Reconciling Causality and Non-Equilibrium Thermodynamics with Hamiltonian Causal Models

Dario Rancati[1] **Max Welling**[2,3] **Francesco Locatello**[1]

[1]Institute of Science and Technology Austria [2]CuspAI [3]University of Amsterdam

Correspondence to: dario.rancati@ist.ac.at

Abstract

Causal modeling of physical temporal phenomena must handle interventions that act along trajectories, nonstationary induced laws, path-dependent effects, and feedback mediated by dynamics, all challenging in standard causal models. We introduce *Hamiltonian Causal Models* (HCMs), a trajectory-level framework in which observed variables interact with local environments and interventions act as controls of Hamiltonian mechanisms. HCMs separate immutable equations of motion from intervenable mechanisms and define causal effects as discrepancies between interventional path laws. A key motivation for HCMs is their natural interface with non-equilibrium thermodynamics. Entropy production quantifies the irreversibility of a process and is a central causal observable: it is estimable from data and witnesses causal effects along the system's evolution that are invisible to endpoint and cumulative versions of the standard average treatment effect. As in physics, cause and effect are not primitives of the relation between two random variables but arise from the non-invertibility of the thermodynamic arrow. With this, our paper reconciles the language of statistical causal models and non-stationary thermodynamics, offering new tools to describe causality in a wide range of physical systems.

1 Introduction

Statistical approaches to causality aim to quantify how systems respond to interventions rather than merely describing observed associations. While correlations between variables are ubiquitous, causal relationships encode how a joint distribution would change under external manipulations [Spirtes et al., 2000, Pearl, 2009, Imbens and Rubin, 2015]. This interventional perspective has driven substantial developments across fields including econometrics [Imbens and Rubin, 2015], epidemiology [Miguel et al., 2023], machine learning [Schölkopf et al., 2021, Peters et al., 2017], and, most recently, AI for science [Schölkopf et al., 2016, Feuerriegel et al., 2024, Cadei et al., 2025, Mencattini et al., 2026]. To accommodate the diverse needs of these domains, multiple complementary frameworks have emerged, including potential outcomes [Imbens and Rubin, 2015], causal graphical models [Pearl, 2009], and structural causal models [Spirtes et al., 2000], but also temporal approaches such as Granger causality [Granger, 1969] and differential equation–based formulations [Peters et al., 2022, Lorch et al., 2024, Boeken and Mooij, 2024]. Despite this progress, no single existing framework natively captures the combination of properties that arise naturally in physical dynamical systems without ad-hoc constructions: interventions may be time-dependent controls acting along trajectories, induced laws may be nonstationary, causal effects may be complex and path-dependent, and feedback or cyclic dependence may unfold through time.

These modeling challenges point naturally to the gap between statistical causal models and how causation emerges in physical systems. As emphasized by Rovelli [2023], in physical systems, causal direction and temporal evolution are both tied to the notion of irreversibility. Indeed,

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

many laws of physics are time-reversal symmetric: a solution to Newton's equations, for instance, remains a valid solution if all the velocities are reversed. Causal experience is not symmetric in this way: if a stone is thrown into a pond, we observe outgoing ripples; the time-reversed movie, with ripples converging inward to the stone's entry point, is compatible with the microscopic equations of motion but overwhelmingly unlikely at the macroscopic level. This suggests that the causal influence of the rock on the pond is not simply encoded in the equations of motion but is tied to the thermodynamic arrow of time.

Non-equilibrium thermodynamics provides quantitative tools for measuring these notions, most notably entropy production, which measures the irreversibility of a trajectory by comparing it with its time reversal.

This connection becomes even sharper if we want to accommodate feedback and time-dependent interventions. In standard causal models, an intervention is idealized as an arbitrary external operation that fixes, replaces, or perturbs a mechanism. At the same time, a causal model is only considered true when it aligns with randomized studies [Peters et al., 2014]. In a physical system, such an operation must be implemented by a controller or protocol acting along the trajectory, and may exchange information with the system doing so. Maxwell's demon [Sagawa and Ueda, 2012b] is the canonical example of a feedback intervention in thermodynamics. The demon observes microscopic fluctuations and acts conditionally on this information, apparently reducing the entropy of the controlled system and thus contradicting the Second Law of Thermodynamics. The resolution is not that the intervention is impossible, but that its thermodynamic analysis is incomplete unless the controller's information (and hence entropic) cost is included. The lesson for causal modeling is that time-dependent or adaptive interventions are not mere assignments: they process information from the trajectory and feed it back into the dynamics, thereby carrying both causal and thermodynamic content.

In this paper, we introduce *Hamiltonian Causal Models* (HCMs), a framework for causal modeling of controlled Hamiltonian and nonequilibrium processes. In HCMs, trajectories are the primitive causal objects, and interventions are represented as time-dependent control variables acting through the dynamics whose thermodynamic impact can be measured with non-equilibrium thermodynamic techniques. The model class we can describe is broad but structured. It includes familiar stochastic dynamics such as controlled Langevin diffusions $dX_t = -\nabla U(X_t, \lambda_t)\mathrm{d}t + \sigma(\lambda_t)\mathrm{d}W_t$ while also allowing more general controlled Hamiltonian and nonequilibrium evolutions. Overall, HCMs follow familiar causal assumptions such as independent mechanisms and markovianity. At the same time, they make it easy to model physical systems, without an ad-hoc characterization of time and dynamics and give us a language that naturally supports enticing properties:

Key Differentiating Properties of HCMs

HCMs provide a language in which one can natively study causal effects over finite time horizons, for stationary or nonstationary trajectories, in discrete or continuous time. Within this language, interventions may be time-dependent, ordered, adaptive, and mutually interacting; feedback loops and cyclic dependence enter through temporal evolution; assumptions such as equilibrium initial conditions or a fully specified static graphical structure are not imposed at the outset; and the distinction between causal interventions and physical laws can be made explicit.

We summarize our contributions as follows:

- **Hamiltonian Causal Models.** We introduce Hamiltonian Causal Models (HCMs), a trajectory-level causal framework for controlled Hamiltonian systems coupled to local environments. In HCMs, interventions are time-dependent policies acting on local Hamiltonian mechanisms, while causal effects are defined as discrepancies between interventional path laws.
- **Entropy production as a causal observable.** We develop the thermodynamic theory of HCMs. We show that work rates identify local causal interactions when the Hamiltonian is known, and that entropy production provides a path-space, data-estimable witness of causal effect. We further show how local entropy production rates can be used to obtain a local causal influence criterion for nonstationary temporal dynamics.
- **Experimental validation.** We validate the theory on controlled dynamical systems. Our experiments show that entropy production detects path-level causal effects that can be invisible to

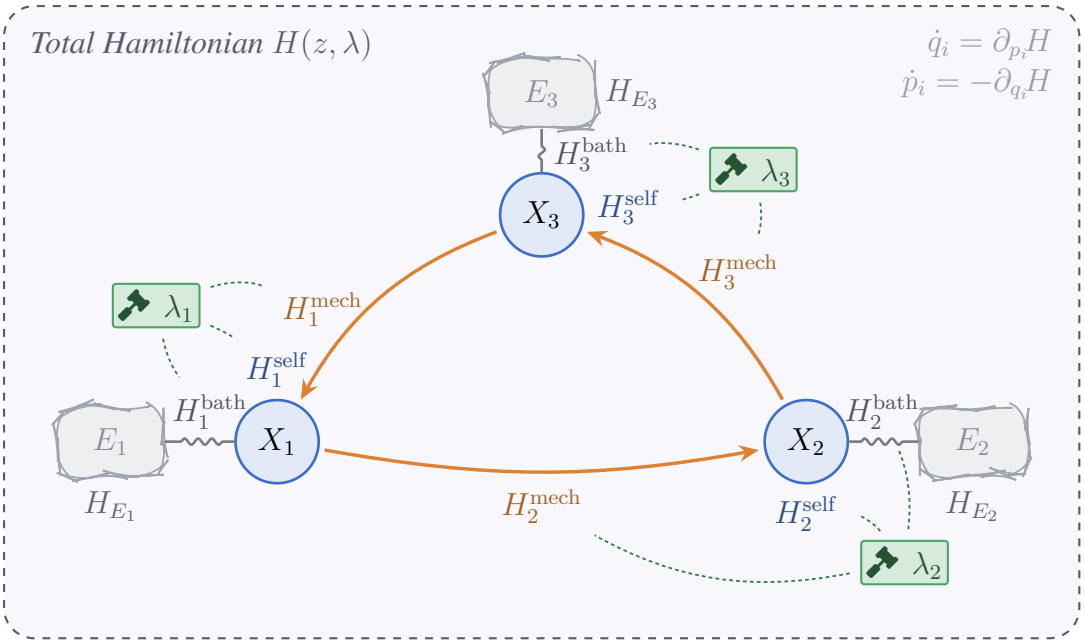

Figure 1: Schematic Representation of an HCM: appropriate Hamiltonians are associated with every component and interaction among them, representing the energy stored in that block. Interventions are modelled to modify the Hamiltonian of each observable X_i and of the incoming interactions.

endpoint or cumulative treatment-effect summaries, and that nodewise entropy-production rates can reveal the mechanism dependence structure induced by HCM interventions in cyclic dynamical systems.

2 Hamiltonian Causal Models

This section introduces Hamiltonian Causal Models (HCMs) and their notion of causal effect. We give the formal definitions and show that HCMs describe a wide range of temporal phenomena, recovering several existing trajectory-level causal models as special cases. Detailed conditions and regularity hypotheses are deferred to Appendix A.

Intuition

A Hamiltonian Causal Model describes an **observed system** of n variables, each paired with an **unobserved thermal bath** with which it exchanges energy and information. A deterministic scalar function H, the *Hamiltonian*, drives the joint evolution of system and environments through a prescribed set of microscopic equations of motion. Stochasticity in the observed trajectories arises from two sources, both extrinsic to the dynamics: randomness in the initial conditions, and the fact that the environments are unobserved. The environments thus play the role of **exogenous noise** variables of classical Pearl causality. The Hamiltonian decomposes into terms associated with each component of the system: an intrinsic potential per variable, couplings to other observed nodes, couplings to the baths, and the baths' own internal dynamics. In this sense, the Hamiltonian's decomposition plays the role of the **causal Markov factorization** from classical structural causal models. **Interventions** modify H, and through it the mechanisms encoded in the decomposition, without altering the equations of motion themselves.

Definition 2.1 (Hamiltonian Causal Model)**.** Fix an horizon $T \in \mathbb{R}_+ \cup \{+\infty\}$. An *Hamiltonian Causal Model* is a tuple $\mathcal{M} = (\mathcal{P}, z_0, \Gamma, \omega, \mathcal{L}, \mathcal{A}, H)$ consisting of:

(1) A probability space $\mathcal{P} = (\Omega, \mathcal{F}, \mathbb{P})$ and a random initial state $z_0 : \Omega \to \Gamma$.

(2) A symplectic manifold (Γ, ω) with a node-wise factorization $\Gamma = \Gamma_X \times \Gamma_E$, $\Gamma_X = \prod_{i=1}^n \Gamma_{X_i}$, $\Gamma_E = \prod_{i=1}^n \Gamma_{E_i}$, where Γ_{X_i} is the state space of the observed variable X_i and Γ_{E_i} the state space of its bath.

(3) An intervention manifold $\mathcal{L} = \prod_{i=1}^n \mathcal{L}_i$ and a class $\mathcal{A}$ of admissible policies $\lambda : [0, T] \times \Omega \to \mathcal{L}$.

(4) A smooth Hamiltonian $H : \Gamma \times \mathcal{L} \to \mathbb{R}$ admitting the node-wise decomposition:

$$H(z, \lambda) = \underbrace{H_E(e)}_{\text{free environments}} + \sum_{i=1}^{n} \Big(\underbrace{H_i^{\text{self}}(x_i, \lambda_i)}_{\text{intrinsic}} + \underbrace{H_i^{\text{mech}}(x_i, x_{-i}, \lambda_i)}_{\text{coupling to other nodes}} + \underbrace{H_i^{\text{bath}}(x_i, e_i, \lambda_i)}_{\text{coupling to environment}} \Big) \quad (1)$$

(5) Dynamics generated by H: setting $H_\lambda(t, z) := H(z, \lambda_t)$, the system evolves under the controlled flow $\Phi_t^\lambda : \Gamma \to \Gamma$ defined by the unique vector field Y_{H_λ} satisfying $\omega(Y_{H_\lambda}(t, \cdot), \cdot) = \mathrm{d}H_\lambda(t, \cdot)$. The *forward observed path law* is $\mathbb{P}_F^\lambda := (\Pi_X \circ \Psi^\lambda)^{\#}\mathbb{P}$, where $\Psi^\lambda(\omega)(t) := \Phi_t^\lambda(z_0(\omega))$ and Π_X projects onto observed coordinates.

We assume regularity sufficient for the controlled flow to be globally defined on $[0, T]$, factorization of the initial environment law across nodes, and existence of a time-reversal compatible with H; precise hypotheses are stated in Appendix A.3.

Causal semantics. As anticipated, the baths $E_1, \ldots, E_n$ play the role of exogenous noise. On interventions, we adopt the point of view of Phenomenological Causality [Janzing and Mejia, 2022], consistent with other works in causality indexing interventions as part of the causal model [Mooij et al., 2020]: causal quantities are defined relative to a set of actions available to an operator, formalized here as the admissible class $\mathcal{A}$ of adapted intervention policies. Through λ_i, the operator can modify any of the three node-i contributions in (1). For example, by reshaping the strength of X_i's couplings to the other observed nodes through H_i^{mech}, or by decoupling X_i from its bath through H_i^{bath}, effectively detaching the subsystem from its exogenous noise. Crucially, interventions modify the Hamiltonian function but never the form of Hamilton's equations, separating intervenable mechanisms from immutable dynamical law.

The Hamiltonian decomposition therefore plays a role analogous to the independent-causal-mechanisms assumption in structural causal modeling, with interventions affecting only one variable. However, there is a striking difference with the usual definition of causal parenthood in DAGs and SCMs. In Hamiltonians, the causal semantics is endowed by the interventional policies and not just by the additive decomposition of the Hamiltonian, unlike, e.g., the structural assignment of a SCM. That two or more variables exchange information via H_i^{mech} can be, on its own, insufficient to determine the direction of causality.

Definition 2.2 (Mechanism dependence)**.** The mechanism dependence set of node i is $\mathcal{D}_i := \{j \neq i : \partial_{x_j} H_i^{\text{mech}} \not\equiv 0\}$. Nodes X_i and X_j are *mechanically paired* if $j \in \mathcal{D}_i$ or $i \in \mathcal{D}_j$.

The following result, proven in Appendix B, shows that HCMs describe a wide class of stochastic dynamics and recover classical structural causal models as a special case.

Theorem 2.3 (Realization, informal)**.** *Let $U : \mathbb{R}^n \times \mathcal{L} \to \mathbb{R}$ be a smooth potential and $\sigma : \mathcal{L} \to \mathbb{R}^{n \times n}$ a smooth, diagonal, positive diffusion coefficient. Then there exists a sequence of HCMs whose observed coordinates have, in the limit of large bath size and small inertial parameter, path law on $C([0, T]; \mathbb{R}^n)$ equal to that of the Itô diffusion*

$$\mathrm{d}X_t \;=\; -\nabla_x U(X_t, \lambda_t)\,\mathrm{d}t \;+\; \sigma(\lambda_t)\,\mathrm{d}W_t, \quad (2)$$

where W is a standard n-dimensional Brownian motion independent of the initial state X_0.

Static causal models as a stationary special case. Choosing $\sigma(\lambda_t) = sI$, $U = -\frac{1}{2}s^2 \log p$ for a smooth strictly positive density p and initializing $X_0 \sim p$ and $\lambda \equiv 0$, the HCM realizes a stationary diffusion with observational distribution p. We thus recover classical structural causal models, together with their cyclic generalizations as defined in Lorch et al. [2024], as the stationary regime of an HCM. Of note, by tuning the set of admissible policies $\mathcal{A}$ one can also define a notion of interventions that takes the dynamic response to λ into account.

With HCMs in place, we turn to the notion of causal effect they support. Let $\mathbb{P}_i^{\lambda} := (\Pi_i)_{\#}\mathbb{P}_F^{\lambda}$ denote the *node-wise marginal path law* of X_i under policy λ, obtained by projecting $\mathbb{P}_F^{\lambda}$ onto the i-th coordinate.

Definition 2.4 (Causal effect)**.** Fix $i \neq j$ and a baseline policy λ_{-j} for all intervention coordinates except λ_j. We say that X_j has a *causal effect* on X_i over $[0,T]$ if there exist two admissible policies $\lambda_j^{(1)}, \lambda_j^{(2)}$ such that $\mathbb{P}_i^{(\lambda_j^{(1)};\lambda_{-j})} \neq \mathbb{P}_i^{(\lambda_j^{(2)};\lambda_{-j})}$.

Combining Definition 2.4 and the notion of mechanical pairing of Definition 2.2, we can define what parenthood is for an HCM.

Definition 2.5 (Operational parents)**.** For $i \neq j$, X_j is a *parent* of X_i, written $j \to i$, if the nodes are mechanically paired according to Definition 2.2 and X_j has a causal effect on X_i in the sense of Definition 2.4.

This graph records intervention-induced influence between mechanistically paired nodes. Such a pair can have both arrows, one arrow, or neither. Definition 2.4 is intrinsically trajectory-level: causal effect is a change in the entire path law of X_i, not in a single-time observable. It is natural to quantify it through discrepancies between the induced path laws.

Definition 2.6 (Generalized trajectory ATE)**.** Let $\mathfrak{D} : \mathcal{P}(C([0,T],\Gamma_{X_i})) \times \mathcal{P}(C([0,T],\Gamma_{X_i})) \to \mathbb{R}$ be a measurable discrepancy on pairs of path laws. For two admissible policies $\lambda^{(1)}, \lambda^{(2)} \in \mathcal{A}$, the *generalized trajectory average treatment effect* on node X_i is

$$\mathrm{ATE}_{0:T}^{(i)}(\mathfrak{D};\lambda^{(1)},\lambda^{(2)}) := \mathfrak{D}\Big(\mathbb{P}_i^{\lambda^{(1)}}, \mathbb{P}_i^{\lambda^{(2)}}\Big).$$

When $\mathfrak{D}$ separates measures ($\mathfrak{D}(\mu,\nu) = 0 \iff \mu = \nu$), X_j has a causal effect on X_i in the sense of Definition 2.4 if and only if $\mathrm{ATE}_{0:T}^{(i)} \neq 0$ for some pair of admissible policies; we record this formally in Proposition B.6 of Appendix B.

Relation to existing notions of causal effect. Definition 2.6 unifies several existing notions of temporal causal effect through suitable choices of $\mathfrak{D}$. *Single-time effects* $\mathbb{E}[f_t(X_i^{\lambda}(t))] - \mathbb{E}[f_t(X_i^{\lambda'}(t))]$ appear in two well-developed strands of the literature: fixed-horizon and regime-comparison effects, as in longitudinal g-methods, marginal structural models, dynamic treatment regimes, and target-trial formulations [Robins et al., 2000, Hernán et al., 2000, Naimi et al., 2017, Murphy, 2003, Hernán and Robins, 2016, Hernán et al., 2022]; and local or lagged effects in structural nested mean models, causal excursion effects, single-time-series experiments, and direct potential-outcome formulations for observational time series [Robins, 1994, Qian et al., 2021, Bojinov and Shephard, 2019, Rambachan and Shephard, 2021]. They are recovered by

$$\mathfrak{D}_{F_t}(\mu,\nu) := \int F_t\, d\mu - \int F_t\, d\nu, \qquad F_t(x) := f_t(x(t)).$$

Integrated effects appearing in history-restricted and history-adjusted marginal structural models, dynamic treatment regimes, and sustained-strategy formulations [Neugebauer et al., 2007, Petersen et al., 2007, Murphy, 2003, Hernán and Robins, 2016] are recovered by

$$\mathfrak{D}_{F_{\mathrm{int}}}(\mu,\nu) := \int F_{\mathrm{int}}\, d\mu - \int F_{\mathrm{int}}\, d\nu, \qquad F_{\mathrm{int}}(x) := \int_0^T f_t(x(t))\, dt.$$

Genuinely path-level effects central to continuous-time g-computation and continuous-time marginal structural models [Gill and Robins, 2001, Røysland, 2011], local-independence and dynamic path-analysis [Didelez, 2008, Aalen et al., 2012], dynamic structural causal models and causal interpretations of stochastic differential equations [Rubenstein et al., 2018, Boeken and Mooij, 2024, Sokol and Hansen, 2014], and recent functional-longitudinal causal inference [Ying, 2024, Sun and Crawford, 2022] are recovered by

$$\mathfrak{D}_{\Phi}(\mu,\nu) := \int \Phi\, d\mu - \int \Phi\, d\nu, \qquad \Phi : C([0,T],\Gamma_{X_i}) \to \mathbb{R},$$

or, more generally, by any law-level discrepancy on path measures (e.g., KL divergence).

A separate, adjacent family studies temporal influence through *Granger-style* or *information-flow* quantities, including Granger-causal predictability [White and Lu, 2010, Barnett and Seth, 2015], transfer entropy [Schreiber, 2000, Barnett et al., 2009], directed information [Massey, 1990, Kramer, 1998, Amblard and Michel, 2013], graphical or structural time-series causality [Eichler and Didelez, 2007, Peters et al., 2013, Runge et al., 2023], and causation-entropy criteria [Sun and Bollt, 2014, Sun et al., 2015, 2014, Surasinghe and Bollt, 2020, Runge et al., 2019]. These quantities measure directed predictability rather than intervention effect, and are therefore outside the scope of Definitions 2.4 and 2.6. They are nonetheless central to our purposes: as we show in the next section, the thermodynamic quantities along an HCM trajectory provide a natural account of such information-flow measures [Prokopenko et al., 2013], bridging the interventional and information-flow perspectives within a single framework.

3 The Statistical Footprint of Causality in Non-equilibrium Thermodynamics

We now pass from the definition of HCMs to the thermodynamic quantities they naturally generate and their relationship with the HCM causal structure. For every intervention policy and every induced trajectory, the model assigns observables with a direct physical meaning. These quantities provide a measurable description of the effect of interventions on a system, and they will allow us to connect the Hamiltonian structure introduced in Section 2 to causal notions such as parenthood and causal effect.

Thermodynamic preliminaries. Fix an admissible intervention policy $\lambda \in \mathcal{A}$, and let $z(\cdot)$ be a microscopic trajectory generated by the controlled Hamiltonian $H_\lambda(t, z) := H(z, \lambda_t)$.

Definition 3.1 (Work)**.** The *work* performed by the intervention policy λ along the trajectory $z(\cdot)$ is

$$W(z;\lambda) \;:=\; \int_0^T \partial_t H_\lambda(t, z(t))\, dt \;=\; \int_0^T \langle \partial_\lambda H(z(t), \lambda(t)),\, \partial_t \lambda(t) \rangle\, dt,$$

where the second equality holds whenever λ is differentiable and $\langle \cdot, \cdot \rangle$ denotes the standard scalar product.

Definition 3.2 (Heat)**.** Let $z(t) = (x(t), e(t))$ be a controlled HCM trajectory and $e_i(t)$ the bath attached to node X_i. The *heat* transferred from bath E_i into the observed system over $[0, T]$ is

$$Q_i[z] \;:=\; -\big(H_{E_i}(e_i(T)) - H_{E_i}(e_i(0))\big).$$

With this sign convention, $Q_i[z] > 0$ means net energy flows from bath E_i into the observed system.

Phenomenological causality and work. The thermodynamic quantities introduced above connect to several aspects of causality. We begin with the most immediate one: under the Phenomenological Causality stance [Janzing and Mejia, 2022] described in Section 2, causal effect is whatever an intervention can produce, and *work* is the natural measure of the energy footprint of an intervention on a system (note in particular that, if an intervention λ is constant in time, then $W[z;\lambda] = 0$). The following result formalizes this: intervention-induced work detects which variables enter the mechanism of a given node, providing a thermodynamic characterization of mechanism dependence.

Proposition 3.3 (Work rate identifies mechanism dependence)**.** *For a given HCM and a node i, assume the admissible interventions class $\mathcal{A}$ contains a λ that is differentiable at a fixed time $t^* \in (0, T)$ and such that $\partial_t \lambda_i(t^*) \neq 0$ and $\partial_t \lambda_k(t^*) = 0$ for $k \neq i$, which is locally faithful at $\lambda(t^*)$ in the sense of Assumption B.1. Define the work rate of λ at t^* by $\dot{W}(t, z; \lambda) := \langle \partial_\lambda H(z, \lambda(t)), \partial_t \lambda(t) \rangle$. Then, for every $j \neq i$, $j \in \mathcal{D}_i$ if and only if there exists a state z^* such that:*

$$\Big[\nabla_x \dot{W}(t^*, z^*; \lambda)\Big]_j \neq 0.$$

This proposition is appealing because work directly tracks which variables participate in a local mechanism. Its limitation is practical: computing $\dot{W}$ requires explicit access to the Hamiltonian, just as computing a mechanism-level causal effect in an SCM would require access to the structural

equations. We therefore turn to a thermodynamic quantity that can be estimated from trajectories: *entropy production*.

The idea behind entropy production is to separate genuine temporal irreversibility from the effect due to the state and the intervention protocol changing over time. An intervened trajectory is compared to the corresponding reversed experiment: the microscopic state z_t is time-reversed and the control schedule λ_t is run backward. Entropy production is the log-likelihood asymmetry between these two ensembles. In this sense, it removes the trivial asymmetry due to the changing controls and the state responding to them and asks: how statistically distinguishable is the forward experiment from the same experiment recorded backward in time? This residual arrow-of-time signal is the thermodynamic footprint of performing λ on a given state, and below we show that its local forms retain causal information while being accessible from trajectory data [Otsubo et al., 2020, 2022].

Definition 3.4 (Backward path law)**.** Let Θ be the time-reversal operator from Definition 2.1. Given $\lambda \in \mathcal{A}$ define the *time-reversed protocol* by $(\Theta\lambda)(t) := \lambda(T-t)$ and pathwise time reversal $\mathcal{R} : \mathcal{Z} \to \mathcal{Z}$ by $(\mathcal{R}z)(t) := \Theta z(T-t)$, where $\mathcal{Z} := C([0,T],\Gamma)$ is the microscopic system-bath path space. Assume the forward initial law is $\rho_0 \otimes \gamma_E$. Let ρ_T^λ be the forward final law of X and Θ_X the restriction of Θ to X. Let $\mathbb{Q}_R^\lambda$ denote the microscopic path law generated under $\Theta\lambda$ from the initial law $(\Theta_X)^\# \rho_T^\lambda \otimes \gamma_E$. The *backward observed path law* is $\mathbb{P}_B^{\Theta\lambda} := (\Pi_X \circ \mathcal{R})^\# \mathbb{Q}_R^\lambda$.

Definition 3.5 (Pathwise entropy production)**.** Assume $\mathbb{P}_F^\lambda$ is absolutely continuous with respect to $\mathbb{P}_B^{\Theta\lambda}$. The *pathwise entropy production* of the controlled process is: $\Sigma^\lambda(x) := \log \frac{d\mathbb{P}_F^\lambda}{d\mathbb{P}_B^{\Theta\lambda}}(x)$. Its expectation $\mathbb{E}_{\mathbb{P}_F^\lambda}[\Sigma^\lambda] = D_{\mathrm{KL}}\big(\mathbb{P}_F^\lambda \,\big\|\, \mathbb{P}_B^{\Theta\lambda}\big)$ is called the *mean entropy production*.

Unlike Shannon entropy, which is a state function of the instantaneous law, Σ^λ is a path-space quantity: it distinguishes processes with the same endpoint distributions but different transient evolution. Under the assumptions stated in Appendix A, this pathwise asymmetry coincides with the total entropy change of the observed system and its baths

$$\Sigma^\lambda \;=\; \Delta s_{\mathrm{sys}}^\lambda \;-\; \sum_i \beta_i Q_i^\lambda, \tag{3}$$

where $\Delta s_{\mathrm{sys}}^\lambda = -\log p_T^\lambda(x(T)) + \log p_0^\lambda(x(0))$ is the standard entropy change of the observed system under λ. Equipped with this interpretation, the inequality $\mathbb{E}_{\mathbb{P}_F^\lambda}[\Sigma^\lambda] \geq 0$ that follows from $\mathbb{E}_{\mathbb{P}_F^\lambda}[\Sigma^\lambda] = D_{\mathrm{KL}}\big(\mathbb{P}_F^\lambda \,\big\|\, \mathbb{P}_B^{\Theta\lambda}\big)$ is the derivation of the *Second Law of Thermodynamics* for Hamiltonian systems [Jarzynski, 1997]. Thus, entropy production is both a statistical measure of time-reversal asymmetry and a physical measure of dissipation.

The quantities introduced here are standard objects of stochastic and non-equilibrium thermodynamics. Three traditions feed into the formulation we adopt: trajectory-level work and heat in stochastic energetics [Sekimoto, 1998, 2010], fluctuation theorems and entropy production as a path-likelihood ratio [Jarzynski, 1997, Crooks, 1999, Seifert, 2012], and information thermodynamics for coupled subsystems [Sagawa and Ueda, 2012a, Ito and Sagawa, 2013, Parrondo et al., 2015], whose tradition expands well beyond the works cited here. Our aim here is not to redefine these objects, but to localize them at the node level inside an explicitly intervention-based causal model, so that entropy production becomes a witness of causal effect.

Projected local entropy production. The entropy production Σ^λ above is global: it compares the forward and backward laws of the whole observed trajectory. Causal effect, however, was defined in Section 2 through changes in the marginal path law of a single variable. The first localization of entropy production is therefore obtained by projecting the path-space asymmetry onto the trajectory of one node. For a node i, let $\mathbb{P}_i^\lambda := (\Pi_i)_\# \mathbb{P}_F^\lambda$ be the marginal observed path law of $X_i(\cdot)$, and let $\mathcal{R}_i$ denote time reversal on X_i-paths. The projected entropy production of node i is the time-reversal asymmetry visible from observing only X_i.

Definition 3.6 (Projected local entropy production)**.** The *projected local entropy production* of node i under the appropriate absolute continuity assumption for the marginal measures is

$$\Sigma_i^{\mathrm{proj},\lambda}(x_i) := \log \frac{\mathrm{d}\mathbb{P}_i^\lambda}{\mathrm{d}(\mathcal{R}_i)_\# \mathbb{P}_i^\lambda}(x_i).$$

Its mean with respect to $\mathbb{P}_i^\lambda$ is $\mathbb{E}_{\mathbb{P}_i^\lambda}[\Sigma_i^{\mathrm{proj},\lambda}(x_i)] := D_{\mathrm{KL}}\big(\mathbb{P}_i^\lambda \,\big\|\, (\mathcal{R}_i)_\# \mathbb{P}_i^\lambda\big)$.

Projected entropy production is local in exactly the same sense as our causal effect definition: it is a functional of the marginal path law of X_i. Consequently, changes in projected entropy production provide a one-way witness of causal effect.

Proposition 3.7 (Projected entropy production witnesses causal effect)**.** *Fix $i \neq j$ and a baseline policy λ_{-j}. For $k = 1, 2$, let $\lambda^{(k)} := (\lambda_j^{(k)}, \lambda_{-j}) \in \mathcal{A}$, and assume $\Sigma_i^{\mathrm{proj},\lambda^{(k)}}$ is well-defined and integrable. If*

$$\mathbb{E}_{\mathbb{P}_i^{\lambda^{(1)}}}\left[\Sigma_i^{\mathrm{proj},\lambda^{(1)}}\right] \neq \mathbb{E}_{\mathbb{P}_i^{\lambda^{(2)}}}\left[\Sigma_i^{\mathrm{proj},\lambda^{(2)}}\right],$$

then X_j has a causal effect on X_i over $[0, T]$ in the sense of Definition 2.4.

The projected quantity above is the right entropy-production object for witnessing causal effect in the sense of Definition 2.4: in the Langevin specialization, its projected-current rate is estimable from realizations of x_i alone via statistical techniques such as those presented in [Otsubo et al., 2020, 2022]. In this sense, projected entropy production detects whether j has an effect on i via a directed path, but not yet whether j is a paired node. This is consistent with several results in causality, where soft interventions identify ancestral relations in DAGs [Tian and Pearl, 2001] and directed connectivity in cyclic models [Mooij and Heskes, 2013].

Detecting Mechanism Dependence in Langevin Dynamics. We next study a complementary local entropy-production response that identifies mechanism dependence in the controlled Langevin setting of Theorem 2.3. In this setting the probability evolution follows the Fokker-Planck equation: $\partial_t p^\lambda(x,t) = -\sum_{k=1}^n \partial_{x_k} J_k^\lambda(x,t)$, where J_k^λ is the probability current associated with coordinate X_k. Inspired by the energetic characterization of global entropy production in (3), and following the information-flow formulation of subsystem thermodynamics [Horowitz and Esposito, 2014, Ito and Sagawa, 2013], we define the local entropy production of node i as the marginal entropy balance of X_i, corrected by the heat exchanged with its bath and by the information flow between X_i and the rest of the system.

Definition 3.8 (Local entropy production)**.** Fix node i with bath E_i and let $s_i^{\mathrm{marg},\lambda}(t) := -\int p_{i,t}^\lambda(x_i) \log p_{i,t}^\lambda(x_i)\, dx_i$ be the marginal Shannon entropy of $X_i(t)$, and let Q_i^λ denote the heat exchanged between bath E_i into the observed system over $[0, T]$. Denote the contribution of the i-th current to the joint system entropy rate by $\dot{s}_{\mathrm{sys}}^{(i),\lambda}(t) := -\int_{\mathbb{R}^n} J_i^\lambda(x,t)\, \partial_{x_i} \log p^\lambda(x,t)\, dx$. We define the information-flow rate associated with node i by:

$$\dot{I}_i^{\mathrm{flow},\lambda}(t) := \frac{d}{dt}\left(s_i^{\mathrm{marg},\lambda}(t)\right) - \dot{s}_{\mathrm{sys}}^{(i),\lambda}(t).$$

The *local entropy production* of node X_i over $[0, T]$ is

$$\Sigma_i^{\mathrm{loc},\lambda} := \Delta s_i^{\mathrm{marg},\lambda} - \beta_i \mathbb{E}_\lambda[Q_i^\lambda] - \int_0^T \dot{I}_i^{\mathrm{flow},\lambda}(t)\, dt.$$

Σ_i^{loc} is still a trajectory-based object: in the Langevin case it can be estimated from samples of the full observed trajectory $x(\cdot)$ using variational estimators [Otsubo et al., 2020, 2022]. The price for moving from projected to channel-local entropy production is therefore additional observational information: projected EP requires only $x_i(\cdot)$, whereas $\dot{\Sigma}_i^{\mathrm{loc}}$ uses the full state as context. The gain is structural: while projected EP witnesses finite-time causal effects, the response of the channel-local EP rate to a right-local intervention identifies mechanism dependence.

Theorem 3.9 (Mechanism dependence from local entropy production)**.** *Consider a controlled Langevin diffusion of the kind described in Theorem 2.3. Fix an intervention time $t^* \in (0, T)$, and write*

$$p_* := p(\cdot, t^*), \qquad \lambda_* := \lambda(t^*).$$

Assume the admissible policy class $\mathcal{A}$ contains two right-local protocols initialized at (p_*, λ_*): a baseline protocol λ^0 with $\lambda^0(s) = \lambda_*$, and a j-probe protocol λ^v with

$$\lambda^v(0) = \lambda_*, \qquad \lambda^v(s) = \lambda_* + vse_j + o(s) \quad \text{as } s \downarrow 0,$$

where $v \in \mathbb{R}$ is the intervention velocity. Let $p^0(\cdot, s)$ and $p^v(\cdot, s)$ be the corresponding laws, both initialized at p_*. For $i \neq j$, define the local entropy-production response

$$R_{i|j}(t^*) := \lim_{v \to 0} \lim_{s \downarrow 0} \frac{\dot{\Sigma}_i^{\mathrm{loc}}[p^v(\cdot, s), \lambda^v(s)] - \dot{\Sigma}_i^{\mathrm{loc}}[p^0(\cdot, s), \lambda^0(s)]}{vs},$$

where $\dot{\Sigma}_i^{\mathrm{loc}}[p, \lambda]$ denotes the local entropy-production rate functional. Then, under Assumption B.2,

$$\boxed{i \in \mathcal{D}_j \quad \Longleftrightarrow \quad R_{i|j}(t^*) \neq 0.}$$

Remark 3.10 (Schwarz symmetry and the role of intervention). Uncontrolled gradient dynamics has $\partial_{x_i}\partial_{x_j} U = \partial_{x_j}\partial_{x_i} U$, so observational data alone identifies only an undirected dependency structure. Soft interventions break this symmetry through the asymmetric mixed partials $\partial_{x_i}\partial_{\lambda_j} U$, which is precisely what $R_{i|j}$ probes.

4 Experimental Results

The goal of these experiments is to demonstrate that entropy production is not just a theoretical descriptor of causal models, but can be *measured* from trajectories and used as a working causal estimand.

ATE experiments: validating Proposition 3.7. In light of Proposition 3.7 we use the projected entropy production $\mathbb{E}_{\mathbb{P}_i^{\lambda^{(1)}}}\left[\Sigma_i^{\mathrm{proj},\lambda^{(1)}}\right]$ as a measure of path-level causal effect, estimated with the current-restricted variant of the NEEP estimator [Kim et al., 2020, Otsubo et al., 2020]. We consider two qualitatively distinct settings, each designed so that a standard observable-level estimand is zero by construction even though the two protocols drive the system along visibly different paths:

- **(A) Same endpoints, different paths.** A linear chain of overdamped oscillators is driven on its root node by two protocols that reach the same final value but follow different trajectories (a smooth ramp vs. an excursion that overshoots and returns). Because both trajectories have the same terminal values, the *endpoint* ATE on the driven node is zero by design.
- **(B) Same interventions, reversed in time.** A small system with a hidden nonlinear mediator receives the same two pulses applied in opposite order. On a symmetric observable summing all coordinates, a time-translation argument forces the *cumulative* ATE $cATE(Y) = \int_0^T \mathbb{E}[Y_t \mid \mathrm{do}(A \circ B)] - \mathbb{E}[Y_t \mid \mathrm{do}(B \circ A)]\, dt$ to vanish.

We compare the *projected EP contrast* $ATE(\Sigma_i^{\mathrm{proj}}) := \mathbb{E}_{\mathbb{P}_i^{\lambda^{(1)}}}\left[\Sigma_i^{\mathrm{proj},\lambda^{(1)}}\right] - \mathbb{E}_{\mathbb{P}_i^{\lambda^{(2)}}}\left[\Sigma_i^{\mathrm{proj},\lambda^{(2)}}\right]$ against *endpoint* ATE in case (A) and *cumulative* ATE in case (B), with bootstrap-generated 95% confidence intervals on the ATEs and a one-sample t-test for the projected entropy production across independent seeds for NEEP. Table 1 reports the headline numbers: in both settings the observable ATE that the design forces to vanish is statistically indistinguishable from 0, while entropy production separates the two protocols confidently. This shows empirically that the projected entropy production is able to detect path-level causal effects in the sense of Definition 2.4 which are invisible to common versions of the ATE. While other path-level discrepancies could also detect such differences, we emphasize that entropy production provides a physically meaningful and estimable discrepancy tied to irreversibility.

Discovery experiments: validating Theorem 3.9. To test the mechanism dependence identification condition induced by Theorem 3.9 we generate random cyclic Erdős–Rényi directed graphs

Setting	Designed-zero observable ATE (95% CI)	$ATE(\Sigma_i^{\mathrm{proj}})$ (NEEP, mean$\pm$SD, p)
(A) Same endpoints	Endpoint ATE $= -0.012$ $[-0.037, +0.012]$	-0.366 ± 0.001, $p = 2.3\times10^{-6}$
(B) Reversed in time	Cumulative ATE $= +0.002$ $[-0.062, +0.067]$	-0.417 ± 0.017, $p = 6.8\times10^{-7}$

Table 1: Marginal/temporal ATE is zero by construction, so specific choices like Endpoint and Cumulative ATE miss this causal effect; $ATE(\Sigma_i^{\mathrm{proj}})$ resolves the path-level effect and identifies a non-zero causal effect.

($n=15$, edge probability 0.10) and instantiate three families of overdamped Langevin HCMs with potentials: $U(x,\lambda) = \frac{1}{2}\sum_i a_i x_i^2 + U_{\mathrm{self}}(x) + \sum_j \lambda_j[b_j x_j + \sum_{i\in\mathcal{D}_j} c_{ij}\,\phi(x_i,x_j)]$ shared $\sigma_0=0.5$ and the following kernels ϕ: Bilinear ($U_{\mathrm{self}}=0,\ \phi=x_ix_j$); Sigmoid ($U_{\mathrm{self}}=0,\ \phi=\tanh(\beta x_i)\,x_j$, $\beta=2.5$); Cubic ($U_{\mathrm{self}}=\frac{\gamma}{4}\sum_i x_i^4,\ \phi=x_ix_j,\ \gamma=0.5$). For each graph we pick a target j randomly and estimate $R_{i|j}$ as a finite difference of $\dot{\Sigma}_i^{\mathrm{loc}}$ estimated with the full-system version of the NEEP estimator [Otsubo et al., 2020], fixing $\lambda_*=2$, $\Delta t=0.02$ and $N=8000$. Full configurations for NEEP's training and the hyperparameters of the potential U are provided in Appendix C. The mechanism dependence set is selected by thresholding values of the estimated $R_{i|j}$ via $\widehat{\mathcal{D}}_j = \{i : |\widehat{R}_{i|j}| > 0.10\cdot\max_i|\widehat{R}_{i|j}|\}$. We report mean F1, precision, recall, and local SHD $\left|\widehat{\mathcal{D}}_j \triangle \mathcal{D}_j)\right|$ across 10 random graphs per potential, along with 95% confidence intervals in table 2. We observe near-perfect recall paired with strong precision in the recovery results.

Potential	F1	Precision	Recall	local SHD $\downarrow$
Bilinear	$0.839_{\,[0.77,\,0.91]}$	$0.775_{\,[0.66,\,0.90]}$	$0.967_{\,[0.90,\,1.00]}$	$1.10_{\,[0.5,\,1.8]}$
Sigmoid	$0.858_{\,[0.79,\,0.93]}$	$0.800_{\,[0.69,\,0.91]}$	$0.955_{\,[0.89,\,1.00]}$	$0.90_{\,[0.4,\,1.5]}$
Cubic	$0.825_{\,[0.70,\,0.93]}$	$0.745_{\,[0.59,\,0.89]}$	$0.971_{\,[0.91,\,1.00]}$	$1.40_{\,[0.4,\,2.7]}$

Table 2: Recovery on random cyclic ER graphs ($n=15$, 10 graphs per row) using the results of Theorem 3.9. All metrics with 95% bootstrap CIs. Performance is very strong despite the challenging non-stationary and cyclic setting.

5 Conclusions

In this paper, we proposed an alternative language for statistical causality that is closely related to how causation is described in physical systems [Rovelli, 2023]. This is appealing because concepts that are traditionally difficult to model with existing frameworks (a notable example is the definition of interventional distributions in cyclic SCMs [Bongers et al., 2021]) become natural. In simple numerical experiments, we have empirically demonstrated that entropy production is an effective witness of causal effects that standard treatment effect formulations struggle to capture and that local entropy-production responses support mechanism-dependence recovery. Clearly, the SCM framework is flexible, and several phenomena described in our paper can also be expressed ad-hoc in that language. At the same time, we have shown in Theorem 2.3 that HCMs generalize SCMs and more naturally model the temporal phenomena arising in physical systems.

In this new language, we hope that many of the concepts already existing in causality will be rediscovered, now capable of modeling broader classes of physical systems. A notable first example is the notion of counterfactuals, which can be naturally posed by noise abduction to the bath trajectories. However, how to perform this abduction step is non-trivial as the noise is no longer just a distribution but a probability flow. Secondly, it must be stated that despite describing many processes such as Langevin diffusions, general temporal SCMs [Peters et al., 2013] with non-gradient drifts are not immediately described by HCMs, and extending the present theory to said framework is a prominent next step we reserve for the future. Finally, the learning of the Hamiltonians and its connections to causal discovery. Classically, causal discovery from observational data rests on structural assumptions such as non-linear additive noise models, and we currently do not know what the HCM analogue is. At the same time, we highlight an interesting connection with the score-matching literature in causal discovery [Rolland et al., 2022, Montagna et al., 2023], which appears in our models in the continuity equation of Langevin processes. The intersection with causal representation learning [Schölkopf et al., 2021], perhaps via diffusion models, is also an interesting, open-ended future direction.

Acknowledgments and Funding

The authors thank Sosuke Ito, Dominik Janzing, Sara Magliacane and Joris Mooij for providing useful insights on the development of the process. DR is financed by the project Building Energy Systems on causal reasoning (BOSS), funded within the Technologies and Innovations for the Climate-Neutral City (TIKS) Programme of the Austrian Research Promotion Agency (FFG).

References

Odd O. Aalen, Kjetil Røysland, Jon Michael Gran, and Bruno Ledergerber. Causality, mediation and time: A dynamic viewpoint. *Journal of the Royal Statistical Society: Series A*, 175(4):831–861, 2012. doi: 10.1111/j.1467-985X.2011.01030.x.

Pierre-Olivier Amblard and Olivier J. J. Michel. The relation between granger causality and directed information theory: A review. *Entropy*, 15(1):113–143, 2013. doi: 10.3390/e15010113.

Lionel Barnett and Anil K. Seth. Granger causality for state-space models. *Physical Review E*, 91 (4):040101, 2015. doi: 10.1103/PhysRevE.91.040101.

Lionel Barnett, Adam B. Barrett, and Anil K. Seth. Granger causality and transfer entropy are equivalent for gaussian variables. *Physical Review Letters*, 103(23):238701, 2009. doi: 10.1103/PhysRevLett.103.238701.

Philip Boeken and Joris M. Mooij. Dynamic structural causal models. *arXiv preprint arXiv:2406.01161*, 2024.

Philip Boeken and Joris M. Mooij. Causal graphs, markov properties and do-calculus for stochastic differential equations, 2026. URL `https://arxiv.org/abs/2607.12140`.

Iavor Bojinov and Neil Shephard. Time series experiments and causal estimands: Exact randomization tests and trading. *Journal of the American Statistical Association*, 114(528):1665–1682, 2019. doi: 10.1080/01621459.2018.1527225.

Stephan Bongers, Patrick Forré, Jonas Peters, and Joris M Mooij. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):2885–2915, 2021.

Riccardo Cadei, Ilker Demirel, Piersilvio De Bartolomeis, Lukas Lindorfer, Sylvia Cremer, Cordelia Schmid, and Francesco Locatello. Prediction-powered causal inferences. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

A. O. Caldeira and A. J. Leggett. Path integral approach to quantum brownian motion. *Physica A: Statistical Mechanics and its Applications*, 121(3):587–616, 1983. doi: 10.1016/0378-4371(83)90013-4.

Gavin E. Crooks. Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Physical Review E*, 60(3):2721–2726, 1999. doi: 10.1103/PhysRevE.60.2721.

Vanessa Didelez. Graphical models for marked point processes based on local independence. *Journal of the Royal Statistical Society: Series B*, 70(1):245–264, 2008. doi: 10.1111/j.1467-9868.2007.00634.x.

Michael Eichler and Vanessa Didelez. Causal reasoning in graphical time series models. In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 109–116, 2007.

Stefan Feuerriegel, Dennis Frauen, Valentyn Melnychuk, Jonas Schweisthal, Konstantin Hess, Alicia Curth, Stefan Bauer, Niki Kilbertus, Isaac S Kohane, and Mihaela van der Schaar. Causal machine learning for predicting treatment outcomes. *Nature Medicine*, 30(4):958–968, 2024.

Richard D. Gill and James M. Robins. Causal inference for complex longitudinal data: The continuous time g-computation formula. *arXiv preprint math/0409436*, 2001.

Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.

Miguel A. Hernán and James M. Robins. Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology*, 183(8):758–764, 2016. doi: 10.1093/aje/kwv254.

Miguel A. Hernán, Babette Brumback, and James M. Robins. Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men. *Epidemiology*, 11(5):561–570, 2000. doi: 10.1097/00001648-200009000-00012.

Miguel A. Hernán, Wei Wang, and David E. Leaf. Target trial emulation: A framework for causal inference from observational data. *JAMA*, 328(24):2446–2447, 2022. doi: 10.1001/jama.2022.21383.

Jordan M. Horowitz and Massimiliano Esposito. Thermodynamics with continuous information flow. *Physical Review X*, 4(3):031015, 2014. doi: 10.1103/PhysRevX.4.031015.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015.

Sosuke Ito and Takahiro Sagawa. Information thermodynamics on causal networks. *Physical Review Letters*, 111(18):180603, 2013. doi: 10.1103/PhysRevLett.111.180603.

Sosuke Ito and Takahiro Sagawa. Maxwell's demon in biochemical signal transduction with feedback loop. *Nature Communications*, 6:7498, 2015. doi: 10.1038/ncomms8498.

Dominik Janzing and Sergio Hernan Garrido Mejia. Phenomenological causality, 2022. URL `https://arxiv.org/abs/2211.09024`.

Christopher Jarzynski. Nonequilibrium equality for free energy differences. *Physical Review Letters*, 78(14):2690–2693, 1997. doi: 10.1103/PhysRevLett.78.2690.

Dong-Kyum Kim, Youngkyoung Bae, Sangyun Lee, and Hawoong Jeong. Learning entropy production via neural networks. *Physical Review Letters*, 125(14), October 2020. ISSN 1079-7114. doi: 10.1103/physrevlett.125.140604. URL `http://dx.doi.org/10.1103/PhysRevLett.125.140604`.

Gerhard Kramer. *Directed Information for Channels with Feedback*. PhD thesis, Swiss Federal Institute of Technology Zürich, 1998.

Lars Lorch, Andreas Krause, and Bernhard Schölkopf. Causal modeling with stationary diffusions. In *International Conference on Artificial Intelligence and Statistics*, pages 1927–1935. PMLR, 2024.

James L. Massey. Causality, feedback and directed information. In *Proceedings of the 1990 International Symposium on Information Theory and its Applications*, pages 303–305, 1990.

Tommaso Mencattini, Riccardo Cadei, and Francesco Locatello. Exploratory causal inference in saence. *International Conference on Learning Representations*, 2026.

A Miguel, ROBINS HERNAN, and M James. *Causal inference: what if*. CRC PRESS, 2023.

Francesco Montagna, Nicoletta Noceti, Lorenzo Rosasco, Kun Zhang, and Francesco Locatello. Causal discovery with score matching on additive models with arbitrary noise. In *Conference on Causal Learning and Reasoning*, pages 726–751. PMLR, 2023.

Joris M Mooij and Tom Heskes. Cyclic causal discovery from continuous equilibrium data. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 431–439, 2013.

Joris M Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. *Journal of machine learning research*, 21(99):1–108, 2020.

Susan A. Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B*, 65(2):331–355, 2003. doi: 10.1111/1467-9868.00389.

Ashley I. Naimi, Stephen R. Cole, and Edward H. Kennedy. An introduction to g methods. *International Journal of Epidemiology*, 46(2):756–762, 2017. doi: 10.1093/ije/dyw323.

Romain Neugebauer, Mark J. van der Laan, Marshall M. Joffe, and Ira B. Tager. Causal inference in longitudinal studies with history-restricted marginal structural models. *Electronic Journal of Statistics*, 1:119–154, 2007. doi: 10.1214/07-EJS050.

Shun Otsubo, Sosuke Ito, Andreas Dechant, and Takahiro Sagawa. Estimating entropy production by machine learning of short-time fluctuating currents. *Physical Review E*, 101(6), 2020. ISSN 2470-0053. doi: 10.1103/physreve.101.062106. URL `http://dx.doi.org/10.1103/PhysRevE.101.062106`.

Shun Otsubo, Sreekanth K. Manikandan, Takahiro Sagawa, and Supriya Krishnamurthy. Estimating time-dependent entropy production from non-equilibrium trajectories. *Communications Physics*, 5(1), January 2022. ISSN 2399-3650. doi: 10.1038/s42005-021-00787-x. URL `http://dx.doi.org/10.1038/s42005-021-00787-x`.

Juan M. R. Parrondo, Jordan M. Horowitz, and Takahiro Sagawa. Thermodynamics of information. *Nature Physics*, 11(2):131–139, 2015. doi: 10.1038/nphys3230.

Judea Pearl. *Causality: Models, Reasoning, and Inference.* Cambridge university press, 2009.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on time series using restricted structural equation models. In *Advances in Neural Information Processing Systems 26*, pages 154–162, 2013.

Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15:2009–2053, 2014.

Jonas Peters, Dominik Janzing, and Bernhard Scholkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

Jonas Peters, Stefan Bauer, and Niklas Pfister. Causal models for dynamical systems. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 671–690. 2022.

Maya L. Petersen, Steven G. Deeks, Jeffrey N. Martin, and Mark J. van der Laan. History-adjusted marginal structural models for estimating time-varying effect modification. *American Journal of Epidemiology*, 166(9):985–993, 2007. doi: 10.1093/aje/kwm232.

Mikhail Prokopenko, Joseph T. Lizier, and Don C. Price. On thermodynamic interpretation of transfer entropy. *Entropy*, 15(2):524–543, 2013. doi: 10.3390/e15020524.

Tianchen Qian, Hyesun Yoo, Predrag Klasnja, Daniel Almirall, and Susan A. Murphy. Estimating time-varying causal excursion effects in mobile health with binary outcomes. *Biometrika*, 108(3): 507–527, 2021. doi: 10.1093/biomet/asaa070.

Ashesh Rambachan and Neil Shephard. When do common time series estimands have nonparametric causal meaning? *arXiv preprint arXiv:1903.01637*, 2021. revised 2025.

James M. Robins. Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics — Theory and Methods*, 23(8):2379–2412, 1994. doi: 10.1080/03610929408831393.

James M. Robins, Miguel Á. Hernán, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000. doi: 10.1097/00001648-200009000-00011.

Paul Rolland, Volkan Cevher, Matthäus Kleindessner, Chris Russell, Dominik Janzing, Bernhard Schölkopf, and Francesco Locatello. Score matching enables causal discovery of nonlinear additive noise models. In *International Conference on Machine Learning*, pages 18741–18753. PMLR, 2022.

Carlo Rovelli. How oriented causation is rooted into thermodynamics. *Philosophy of Physics*, 1(1), 2023.

Kjetil Røysland. A martingale approach to continuous-time marginal structural models. *Bernoulli*, 17(3):895–915, 2011. doi: 10.3150/10-BEJ303.

Paul K. Rubenstein, Stephan Bongers, Joris M. Mooij, and Bernhard Schölkopf. From deterministic odes to dynamic structural causal models. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 114–123, 2018.

Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel D. Mahecha, Jordi Muñoz-Marí, Egbert H. van Nes, Jonas Peters, Rick Quax, Markus Reichstein, Marten Scheffer, Bernhard Schölkopf, Peter Spirtes, George Sugihara, Jie Sun, Kun Zhang, and Jakob Zscheischler. Inferring causation from time series in earth system sciences. *Nature Communications*, 10:2553, 2019. doi: 10.1038/s41467-019-10105-3.

Jakob Runge, Andreas Gerhardus, Gherardo Varando, Veronika Eyring, and Gustau Camps-Valls. Causal inference for time series. *Nature Reviews Earth & Environment*, 4:487–505, 2023. doi: 10.1038/s43017-023-00431-y.

Takahiro Sagawa and Masahito Ueda. Nonequilibrium thermodynamics of feedback control. *Physical Review E*, 85(2):021104, 2012a. doi: 10.1103/PhysRevE.85.021104.

Takahiro Sagawa and Masahito Ueda. Information thermodynamics: Maxwell's demon in nonequilibrium dynamics, 2012b. URL `https://arxiv.org/abs/1111.5769`.

Bernhard Schölkopf, David W Hogg, Dun Wang, Daniel Foreman-Mackey, Dominik Janzing, Carl-Johann Simon-Gabriel, and Jonas Peters. Modeling confounding by half-sibling regression. *Proceedings of the National Academy of Sciences*, 113(27):7391–7398, 2016.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Thomas Schreiber. Measuring information transfer. *Physical Review Letters*, 85(2):461–464, 2000. doi: 10.1103/PhysRevLett.85.461.

Udo Seifert. Stochastic thermodynamics, fluctuation theorems and molecular machines. *Reports on Progress in Physics*, 75(12):126001, 2012. doi: 10.1088/0034-4885/75/12/126001.

Ken Sekimoto. Langevin equation and thermodynamics. *Progress of Theoretical Physics Supplement*, (130):17–27, 1998. doi: 10.1143/PTPS.130.17.

Ken Sekimoto. *Stochastic Energetics*, volume 799 of *Lecture Notes in Physics*. Springer, Berlin, Heidelberg, 2010. doi: 10.1007/978-3-642-05411-2.

Alexander Sokol and Niels Richard Hansen. Causal interpretation of stochastic differential equations. *Electronic Journal of Probability*, 19(100):1–24, 2014. doi: 10.1214/EJP.v19-2891.

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

Jie Sun and Erik M. Bollt. Causation entropy identifies indirect influences, dominance of neighbors and anticipatory couplings. *Physica D: Nonlinear Phenomena*, 267:49–57, 2014. doi: 10.1016/j.physd.2013.07.001.

Jie Sun, Carlo Cafaro, and Erik M. Bollt. Identifying the coupling structure in complex systems through the optimal causation entropy principle. *Entropy*, 16(6):3416–3433, 2014. doi: 10.3390/e16063416.

Jie Sun, Dane Taylor, and Erik M. Bollt. Causal network inference by optimal causation entropy. *SIAM Journal on Applied Dynamical Systems*, 14(1):73–106, 2015. doi: 10.1137/140956166.

Jinghao Sun and Forrest W. Crawford. Causal identification for continuous-time stochastic processes. *arXiv preprint arXiv:2211.15934*, 2022.

Sudam Surasinghe and Erik M. Bollt. On geometry of information flow for causal inference. *Entropy*, 22(4):396, 2020. doi: 10.3390/e22040396.

Jin Tian and Judea Pearl. Causal discovery from changes. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 512–521, 2001.

Halbert White and Xun Lu. Granger causality and dynamic structural systems. *Journal of Financial Econometrics*, 8(2):193–243, 2010. doi: 10.1093/jjfinec/nbq006.

Andrew Ying. Causality for functional longitudinal data. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pages 665–687, 2024.

A Foundations of Hamiltonian Causal Models

In this section we expand the context on Definition 2.1. We discuss precisely the role of our assumptions and give context to related assumptions in similar settings in both Physics and Causality.

Definition A.1 (Hamiltonian Causal Model, detailed)**.** Fix a time horizon $T \in \mathbb{R}_+ \cup \{+\infty\}$. A *Hamiltonian Causal Model* is a tuple

$$\mathcal{M} = (\mathcal{P}, z_0, \Gamma, \omega, \mathcal{L}, \mathcal{A}, H)$$

with the following components.

(1) **Probability space and initial condition.** $\mathcal{P} = (\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and $z_0 : \Omega \to \Gamma$ is a random initial state.

(2) **Phase space.** (Γ, ω) is a symplectic manifold with node-wise product factorization

$$\Gamma = \Gamma_X \times \Gamma_E, \qquad \Gamma_X = \prod_{i=1}^{n} \Gamma_{X_i}, \qquad \Gamma_E = \prod_{i=1}^{n} \Gamma_{E_i}.$$

The coordinates X_i are the observed system variables, while E_i is the unobserved local environment, or bath, attached to X_i.

(3) **Intervention space and admissible policies.** The intervention manifold factorizes as

$$\mathcal{L} = \prod_{i=1}^{n} \mathcal{L}_i.$$

We fix an information filtration $(\mathcal{G}_t)_{t\in[0,T]}$, representing the information available to the operator up to time t. The admissible class $\mathcal{A}$ consists of nonanticipative policies

$$\lambda : [0, T] \times \Omega \to \mathcal{L}, \qquad \lambda_t = (\lambda_{1,t}, \ldots, \lambda_{n,t}),$$

such that λ_t is $\mathcal{G}_t$-measurable for every t. Deterministic open-loop protocols are included as the special case in which λ_t is non-random. When a result involves work rates, time-reversal protocols, or infinitesimal probes, we restrict to the relevant subclass of policies whose sample paths are sufficiently regular, typically piecewise C^1 or deterministic C^1.

(4) **Hamiltonian and mechanism decomposition.** The Hamiltonian

$$H : \Gamma \times \mathcal{L} \to \mathbb{R}$$

is smooth and decomposes as

$$H(z, \lambda) = \sum_{i=1}^{n} H_{E_i}(e_i) + \sum_{i=1}^{n} \left[H_i^{\mathrm{self}}(x_i, \lambda_i) + H_i^{\mathrm{mech}}(x_i, x_{-i}, \lambda_i) + H_i^{\mathrm{bath}}(x_i, e_i, \lambda_i) \right]. \quad (4)$$

Here H_i^{self} is the intrinsic term of node i, H_i^{mech} encodes the influence of other observed variables on X_i, and H_i^{bath} encodes the interaction between X_i and its local bath E_i. The intervention coordinate λ_i acts only on the node-i mechanism and its bath coupling.

(5) **Hamiltonian dynamics.** For a policy $\lambda \in \mathcal{A}$, define the time-dependent Hamiltonian

$$H_\lambda(t, z) := H(z, \lambda_t).$$

The controlled Hamiltonian vector field Y_{H_λ} is defined by

$$\omega(Y_{H_\lambda}(t, \cdot), \cdot) = \mathrm{d}H_\lambda(t, \cdot).$$

The corresponding controlled flow is denoted by $\Phi^\lambda_{s,t} : \Gamma \to \Gamma$, and $\Phi^\lambda_t := \Phi^\lambda_{0,t}$. The microscopic path map is

$$\Psi^\lambda(\omega)(t) := \Phi^\lambda_t(z_0(\omega)).$$

The forward observed path law is

$$\mathbb{P}^\lambda_F := (\Pi_X \circ \Psi^\lambda)^\# \mathbb{P},$$

where Π_X projects a full system–bath trajectory onto its observed coordinates.

We recall the definition of mechanism dependence:

Definition A.2 (Mechanism dependence)**.** As in Definition 2.2, define

$$\mathcal{D}_i := \{j \neq i : \partial_{x_j} H_i^{\mathrm{mech}} \not\equiv 0\}.$$

Assumption A.3 (Standing HCM hypotheses)**.** Unless stated otherwise, the following hypotheses are assumed.

(H1) **Initial bath factorization.** The initial environment law factorizes across nodes:

$$\mu^0_E := (\pi_E \circ z_0)^\# \mathbb{P} = \bigotimes_{i=1}^n \mu^0_{E_i}.$$

Thus the baths provide independent exogenous randomness at the initial time.

(H2) **Reservoir initialization and local equilibrium.** Each bath is initialized at inverse temperature $\beta_i > 0$, typically in the Gibbs state

$$\mu^0_{E_i}(\mathrm{d}e_i) = Z_i^{-1} \exp\big(-\beta_i H_{E_i}(e_i)\big)\, \mathrm{d}e_i.$$

We work in the standard reservoir regime in which the bath remains an equilibrium noise source at the thermodynamic level relevant for entropy bookkeeping. Equivalently, heat exchanged with bath E_i satisfies the local detailed-balance convention

$$\Delta s_{E_i} = -\beta_i Q_i, \qquad Q_i := -\big(H_{E_i}(e_i(T)) - H_{E_i}(e_i(0))\big).$$

For finite Hamiltonian baths this is an idealization; it is recovered in weak-coupling, large-reservoir, or thermodynamic-limit regimes.

(H3) **Microscopic reversibility.** There exists an anti-symplectic involution $\Theta : \Gamma \to \Gamma$ such that

$$\Theta^2 = \mathrm{id}, \qquad \Theta^* \omega = -\omega,$$

and Θ does not mix observed and environmental coordinates:

$$\pi_X \circ \Theta = \Theta_X \circ \pi_X, \qquad \pi_E \circ \Theta = \Theta_E \circ \pi_E.$$

For each fixed control value,

$$H(\Theta z, \lambda) = H(z, \lambda).$$

Writing the time-reversed protocol as

$$(\Theta\lambda)(t) := \lambda(T - t),$$

the controlled flows satisfy

$$\Theta \circ \Phi^\lambda_{s,t} = \Phi^{\Theta\lambda}_{T-s,T-t} \circ \Theta.$$

This is the microscopic reversibility condition underlying the forward/backward path-law comparison used to define entropy production.

(H4) **Well-posed controlled evolution.** For every $\lambda \in \mathcal{A}$, the controlled Hamiltonian equation

$$\dot{z}(t) = Y_{H(\cdot,\lambda_t)}(z(t)), \qquad z(0) = z_0,$$

admits a unique global nonanticipative solution on $[0, T]$. The induced microscopic path map is denoted by Ψ^λ. For deterministic open-loop protocols, this solution is generated by a controlled Hamiltonian flow $\Phi^\lambda_{s,t} : \Gamma \to \Gamma$, so that

$$\Psi^\lambda(\omega)(t) = \Phi^\lambda_{0,t}(z_0(\omega)).$$

(H5) **Interventional faithfulness.** The intervention coordinates are faithful to the mechanism graph: for every $i \neq j$,

$$j \in \mathcal{D}_i \quad \Longleftrightarrow \quad \partial_{x_j}\partial_{\lambda_i} H(z, \lambda) \not\equiv 0.$$

Equivalently, every node paired to X_i leaves a nontrivial infinitesimal signature in the response of the node-i intervention channel, and non-paired leave no such signature.

Remark A.4 (No hidden common bath is structural). The decomposition in Definition A.1 contains only local bath couplings of the form

$$H_i^{\mathrm{bath}}(x_i, e_i, \lambda_i).$$

Hence no bath variable E_i directly couples to several observed nodes. This is not an additional assumption but part of the HCM architecture. Causally, it plays the role of excluding latent common causes that enter through a shared environmental degree of freedom.

Remark A.5 (Independent mechanisms are encoded by the decomposition). The node-wise Hamiltonian decomposition

$$H = \sum_i H_{E_i} + \sum_i \left(H_i^{\mathrm{self}} + H_i^{\mathrm{mech}} + H_i^{\mathrm{bath}}\right)$$

is the HCM analogue of an independent-mechanisms or causal Markov factorization. The mechanism of node i consists of its intrinsic term, its coupling to observed nodes, and its coupling to its local bath. Intervening on λ_i modifies this node-i mechanism, but does not modify the symplectic form, the Hamilton equations, or the mechanisms assigned to other nodes.

Remark A.6 (Faithfulness is not implied by the decomposition). Mechanism dependence is defined by whether X_j appears in H_i^{mech}. Interventional faithfulness additionally requires that the available intervention coordinate λ_i probes the part of the node-i mechanism in which that variable appears. Without this condition, a mechnism dependence could be present in the Hamiltonian but invisible to work-rate or local-response tests because the chosen intervention parametrization does not affect the relevant coupling.

Remark A.7. The decomposition of H is part of the causal model specification. As in SCMs, the causal graph is not inferred from an observational object alone; it is encoded by the modular decomposition of the mechanisms and by the associated intervention coordinates.

Causal role of the construction. The HCM decomposition is the causal content of the model. The local bath structure plays the role of node-wise exogenous noise; the absence of common bath terms excludes hidden common environmental causes; and the node-wise Hamiltonian decomposition plays the role of a causal mechanism factorization. The distinction from a classical SCM is that mechanisms are not assignment functions. They are Hamiltonian energy terms which, together with the fixed symplectic law, generate trajectories. Interventions therefore modify Hamiltonian mechanisms, but not the background equations of motion.

Physical role of the assumptions. The standing hypotheses are the physical regularity conditions needed to turn the Hamiltonian construction into well-defined thermodynamics. Well-posedness ensures that every admissible intervention induces a path law. Reservoir initialization and local detailed balance identify bath energy changes with entropy exchange. Microscopic reversibility supplies the backward experiment against which the forward path law is compared. Together, these are the standard ingredients behind entropy production as a path-likelihood ratio and the fluctuation-theorem formulation of the second law [Jarzynski, 1997, Crooks, 1999, Seifert, 2012].

Remark A.8 (Canonical coordinates). By Darboux's theorem, every point of a $2m$-dimensional symplectic manifold has local coordinates

$$(q_1, \dots, q_m, p_1, \dots, p_m)$$

such that

$$\omega = \sum_{k=1}^{m} dq_k \wedge dp_k.$$

In these coordinates the Hamiltonian vector-field equation $\omega(Y_{H_\lambda}, \cdot) = dH_\lambda$ becomes the familiar Hamilton equations

$$\dot{q}_k = \frac{\partial H_\lambda}{\partial p_k}, \qquad \dot{p}_k = -\frac{\partial H_\lambda}{\partial q_k}.$$

Thus the abstract symplectic formulation used in Definition A.1 reduces locally to standard Hamiltonian mechanics.

Remark A.9 (Reservoir stationarity and system–bath correlations). The reservoir condition should be understood at the thermodynamic scale at which the bath acts as an equilibrium noise source. It does not require the joint law of $(X_i(t), E_i(t))$ to factorize for $t > 0$. In fact, system–bath coupling generally creates correlations along trajectories, and these correlations are precisely what mediate friction, noise, and heat exchange. What is assumed is that the bath is large enough, or weakly enough coupled, that its marginal thermodynamic state remains effectively fixed by β_i over the time horizon of interest.

Remark A.10 (Adaptive interventions and nonanticipativity). The measurability condition on λ_t is the continuous-time analogue of requiring interventions to depend only on information available before they are applied. It allows both open-loop protocols and feedback interventions, such as controllers whose action depends on the observed history $X_{[0,t]}$. The additional smoothness assumptions imposed in later results are theorem-specific: they are needed only when differentiating the protocol, defining an instantaneous work rate, or constructing infinitesimal ramp interventions. Thus adaptive interventions belong to the HCM language, while particular thermodynamic identities may be stated for the smooth subclass where the calculus is well-defined.

B Proofs

B.1 Proof of Proposition 3.3

Assumption B.1 (Local work-rate faithfulness)**.** Fix a node i, a probing time t^*, and write

$$\lambda_* := \lambda(t^*), \qquad u_i := \dot{\lambda}_i(t^*) \in T_{\lambda_{*,i}}\mathcal{L}_i.$$

We say that the i-probe is locally work-rate faithful at (t^*, λ_*) if, for every $j \neq i$,

$$j \in \mathcal{D}_i \quad \Longleftrightarrow \quad \left\langle \partial_{x_j}\partial_{\lambda_i} H(\cdot, \lambda_*), u_i \right\rangle \not\equiv 0.$$

Here $\not\equiv 0$ means nonzero as a function of the phase-space variables. When $\mathcal{L}_i$ is one-dimensional and $u_i \neq 0$, this reduces to

$$j \in \mathcal{D}_i \quad \Longleftrightarrow \quad \partial_{x_j}\partial_{\lambda_i} H(\cdot, \lambda_*) \not\equiv 0.$$

Proof. At the probing time t^*, only the i-th intervention coordinate has nonzero velocity. Therefore

$$\dot{W}(t^*, z; \lambda) = \left\langle \partial_{\lambda_i} H(z, \lambda_*), \dot{\lambda}_i(t^*) \right\rangle.$$

Consequently, for $j \neq i$,

$$\left[\nabla_x \dot{W}(t^*, z; \lambda)\right]_j = \left\langle \partial_{x_j}\partial_{\lambda_i} H(z, \lambda_*), \dot{\lambda}_i(t^*) \right\rangle.$$

Thus there exists z^* such that

$$\left[\nabla_x \dot{W}(t^*, z^*; \lambda)\right]_j \neq 0$$

if and only if

$$\left\langle \partial_{x_j}\partial_{\lambda_i} H(\cdot, \lambda_*), \dot{\lambda}_i(t^*) \right\rangle \not\equiv 0.$$

By local work-rate faithfulness, Assumption B.1, this is equivalent to $j \in \mathcal{D}_i$). □

B.2 Proof of Proposition 3.7

Proof. Assume, for contradiction, that X_j has no causal effect on X_i under the two policies. Then

$$\mathbb{P}_i^{\lambda^{(1)}} = \mathbb{P}_i^{\lambda^{(2)}}.$$

Since $\mathcal{R}_i$ is fixed, this implies

$$(\mathcal{R}_i)_{\#}\mathbb{P}_i^{\lambda^{(1)}} = (\mathcal{R}_i)_{\#}\mathbb{P}_i^{\lambda^{(2)}}.$$

Therefore

$$D_{\mathrm{KL}}\left(\mathbb{P}_i^{\lambda^{(1)}} \,\middle\|\, (\mathcal{R}_i)_{\#}\mathbb{P}_i^{\lambda^{(1)}}\right) = D_{\mathrm{KL}}\left(\mathbb{P}_i^{\lambda^{(2)}} \,\middle\|\, (\mathcal{R}_i)_{\#}\mathbb{P}_i^{\lambda^{(2)}}\right),$$

that is,

$$\mathbb{E}_{\mathbb{P}_i^{\lambda^{(1)}}}\left[\Sigma_i^{\mathrm{proj},\lambda^{(1)}}\right] = \mathbb{E}_{\mathbb{P}_i^{\lambda^{(2)}}}\left[\Sigma_i^{\mathrm{proj},\lambda^{(2)}}\right].$$

This contradicts the assumed strict inequality. Hence

$$\mathbb{P}_i^{\lambda^{(1)}} \neq \mathbb{P}_i^{\lambda^{(2)}},$$

so X_j has a causal effect on X_i. □

B.3 Proof of Theorem 3.9

We state the regularity and faithfulness assumptions used by the local discovery result. The first two assumptions are analytic regularity conditions ensuring that the channel entropy-production rate is differentiable at the intervention point. The last two assumptions are the HCM faithfulness and non-degeneracy requirements specialized to the Langevin setting.

Assumption B.2 (Regularity and faithfulness for local discovery)**.** Fix $t^* \in (0, T)$, and write

$$p_* := p(\cdot, t^*), \qquad \lambda_* := \lambda(t^*).$$

We assume the following conditions hold.

(D1) **Regular overdamped Langevin dynamics.** The observed process is

$$dX_t = -\nabla_x U(X_t, \lambda(t))\, dt + \sqrt{2D(\lambda(t))}\, dW_t, \qquad D(\lambda) = \mathrm{diag}(D_1(\lambda_1), \ldots, D_n(\lambda_n)),$$

with $U \in C^3(\mathbb{R}^n \times \Lambda)$. Each $D_i \in C^2(\Lambda_i)$ is bounded above and away from zero on the protocol range.

(D2) **Density and current regularity.** The law of X_{t^*} has a strictly positive density p_*. For the right-local continuations used below, the densities $p^v(\cdot, s)$ solve the Fokker–Planck equation classically near $s = 0^+$, with probability current

$$J_i[p, \lambda](x) = -\partial_{x_i} U(x, \lambda) p(x) - D_i(\lambda_i)\partial_{x_i} p(x).$$

The channel entropy-production functional

$$F_i[p, \lambda] := \int_{\mathbb{R}^n} \frac{J_i[p, \lambda](x)^2}{D_i(\lambda_i) p(x)}\, dx$$

is finite and differentiable near (p_*, λ_*), and differentiation under the integral sign is valid for frozen-law derivatives in λ.

(D3) **Right-local probing intervention.** For each probe node j, the admissible policy class contains two right-local continuations initialized at (p_*, λ_*):

$$\lambda^0(s) = \lambda_*, \qquad \lambda^v(s) = \lambda_* + vse_j + o(s) \quad \text{as } s \downarrow 0,$$

with $p^0(\cdot, 0) = p^v(\cdot, 0) = p_*$. The two continuations are compared as ensemble laws; no pathwise coupling is assumed.

(D4) **Local form of HCM faithfulness.** The mechanistic dependence induced by the HCM decomposition is locally faithful to the Langevin potential at λ_*: for every $i \neq j$,

$$i \in \mathcal{D}_j \quad \Longleftrightarrow \quad \partial_{x_i}\partial_{\lambda_j} U(\cdot, \lambda_*) \not\equiv 0 \quad \text{in } L^2(dx).$$

Equivalently, non-paired have no mixed structural response to the j-intervention channel, while every true parent has a nontrivial mixed response.

(D5) **Nonequilibrium non-degeneracy.** For every $i \in \mathcal{D}_j$ with $i \neq j$,

$$J_i[p_*, \lambda_*] \in L^2(dx), \qquad \partial_{x_i}\partial_{\lambda_j} U(\cdot, \lambda_*) \in L^2(dx),$$

and

$$\big\langle J_i[p_*, \lambda_*], \partial_{x_i}\partial_{\lambda_j} U(\cdot, \lambda_*)\big\rangle_{L^2(dx)} \neq 0.$$

Proof. Throughout the proof, all spatial integrals are over $\mathbb{R}^n$. We use the local entropy-production rate identity from Lemma B.4, namely

$$\dot{\Sigma}_i^{\mathrm{loc}}[p, \lambda] = F_i[p, \lambda] := \int \frac{J_i[p, \lambda](x)^2}{D_i(\lambda_i)p(x)}\, dx,$$

where

$$J_i[p, \lambda](x) = -\partial_{x_i} U(x, \lambda)p(x) - D_i(\lambda_i)\partial_{x_i} p(x).$$

Recall that the local response is the right-local contrast

$$R_{i|j}(t^*) := \lim_{v\to 0} \lim_{s\downarrow 0} \frac{F_i[p^v(\cdot, s), \lambda^v(s)] - F_i[p^0(\cdot, s), \lambda^0(s)]}{vs},$$

whenever the iterated limit exists. Fix $i \neq j$. The baseline and probe continuations start from the same ensemble state and protocol value:

$$p^0(\cdot, 0) = p^v(\cdot, 0) = p_*, \qquad \lambda^0(0) = \lambda^v(0) = \lambda_*.$$

Moreover,

$$\lambda^0(s) = \lambda_*, \qquad \lambda^v(s) = \lambda_* + vse_j + o(s).$$

At $s = 0$, the Fokker–Planck velocity of the density depends on the protocol level λ_*, but not on the outgoing protocol velocity. Hence

$$\partial_s p^v(\cdot, s)\big|_{s=0^+} = \partial_s p^0(\cdot, s)\big|_{s=0^+}.$$

By differentiability of F_i, the density-variation terms therefore cancel in the first-order contrast defining $R_{i|j}(t^*)$. The only surviving term is the frozen-law derivative with respect to λ_j:

$$R_{i|j}(t^*) = \partial^{\mathrm{fr}}_{\lambda_j} F_i[p_*, \lambda_*],$$

where $\partial^{\mathrm{fr}}_{\lambda_j}$ denotes differentiation in λ_j while keeping p fixed.

It remains to compute this derivative. Since $i \neq j$, the diagonal diffusion coefficient $D_i(\lambda_i)$ is independent of λ_j. Thus

$$\partial^{\mathrm{fr}}_{\lambda_j} F_i[p, \lambda] = \int \frac{2J_i[p, \lambda](x)}{D_i(\lambda_i)p(x)}\, \partial^{\mathrm{fr}}_{\lambda_j} J_i[p, \lambda](x)\, dx.$$

Holding p fixed,

$$\partial^{\mathrm{fr}}_{\lambda_j} J_i[p, \lambda](x) = -\partial_{x_i}\partial_{\lambda_j} U(x, \lambda)\, p(x),$$

and therefore

$$\partial^{\mathrm{fr}}_{\lambda_j} F_i[p, \lambda] = -\frac{2}{D_i(\lambda_i)} \int J_i[p, \lambda](x)\partial_{x_i}\partial_{\lambda_j} U(x, \lambda)\, dx.$$

Evaluating at (p_*, λ_*) gives

$$R_{i|j}(t^*) = -\frac{2}{D_i(\lambda_{*,i})} \int J_i[p_*, \lambda_*](x)\partial_{x_i}\partial_{\lambda_j} U(x, \lambda_*)\, dx.$$

If $i \notin \mathcal{D}_j$, local HCM faithfulness gives

$$\partial_{x_i}\partial_{\lambda_j} U(\cdot, \lambda_*) \equiv 0,$$

and hence $R_{i|j}(t^*) = 0$. Conversely, if $i \in \mathcal{D}_j$, then local HCM faithfulness gives

$$\partial_{x_i}\partial_{\lambda_j} U(\cdot, \lambda_*) \not\equiv 0,$$

and the nonequilibrium non-degeneracy assumption makes the displayed inner product nonzero. Since $D_i(\lambda_{*,i}) > 0$, this implies $R_{i|j}(t^*) \neq 0$. Therefore

$$i \in \mathcal{D}_j \quad \Longleftrightarrow \quad R_{i|j}(t^*) \neq 0.$$

□

Remark B.3 (Genericity of response non-degeneracy). For fixed nonzero

$$g_{ij} := \partial_{x_i}\partial_{\lambda_j} U(\cdot, \lambda_*),$$

the set

$$\left\{ J_i \in L^2(dx) : \langle J_i, g_{ij} \rangle_{L^2(dx)} = 0 \right\}$$

is a closed codimension-one hyperplane in $L^2(dx)$. Thus Assumption B.2(D5) is a generic nonequilibrium condition: it fails only when the current is orthogonal to the structural mixed derivative, for instance when $J_i(\cdot, t^*) \equiv 0$.

B.4 Proof of Theorem 2.3

The realization result used in Theorem 2.3 is a standard consequence of the oscillator-bath derivation of generalized Langevin equations and of the Markovian limits. We recall the general construction only to make explicit that it is compatible with the HCM factorization. For an example of proof along with a contextualization of the terminology used here, we refer the reader to [Sekimoto, 2010].

Consider the Hamiltonian

$$H^{N,m}(z, \lambda) = \sum_i \frac{p_i^2}{2m} + U(x, \lambda) + \sum_i \left[H^N_{E_i}(e_i) + H^N_{X_i E_i}(x_i, e_i, \lambda_i) \right],$$

where each E_i is an independent oscillator bath initialized at a Gibbs law. For Ohmic spectral densities, the marginal dynamics of x converges, in the Markovian bath limit and the overdamped limit, to the controlled Langevin diffusion

$$dX_t = -\nabla_x U(X_t, \lambda_t)\, dt + \sigma(\lambda_t)\, dW_t,$$

after choosing unit mobility and bath temperatures satisfying $D_i(\lambda_i) = \sigma_i^2(\lambda_i)/2$. This is the classical Caldeira-Leggett construction of Langevin dynamics from Hamiltonian system–bath dynamics.

Proof sketch of Theorem 2.3. The construction above satisfies the HCM axioms: the initial bath law factorizes across nodes, each bath couples only to its corresponding observed coordinate, and the intervention coordinate λ_i enters only the i-th local Hamiltonian terms. Standard oscillator-bath theory gives a generalized Langevin equation with memory kernel determined by the bath spectral density. In the Ohmic limit the memory kernel converges to a delta kernel and the random bath force converges to white noise satisfying the fluctuation–dissipation relation. The subsequent limit [Caldeira and Leggett, 1983] yields the overdamped Itô diffusion in Theorem 2.3. □

B.5 Entropy Production in Langevin Dynamics

The identity below is standard in stochastic thermodynamics; see, e.g., Seifert [2012], Horowitz and Esposito [2014], Otsubo et al. [2020]. We include the short proof only for completeness and to make explicit the quantity used in Theorem 3.9 and in the experiments.

Lemma B.4 (Channel-local entropy-production rate)**.** *Consider the controlled overdamped Langevin diffusion*

$$dX_t = b(X_t, \lambda_t)\, dt + \sqrt{2D(\lambda_t)}\, dW_t, \qquad b_i(x, \lambda) = -\partial_{x_i} U(x, \lambda),$$

with diagonal diffusion

$$D(\lambda) = \mathrm{diag}(D_1(\lambda_1), \ldots, D_n(\lambda_n)).$$

Assume the joint law of X_t has a smooth strictly positive density $p^\lambda(x, t)$ with sufficient decay at infinity. Let

$$J_i^\lambda(x, t) := b_i(x, \lambda_t) p^\lambda(x, t) - D_i(\lambda_i(t))\, \partial_{x_i} p^\lambda(x, t)$$

be the i-th probability current. Then the channel-local entropy-production rate of node i is

$$\dot\Sigma_i^{\mathrm{loc}}[p^\lambda(\cdot, t), \lambda_t] = \int_{\mathbb{R}^n} \frac{\left(J_i^\lambda(x, t)\right)^2}{D_i(\lambda_i(t))\, p^\lambda(x, t)}\, dx.$$

Proof. We suppress the superscript λ and the time argument. The Fokker–Planck equation is

$$\partial_t p(x,t) = -\sum_{k=1}^{n} \partial_{x_k} J_k(x,t).$$

The contribution of the i-th current to the joint system-entropy rate is

$$\dot{s}^{(i)}_{\mathrm{sys}}(t) = -\int_{\mathbb{R}^n} J_i(x,t)\, \partial_{x_i} \log p(x,t)\, dx.$$

With the sign convention that Q_i denotes heat transferred from bath E_i into the observed system, local detailed balance gives the corresponding medium-entropy contribution in the form

$$-\beta_i \dot{Q}_i(t) = \int_{\mathbb{R}^n} \frac{b_i(x,\lambda_t)}{D_i(\lambda_i(t))} J_i(x,t)\, dx.$$

This is the standard overdamped thermodynamic-force expression; equivalently, one may view D_i as the mobility-temperature product in units where the mobility has been absorbed into the drift. By Definition 3.8, the information-flow correction cancels the difference between the marginal entropy rate and the i-th contribution to the joint system-entropy rate, hence

$$\dot{\Sigma}^{\mathrm{loc}}_i = \dot{s}^{(i)}_{\mathrm{sys}} - \beta_i \dot{Q}_i.$$

Substituting the two displayed expressions gives

$$\dot{\Sigma}^{\mathrm{loc}}_i = \int_{\mathbb{R}^n} J_i(x,t) \left[\frac{b_i(x,\lambda_t)}{D_i(\lambda_i(t))} - \partial_{x_i} \log p(x,t) \right] dx.$$

Since

$$\frac{b_i}{D_i} - \partial_{x_i} \log p = \frac{b_i p - D_i \partial_{x_i} p}{D_i p} = \frac{J_i}{D_i p},$$

we obtain

$$\dot{\Sigma}^{\mathrm{loc}}_i = \int_{\mathbb{R}^n} \frac{J_i(x,t)^2}{D_i(\lambda_i(t)) p(x,t)}\, dx,$$

as claimed. □

Remark B.5 (Projected entropy-production estimation). For the projected entropy-production experiments, we feed only the partial trajectory $x_i(\cdot)$ to the variational entropy-production estimator of Otsubo et al. [2020, 2022]. This changes the target of estimation relative to the full-state channel-local quantity in Lemma B.4: the estimator can only use currents and test functions measurable with respect to the observed marginal trajectory of X_i. Consequently, it estimates the time-reversal asymmetry visible after projecting the dynamics onto X_i, or equivalently the projection of the thermodynamic force onto the restricted function class available from $x_i(\cdot)$. This is precisely the object needed for the path-level causal effect witness in Proposition 3.7, since that proposition is formulated in terms of the marginal path law $\mathbb{P}^{\lambda}_i$. The statistical and thermodynamic consequences of estimating entropy production from partial observations, including the resulting coarse-grained or projected interpretation, are discussed in the original estimation work of Otsubo et al. [2020, 2022].

B.6 Separation of Measures

Proposition B.6 (Causal effect and separating path-law discrepancies)**.** *Fix $i \neq j$, a baseline policy λ_{-j}, and a discrepancy*

$$\mathfrak{D} : \mathcal{P}(C([0,T], \Gamma_{X_i})) \times \mathcal{P}(C([0,T], \Gamma_{X_i})) \to \mathbb{R}.$$

Assume that $\mathfrak{D}$ separates probability measures, in the sense that

$$\mathfrak{D}(\mu,\nu) = 0 \quad \Longleftrightarrow \quad \mu = \nu.$$

Then X_j has a causal effect on X_i over $[0,T]$ in the sense of Definition 2.4 if and only if there exist two admissible policies

$$\lambda^{(1)} = (\lambda^{(1)}_j, \lambda_{-j}), \qquad \lambda^{(2)} = (\lambda^{(2)}_j, \lambda_{-j}),$$

such that

$$\mathrm{ATE}^{(i)}_{0:T}\big(\mathfrak{D}; \lambda^{(1)}, \lambda^{(2)}\big) \neq 0.$$

Proof. By Definition 2.4, X_j has a causal effect on X_i over $[0, T]$ if and only if there exist two admissible policies

$$\lambda^{(1)} = (\lambda_j^{(1)}, \lambda_{-j}), \qquad \lambda^{(2)} = (\lambda_j^{(2)}, \lambda_{-j}),$$

such that the corresponding marginal path laws of X_i are different:

$$\mathbb{P}_i^{\lambda^{(1)}} \neq \mathbb{P}_i^{\lambda^{(2)}}.$$

Since $\mathfrak{D}$ separates measures, this is equivalent to

$$\mathfrak{D}\left(\mathbb{P}_i^{\lambda^{(1)}}, \mathbb{P}_i^{\lambda^{(2)}}\right) \neq 0.$$

By Definition 2.6, the left-hand side is exactly

$$\mathrm{ATE}_{0:T}^{(i)}\big(\mathfrak{D}; \lambda^{(1)}, \lambda^{(2)}\big).$$

Hence X_j has a causal effect on X_i if and only if the generalized trajectory ATE is nonzero for some pair of admissible policies. □

C Experimental Details

C.1 Comparison of Entropy Production and other ATEs

We test the central claim that the entropy production difference:

$$ATE(\Sigma_i^{\mathrm{proj}}) := \mathbb{E}_{\mathbb{P}_i^{\lambda^{(1)}}}\left[\Sigma_i^{\mathrm{proj},\lambda^{(1)}}\right] - \mathbb{E}_{\mathbb{P}_i^{\lambda^{(2)}}}\left[\Sigma_i^{\mathrm{proj},\lambda^{(2)}}\right]$$

is a causal estimand that detects effects which are *invisible* to the two standard observable-level estimands: the endpoint average treatment effect $ATE = \mathbb{E}[Y(T) \mid \mathrm{do}(1)] - \mathbb{E}[Y(T) \mid \mathrm{do}(2)]$ and the cumulative ATE $cATE = \int_0^T \mathbb{E}[Y(t) \mid \mathrm{do}(1)] - \mathbb{E}[Y(t) \mid \mathrm{do}(2)]\, dt$. The two systems below are *designed* so that one of these standard estimands is provably zero under the null (same-endpoint and time-translation symmetry, respectively), while the underlying dynamics are demonstrably non-equivalent in a global sense.

C.1.1 Experiment A: Same-Endpoint Protocols on a 10-Node DAG

Data-generating process Ten coupled overdamped Langevin oscillators on the linear chain $0 \to 1 \to \cdots \to 9$

$$dq_i = \Big[-k\big(q_i - a_i(t)\big) + \kappa \sum_{j \in \mathrm{adj}(i)} (q_j - q_i)\Big]dt + \sigma\, dW_i,$$

with stiffness $k = 5$, nearest-neighbour coupling $\kappa = 1.5$, and noise amplitude $\sigma = 1$. The relaxation time of the driven node is $\tau_0 = 1/(k + \kappa) \approx 0.15$, much smaller than the protocol horizon $T = 10$, so all nodes are in the quasi-static regime with respect to the control. Initial conditions are drawn $q(0) \sim \mathcal{N}(\mathbf{0}, \sigma_q^2 I)$ with $\sigma_q = 0.3$. We integrate $N = 1000$ trajectories with Euler–Maruyama at $\Delta t = 0.01$ ($n_{\mathrm{steps}} = 1000$).

Two same-endpoint protocols. Both policies act on node 0 only, with $a_0(T) = a_{\mathrm{end}} = 1$:

$$\lambda_1 \text{ (ramp)}\colon\ a_0(t) = t/T, \qquad \lambda_2 \text{ (excursion)}\colon\ a_0(t) = \begin{cases} 4 \cdot 2t/T & t \le T/2, \\ 4 + (1-4) \cdot \frac{2(t - T/2)}{T} & t > T/2. \end{cases}$$

Because $T \gg \tau_0$ and $a_0(T)$ matches across protocols, the endpoint distribution μ_T is identical in the limit $T/\tau_0 \to \infty$ for every node, hence the endpoint ATE on the observable $Y(t) = q_0(t)$ vanishes by construction.

Estimation of Entropy Production We use NEEP [Kim et al., 2020, Otsubo et al., 2020]. Architecture: $L = 3$ fully-connected hidden layers of width $H = 64$; optimiser Adam, learning rate 10^{-3}, $n_{\mathrm{step}} = 1000$ gradient steps; trained on the single-coordinate marginal $\{q_0(t)\}$ (the directly forced node). We run $K = 3$ independent seeds and report mean $\pm$ standard deviation.

C.1.2 Experiment B: Protocol-Order Effect on a 3-Node Hidden-Mediator System

Data-generating process Three overdamped Langevin nodes $q = [q_1, q_2, s]$ with a hidden nonlinear mediator s, defined by:

$$\begin{aligned} dq_1 &= \big[-k(q_1 - \lambda_a(t)) - \gamma_c\, s\, q_2 - \delta_c\, s\big]\, dt + \sigma\, dW_1, \\ dq_2 &= \big[-k(q_2 - \lambda_b(t)) - \gamma_c\, s\, q_1\big]\, dt + \sigma\, dW_2, \\ ds &= \big[-k_s\, s - \gamma_c\, q_1 q_2 - \delta_c\, q_1\big]\, dt + \sigma_s\, dW_s. \end{aligned}$$

Parameters: $k = k_s = 4$ (so $\tau_q = \tau_s = 1/4 = 0.25$), symmetric coupling $\gamma_c = 0.05$ (small, so non-cancelling nonlinear terms are negligible), asymmetric coupling $\delta_c = 1.0$ (the primary driver of entropy production asymmetry, which scales as δ_c^2), $\sigma = \sigma_s = 1$. Initial conditions $q_1(0), q_2(0), s(0) \sim \mathcal{N}(0, 0.3^2)$. We integrate $N = 2000$ trajectories at $\Delta t = 0.01$, $T = 8$ ($n_{\text{steps}} = 800$).

Two protocol orderings. Two raised-cosine pulses of amplitude $A = 3$, each of duration $T/2$, applied in opposite order:

$$AB\colon\ \lambda_a \text{ on } [0, T/2],\ \lambda_b \text{ on } [T/2, T]; \qquad BA\colon\ \lambda_b \text{ on } [0, T/2],\ \lambda_a \text{ on } [T/2, T].$$

The total control effort and the marginal time-budget per channel are identical; only the order differs. For the observable $Y = q_1 + q_2 + s$ a time-translation argument gives $cATE(Y) = \int_0^T \mathbb{E}[Y_t \mid AB] - \mathbb{E}[Y_t \mid BA]\, dt = 0$ up to $O(\gamma_c)$ corrections.

Estimator NEEP as in Experiment A, but trained on the full 3-D state $q = (q_1, q_2, s)$ to capture the mediator-driven asymmetry. Architecture and optimiser as before; $K = 5$ independent seeds.

C.1.3 Hyperparameters and Reproducibility

We provide detailed hyperparameters for reproducibility in the attached code.

Compute Each experiment fits in under 10 minutes wall-clock on a single GPU (mostly in NEEP training).

Limitations and scope Both designs are deliberately small physical systems chosen to make the analytic ATE-cancellation argument transparent; we leverage the quasi-static regime ($T \gg \tau$) in Exp. A and the time-translation symmetry of the integrated observable in Exp. B. The point of these experiments is *not* scale, but to demonstrate that entropy production resolves causal asymmetries that any observable-level estimand built on Y will miss *by construction*. This can be taken as a lesson in selecting the correct quantities to estimate in temporal notions of causality: while marginal and cumulative notions of the ATE might be correct for certain applications, they are not fully descriptive of path-level constructions, where instead entropy production can help.

C.2 Mechanism Identification

We evaluate our method for mechanism identification on *random cyclic* interaction graphs under three families of λ-gated coupling potentials of increasing nonlinearity: a **bilinear** reference model, a saturating **tanh-sigmoid** model, and a quartic-confined **cubic** model. All three families share the same two-intervention response protocol, the same neural-network EP estimator, and the same fixed thresholding rule (§C.2.3).

C.2.1 Data-Generating Processes

Random graphs (cyclic) For each system family we draw $G = 10$ independent Erdős–Rényi *directed* graphs on $n = 15$ nodes with edge probability $p_{\text{edge}} = 0.10$, self-loops removed. In the realized draws, each graph has between 20 and 28 directed edges and typically contains at least one directed cycle (>90%). For each graph we pick a target node j with $|\mathcal{D}_j| \geq 2$ if possible (else ≥ 1) by uniform sampling among candidate nodes with the required set size, using a deterministic

seed schedule (base seed 20240501, increment 7919 per graph) so that the same underlying graph topology and target j are used across the three system families. All three families share the structure

$$dX_i \;=\; -\nabla_x U\, dt \;+\; \sigma_0\, dW_i, \qquad i = 1, \ldots, n, \tag{5}$$

with $\sigma_0 = 0.5$ and self-stiffness coefficients $a_i \sim \mathrm{Unif}(a_{\mathrm{lo}}, a_{\mathrm{hi}})$. The drift $-\nabla_x U$ comes from a λ-forced potential

$$U(\mathbf{x}, \lambda) \;=\; \tfrac{1}{2}\sum_i a_i x_i^2 \;+\; U_{\mathrm{self}}(\mathbf{x}) \;+\; \sum_j \lambda_j \left[b_j x_j + \sum_{i \in \mathcal{D}_j} c_{ij}\, \phi(x_i, x_j) \right], \tag{6}$$

with own-coupling $b_j \sim \mathrm{Unif}(0.5, 1.5) \cdot \{\pm 1\}$ and edge weights $c_{ij} = \frac{1}{2} \cdot \mathrm{Unif}(1.5, 3.5) \cdot \{\pm 1\}$ on graph edges (zero off-edges). The three families differ in the coupling kernel ϕ and the self-confinement U_{self}:

- **Bilinear** (`lambda_coupled`): $U_{\mathrm{self}} = 0$, $\phi(x_i, x_j) = x_i x_j$. Drift remains linear in $\mathbf{x}$ at any fixed λ; marginals are Gaussian.
- **Sigmoid** (`lambda_sigmoid`, $\beta = 2.5$): $U_{\mathrm{self}} = 0$, $\phi(x_i, x_j) = \tanh(\beta x_i)\, x_j$. The input saturates, yielding a strongly nonlinear response of the child to large parent excursions.
- **Cubic** (`lambda_cubic`, $\gamma = 0.5$): $U_{\mathrm{self}} = \frac{\gamma}{4}\sum_i x_i^4$, $\phi(x_i, x_j) = x_i x_j$. The quartic self-confinement makes the marginal stationary distribution distinctly non-Gaussian.

Two-intervention response protocol We estimate the response matrix $R_{i|j}$ via the corrected two-intervention design as implied by 3.9. Starting from a common burn-in snapshot of the null system ($T_{\mathrm{burn}} = 4.0$, $\Delta t_{\mathrm{burn}} = 0.02$, N trajectories), we simulate two ensembles on a short window of length $s_{\mathrm{window}} = 0.4$ (with $n_{\mathrm{steps}} = 40$ Euler-Maruyama substeps, $\Delta t = 0.01$) under

$$\lambda^0(s) = \lambda^\star, \qquad \text{(frozen)} \tag{7}$$

$$\lambda^v(s) = \lambda^\star + v\, s\, \mathbf{e}_j, \qquad \text{(probe)} \tag{8}$$

with response velocity $v = 0.5$ and *independent* Brownian noise streams. Under this single-coordinate intervention the analytic ground-truth response $R^{\mathrm{true}}_{i|j}$ of every non-paired node vanishes *exactly*, giving a clean binary classification problem. For each candidate $i \neq j$ we fit a single channel-EP estimator on the concatenated frozen+probe trajectories and read

$$\hat{R}_{i|j} \;=\; \mathrm{slope}_s\big[\hat{\sigma}_i^v(s) - \hat{\sigma}_i^0(s)\big] \,/\, v, \tag{9}$$

where the slope is a least-squares fit through the origin over the n_{steps} midpoints of the short window.

C.2.2 NEEP Training

The per-coordinate, time-resolved EP rate $\hat{\sigma}_i(t)$ is obtained with the non-stationary local NEEP estimator [Otsubo et al., 2020] implemented as a feed-forward network with a temporal kernel basis (`FNNKt`) using the public `LearnEntropy` library: the input is the joint state $x \in \mathbb{R}^n$, the output is the time-resolved scalar EP rate of coordinate i. We use the same architecture and schedule for every node:

- Architecture: $L = 2$ hidden layers, width $H = 64$, $n_{\mathrm{output}} = 20$ temporal kernel centres [Kim et al., 2020].
- Optimiser: Adam, learning rate 5×10^{-4}, weight decay 10^{-3}, 2000 gradient-ascent steps.
- Train/validation split: $70/30$ along the trajectory axis; best-on-validation checkpoint selected.

Each graph took 155–294 s on a single GPU. The full sweep (3 systems $\times$ 10 graphs) fits in roughly one GPU-hour.

C.2.3 Metrics

For each graph we report:

- **Headline F1, precision, recall.** A node $i \neq j$ is predicted as paired to j iff

$$|\hat{R}_{i|j}| \;>\; \tau \max_{i' \neq j} |\hat{R}_{i'|j}|, \qquad \tau = 0.10. \tag{10}$$

  Precision, recall, and F1 are computed against $\mathcal{D}_j$.
- **Localized Structural Hamming Distance.** We define the localized SHD at the target j as the symmetric difference between the predicted set and the true dependent set,

$$\mathrm{SHD}_j \;:=\; \left|\widehat{\mathcal{D}}_j \triangle \mathcal{D}_j\right| \;=\; \mathrm{FP}_j + \mathrm{FN}_j, \tag{11}$$

  scored with the same threshold rule (10). This is the SHD of the j-th in-edge column of the adjacency matrix and is the finest-grain SHD compatible with our recovery setup.

We aggregate across the $G = 10$ graphs of each system family by reporting the mean and a 95% percentile bootstrap interval ($B = 20{,}000$ resamples). The bootstrap samples are over graph indices, so the intervals reflect graph-to-graph variability at fixed estimator hyperparameters.

D Relationship to Dynamic Structural Causal Models

As mathematical objects, dynamic structural causal models (DSCMs) describe assignments between entire trajectories and can represent dynamics beyond the gradient-drift Langevin class studied here, including SDEs with non-gradient drifts [Boeken and Mooij, 2024, 2026]. Such generality requires conditions ensuring that observational and interventional distributions are well-defined, a non-trivial issue already for cyclic SCMs without an explicit temporal interpretation [Bongers et al., 2021]. DSCMs do come equipped with interventional semantics, and their graphical results should therefore not be characterized as merely observational. Our emphasis is instead on the additional physical structure and its statistical implications. The recent work of Boeken and Mooij [2026] provides graphical Markov properties and do-calculus for systems of causal SDEs, including cyclic and non-Markovian systems. Its constraint-based discovery illustrations rely on a conditional-independence oracle, while statistical aspects of causal effect estimation are explicitly identified as future work. HCMs provide a complementary contribution in the Hamiltonian setting: the connection to thermodynamics supplies physically meaningful, path-level quantities whose contrasts can witness causal effects under the conditions stated in this paper, together with trajectory-based estimation procedures in the Langevin settings we investigate. The ease of defining interventional quantities is also a central advantage of the HCM formulation: interacting, time-dependent controls enter directly through the Hamiltonian, and the admissible class of policies specifies the available interventions. The framework permits cyclic interactions, non-stationary evolution, and memory in the observed subsystem induced by the baths. These features are not exclusive to HCMs: DSCMs can also represent parameter interventions by including the parameters as endogenous processes [Boeken and Mooij, 2026]. HCMs instead make the physical implementation of controls and their thermodynamic consequences explicit within the model. Thus, the contribution is additional physical and statistical structure for the dynamics covered by our assumptions, rather than greater representational generality than generic DSCMs.

E Autonomous System–Demon Coupling

An HCM node comprises an observed subsystem X_i, its associated bath E_i, and its control λ_i. The spaces Γ_X, Γ_E, and $\mathcal{L}$ have corresponding factorizations, with the Hamiltonian coupling these components. To illustrate this structure in the setting of autonomous Maxwell demons [Ito and Sagawa, 2015], consider two nodes, a system S and a demon D. At the microscopic level, take $X_i = (x_i, p_i) \in \mathbb{R}^2$, with a Caldeira–Leggett bath $E_i \in \mathbb{R}^{2N_i}$ prepared at inverse temperature $\beta_i > 0$, as in the realization construction. Let $\mathcal{L}_S = \mathcal{L}_D = \mathbb{R}$ and choose

$$H_S^{\mathrm{self}} = \frac{p_S^2}{2m_S} + \frac{\kappa_S}{4}(x_S^2 - 1)^2, \qquad H_D^{\mathrm{self}} = \frac{p_D^2}{2m_D} + \frac{\kappa_D}{2}x_D^2,$$

$$H_S^{\mathrm{mech}} = \lambda_S x_S x_D, \qquad H_D^{\mathrm{mech}} = \lambda_D x_S x_D,$$

with $m_i, \kappa_i > 0$. The system has a double-well potential, while the demon's preferred position depends on the system coordinate and its state exerts a force back on the system. Both controls

modify the same reciprocal coupling, whose total strength is $\lambda_S + \lambda_D$; they are not independent directional gates for measurement and feedback.

Taking $\mathcal{A} = \{\lambda : \lambda(t) \equiv (1, 1)\}$ gives autonomous dynamics without external modulation. The mechanism-dependence sets are $\mathcal{D}_S = \{D\}$ and $\mathcal{D}_D = \{S\}$, so the nodes are mechanically paired. With this singleton admissible class, however, there are no intervention contrasts and hence no operational parents under Definition 2.5. Enlarging $\mathcal{A}$ to include varying control protocols allows interventional comparisons and externally driven operation.

In the overdamped Langevin limit with unit mobilities and independent thermal noises, let U denote the total potential, p the joint density, and $J_i = -(\partial_{x_i} U)p - \beta_i^{-1}\partial_{x_i} p$. Under the usual regularity and decay conditions, the local entropy production of Definition 3.8 satisfies

$$\Sigma_i^{\mathrm{loc}} = \Delta s_i^{\mathrm{marg}} - \beta_i \mathbb{E}[Q_i] - \int_0^T \dot{I}_i^{\mathrm{flow}}(t)\, dt = \int_0^T \int_{\mathbb{R}^2} \frac{\beta_i J_i(x,t)^2}{p(x,t)}\, dx\, dt \geq 0.$$

Consequently, the system's entropy balance before the information correction obeys

$$\Delta s_S^{\mathrm{marg}} - \beta_S \mathbb{E}[Q_S] \geq \int_0^T \dot{I}_S^{\mathrm{flow}}(t)\, dt.$$

When the system consumes correlations, the right-hand side can be negative: its entropy balance must therefore be interpreted together with the information exchanged with the demon [Horowitz and Esposito, 2014]. Summing the two local productions recovers the joint second law,

$$\Sigma_S^{\mathrm{loc}} + \Sigma_D^{\mathrm{loc}} = \Delta s_{S,D} - \beta_S \mathbb{E}[Q_S] - \beta_D \mathbb{E}[Q_D] \geq 0.$$

This construction illustrates the coupling and information balance underlying autonomous demon models; demonstrating a work-extracting operating regime requires additional analysis.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: We present Hamiltonian Causal Models, a model that allows to quantify causal phenomena for a wide range of physical processes via non-equilibrium Thermodynamic tools, supported by emperimental evidence through our estimation of entropy production

   Guidelines:

   - The answer [N/A] 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 [N/A] 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: We address what we see as the most important limitations of our approach in the conclusions. We cite the problem of learning H, which we deliberately ignore as we consider orthogonal to the contributions presented here; the limited discussion of the counterfactual framework as opposed to the interventional one discussed here; the existence of temporal series without an Hamiltonian description

   Guidelines:

   - The answer [N/A] 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: We include Proofs and precise assumptions in the Appendix.

   Guidelines:

   - The answer [N/A] 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 include anonimized code to reproduce the reported experiments as supplementary material, as well as experimental specifications in the appendix.

   Guidelines:

   - The answer [N/A] 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: We release the code as supplementary material.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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) necessary to understand the results?

   Answer: [Yes]

   Justification: We report them in the appendix

   Guidelines:

   - The answer [N/A] 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: We report statistical oscillation measures for the parent influence experiments, the one setting where this question is applicable.

   Guidelines:

   - The answer [N/A] 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: In the Appendix.

   Guidelines:

   - The answer [N/A] 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 did.

   Guidelines:

   - The answer [N/A] 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: [N/A]

    Justification: As a method primarily designed to be theoretical and applicable to physical processes, we do not see a particular risk of negative (or absolute) societal impact different from that of any machine-learning model that is applicable on arbitrary data for user-oriented purposes.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: See Above: we do not see a particular risk for the present paper.

    Guidelines:

    - The answer [N/A] 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: [Yes]

    Justification: The main external artifact leveraged, the code for Entropy production estimation, is properly referenced to the original publications thatintroduced them.

    Guidelines:

    - The answer [N/A] 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: We provide anonimied instructions in the attached code repository on how to utilize the model.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] 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.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.