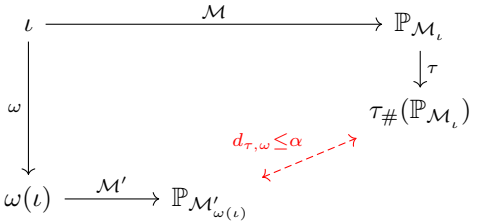

Figure 3: The abstraction error compares the distributions $\tau_{\#}(\mathbb{P}_{\mathcal{M}_\iota})$ and $\mathbb{P}_{\mathcal{M}'_{\omega(\iota)}}$ for each intervention ι using the divergence $d_{\tau,\omega}$, as indicated by the red dotted arrow. If the divergence is zero then we recover the commutative diagram in Figure 2.

Definition 4 differs from previously defined notions of abstraction error in the causal abstraction literature. Whilst Beckers et al. [2020] employ a maximum over interventions, we instead take an expectation over a fixed interventional distribution η. This is motivated by the fact that policymakers will often hold preferences over possible interventions, which may, for example, reflect the cost or feasibility of implementing each intervention in the real world. Through the specification of η, one may implicitly favour surrogates which perform well with respect to interventions of high interest. Further discussion is provided in Appendix B.

4 Method

The definitions of abstraction and abstraction error provide us with a framework for learning surrogates, and, in the remainder, we assume that the base model $\mathcal{M}$ is implicitly represented by a

simulation model of a complex socio-technical system. Our goal is then to identify a surrogate model which is interventionally consistent with this simulator. Specifically, from a set of candidate surrogate models $\mathfrak{M}$, we seek a surrogate and a τ-ω transformation that minimises the abstraction error.

To proceed, we assume that $\mathfrak{M}$ is a parameterised family $\mathfrak{M}^{\Psi} := \{\mathcal{M}^{\psi} : \psi \in \Psi\}$ of differentiable surrogate simulators with tractable probability mass or density function q^{ψ}. Here, $\mathcal{M}^{\psi}$ denotes the causal model induced by a surrogate whose structural equations are parameterised by ψ, and Ψ denotes the set of feasible parameter values. Such a family of surrogate models can be constructed through a composition of differential equation- or deep learning-based modelling, in combination with probability distributions with reparameterisable sampling procedures; an example is a latent neural ordinary differential equation model [Rubanova et al., 2019], which we use in the experiments in Section 5. We further assume only the ability to sample from $\tau_{\#}(\mathbb{P}_{\mathcal{M}})$, amounting to running the simulator and applying τ to the output.

Generally speaking, policymakers know what macroscopic quantities are of interest when modelling a complex system, and how to aggregate the microscopic variables into global statistics. For example, in macroeconomic settings, policymakers will often be concerned with aggregate quantities such as unemployment rates or aggregate demand, which can be derived from the state of the agents. Further specific examples are discussed in Appendix C.1. We thus assume that the map τ, which defines the aggregate, emergent quantities of interest to the policymaker, is pre-specified.

Hence, to find an appropriate τ-ω transformation, we need only to identify an intervention map $\omega^{\star}$ between $\mathcal{I}$ and $\mathcal{I}'$. For computational tractability, we select $\omega^{\star}$ from a parameterised family $\Omega^{\Phi} := \{\omega^{\phi} : \phi \in \Phi\}$ with parameters ϕ ranging over the set Φ. For example, ϕ may be the weights of a neural network. We then select $\phi^{\star}$ and $\psi^{\star}$ jointly by minimising $d_{\tau,\omega}(\mathcal{M}, \mathcal{M}')$ over $\Omega^{\Phi} \times \mathfrak{M}^{\Psi}$. Since each element of $\mathfrak{M}$ has a differentiable and tractable distribution, a convenient choice of discrepancy d is the Kullback-Leibler (KL) divergence, such that out problem becomes:

$$\phi^{\star}, \psi^{\star} = \underset{\phi \in \Phi, \psi \in \Psi}{\arg\min}\, d_{\tau,\omega^{\phi}}(\mathcal{M}, \mathcal{M}^{\psi}) \quad \text{with} \quad d_{\tau,\omega^{\phi}}(\mathcal{M}, \mathcal{M}^{\psi}) = \mathbb{E}_{\eta}\mathbb{E}_{\mathbb{P}^{\#}_{\mathcal{M}_{\iota}}}\left[\log \frac{\mathrm{d}\mathbb{P}^{\#}_{\mathcal{M}_{\iota}}}{\mathrm{d}\mathbb{P}_{\mathcal{M}^{\psi}_{\omega^{\phi}(\iota)}}}\right]. \quad (8)$$

The KL divergence can be minimised using Monte Carlo estimates of the gradient

$$G(\phi, \psi) = \nabla_{\phi,\psi}\, d_{\tau,\omega^{\phi}}(\mathcal{M}, \mathcal{M}^{\psi}) \approx \frac{1}{B}\sum_{b=1}^{B} -\nabla_{\phi,\psi} \log q^{\psi}_{\omega^{\phi}(\iota^{(b)})}(\mathbf{y}^{(b)}) \quad (9)$$

where $\iota^{(b)} \sim \eta$, $\mathbf{y}^{(b)} \sim \tau_{\#}(\mathbb{P}_{\mathcal{M}_{\iota^{(b)}}})$, $q^{\psi}_{\omega^{\phi}(\iota)}$ is the probability mass/density function for $\mathcal{M}^{\psi}_{\omega^{\phi}(\iota)}$, and $B \geq 1$ is the size of a batch drawn from $R \geq B$ training examples from the joint distribution over the $\iota^{(b)}$ and $\mathbf{y}^{(b)}$. Once $(\phi^{\star}, \psi^{\star})$ has been selected, we may generate data from the macromodel for `ABM` intervention ι by sampling from $\mathbb{P}_{\mathcal{M}^{\psi^{\star}}_{\omega^{\phi^{\star}}(\iota)}}$. Algorithm 1 summarises the training procedure.

4.1 Theory

Definition 4 is closely related to exact transformations:

Proposition 1. *Let η be an interventional distribution, d be a statistical divergence, and (τ, ω) be a τ-ω transformation between `SCMs` $\mathcal{M}$ and $\mathcal{M}'$. If τ-ω is 0-approximate ($d_{\tau,\omega}(\mathcal{M}, \mathcal{M}') = 0$), then we have η-almost-surely*

$$\tau_{\#}(\mathbb{P}_{\mathcal{M}_{\iota}}) = \mathbb{P}_{\mathcal{M}'_{\omega(\iota)}}.$$

The proof is in Appendix D.1. In particular, when $\mathcal{I}$ is finite and $\eta(\iota) > 0\ \forall \iota \in \mathcal{I}$, then any 0-approximate τ-ω transformation is an exact τ-ω transformation between $\mathcal{M}$ and $\mathcal{M}'$. This motivates our own choice of loss

Algorithm 1: Summary of the training procedure.

Input: Budget R; batch size $B \in [\![1, R-1]\!]$; `ABM` $\mathcal{M}$; intervention distribution η; surrogate family $\mathfrak{M}^{\Psi}$; abstraction map family Ω^{Φ}

Result: Trained surrogate and abstraction map parameters, ψ^{*} and ϕ^{*}

Set $\mathcal{D} = \emptyset$;
for $r = 1$ **to** R **do**
  Sample $\iota^{(r)} \sim \eta$, $\mathbf{x}^{(r)} \sim \mathbb{P}_{\mathcal{M}_{\iota^{(r)}}}$;
  $\mathcal{D} \leftarrow \mathcal{D} \cup (\iota^{(r)}, \tau(\mathbf{x}^{(r)}))$
end
while *not converged* **do**
  Sample minibatch $\{(\iota^{(b)}, \tau(\mathbf{x}^{(b)}))\}_{b=1}^{B}$ uniformly from $\mathcal{D}$;
  Take gradient step in ϕ, ψ using Equation 9
end

function: minimising Equation 8 induces ϕ and ψ to produce a surrogate that behaves the same way as the simulator under interventions of interest.

Definition 4 employs an expectation over an interventional distribution η. As a result, even when the abstraction error is low, there may still be a large discrepancy between $\tau_{\#}(\mathbb{P}_{\mathcal{M}_\iota})$ and $\mathbb{P}_{\mathcal{M}'_{\omega(\iota)}}$ for some fixed intervention $\iota \in \mathcal{I}$. Proposition 2 provides an upper bound on the error associated with any intervention sampled from η when d is the KL divergence and the simulator state space is finite:

Proposition 2. *Let d be the KL divergence and $\texttt{CE}_\iota = \mathbb{E}_{\mathbf{Y}\sim\tau_{\#}(\mathbb{P}_{\mathcal{M}_\iota})}\left[-\log q_{\omega(\iota)}(\mathbf{Y})\right]$ denote the cross-entropy of $\mathbb{P}_{\mathcal{M}'_{\omega(\iota)}}$ with respect to $\tau_{\#}(\mathbb{P}_{\mathcal{M}_\iota})$. Assume $\mathrm{dom}[\mathbf{X}]$ is finite. Then for all $\varepsilon > 0$,*

$$\mathbb{P}_\eta\left(d\left(\tau_{\#}(\mathbb{P}_{\mathcal{M}_\iota}), \mathbb{P}_{\mathcal{M}'_{\omega(\iota)}}\right) \geq \varepsilon\right) \leq \frac{\mathbb{E}_{\iota\sim\eta}[\texttt{CE}_\iota]}{\varepsilon}.$$

The proof is in Appendix D.2. This shows that it is only with low probability that the effects of individual interventions are captured poorly by the surrogate when the surrogate and abstraction map parameters, ψ and ϕ, are found by minimising Equation 8.

5 Case study

Here, we outline a case study[2] in which we learn interventionally consistent surrogates for the spatial SIRS ABM from Example 1, allowing us to experiment more rapidly with policy interventions while remaining confident that the causal behaviour of the original SIRS ABM is approximately preserved. Further experimental details and results are given in Appendix E. We consider three families of surrogate models with endogenous variables $\tilde{I}_0 \in [0,1]$, $\tilde{\boldsymbol{\theta}}_t \in \mathbb{R}^3_{\geq 0}$ for $t \in [\![1, T]\!]$, and $\tilde{\mathbf{y}}_t \in \{(a,b,c) \mid a,b,c \in [\![0,N]\!], a+b+c = N\}$ for $t \in [\![0,T]\!]$, where a, b, c denote, respectively, the number of susceptible, infected, and recovered individuals in the population. The DAGs underlying the SCMs of each of these families are as in Figure 4, and the three families differ only in the form of the structural equations mapping from $\tilde{I}_0$ and $\tilde{\boldsymbol{\theta}}_{0:t}$ to the $\tilde{\mathbf{y}}_t$. Throughout, we let q^ψ be a Multinomial emission distribution and ψ be trainable parameters of these structural equations.

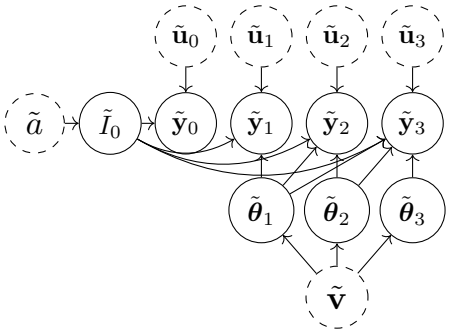

Figure 4: The DAG induced by the SCMs corresponding to the surrogate families for $T = 3$.

Surrogate family 1 consists of a latent ODE (LODE) built by feeding the classical SIRS ODE's three state variables (which take values in the two-simplex) in as the class probabilities of q^ψ. Here, $\psi = \emptyset$.

Surrogate family 2 consists of a latent ODE-RNN (LODE-RNN), where we run a recurrent network (RNN) with parameters ψ over the output of the SIRS ODE. The RNN outputs the class logits of q^ψ.

Surrogate family 3 consists of a latent RNN (LRNN) constructed by running an RNN with trainable parameters ψ over the $\tilde{\boldsymbol{\theta}}_t$, and the output of the RNN at each $t \in [\![0,T]\!]$ indexes the class logits of q^ψ.

Given $\tilde{\boldsymbol{\theta}}_{1:T}, \tilde{I}_0$, these surrogates enjoy tractable likelihood functions, which factorise as $q^\psi(\tilde{\mathbf{y}}_{0:T} \mid \tilde{\boldsymbol{\theta}}_{1:T}, \tilde{I}_0) = q^\psi(\tilde{\mathbf{y}}_0 \mid \tilde{I}_0) \prod_{t=1}^T q^\psi(\tilde{\mathbf{y}}_t \mid \tilde{\boldsymbol{\theta}}_{1:t}, \tilde{I}_0)$.

Interventions & the τ-ω transformation. Denoting

$$\begin{aligned} \iota_{\mathbf{v},a} &= \mathrm{do}\left(\boldsymbol{\theta}_{1:T} = \mathbf{v}, I_0 = a\right), \\ \iota_{\mathbf{v},a,t_l} &= \mathrm{do}\left(\boldsymbol{\theta}_{1:t_l-1} = \boldsymbol{\theta}_{t_l+6:T} = \mathbf{v}, \boldsymbol{\theta}_{t_l:t_l+5} = \mathbf{v} \odot (0,1,1), I_0 = a\right), \end{aligned} \tag{10}$$

we define two subsets $\mathcal{I} = \mathcal{I}_{\mathrm{init}} \cup \mathcal{I}_{\mathrm{init,\,lock}}$ of interventions for the ABM:

$$\mathcal{I}_{\mathrm{init}} = \{\iota_{\mathbf{v},a} \mid (\mathbf{v}, a) \in [0,1]^4\} \quad \text{and} \quad \mathcal{I}_{\mathrm{init,\,lock}} = \{\iota_{\mathbf{v},a,t_l} \mid (\mathbf{v}, a, t_l) \in [0,1]^4 \times [\![5,10]\!]\}. \tag{11}$$

[2]Code for reproducing the experimental results is available at `https://github.com/joelnmdyer/neurips_ics4csm`.

Table 1: Metrics for interventionally (**I**) & observationally (**O**) trained surrogates on interventional ($\mathbf{I}'$) & observational ($\mathbf{O}'$) test sets ($\text{median}^{\text{third quartile}}_{\text{first quartile}}$ from 5 repeats). **Bold** denotes best performance.

Test	**Model**	LRNN		LODE-RNN		LODE	
	Train	**I**	**O**	**I**	**O**	**I**	**O**
$\mathbf{I}'$	AMSE ($\times 10^{-1}$)	$\mathbf{3.48^{3.91}_{3.41}}$	$49.4^{52.6}_{46.7}$	$\mathbf{3.35^{3.41}_{3.18}}$	$18.5^{21.9}_{17.1}$	$\mathbf{8.15^{8.24}_{8.06}}$	$22.4^{22.7}_{22.1}$
	ANLL ($\times 10^{3}$)	$\mathbf{2.09^{2.16}_{2.03}}$	$21.8^{22.9}_{20.1}$	$\mathbf{1.99^{2.00}_{1.98}}$	$8.40^{9.89}_{8.27}$	$\mathbf{4.01^{4.02}_{4.00}}$	$10.0^{10.1}_{9.91}$
$\mathbf{O}'$	AMSE ($\times 10^{-1}$)	$4.13^{4.26}_{4.11}$	$\mathbf{2.95^{3.16}_{2.62}}$	$3.59^{3.68}_{3.54}$	$\mathbf{2.52^{2.78}_{2.16}}$	$18.4^{18.7}_{18.1}$	$\mathbf{4.36^{4.40}_{4.32}}$
	ANLL ($\times 10^{3}$)	$2.22^{2.23}_{2.16}$	$\mathbf{1.64^{1.71}_{1.43}}$	$1.86^{1.97}_{1.85}$	$\mathbf{1.43^{1.53}_{1.27}}$	$7.63^{7.74}_{7.52}$	$\mathbf{2.15^{2.17}_{2.13}}$

The first of these is a subset of interventions that fix the initial proportion of infected individuals in the ABM, as well as its parameter values. The second subset of interventions is the set of interventions that fix (a) the initial proportion of infected individuals in the ABM, (b) the values of the ABM's parameters before, during, and beyond a lockdown beginning at time $t_l \in [\![5, 10]\!]$ with duration equal to 5 time steps, and (c) the value of t_l. Similarly defining

$$\iota'_{\tilde{\mathbf{v}},\tilde{a}} = \mathrm{do}\left(\tilde{\boldsymbol{\theta}}_{1:T} = \tilde{\mathbf{v}}, \tilde{I}_0 = \tilde{a}\right), \tag{12}$$
$$\iota'_{\tilde{\mathbf{v}},\tilde{a},\tilde{t}_l} = \mathrm{do}\left(\tilde{\boldsymbol{\theta}}_{1:\tilde{t}_l-1} = \tilde{\boldsymbol{\theta}}_{\tilde{t}_l+6:T} = \tilde{\mathbf{v}}, \tilde{\boldsymbol{\theta}}_{\tilde{t}_l:\tilde{t}_l+5} = \tilde{\mathbf{v}} \odot (0,1,1), = \tilde{\mathbf{v}}, \tilde{I}_0 = \tilde{a}\right),$$

we define $\mathcal{I}' = \mathcal{I}'_{\text{init}} \cup \mathcal{I}'_{\text{init, lock}}$ for the surrogates, where, letting $\mathbb{D} = \mathbb{R}^3_{\geq 0} \times [0,1]$, we have

$$\mathcal{I}'_{\text{init}} = \{\iota'_{\tilde{\mathbf{v}},\tilde{a}} \mid (\tilde{\mathbf{v}}, \tilde{a}) \in \mathbb{D}\} \quad \text{and} \quad \mathcal{I}'_{\text{init, lock}} = \{\iota'_{\tilde{\mathbf{v}},\tilde{a},\tilde{t}_l} \mid (\tilde{\mathbf{v}}, \tilde{a}, \tilde{t}_l) \in \mathbb{D} \times [\![5, 10]\!]\}. \tag{13}$$

The map τ is taken to map: $\boldsymbol{\theta}_t$ identically to $\tilde{\boldsymbol{\theta}}_t$ for each $t \in [\![1, T]\!]$; the microstate $\mathbf{x}_t$ of the ABM at each time step to the $\tilde{\mathbf{y}}_t$ through an aggregation map that counts the number of agents in $\mathbf{x}_t$ in each of the three states (susceptible, infectious, and recovered); and the initial proportion I_0 of infected agents in the ABM identically to $\tilde{I}_0$. Further, for a neural network $f^\phi : [0,1]^3 \to \mathbb{R}^3_{\geq 0}$, we take

$$\omega^\phi : \quad \iota_{\mathbf{v},a} \mapsto \iota'_{f^\phi(\mathbf{v}),a} \quad , \quad \iota_{\mathbf{v},a,t_l} \mapsto \iota'_{f^\phi(\mathbf{v}),a,t_l}. \tag{14}$$

The benefits of training for interventional consistency We use Algorithm 1 to jointly learn the parameters ϕ, ψ of the surrogates and the map ω^ϕ described above in two different ways: training the surrogate models with η taken to be a uniform distribution $\mathcal{U}(\mathcal{I}_{\text{init}})$ over $\mathcal{I}_{\text{init}}$, which entails comparing the behaviour of the surrogate and ABM without lockdowns at different parameters; and training with η instead taken to be a uniform distribution $\mathcal{U}(\mathcal{I})$ over $\mathcal{I}$, which entails comparing the behaviour of the surrogate and ABM under different lockdowns, or no lockdowns at all, at different parameters. We indicate the two approaches to training the surrogates with, respectively, bold uppercase **O** and **I**. Appendix E details the training procedure and network architectures. We assess the interventional consistency of the surrogates trained in these two ways by computing error metrics on a hold-out test dataset $\mathbf{I}' = \{(\iota^{(r')}, \mathbf{y}^{(r')}_{0:T})\}_{r'=1}^{R'}$ of size $R' = 1000$, generated as $\iota^{(r')} \sim \eta = \mathcal{U}(\mathcal{I})$, $\mathbf{y}^{(r')}_{0:T} \sim \tau_{\#}\left(\mathbb{P}_{\mathcal{M}_{\iota^{(r')}}}\right)$. Specifically, we inspect the average mean squared error (AMSE) between trajectories from the trained surrogates and $\mathbf{y}^{(r')}_{0:T}$, and the average negative log-likelihood (ANLL) of this test data under the likelihood of the learned surrogates. Observational consistency is checked on a different hold-out test set $\mathbf{O}'$, generated by instead taking $\eta = \mathcal{U}(\mathcal{I}_{\text{init}})$.

Table 1 shows these performance metrics evaluated on $\mathbf{I}'$ and $\mathbf{O}'$ for all surrogate families and training schemes. We observe that far lower values of the error metrics are obtained by the interventionally, rather than observationally, trained surrogates when assessing interventional consistency. This suggests that training on interventional data can result in more accurate predictions about the effect of interventions in the ABM, and that data drawn from the relevant interventional distributions associated with the ABM should be included during training if the policy-maker intends to perform policy experiments with the surrogate. We also report a minor drop in observational consistency when training with data from the combined intervention set $\mathcal{I}$ instead of $\mathcal{I}_{\text{init}}$, which can be explained by the overfit of the observationally-trained model on the observational distribution. We also observe

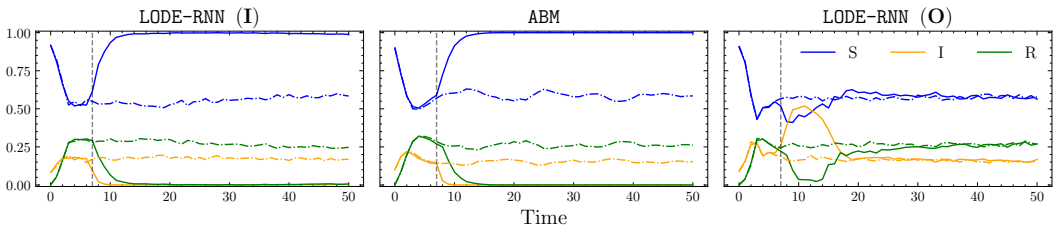

Figure 5: Example trajectories from the ABM (middle) and the LODE-RNN trained interventionally (left) and observationally (right). A lockdown is imposed at the dashed vertical line. Solid (resp. dot-dash) lines show trajectories under (resp. without) the lockdown. The transmission-inhibiting effect of the lockdown is vastly underestimated in the observationally trained surrogate, while the interventionally trained surrogate accurately predicts a reduction in disease transmission.

that the LODE-RNN – which combines the "mechanistic" SIRS ODE with a flexible RNN – achieves the best interventional and observational consistencies of all surrogates, suggesting that such hybrid approaches to constructing flexible surrogates are promising choices under our proposed method.

In Figure 5, we show an example of a possible negative consequence of failing to train a surrogate on data drawn from the appropriate ABM interventional distributions. In the middle panel, we show the change in the ABM trajectory induced by imposing a lockdown at time $t_l = 7$, while in the left (resp. right) panel we show corresponding trajectories from the interventionally (resp. observationally) trained surrogates under the equivalent intervention learned through our training procedure. While the interventionally trained LODE-RNN correctly predicts that the lockdown effectively impedes the spread of the disease in the ABM, the observationally trained surrogate predicts that the lockdown will temporarily *increase* infections, before approximately reverting to the behaviour of the model without a lockdown.

The use of such a surrogate model in policy experiments when limited computational resources do not permit use of the accurate, high-fidelity ABM of the underlying complex system may therefore have misdirected policy-makers towards suboptimal, and away from effective, interventions. Indeed, while the SIRS ABM predicts that any lockdown is better than no lockdown at all for reducing the number of infections occurring over the simulated time horizon, we see that the observationally trained surrogates often do not predict that no lockdown is the worst intervention in this respect, and in some cases mistakenly predict that no lockdown is the *best* intervention. For example, the observational LRNN predicts that no lockdown was the best intervention in 1 of 5 training repeats, and was not the worst option in all 5 of 5 training repeats. In contrast, none of the interventionally trained surrogates predict that no lockdown is the best intervention, and only the interventional LODE model predicts that no lockdown is not the worst option (in only 2 out of 5 training repeats). This highlights the potential importance of training surrogate models for interventional consistency when their purpose is to help inform downstream decision-making tasks. Furthermore, this suggests that a possible benchmark criterion in further research on interventional surrogates could be the degree to which different surrogates preserve the ordering of interventions with respect to those downstream tasks of interest.

6 Related work

Surrogates are often used to expedite simulation-based inference when modelling complex systems [Heppenstall et al., 2021]. Modern approaches rely on established machine learning methods such as random forests [Lamperti et al., 2018, De Leeuw et al., 2023], artificial neural networks [Anirudh et al., 2022, De Leeuw et al., 2023], support vector machines [ten Broeke et al., 2021], kriging [Salle and Yıldızoğlu, 2014], and mixture density networks (MDNs) [Platt, 2022]. Our experiments also rely on established machine learning methods to construct surrogates; for example, our LRNN surrogate family resembles that of Platt [2022], in which MDNs are used to approximate an ABM's transition density. However, in such works, the *causal/interventional* consistency of the surrogate with respect to the simulator and policy interventions of interest is not considered. In contrast to prior work, our work explicitly details the causal relation between the surrogate and the underlying simulator via

causal abstraction, which broadens the scope of surrogate modelling beyond its current use case of expediting calibration to also enable the use of surrogates for policy experimentation.

Causal abstraction and exact transformations were introduced by Rubenstein et al. [2017]. Beckers and Halpern [2019] extended this work by proposing stricter definitions of causal abstraction, and in Beckers et al. [2020], where approximate abstractions are introduced to account for uncertainty and simplification. Causal abstraction found practical application in Geiger et al. [2021] for learning interpretable neural networks. Rischel and Weichwald [2021] discusses an alternative category-theoretical definition of abstraction; this was used to learn abstractions to transfer data between models at different levels of abstraction in Zennaro et al. [2023a]. Further related work includes a multi-marginal Optimal Transport solution to the abstraction learning problem [Felekis et al., 2024], as well as constructive abstraction learning in neural causal models [Xia et al., 2023] and cluster DAGs [Anand et al., 2023]. However, none of these approaches reduce the state space of the SCM or the cost of simulation, as our approach does.

7 Conclusion

We propose a rigorous framework for learning interventionally consistent surrogates for complex simulation models, formalised with casual abstraction. This is the first application of causal abstraction to surrogate modelling. Our approach applies to any simulator corresponding to any DAG, and does not require explicit knowledge of the simulator's SCM. Through experiments, we highlight the efficacy of our method against purely observational surrogates that do not learn to match interventional data under equivalent interventions. Using our framework, policy-makers may be able to more rapidly draw insights from complex simulators about the possible effects of interventions – in our experiments, our surrogates simulate approximately three times faster than the original complex simulators – and swiftly prepare effective responses to future crises.

Our work naturally suffers limitations. Investigating the sample complexity of abstraction learning would be desirable in future work. Our definition of abstraction error involves an expectation over interventions rather than a maximum as in Beckers et al. [2020]; this produces a computationally tractable optimisation problem, but introduces the possibility that one or more interventions is captured poorly by the learned abstraction map, even for a low abstraction error. In our experiments, we have assumed surrogate models with tractable and differentiable density functions, permitting us to use a KL divergence within our definition of abstraction error; future work might extend our approach by considering different surrogate families with these properties, such as families based on normalising flows [Tabak and Vanden-Eijnden, 2010], or alternative divergences that relax the requirement for tractable densities, such as maximum mean discrepancies [Gretton et al., 2012]. Finally, our method does not directly exploit knowledge of the simulators' causal graphs to accelerate abstraction learning. It is possible that exploiting access to the base SCM/DAG may expedite abstraction by allowing us to focus on minimal intervention sets [Aglietti et al., 2020, Lee and Bareinboim, 2018], or leverage the identifiability of interventional distributions to reduce the number of simulations required from the base model [Lattimore et al., 2016, Bilodeau et al., 2022]. However, it is unclear whether or not applying the do-calculus on large causal graphs is more efficient than simulating interventions directly. The "black-box" nature of our approach may be beneficial for this reason, and since it does not require the modeller to explicitly write their simulator as an SCM, making it generically applicable.

Acknowledgments and Disclosure of Funding

JD, NB, AC, and MW acknowledge funding from a UKRI AI World Leading Researcher Fellowship awarded to Wooldridge (grant EP/W002949/1). MW and AC also acknowledge funding from Trustworthy AI - Integrating Learning, Optimisation and Reasoning (TAILOR), a project funded by European Union Horizon2020 research and innovation program under Grant Agreement 952215. YF: This scientific paper was supported by the Onassis Foundation - Scholarship ID: F ZR 063-1/2021-2022. TD acknowledges support from a UKRI Turing AI Acceleration Fellowship [EP/V02678X/1].

References

Virginia Aglietti, Xiaoyu Lu, Andrei Paleyes, and Javier González. Causal bayesian optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 3155–3164. PMLR, 2020.

Tara V Anand, Adele H Ribeiro, Jin Tian, and Elias Bareinboim. Causal effect identification in cluster dags. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12172–12179, 2023.

Rushil Anirudh, Jayaraman J. Thiagarajan, Peer-Timo Bremer, Timothy Germann, Sara Del Valle, and Frederick Streitz. Accurate calibration of agent-based epidemiological models with neural network surrogates. In Peng Xu, Tingting Zhu, Pengkai Zhu, David A. Clifton, Danielle Belgrave, and Yuanting Zhang, editors, *Proceedings of the 1st Workshop on Healthcare AI and COVID-19, ICML 2022*, volume 184 of *Proceedings of Machine Learning Research*, pages 54–62. PMLR, 22 Jul 2022. URL `https://proceedings.mlr.press/v184/anirudh22a.html`.

Rafa Baptista, J Doyne Farmer, Marc Hinterschweiger, Katie Low, Daniel Tang, and Arzu Uluc. Staff Working Paper No. 619 Macroprudential policy in an agent-based model of the UK housing market. 2016.

Marco Bardoscia, Adrian Carro, Marc Hinterschweiger, Mauro Napoletano, Andrea Roventini, and Arzu Uluc. The impact of prudential regulations on the uk housing market and economy: insights from an agent-based model. *Bank of England Working Paper*, 2024.

Sander Beckers and Joseph Y Halpern. Abstracting causal models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2678–2685, 2019.

Sander Beckers, Frederick Eberhardt, and Joseph Y Halpern. Approximate causal abstractions. In *Uncertainty in Artificial Intelligence*, pages 606–615. PMLR, 2020.

Blair Bilodeau, Linbo Wang, and Dan Roy. Adaptively exploiting d-separators with causal bandits. *Advances in Neural Information Processing Systems*, 35:20381–20392, 2022.

Rama Cont. Volatility clustering in financial markets: empirical facts and agent-based models. *Long memory in economics*, pages 289–309, 2007.

Benyamin De Leeuw, S. Sahand Mohammadi Ziabari, and Alexei Sharpanskykh. Surrogate modeling of agent-based airport terminal operations. In Fabian Lorig and Emma Norling, editors, *Multi-Agent-Based Simulation XXIII*, pages 82–94, Cham, 2023. Springer International Publishing. ISBN 978-3-031-22947-3.

Florent Delgrange, Ann Nowé, and Guillermo A. Pérez. Distillation of rl policies with formal guarantees via variational abstraction of markov decision processes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(6):6497–6505, Jun. 2022. doi: 10.1609/aaai.v36i6.20602. URL `https://ojs.aaai.org/index.php/AAAI/article/view/20602`.

Joel Dyer, Patrick Cannon, J. Doyne Farmer, and Sebastian M Schmon. Calibrating agent-based models to microdata with graph neural networks. In *ICML 2022 Workshop AI for Agent-Based Modelling*, 2022.

Joel Dyer, Patrick Cannon, J Doyne Farmer, and Sebastian M Schmon. Black-box Bayesian inference for agent-based models. *Journal of Economic Dynamics and Control*, 161:104827, 2024.

Annalisa Fabretti. On the problem of calibrating an agent based model for financial markets. *Journal of Economic Interaction and Coordination*, 8(2):277–293, Oct 2013. ISSN 1860-7128. doi: 10.1007/s11403-012-0096-3. URL `https://doi.org/10.1007/s11403-012-0096-3`.

Nabil T Fadai, Ruth E Baker, and Matthew J Simpson. Accurate and efficient discretizations for stochastic models providing near agent-based spatial resolution at low computational cost. *Journal of the Royal Society Interface*, 16(159):20190421, 2019.

Yorgos Felekis, Fabio Massimo Zennaro, Nicola Branchini, and Theodoros Damoulas. Causal optimal transport of abstractions. In Francesco Locatello and Vanessa Didelez, editors, *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pages 462–498. PMLR, 01–03 Apr 2024. URL https://proceedings.mlr.press/v236/felekis24a.html.

Linda Geaves, Jim Hall, and Edmund Penning-Rowsell OBE. Integrating irrational behavior into flood risk models to test the outcomes of policy interventions. *Risk Analysis*, 44(5):1067–1083, 2024.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34:9574–9586, 2021.

Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G. Bellemare. DeepMDP: Learning continuous latent space models for representation learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2170–2179. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/gelada19a.html.

Maziar Ghorbani, Diana Suleimenova, Alireza Jahani, Arindam Saha, Yani Xue, Kate Mintram, Anastasia Anagnostou, Auke Tas, William Low, Simon JE Taylor, et al. Flee 3: Flexible agent-based simulation for forced migration. *Journal of Computational Science*, 81:102371, 2024.

M. Gilli and P. Winker. A global optimization heuristic for estimating agent based models. *Computational Statistics and Data Analysis*, 42(3):299–312, 2003. ISSN 0167-9473. doi: https://doi.org/10.1016/S0167-9473(02)00214-1. URL https://www.sciencedirect.com/science/article/pii/S0167947302002141. Computational Ecomometrics.

Jakob Grazzini and Matteo Richiardi. Estimation of ergodic agent-based models by simulated minimum distance. *Journal of Economic Dynamics and Control*, 51:148–165, 2015. ISSN 0165-1889. doi: https://doi.org/10.1016/j.jedc.2014.10.006. URL https://www.sciencedirect.com/science/article/pii/S0165188914002814.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

AG Haldane and AE Turrell. An interdisciplinary model for macroeconomics. *Oxford Review of Economic Policy*, 34(1-2):219–251, 2018.

Alison Heppenstall, Andrew Crooks, Nick Malleson, Ed Manley, Jiaqi Ge, and Michael Batty. Future developments in geographical agent-based models: Challenges and opportunities. *Geographical Analysis*, 53(1):76–91, 2021.

Nick Jagiella, Dennis Rickert, Fabian J Theis, and Jan Hasenauer. Parallelization and high-performance computing enables automated statistical inference of multi-scale models. *Cell systems*, 4(2):194–206, 2017.

Armin Kekić, Bernhard Schölkopf, and Michel Besserve. Targeted reduction of causal models. *arXiv preprint arXiv:2311.18639*, 2023.

Cliff C. Kerr, Robyn M. Stuart, Dina Mistry, Romesh G. Abeysuriya, Katherine Rosenfeld, Gregory R. Hart, Rafael C. Núñez, Jamie A. Cohen, Prashanth Selvaraj, Brittany Hagedorn, Lauren George, Michał Jastrzębski, Amanda S. Izzo, Greer Fowler, Anna Palmer, Dominic Delport, Nick Scott, Sherrie L. Kelly, Caroline S. Bennette, Bradley G. Wagner, Stewart T. Chang, Assaf P. Oron, Edward A. Wenger, Jasmina Panovska-Griffiths, Michael Famulare, and Daniel J. Klein. Covasim: An agent-based model of covid-19 dynamics and interventions. *PLOS Computational Biology*, 17(7):1–32, 07 2021. doi: 10.1371/journal.pcbi.1009149. URL https://doi.org/10.1371/journal.pcbi.1009149.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Francesco Lamperti, Andrea Roventini, and Amir Sani. Agent-based model calibration using machine learning surrogates. *Journal of Economic Dynamics and Control*, 90:366–389, 2018. ISSN 0165-1889. doi: https://doi.org/10.1016/j.jedc.2018.03.011. URL https://www.sciencedirect.com/science/article/pii/S0165188918301088.

Finnian Lattimore, Tor Lattimore, and Mark D Reid. Causal bandits: Learning good interventions via causal inference. *Advances in neural information processing systems*, 29, 2016.

Sanghack Lee and Elias Bareinboim. Structural causal bandits: Where to intervene? *Advances in neural information processing systems*, 31, 2018.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: Foundations and learning algorithms*. MIT Press, 2017.

Donovan Platt. Bayesian estimation of economic simulation models using neural networks. *Computational Economics*, 59(2):599–650, 2022.

Christopher Rackauckas, Yingbo Ma, Julius Martensen, Collin Warner, Kirill Zubov, Rohit Supekar, Dominic Skinner, Ali Ramadhan, and Alan Edelman. Universal differential equations for scientific machine learning, 2021.

Eigil F Rischel and Sebastian Weichwald. Compositional abstraction error and a category of causal models. In *Uncertainty in Artificial Intelligence*, pages 1013–1023. PMLR, 2021.

Eigil Fjeldgren Rischel. The category theory of causal models. Master's thesis, University of Copenhagen, 2020.

Yulia Rubanova, Ricky TQ Chen, and David K Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. *Advances in neural information processing systems*, 32, 2019.

Paul K Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Causal consistency of structural equation models. In *33rd Conference on Uncertainty in Artificial Intelligence (UAI 2017)*, pages 808–817. Curran Associates, Inc., 2017.

John L Sabo. Stochasticity, predator–prey dynamics, and trigger harvest of nonnative predators. *Ecology*, 86(9):2329–2343, 2005.

Isabelle Salle and Murat Yıldızoğlu. Efficient sampling and meta-modeling for computational economic models. *Computational Economics*, 44(4):507–536, Dec 2014. ISSN 1572-9974. doi: 10.1007/s10614-013-9406-7. URL https://doi.org/10.1007/s10614-013-9406-7.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

Esteban G Tabak and Eric Vanden-Eijnden. Density estimation by dual ascent of the log-likelihood. *Communications in Mathematical Sciences*, 8(1):217–233, 2010.

Guus ten Broeke, George van Voorn, Arend Ligtenberg, and Jaap Molenaar. The use of surrogate models to analyse agent-based models. *Journal of Artificial Societies and Social Simulation*, 24 (2):3, 2021. ISSN 1460-7425. doi: 10.18564/jasss.4530. URL http://jasss.soc.surrey.ac.uk/24/2/3.html.

Samuel Wiese, Jagoda Kaszowska-Mojsa, Joel Dyer, Jose Moran, Marco Pangallo, Francois Lafond, John Muellbauer, Anisoara Calinescu, and J Doyne Farmer. Forecasting macroeconomic dynamics using a calibrated data-driven agent-based model. *arXiv preprint arXiv:2409.18760*, 2024.

Uri Wilensky and Kenneth Reisman. Thinking like a wolf, a sheep, or a firefly: Learning biology through constructing and testing computational theories—an embodied modeling approach. *Cognition and instruction*, 24(2):171–209, 2006.

The World Health Organization, The Organisation for Economic Co-operation & Development, and The International Bank for Reconstruction & Development/The World Bank. Strengthening pandemic preparedness and response through integrated modelling, 2024. URL https://www.who.int/publications/i/item/9789240090880. Licence: CC BY-NC-SA 3.0 IGO.

Louise Wright and Stuart Davidson. How to tell the difference between a model and a digital twin. *Advanced Modeling and Simulation in Engineering Sciences*, 7(1):1–13, 2020.

Kevin Muyuan Xia, Yushu Pan, and Elias Bareinboim. Neural causal models for counterfactual identification and estimation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=vouQcZS8KfW.

Fabio Massimo Zennaro, Máté Drávucz, Geanina Apachitei, W. Dhammika Widanage, and Theodoros Damoulas. Jointly learning consistent causal abstractions over multiple interventional distributions. In *2nd Conference on Causal Learning and Reasoning*, 2023a. URL https://openreview.net/forum?id=RNs7aMS6zDq.

Fabio Massimo Zennaro, Paolo Turrini, and Theo Damoulas. Quantifying consistency and information loss for causal abstraction learning. In *Proceedings of the Thrity-Second International Conference on International Joint Conferences on Artificial Intelligence*, 2023b.

Fabio Massimo Zennaro, Nicholas Bishop, Joel Dyer, Yorgos Felekis, Anisoara Calinescu, Michael Wooldridge, and Theodoros Damoulas. Causally abstracted multi-armed bandits. In Negar Kiyavash and Joris M. Mooij, editors, *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pages 4109–4139. PMLR, 15–19 Jul 2024. URL https://proceedings.mlr.press/v244/zennaro24a.html.

A Assumptions Underlying Markovian SCMs

Definition 1 implies the standard assumptions of (i) *acyclicity* of the DAG $\mathcal{G}_{\mathcal{M}}$ and (ii) *causal sufficiency*, meaning that there are no unobserved confounders [Pearl, 2009, Peters et al., 2017]. These two assumptions entail that our SCMs are Markovian.

We also assume *faithfulness*, guaranteeing that independencies in the data are captured in the graphical model Spirtes et al. [2000].

B Other Notions of Abstraction Error

As discussed in Section 3, Definition 4 is closely related to the notion of abstraction error introduced by Beckers et al. [2020]. In contrast to Definition 4, Beckers et al. [2020] employ a maximum over the intervention set $\mathcal{I}$ instead of an expectation. Hence, the abstraction error introduced by Beckers et al. [2020] may be viewed as a worst-case version of Definition 4.

In addition, Beckers and Halpern [2019] assume the intervention map ω can be implicitly defined by the map τ, and require the abstraction map τ to be consistent. That is, the image of $\mathcal{I}$ under the intervention map induced by τ must equal $\mathcal{I}'$. Since we do not couple the maps τ and ω we enforce no such condition. Additionally, Beckers et al. [2020] enforce surjectivity of τ. Since this makes no practical difference in a surrogate's use in downstream tasks, we dispense with this assumption.

Alternative notions of abstraction error have been introduced by Zennaro et al. [2023b], building upon the notion of exact transformations introduced by Rischel [2020]. Such alternative notions of abstraction have recently been used to define transfer learning protocols between SCMs related by an approximate causal abstraction in Zennaro et al. [2024]. We conjecture that an analogous version of our framework may be developed for this setting, wherein the aggregation function over the intervention set is again chosen to be an expectation over an interventional distribution η instead of a maximum, and we leave this as a direction for future work.

C Additional Related Work

Surrogate modelling of complex simulators is closely related to the problem of simulation-based inference. Inference involves tuning model parameters so that data generated by the simulator matches that generated by the real world system being modelled. Analogously, surrogate modelling consists of tuning surrogate parameters so that data generated by the surrogate matches data generated by the corresponding simulator. Hence, methods for calibration can naturally be applied to learn surrogates. Several calibration techniques and metrics have been proposed in the literature, including the method of simulated moments [Fabretti, 2013, Gilli and Winker, 2003] and minimum simulation distance [Grazzini and Richiardi, 2015]. We refer the reader to Dyer et al. [2022, 2024] for thorough surveys. Unlike our framework, surrogates trained for the purpose of parameter estimation do not typically account for interventional consistency explicitly.

More generally, our framework bears similarities to latent space modelling of Markov decision processes (MDPs) [Gelada et al., 2019], wherein one attempts to learn a smaller latent MDP from a target MDP, whose size precludes its use in downstream tasks. For downstream tasks such as formal verification of policies, Delgrange et al. [2022] employs the bisimulation metric to measure the consistency of their latent MDPs with respect to the target. Abstraction error plays an analogous role in our framework, where the original MDP corresponds to the simulator, and the latent MDP the surrogate. Likewise, the surrogates we propose in Section 5 are implicitly connected to the scientific modelling framework of Rackauckas et al. [2021], who embed prior information regarding system dynamics into systems of universal differential equations represented by neural architectures such as neural ODEs. We embed the underlying dynamics of the classical SIRS ODE into several surrogates in an attempt to learn better causal abstractions. Our work also bears some similarities to, yet differs substantially in several key ways from, Kekić et al. [2023], who also use an abstraction error to learn reduced causal models from larger SCMs. While their approach focuses on a single target variable at a fixed time horizon, assumes Gaussian noise and linear structural functions, and focuses on explainability of outcomes, we track multiple interdependent variables over the entire time horizon with a focus on accurate simulation from interventional distributions. Our approach

is therefore more tailored to large-scale and realistic nonlinear simulators. In contrast, the method presented in Kekić et al. [2023] becomes impractical for large-scale models.

C.1 Examples of τ Maps in Real Modelling Scenarios

We provide here some practical examples, beyond the two case studies we present, that illustrate how the τ map may be chosen for large-scale simulators, as a guideline for practitioners. We consider three exaples from the literature on policy modelling below:

1. Consider the model of forced migration in Ghorbani et al. [2024]. Variables of interest to these modellers are the total number of displaced people by location over time by age, gender, and other demographic characteristics. τ would therefore be defined by counting the number of agents in each of these states at each location, i.e. $\tau_{l,d}(x_t) = \sum_{a \in A} \mathbb{I}[\text{agent } a \text{ has demographic features } d \text{ and is in location } l \text{ at time } t]$ where l indexes locations, d are demographic features, x is the state of the simulation at time t, A is the set of all agents, and $\mathbb{I}$ is the indicator function.
2. Consider the model of flood risk mitigation behaviours proposed in Geaves et al. [2024], which models how households decide to take precautions to protect themselves from floods in high flood risk areas. The modellers are interested in the different precautions households take under different policy interventions, namely whether they: do nothing; purchase insurance; purchase property-level protection; and purchase property-level protection and insurance (see Fig. 3 of Geaves et al. [2024]). τ would count the number of agents taking such actions in this case (as in the example above).
3. Consider the UK housing market model proposed in Bardoscia et al. [2024], in which households consume goods, provide labour and invest in housing, whilst banks assess the credit worthiness of borrowers and set commercial interest rates. Tables 2-6 of Bardoscia et al. [2024] define macroeconomic market statistics such as inflation rate, unemployment rate and real interest rate that are of interest to the modellers. τ would therefore be defined by standard macroeconomic formulas for these quantities.

D Proof

D.1 Proof of Proposition 1

Proof. By non-negativity of the divergence d we have $d\left(\tau_{\#}(\mathbb{P}_{\mathcal{M}_\iota}), \mathbb{P}_{\mathcal{M}'_{\omega(\iota)}}\right) \geq 0$ for all $\iota \in \mathcal{I}$. Hence $d_{\tau,\omega}(\mathcal{M}, \mathcal{M}')$ corresponds to an expectation over a non-negative random variable. Since this expectation is equal to zero, we conclude that $d\left(\tau_{\#}(\mathbb{P}_{\mathcal{M}_\iota}), \mathbb{P}_{\mathcal{M}'_{\omega(\iota)}}\right) = 0$ almost surely with respect to the distribution η. Positivity of the divergence d then implies that $\tau_{\#}(\mathbb{P}_{\mathcal{M}_\iota}) = \mathbb{P}_{\mathcal{M}'_{\omega(\iota)}}$ almost surely with respect to the distribution η. □

D.2 Proof of Proposition 2

Proof. Using Markov's inequality and the fact that $d_{\tau,\omega^\phi}(\mathcal{M}, \mathcal{M}^\psi) = \mathbb{E}_{\iota\sim\eta}\left[d\left(\tau_{\#}(\mathbb{P}_{\mathcal{M}_\iota}), \mathbb{P}_{\mathcal{M}^\psi_{\omega^\phi(\iota)}}\right)\right]$:

$$\mathbb{P}_\eta\left(d\left(\tau_{\#}(\mathbb{P}_{\mathcal{M}_\iota}) \,\|\, \mathbb{P}_{\mathcal{M}^\psi_{\omega^\phi(\iota)}}\right) \geq \epsilon\right) \leq \frac{d_{\tau,\omega^\phi}\left(\mathcal{M}, \mathcal{M}^\psi\right)}{\epsilon}.$$

Since we have a finite domain, the likelihood functions associated with (a) the pushforward measure of the `ABM` under τ and (b) the surrogate macromodel can be written as probability mass functions, whose logarithms are non-positive. Since we have assumed $\mathbb{P}_{\mathcal{M}^\psi_{\omega^\phi(\iota)}} \ll \tau_{\#}\left(\mathbb{P}_{\mathcal{M}_\iota}\right)$, we have that $0 \leq -\log q^\psi_{\omega^\phi(\iota)}(\mathbf{Y}) < \infty$ for any $\mathbf{Y} \sim \tau_{\#}\left(\mathbb{P}_{\mathcal{M}_\iota}\right)$, and therefore

$$0 \leq \mathbb{E}_{\iota\sim\eta}\left[\mathtt{CE}_\iota\right] < \infty. \tag{15}$$

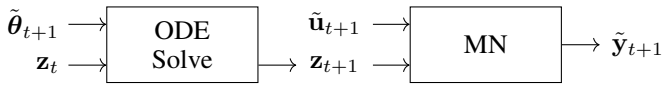

Figure 6: A schematic representation of the LODE surrogate family for a single time step. First, the output of the SIRS ODE for the next time step, $\mathbf{z}_{t+1}$, is computed via ODESolve. Then, $\mathbf{z}_{t+1}$ serves as the logits for a multinomial distribution from which $\tilde{\mathbf{y}}_t$ is sampled. This sampling procedure is denoted by MN in the diagram. The exogenous variables required to reparameterise the multinomial distribution during sampling are denoted by $\tilde{\mathbf{u}}_t$.

We also have that

$$-\mathbb{H}_{\tau_{\#}(\mathbb{P}_{\mathcal{M}_\iota})} \leq 0 \;\Rightarrow\; \mathbb{E}_{\iota\sim\eta}\left[-\mathbb{H}_{\tau_{\#}(\mathbb{P}_{\mathcal{M}_\iota})}\right] \leq 0, \tag{16}$$

where $\mathbb{H}_{\tau_{\#}(\mathbb{P}_{\mathcal{M}_\iota})}$ is the entropy of the probability mass function associated with $\tau_{\#}(\mathbb{P}_{\mathcal{M}_\iota})$, and that

$$d\left(\tau_{\#}(\mathbb{P}_{\mathcal{M}_\iota}), \mathbb{P}_{\mathcal{M}^{\psi}_{\omega^{\phi}(\iota)}}\right) = -\mathbb{H}_{\tau_{\#}(\mathbb{P}_{\mathcal{M}_\iota})} + \texttt{CE}_\iota \geq 0 \tag{17}$$

$$\Rightarrow d_{\tau,\omega^{\phi}}\left(\mathcal{M}, \mathcal{M}^{\psi}\right) = \mathbb{E}_{\iota\sim\eta}\left[-\mathbb{H}_{\tau_{\#}(\mathbb{P}_{\mathcal{M}_\iota})}\right] + \mathbb{E}_{\iota\sim\eta}\left[\texttt{CE}_\iota\right] \leq \mathbb{E}_{\iota\sim\eta}\left[\texttt{CE}_\iota\right]. \tag{18}$$

We write the upper bound above in terms of the cross-entropy, since this can be estimated with finite samples, whereas the full KL-divergence cannot be estimated in general due to the complexity of evaluating the density associated with $\tau_{\#}(\mathbb{P}_{\mathcal{M}_\iota})$ for an arbitrary ABMs. Hence

$$\mathbb{P}_\eta\left(d\left(\tau_{\#}(\mathbb{P}_{\mathcal{M}_\iota}) \,\|\, \mathbb{P}_{\mathcal{M}^{\psi}_{\omega^{\phi}(\iota)}}\right) \geq \epsilon\right) \leq \frac{\mathbb{E}_{\iota\sim\eta}\left[\texttt{CE}_\iota\right]}{\epsilon}. \tag{19}$$

□

E Further Experimental Details

As described in the main text, the three surrogate families we consider have SCMs whose corresponding DAGs can be drawn as in Figure 4. In this section, we fully specify the corresponding SCM for each surrogate. Furthermore, for each surrogate, we provide details on the procedure used to train the parameters ψ and ϕ, which respectively describe the structural equations of each SCM and their corresponding intervention map ω.

E.1 The LODE Surrogate Family

To construct a set $\mathfrak{M}$ of probabilistic SCMs, we define a latent neural ordinary differential equation (LNODE) based on the classical SIRS ODE system. The SIRS ODE system takes the form

$$\begin{aligned} \frac{\mathrm{d}\tilde{S}_t}{\mathrm{d}t} &= \tilde{\gamma}_t\tilde{R}_t - \tilde{\alpha}_t\tilde{I}_t\tilde{S}_t, \qquad \frac{\mathrm{d}\tilde{I}_t}{\mathrm{d}t} = \tilde{\alpha}_t\tilde{I}_t\tilde{S}_t - \tilde{\beta}_t\tilde{I}_t, \\ \frac{\mathrm{d}\tilde{R}_t}{\mathrm{d}t} &= \tilde{\beta}_t\tilde{I}_t - \tilde{\gamma}_t\tilde{R}_t, \end{aligned} \tag{20}$$

where $\tilde{\boldsymbol{\theta}}_t = (\tilde{\alpha}_t, \tilde{\beta}_t, \tilde{\gamma}_t) \in \mathbb{R}^3_{\geq 0}$ are the ODE parameters and $\mathbf{z}_t = (\tilde{S}_t, \tilde{I}_t, \tilde{R}_t) \in \mathcal{S}\ \forall t \in [0, T]$ is the ODE state, where $\mathcal{S}$ is the two-simplex. Note that $\mathbf{z}_t$ represents the proportion of susceptible, infected and recovered individuals in the population according the SIRS ODE. Whilst the parameters $\tilde{\boldsymbol{\theta}}_t$ may change over time – which will permit the experimenter to intervene on the values of the parameters at different time steps – we assume the simplest case of assigning the same vector $\tilde{\mathbf{v}} \in \mathbb{R}^3_{\geq 0}$ to all $\tilde{\boldsymbol{\theta}}_t$ when no interventions are applied:

$$\tilde{\boldsymbol{\theta}}_t = \tilde{\mathbf{v}}, \quad \forall t \in [0, T]. \tag{21}$$

In other words, Equation (21) describes the structural equation $\tilde{f}_{\tilde{\boldsymbol{\theta}}_t}$ for $\tilde{\boldsymbol{\theta}}_t$. Practically speaking, the choice of $\tilde{\mathbf{v}}$ is inconsequential, as we can model any change to $\tilde{\boldsymbol{\theta}}_t$ as an intervention. Given $\tilde{\boldsymbol{\theta}}_t$, the ODE state $\mathbf{z}_t$ evolves according to the following rule:

$$\mathbf{z}_t = \text{ODESolve}(\mathbf{z}_{t-1}, \tilde{\boldsymbol{\theta}}_t), \quad t \in [\![1, T]\!], \tag{22}$$

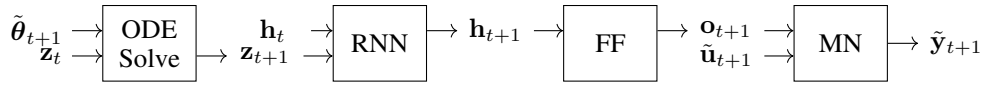

Figure 7: A schematic representation of the LODE-RNN surrogate family for a single time step. First, the output of the SIRS ODE for the next time step, $\mathbf{z}_{t+1}$, is computed via ODESolve. Then, $\mathbf{z}_{t+1}$ is passed through to the hidden state of a recurrent neural network (denoted by RNN in the diagram) that updates its hidden state from $\mathbf{h}_t$ to $\mathbf{h}_{t+1}$. The updated hidden state is passed to a feedforward neural network (denoted by FF in the diagram), which computes the logits $\mathbf{o}_{t+1}$ for a multinomial distribution from which $\tilde{\mathbf{y}}_{t+1}$ is sampled.

where ODESolve denotes numerical integration of System 20 between times $t-1$ and t. In our experiments, we compute this using a Euler scheme with step size $\Delta t = 1$. The initial state of the ODE is taken to be $\mathbf{z}_0 = (1 - \tilde{I}_0, \tilde{I}_0, 0)$. One may change the initial state $\mathbf{z}_0$ through interventions on $\tilde{I}_0$, which is modelled as an endogenous variable.

Given $\mathbf{z}_t$, we draw the endogenous variables $\tilde{\mathbf{y}}_t$ from a multinomial distribution whose class probabilities are given by $\mathbf{z}_t$. Whilst $\mathbf{z}_t$ represents the percentage of susceptible, infected, and recovered individuals predicted by the SIR ODE, $\tilde{\mathbf{y}}_t$ represents the actual counts observed by the experimenter. We write $\tilde{f}_{\tilde{\mathbf{y}}_t}(\tilde{I}_0, \tilde{\boldsymbol{\theta}}_{1:t}, \tilde{\mathbf{u}}_t)$ to denote the structural function associated with $\tilde{\mathbf{y}}_t$, where the dependence on $\tilde{I}_0$ and $\tilde{\boldsymbol{\theta}}_{t'}$ for $t' \leq t$ is mediated by the trajectory followed by the $\mathbf{z}_{t'}$ for $t' \leq t$, and $\tilde{\mathbf{u}}_t$ are the exogenous random variables required to reparameterise the multinomial sampling procedure on each time step.

Note that $\psi = \emptyset$ for this family of surrogates, and hence $\mathfrak{M}$ is a singleton. For the function f^ϕ comprising the intervention map ω^ϕ, we take a feedforward network with layer sizes 3, 32, 64, 64, 64, 32, 3. A ReLU activation is applied after each hidden layer, and a sigmoid activation is applied to the final output layer. The sigmoid activation function ensures that the predicted intervention vector $f^\phi(\mathbf{v})$ on the parameters of the LODE has all of its components in the range $[0, 1]$, which is suitable when forward simulating the ODE with an Euler scheme with $\Delta t = 1$. This feedforward network consists of 12,739 trainable parameters.

E.2 The LODE-RNN Surrogate Family

This surrogate family closely mimics the LODE family described above, and differs only in that the class *logits* of the multinomial distributions are instead indexed by the output of a feedforward network – with layer sizes 32, 32, 64, 32, 16, 3, where all hidden layers are followed by a ReLU activation function – which maps from the hidden state $\mathbf{h}_t \in \mathbb{R}^{32}$ of a GRU recurrent network that is passed over the trajectory $\mathbf{z}_{0:T}$ generated from the SIRS ODE (forward simulated as described above). The combined action of the ODE solver, the GRU-feedforward networks, and the reparameterisation of sampling from the multinomial distributions, define the structural equations $\tilde{f}_{\tilde{\mathbf{y}}_t} : (\tilde{I}_0, \tilde{\boldsymbol{\theta}}_{1:t}, \tilde{\mathbf{u}}_t) \mapsto \tilde{\mathbf{y}}_t$ for each $t \in [\![1, T]\!]$.

For this model, ψ is the collection of trainable parameters comprising these GRU and feedforward networks. For f^ϕ, we use a feedforward network with layer sizes 3, 32, 64, 32, 3, where a ReLU activation is applied after all hidden layers and a sigmoid activation is applied after the final layer. Thus, the total number of trainable parameters from ψ and ϕ combined is 13,798.

E.3 The LRNN Surrogate Family

This surrogate family makes no use of the SIRS ODE model. Instead, the logits of the multinomial distributions for $t \in [\![1, T]\!]$ are indexed by the outputs $(\mathbf{o}_1, \ldots, \mathbf{o}_T), \mathbf{o}_t \in \mathbb{R}^3$ of a feedforward network – with layer sizes 32, 32, 64, 32, 16, 3, and where all hidden layers are followed by a ReLU activation function – that maps from the hidden state $\mathbf{h}_t \in \mathbb{R}^{32}$ of a GRU recurrent network which is passed over the sequence $\tilde{\boldsymbol{\theta}}_{1:T}$. The initial hidden state is chosen to be $\mathbf{h}_0 = (1 - \tilde{I}_0, \tilde{I}_0, \mathbf{0})$, where $\mathbf{0}$ is a vector of 30 zeros. We also take $\mathbf{o}_0 = (\log(1 - \tilde{I}_0), \log(\tilde{I}_0), -\infty)$ which indexes the logits of the multinomial distribution at time $t = 0$. Once again, we may write the structural equations $\tilde{f}_{\tilde{\mathbf{y}}_t}$ for

Figure 8: A schematic representation of the LRNN surrogate family for a single time step. First, the parameters $\boldsymbol{\theta}_{t+1}$ are passed to a recurrent neural network (denoted by RNN in the diagram) that updates its hidden state. The updated hidden state is passed to a feedforward neural network (denoted by FF in the diagram), which computes the logits $\mathbf{o}_{t+1}$ for a multinomial distribution from which $\tilde{\mathbf{y}}_{t+1}$ is sampled.

the $\tilde{\mathbf{y}}_t$ in terms of $\tilde{I}_0$, $\tilde{\boldsymbol{\theta}}_{1:t}$, and the exogenous random variables $\tilde{\mathbf{u}}_t$ required to reparameterise the sampling procedure from the multinomial distribution.

Since we use exactly the same networks in this surrogate family as in the LODE-RNN family, the total number of trainable parameters from ψ and ϕ combined is also 13,798.

E.4 The likelihood function for each of these surrogate families

Having intervened on the $\tilde{I}_0$ and $\tilde{\boldsymbol{\theta}}_t$ with known values, the class probabilities for each multinomial distribution is completely determined given the deterministic dynamics within the structural equations mapping to the $\tilde{\mathbf{y}}_t$.

E.5 Formalising the τ map

Taking $\mathrm{dom}[I_0] = \mathcal{J}_{\mathcal{M}} = [0,1]$, $\mathrm{dom}[\mathbf{X}_{0:T}] = \mathcal{X}^{T+1}$ with $\mathcal{X} = \{0,1,2\}^N$, and $\mathrm{dom}[\boldsymbol{\Theta}_{1:T}] = \mathcal{P}_{\mathcal{M}}^T$ with $\mathcal{P} = [0,1]^3$, we define

$$\tau : \mathcal{J}_{\mathcal{M}} \times \mathcal{X}^{T+1} \times \mathcal{P}_{\mathcal{M}}^T \to \mathcal{J}_{\mathcal{M}'} \times \mathcal{Y}^{T+1} \times \mathcal{P}_{\mathcal{M}'}^T$$

which operates componentwise as

$$\tau\left(I_0, \mathbf{x}_{0:T}, \boldsymbol{\theta}_{0:T}\right) = \left(\tau_i(I_0), \tau_x(\mathbf{x}_{0:T}), \tau_\theta(\boldsymbol{\theta}_{0:T})\right) \tag{23}$$

where

$$\tau_i = \mathrm{id}, \tag{24}$$

$$\tau_x : \mathbf{x}_{0:T} \mapsto \left(\sum_{n=1}^N \mathbb{I}_{\mathbf{x}_{nt}=0}, \sum_{n=1}^N \mathbb{I}_{\mathbf{x}_{nt}=1}, \sum_{n=1}^N \mathbb{I}_{\mathbf{x}_{nt}=2}\right)_{0:T},$$

$$\tau_\theta = \mathrm{id}. \tag{25}$$

In the above, id is the identity map, and τ_x acts by counting the total number of susceptible, infected, and recovered individuals in the ABM at each time step.

E.6 Further experimental details on the training procedure

All models were trained on CPU on a 2022 MacBook Pro, operating on macOS Ventura 13.2.1. Training one surrogate model on this machine took on average approximately 20 minutes, amounting to approximately 600 minutes in total to produce the results reported in Table 1. Initial attempts at experiments while the code was still in development contribute approximately 200 additional minutes. Software dependencies are specified in the GitHub repository containing the code for this paper, which will be made public upon acceptance.

We assume periodic boundary conditions in both spatial dimensions for the ABM presented in Example 1, which is used in all of our experiments.

As suggested by Figures 1, 4, and 6-8, the parameters $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}$ are fed into the models at each time step.

For the LODE and LODE-RNN surrogate families, we forward simulate the SIRS ODE with an Euler scheme with step size $\Delta t = 1$.

For all surrogates, the neural networks comprising the ω^ϕ map and structural equations parameterised by ψ were trained with a learning rate of 10^{-2} for a maximum number of 1000 epochs, batch size $B = 50$, and with the Adam optimiser [Kingma and Ba, 2014]. A total number of $R = 1000$ training samples was generated from the `ABM` for each of the observational and interventional training sets; these were each split 5 times into different training and validation sets of sizes 800 and 200, respectively, with a new surrogate model trained from scratch on each of these splits. We apply an early stopping criterion in which training is ceased if the validation error does not decrease for 20 consecutive epochs.

E.7 Additional Case Study

In this case study, we consider a different policy scenario: reintroducing a species into an ecology, and simulating the ensuing population dynamics. Specifically, we adapt slightly a model from Wilensky and Reisman [2006]: we model an environment initially consisting of grass, sheep, and wolves, in which grass grows and is eaten by sheep, sheep eat grass and reproduce and are eaten by wolves, and wolves eat sheep and reproduce. The intervention we consider entails reintroducing a third animal species – bears, which eat both sheep and wolves, and also reproduce – whose population is originally zero but is made non-zero at some intervention time t. We imagine that t is the variable the policymaker wants to optimise here.

We simulate the interactions between these four species in a spatial model, in which members of each animal species move around the grid and interact with the other species. We are then interested in understanding how the reintroduction of the bears affects the overall population dynamics, i.e., the counts of each animal in each species, along with the quantity of grass over time. As in the epidemic case study, we consider the problem of learning interventionally consistent surrogates for this complex spatiotemporal simulator, and once again examine three possible approaches for constructing surrogate families:

1. a family of deterministic mechanistic models based on a discrete-time Lotka-Volterra model of population dynamics [see, e.g., Sabo, 2005], where (analogously to the `LODE` surrogate family discussed in the epidemic case study) the underlying deterministic dynamics of the population dynamics model index a probability distribution at each time step (in this case, a Binomial distribution for each of the 4 species);
2. an `LRNN` family, exactly mirroring the `LRNN` family considered in the epidemic case study presented already;
3. and a third family considers a hybrid approach, where (as in the `LODE-RNN` family considered in the epidemic case study) we pass a recurrent network over the underlying Lotka-Volterra-type population dynamics model first before taking the output of the recurrent network to index the Binomial distributions for each of the four species.

A table for the results of this additional case study is shown in Table 2, where we see that the results are qualitatively very similar to the epidemic case study already presented: we see that training surrogates using our framework yields significant improvements in the surrogates' interventional consistency over observationally trained baselines, and that interventionally trained surrogates only see a minor decrease in performance on observational data compared to the drop in performance the observational surrogates see on interventional data.

Table 2: Metrics for interventionally (**I**) & observationally (**O**) trained surrogates on interventional ($\mathbf{I}'$) & observational ($\mathbf{O}'$) test sets ($\text{median}^{\text{third quartile}}_{\text{first quartile}}$ from 5 repeats) for the predator-prey case study. AMSE & ANLL measure ability to model counts of each species over time. **Bold** denotes best performance.

Test	Model / Train	LRNN I	LRNN O	LODE-RNN I	LODE-RNN O	LODE I	LODE O
$\mathbf{I}'$	AMSE ($\times 10^2$)	$\mathbf{1.65^{1.80}_{1.62}}$	$3.30^{4.32}_{2.37}$	$\mathbf{1.79^{1.83}_{1.79}}$	$2.23^{2.29}_{2.13}$	$\mathbf{41.84^{44.74}_{40.13}}$	$201.78^{250.80}_{46.11}$
	ANLL ($\times 10^3$)	$\mathbf{0.78^{0.81}_{0.77}}$	$11.20^{11.64}_{10.09}$	$\mathbf{0.93^{0.94}_{0.92}}$	$5.63^{5.96}_{4.72}$	$\mathbf{6.54^{10.14}_{6.44}}$	$47.14^{50.33}_{37.33}$
$\mathbf{O}'$	AMSE ($\times 10^2$)	$1.85^{2.03}_{1.81}$	$\mathbf{1.61^{1.64}_{1.59}}$	$2.18^{2.19}_{2.08}$	$\mathbf{1.56^{1.67}_{1.51}}$	$63.89^{326.33}_{38.25}$	$\mathbf{36.75^{38.25}_{36.08}}$
	ANLL ($\times 10^3$)	$0.76^{0.80}_{0.71}$	$\mathbf{0.69^{0.70}_{0.68}}$	$1.25^{1.25}_{1.24}$	$\mathbf{0.66^{0.70}_{0.65}}$	$32.11^{33.90}_{10.56}$	$\mathbf{5.49^{6.18}_{4.96}}$

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our abstract and introduction motivate our work and its purpose, and indicates the theoretical results we provide in Section 4.1 and empirical results we provide in Section 5.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: These are discussed in the Conclusion (Section 7).

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The assumptions are provided in the Propositions in Section 4.1 (and more generally in Section 2 and Appendix A), and the proofs are located in Appendices D.1 and D.2.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.

- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide details on the method and training algorithm in Section 4. We describe the experimental setup in Section 5, and provide further details on neural network architectures, training hyperparameters etc. in Appendix E. Finally, we also make code for our experiments publicly available, in order for the reader to be able to reproduce the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Open source code is provided on GitHub, along with instructions for how to run the script to reproduce the results.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: These are provided in Appendix E.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Experiments were all repeated 5 times with cross-validation, resulting in first quantile, median, and third quantile error metrics reported in Table 1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: This is provided in Appendix E.6.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have discussed the potential positive societal consequences of this line of work in, e.g., Section 1. While we do not see any reason that our work should be seen as posing any particular risk of threatening harm to society, we recognise that causal simulation modelling in general can be used by malicious actors to improve their own decision making, which could support them to achieve their own malicious goals.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have discussed the potential positive impact of our work in, e.g., Section 1, namely that our methods may assist policymakers to more rapidly identify policy interventions that save lives and mitigate economic costs during pandemics and other crises. As discussed in point 9. of this checklist, we do however recognise that simulation modelling and machine learning in general can be misused by malevolent actors to support their own decision-making.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not see that our method for learning surrogate models poses a particularly high risk for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: We do not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Code for reproducing the results is provided on GitHub. The README contains instructions for how to reproduce the results, while the LICENCE file specifies the licence attached to the code.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.