Intervention-Based Time Series Causal Discovery via Simulator-Generated Interventional Distributions

Tsuyoshi Okita *

Posted Date: 11 May 2026

doi: 10.20944/preprints202605.0617.v1

Keywords: machine learning; time-series causal inference; intervention; physics simulation

 doi:10.20944/preprints202605.0617.v1

Article

Intervention-Based Time Series Causal Discovery via Simulator-Generated Interventional Distributions

Tsuyoshi Okita

Kyushu Institute of Technology, Japan; tsuyoshi.okita@gmail.com

Abstract

In many scientific domains, physics-based simulators—programs that compute system behaviour from governing equations, such as density functional theory for materials or fluid dynamics solvers—encode causal mechanisms and can predict system behaviour under hypothetical interventions. Machine learning extracts patterns from observational time series at scale, but those patterns reflect statistical associations ($P(Y \mid X)$), not causal effects ($P(Y \mid \mathrm{do}(X))$): in the presence of latent confounders, the structural VAR is provably non-identifiable from observational data alone (Fact 3.3), and no amount of statistical sophistication can substitute for genuine interventional data. Bridging these two traditions has so far been limited to using simulators for prediction; no existing framework uses them for *causal structure discovery* in time series. We propose SVAR-FM (Structural VAR with Flow Matching), a framework that treats a physical simulator as a mechanical realization of Pearl's $\mathrm{do}(\cdot)$ operator. Clamping a variable in the simulator physically severs confounding paths, producing interventional data by construction rather than by statistical argument. Conditional Flow Matching then parameterizes the interventional conditionals, enabling nonlinear mechanism learning. This yields four results. (1) The full structural VAR—contemporaneous and lagged edges jointly—becomes identifiable under a *coverage condition* on the simulator-clampable variables, verifiable *a priori* from domain knowledge alone (Theorem 4.1). The argument is intrinsic to the time series setting and has no i.i.d. counterpart. (2) An end-to-end error bound $|\hat{e}_{i\to j} - e^*_{i\to j}| \le O(M^{-1/2}) + O(\delta_{\mathcal{S}}) + O(\varepsilon_{\mathrm{FM}})$ (Theorem 5.2) cleanly separates Monte Carlo sampling, simulator fidelity $\delta_{\mathcal{S}}$, and Flow Matching approximation. A sharp consequence is a *sign-flip regime* (Corollary 5.5): when $\delta_{\mathcal{S}}$ exceeds a threshold set by the signal magnitude, the estimated causal effect reverses sign—a prediction that the prevailing forward-prediction view of simulators cannot produce. (3) The CausalSim benchmark confirms that SVAR-FM recovers the correct causal sign across four scientific domains (macroeconomics, diabetes, cosmic ray physics, and battery degradation) where observational methods produce sign-reversed estimates due to confounding. (4) A case study in ultrafast laser physics tests the sign-flip prediction by physically varying $\delta_{\mathcal{S}}$ through the accuracy level of a first-principles quantum solver: the low-accuracy setting produces a sign-reversed estimate, while the high-accuracy setting recovers the correct positive slope ($R^2 = 0.983$, zero bias relative to ground truth), providing the first experimental demonstration of a simulator-fidelity-dominated failure mode in causal discovery.

Keywords: machine learning; time-series causal inference; intervention; physics simulation

1. Introduction

Two powerful but historically separate traditions exist for understanding complex dynamical systems. On the one hand, mechanistic simulators encode causal mechanisms at the microscopic level and can predict system behaviour under hypothetical interventions. These range from *first-principles solvers*—programs that compute material properties directly from quantum mechanics, a cornerstone of recent AI-for-Science efforts in new materials discovery (Merchant et al., 2023; Zhang et al., 2024) and drug design (Jumper et al., 2021)—to Navier–Stokes solvers for fluid dynamics, kinetic models for chemical reactions, and agent-based models for financial markets. On the other hand, machine

 doi:10.20944/preprints202605.0617.v1

learning methods extract statistical patterns from observational time series at scale, but the patterns they find are associational: they reflect correlations shaped by confounding, feedback, and latent common causes, not the underlying causal structure.

Bridging these two traditions has long been recognised as a key challenge. Physics-informed machine learning (Raissi et al., 2019; Karniadakis et al., 2021) incorporates physical laws as regularization of predictive models, and AI-for-Science surrogate models (Jumper et al., 2021; Lam et al., 2023; Bodnar et al., 2024) learn to replace simulators for faster forward prediction. In both cases, however, the simulator's role is to improve *prediction*, not to identify *causal structure*. The causal graph—which variable drives which, and through what mechanism—remains inaccessible from observational data alone whenever latent confounders are present.

Causal inference (Pearl, 2009) provides the missing language: the $\mathrm{do}(\cdot)$ operator formally distinguishes intervention from observation, and under appropriate conditions makes causal structure identifiable. Yet existing causal discovery methods for time series either assume causal sufficiency (no latent confounders) (Granger, 1969; Runge et al., 2019; Hyvärinen et al., 2010) or operate on i.i.d. data (Wang et al., 2017; Mooij et al., 2020; Brouillard et al., 2020). In the presence of latent confounders, the structural VAR is provably non-identifiable from observational data alone (Fact 3.3). The fundamental difficulty is that no amount of statistical sophistication—non-Gaussianity, heteroscedasticity, score regularity—can substitute for the interventional data that the $\mathrm{do}(\cdot)$ operator requires.

The key observation of this paper is that *the simulators already in routine use in the physical sciences can mechanically realize Pearl's* $\mathrm{do}(\cdot)$ *operator*. When a simulator is run with a variable clamped to a fixed value (e.g., fixing temperature in an Arrhenius kinetics model, or fixing laser intensity in a quantum dynamics solver), confounding paths through that variable are physically severed; the resulting output is interventional by construction, not by statistical argument. This is categorically different from physics-informed ML (which uses physical laws for prediction, not causal identification) and from i.i.d. interventional methods (whose interventions are experimental labels, not simulator-realized manipulations of a time-varying process). Incorporating simulators as causal operators into machine learning therefore requires a framework that integrates causal inference with time series analysis and explicitly accounts for the simulator's imperfection—the gap between the simulator's physics and reality.

The framework we develop, *SVAR-FM* (Structural VAR with Flow Matching), uses a structural VAR (Sims, 1980) as the causal language for time series and conditional Flow Matching (Lipman et al., 2023; Tong et al., 2024) as the distributional learner. The central move is that the simulator is not one more ingredient of the learning objective but the operator that defines what is being identified in the first place. Once the simulator is taken seriously in this role, a new object enters the analysis—its fidelity $\delta_{\mathcal{S}}$, which classical causal theory does not track, and which we will show controls both the sample complexity and the sign of the recovered effects.

We evaluate SVAR-FM on a case study in ultrafast laser physics (Corkum, 1993) where a single first-principles quantum-mechanical solver generates both observational and interventional data, and the solver's physical accuracy can be tuned continuously—providing a direct test of the sign-flip prediction. We also evaluate on the CausalSim benchmark spanning four scientific domains (macroeconomics, diabetes, cosmic ray physics, and battery degradation).

Against this background, the paper addresses four interrelated problems.

1. How should a simulator enter causal inference? Existing uses of simulators in ML—as prediction targets (Jumper et al., 2021; Lam et al., 2023), as supervised data sources (Cranmer et al., 2020), or as tools inside autonomous agents (Boiko et al., 2023)—treat the simulator's forward output as the quantity of interest. None uses the simulator as a *causal operator*. We formalise the simulator-as-do-operator view: when the simulator clamps a variable, confounding paths are physically severed, and the resulting output is interventional by construction.

2. Under what conditions is the causal graph identifiable? In the i.i.d. setting, Eberhardt's theory (Eberhardt et al., 2005,0) shows that $O(\log d)$ interventional experiments suffice; practical algorithms

 doi:10.20944/preprints202605.0617.v1

(IGSP (Wang et al., 2017), JCI (Mooij et al., 2020)) implement this for i.i.d. data. For time series, however, contemporaneous and lagged edges must be handled jointly, and spurious high-frequency effects from unobserved lagged confounders have no i.i.d. counterpart. We prove that the full SVAR is identifiable under a *coverage condition* on the simulator-clampable variables (Theorem 4.1), verifiable *a priori* from domain knowledge alone.

3. What happens when the simulator is imperfect? Every real simulator has finite fidelity $\delta_{\mathcal{S}}$. Classical causal theory does not track this quantity; the AI-for-Science literature treats it as a prediction loss to minimise. We derive an end-to-end error bound (Theorem 5.2) that decomposes the estimation error into Monte Carlo, simulator fidelity, and Flow Matching terms. A sharp consequence is a *sign-flip regime* (Corollary 5.5): when $\delta_{\mathcal{S}}$ exceeds a threshold set by the signal magnitude, the estimated causal effect reverses sign—a prediction that the forward-prediction view of simulators cannot produce.

4. Can the sign-flip prediction be tested? A case study in ultrafast laser physics (§8) provides a setting where $\delta_{\mathcal{S}}$ is physically controllable: switching between two levels of physical accuracy in the same first-principles solver changes $\delta_{\mathcal{S}}$ from large (sign-reversed estimate) to small (correct sign recovered, $R^2 = 0.983$). The CausalSim benchmark confirms the same sign-reversal pattern across four additional domains (macroeconomics, diabetes, cosmic ray physics, battery degradation).

In summary, the main contributions of this paper are:

- A framework (SVAR-FM) that treats physics-based simulators as mechanical realizations of Pearl's do$(\cdot)$ operator for time series causal discovery, bridging mechanistic simulation and causal inference.
- An identifiability theorem (Theorem 4.1) showing that the full structural VAR is recoverable under a coverage condition verifiable *a priori* from domain knowledge, with an argument intrinsic to the time series setting.
- An end-to-end error bound (Theorem 5.2) that decomposes estimation error into Monte Carlo, simulator fidelity $\delta_{\mathcal{S}}$, and Flow Matching terms, and predicts a sign-flip regime (Corollary 5.5).
- The CausalSim benchmark: four cross-domain experiments (macroeconomics, diabetes, cosmic ray physics, battery degradation) demonstrating that SVAR-FM recovers correct causal signs where observational methods fail.
- A case study in ultrafast laser physics providing the first experimental test of the sign-flip prediction by physically varying the simulator's accuracy level.

Sections 2–3 review related work and formalise the non-identifiability of SVARs from observational data. Section 4 introduces the simulator-as-do-operator framework and states the identifiability theorem; Section 5 derives the end-to-end error bound (Theorem 5.2) and the sign-flip corollary, connecting simulator fidelity $\delta_{\mathcal{S}}$ to the sign of the recovered effects. Section 6 describes the algorithm. Section 7 reports the CausalSim benchmark across four domains, and Section 8 presents the ultrafast laser physics case study. Section 9 discusses positioning and limitations, and concludes.

2. Related Works

SVAR-FM proposes a use of physical simulators that has no direct counterpart in the existing causal-discovery or AI-for-Science literature—the simulator as a mechanical realization of Pearl's do$(\cdot)$ operator, with its fidelity $\delta_{\mathcal{S}}$ treated as a first-class object of the error analysis. To make this positioning precise, we contrast the proposal against four research streams whose concerns it touches: (i) time series causal discovery, (ii) intervention-based causal discovery, (iii) causal discovery under latent confounders, and (iv) simulation-based and mechanistic approaches to causal inference. For each stream we identify, in turn, what is and is not shared with SVAR-FM in the choice of data source, the role assigned to the simulator, the form of the identifiability argument, and the object that controls the error. Table 1 summarises the positioning at a glance. Table A11 (Appendix B) gives the full time-series listing; Table A12 (Appendix B) gives the i.i.d. listing; the rest of this section develops the argument in prose.

 doi:10.20944/preprints202605.0617.v1

Table 1. Overview: positioning of representative methods across the assumption space. "Setting" is i.i.d. or time series (TS); "Conf." denotes latent confounders; "Int." denotes use of interventional data; "NL" denotes support for nonlinear mechanisms; "Graph" indicates whether the causal graph is an input (*known*) or an output (*discovered*). Full listings of i.i.d. and TS methods appear in Appendix B (Tables A12 and A11).

Method	Setting	Conf.	Int.	NL	Graph
i.i.d. observational					
ANM (Hoyer et al., 2009)	i.i.d.	×	×	◯	discovered
Nonlinear CD w/ latent (Kaltenpoth and Vreeken, 2023)	i.i.d.	◯	×	◯	discovered
i.i.d. interventional					
DCDI (Brouillard et al., 2020)	i.i.d.	×	◯	◯	discovered
JCI (Mooij et al., 2020)	i.i.d.	◯	◯	◯	discovered
DeCaFlow (Almodóvar et al., 2025)	i.i.d.	◯	◯	◯	known
Time series observational					
PCMCI (Runge et al., 2019)	TS	×	×	◯	discovered
LPCMCI (Gerhardus and Runge, 2020)	TS	◯	×	◯	discovered (PAG)
Rahmani–Frossard (Rahmani and Frossard, 2025)	TS	×	×	◯	discovered
Time series flow-based, graph known					
DoFlow (Wu et al., 2025)	TS	×	◯	◯	known
Time series, graph discovered, latent confounders, interventional (ours)					
SVAR-FM (ours)	**TS**	◯	◯	◯	**discovered**

2.1. Time series causal discovery

Granger causality (Granger, 1969) and its measure-theoretic extensions (Dahlhaus and Eichler, 2003; Eichler, 2012) recover directed relationships from predictive improvement under causal sufficiency and linearity. Constraint-based methods—PCMCI (Runge et al., 2019), PCMCI$^+$ (Runge, 2020), and the decadal synthesis of Runge et al. (2023)—lift linearity via nonparametric tests but retain causal sufficiency. VARLiNGAM (Hyvärinen et al., 2010) and SpinSVAR (Misiakos and Püschel, 2025) exploit non-Gaussianity for linear-SVAR identification.

Score-based and deep-learning approaches—DYNOTEARS (Pamfil et al., 2020), Neural Granger (Tank et al., 2022), Rhino (Gong et al., 2023), CUTS/CUTS$^+$ (Cheng et al., 2023), Amortized CD (Löwe et al., 2022), TS-CausalNN (Assaad et al., 2022), Sun et al. (2023), temporal score matching (Chen et al., 2024), and score-informed neural operators (Kang et al., 2025)—broaden the mechanism class but still operate on observational data under causal sufficiency. Information-theoretic sample-complexity bounds (Yin et al., 2023; Veedu et al., 2023; Zhu et al., 2024) place the observational baseline.

A parallel flow-based line—CAREFL (Khemakhem et al., 2021), Causal NF (Javaloy et al., 2023), OCDaf (Kamkari et al., 2023), Hoang et al. (2024), CASTOR (Rahmani and Frossard, 2025a), and Rahmani and Frossard (2025,2)—parameterizes causal generative processes with flows and extracts identifiability from statistical noise properties (non-Gaussianity, heteroscedasticity), requiring causal sufficiency.

Every method above recovers causal structure by *adding* an assumption to observational data (sparsity, acyclicity, non-Gaussianity, regime structure). SVAR-FM takes a different route: it *drops* these assumptions and takes as input a different kind of data—simulator-generated $\mathrm{do}(\cdot)$ realizations. Identification is handled by the simulator's do-operator (Theorem 4.1); Flow Matching then parameterizes the interventional conditionals, enabling nonlinear mechanism learning while preserving tolerance to latent confounders.

2.2. Intervention-based causal discovery

The closest methodological relatives of SVAR-FM are methods that discover causal structure from interventional data. In the i.i.d. setting, Eberhardt's theory (Eberhardt et al., 2005,0; Eberhardt, 2012) shows that $O(\log d)$ experiments suffice for complete identification; practical algorithms include GIES (Hauser and Bühlmann, 2012), IGSP (Wang et al., 2017), UT-IGSP (Squires et al., 2020), JCI (Mooij et al., 2020), DCDI (Brouillard et al., 2020), ENCO (Lippe et al., 2022), and Bicycle (Rohbeck et al., 2024).

 doi:10.20944/preprints202605.0617.v1

All of these assume i.i.d. data; none addresses structural VARs with contemporaneous and lagged edges. SVAR-FM shares the general principle (use interventions for identifiability) but differs in three respects: (i) it operates on time series and handles the VAR structure natively; (ii) it admits latent confounders (Theorem 4.1 imposes no causal sufficiency); (iii) its interventions come from a physical simulator whose fidelity $\delta_{\mathcal{S}}$ enters the error bound as a first-class term.

2.3. Causal discovery under latent confounders

Latent confounders require moving beyond Markov equivalence classes. The FCI family (Spirtes et al., 2000) identifies partial ancestral graphs (PAGs); tsFCI (Entner and Hoyer, 2010), LPCMCI (Gerhardus and Runge, 2020), and SVAR-GFCI (Malinsky and Spirtes, 2018) extend this to time series. In the i.i.d. setting, JCI (Mooij et al., 2020) and DeCaFlow (Almodóvar et al., 2025) handle latent confounders with interventional data.

SVAR-FM also admits latent confounders, but its identification route is fundamentally different: the FCI methods use conditional-independence constraints to output a PAG (an equivalence class), while SVAR-FM uses physically realized interventions to identify the full SVAR (a single graph, not an equivalence class). In terms of Pearl's do-calculus (Pearl, 2009), the identification strategy of SVAR-FM relies primarily on Rule 2 (the action/observation exchange): under an intervention that severs all backdoor paths, $P(Y \mid \mathrm{do}(X{=}x), Z) = P(Y \mid X{=}x, Z)$, which justifies using the simulator output as if it were experimental data. Settings requiring Rule 1 (insertion/deletion of observations) or Rule 3 (insertion/deletion of actions)—such as front-door identification when no variable can be directly intervened upon—fall outside the current scope of SVAR-FM. Extending the framework to exploit such indirect identification strategies via the simulator is a direction for future work.

2.4. Simulation-based and mechanistic approaches

Several lines of work bring simulators and mechanistic models into contact with causal inference, but with different goals. Simulation-based inference (SBI) (Cranmer et al., 2020; Brehmer et al., 2020; Lueckmann et al., 2021; Radev et al., 2023) estimates posterior distributions over the parameters of a mechanistic model whose causal structure is *fixed*; SVAR-FM *discovers* the structure. Park et al. (Park et al., 2023) (GOBI) use ODE data-reproducibility for causal inference; the causal criterion is the ability to reproduce an observed time series, not a do-operator intervention. Deep SCMs (Pawlowski et al., 2020; Sanchez and Tsaftaris, 2022; Chao et al., 2024) and DoFlow (Wu et al., 2025) learn causal mechanisms under a *known* graph; SVAR-FM discovers the graph. DeCaFlow (Almodóvar et al., 2025) and PO-Flow (Wu et al., 2025) extend this to confounders and potential outcomes, respectively, but still require the graph as input. CaTSG (Xia et al., 2025) generates causal time series under a known graph. Identifiable flow models (Le et al. (Le et al., 2025)) learn mechanisms from a known ordering using conditional Flow Matching—the same technical primitive as SVAR-FM's Phase 4, but solving a different problem (mechanism learning vs. structure discovery).

2.5. AI for Science and the role of simulators

The past five years have seen the emergence of *AI for Science* (AI4S) as a recognised research agenda at major ML venues (Karniadakis et al., 2021; Wang et al., 2023). The broader ecosystem now includes dedicated workshops at NeurIPS, ICML and ICLR (AI4Science, AI4Mat, ML4Science), a growing set of scientific foundation models, and a parallel push towards autonomous research agents. Much of this work can be organised around three recurring uses of simulators.

(a) Simulators as prediction targets. A first line trains ML models *to replace* a first-principles simulator whose output is itself the quantity of scientific interest. Representative examples include AlphaFold and its successors for protein structure (Jumper et al., 2021; Abramson et al., 2024), GraphCast (Lam et al., 2023) and GenCast (Price et al., 2025) for global weather, the Aurora atmospheric foundation model (Bodnar et al., 2024), GNoME (Merchant et al., 2023) and MatterSim (Zhang et al., 2024) for materials discovery, and MACE for interatomic potentials (Batatia et al., 2022). In these works

 doi:10.20944/preprints202605.0617.v1

the simulator (or the experimental ground truth it approximates) is the target, and the ML model's job is to reproduce it faster or more accurately.

(b) Simulators as training-data generators. A second line uses simulators to supply supervised data for tasks where real observations are scarce or expensive: PDE surrogates (Brunton and Kutz, 2024; Li et al., 2021), physics-informed learning (Raissi et al., 2019; Karniadakis et al., 2021), and simulation-based inference with neural density estimators (Cranmer et al., 2020; Brehmer et al., 2020; Lueckmann et al., 2021; Radev et al., 2023). Here the simulator is a teacher; the learned model inherits its inductive biases and is evaluated on held-out simulator states or on downstream tasks.

(c) Simulators inside autonomous research pipelines. A third, newer line treats the simulator as one tool among many inside an LLM-driven research agent: Coscientist (Boiko et al., 2023), the AI Scientist line (Lu et al., 2025), and proposals for AI co-scientists (Gottweis et al., 2025) all chain simulator calls with literature search, code execution, and (optionally) robotic experimentation. The simulator there is one substitutable component of an open-ended search procedure.

SVAR-FM proposes a use of simulators that is not captured by (a)–(c) above. It is not a fourth item in the same list: the simulator here is neither a prediction target, nor a supervised-data source, nor a swappable tool inside an agent. It is an operator—an executable realization of Pearl's $\mathrm{do}(\cdot)$ on the real-world data generating process—and the thing being learned is not the simulator's forward map but the causal structure whose effects the operator makes identifiable. This shift in role changes what the analysis must say about the simulator. Its imperfections cannot be rolled into a prediction loss, because the quantity of interest is the causal effect $e^*_{i\to j}$, which depends nonlinearly on the interventional distribution and can be sign-flipped by a bounded simulator bias δ_S (Theorem 5.2, Corollary 5.5). The bias becomes a first-class object of the analysis, with its own error term and its own threshold for sign reversal—features that none of the three uses (a)–(c) need to produce, because in none of them is the object of interest a causal structure. We therefore regard SVAR-FM not as an application of AI-for-Science methodology to causal discovery, but as a proposal that simulators in AI for Science have a second, distinct role beyond forward prediction.

3. Problem Setting

3.1. Notation

Consider d observed time series variables $\mathbf{X}_t = (X_{1,t}, \ldots, X_{d,t})^\top \in \mathbb{R}^d$ $(t = 1, \ldots, T)$. The true causal structure is represented by a contemporaneous causal matrix $B_0 \in \mathbb{R}^{d\times d}$ and lagged causal matrices $\{B_l\}_{l=1}^p$.

3.2. Structural VAR Model

The structural VAR (SVAR) was introduced by Sims (Sims, 1980) and has become a standard tool in macroeconomics (Kilian and Lütkepohl, 2017).

Definition 3.1 (Structural VAR (SVAR) (Sims, 1980))**.** *The structural VAR model is defined as:*

$$B_0\mathbf{X}_t = \sum_{l=1}^{p} B_l\mathbf{X}_{t-l} + \boldsymbol{\epsilon}_t \tag{1}$$

where B_0 is the contemporaneous causal structure (with unit diagonal entries), B_l represents the causal effects at lag l, and $\boldsymbol{\epsilon}_t$ denotes the structural shocks (mutually independent, $\mathbb{E}[\boldsymbol{\epsilon}_t] = 0$, $\mathrm{Cov}(\boldsymbol{\epsilon}_t) = \Sigma_\epsilon$ is diagonal).

Rearranging Eq. (1) yields the reduced-form VAR:

$$\mathbf{X}_t = \sum_{l=1}^{p} \Phi_l\mathbf{X}_{t-l} + \mathbf{u}_t \tag{2}$$

where $\Phi_l = B_0^{-1}B_l$ and $\mathbf{u}_t = B_0^{-1}\boldsymbol{\epsilon}_t$.

 doi:10.20944/preprints202605.0617.v1

3.3. The Identification Problem

The identification problem of structural VARs has been extensively studied in econometrics (Sims, 1980; Kilian and Lütkepohl, 2017).

[Non-identifiability of Structural VAR (Kilian and Lütkepohl, 2017)] For the error covariance matrix $\Sigma_u = \mathrm{Cov}(\mathbf{u}_t)$, there exist infinitely many pairs (B_0, Σ_ϵ) satisfying

$$\Sigma_u = B_0^{-1}\Sigma_\epsilon (B_0^{-1})^\top \tag{3}$$

For any orthogonal matrix Q, the pair $(B_0 Q, Q^\top \Sigma_\epsilon Q)$ also satisfies Eq. (3).

(Proof: see Appendix A.)

Remark 3.1 (Conventional identification strategies)**.** *Kilian and Lütkepohl (Kilian and Lütkepohl, 2017) systematize the following identification strategies: (1) Short-run restrictions: imposing zeros in B_0 (Sims, 1980); (2) Long-run restrictions: constraining cumulative effects (Blanchard and Quah, 1989); (3) Sign restrictions: imposing sign constraints on effects (Uhlig, 2005); (4) Non-Gaussianity: SVAR-LiNGAM (Hyvärinen et al., 2010). All four strategies recover identifiability by imposing additional statistical or sign restrictions on the SVAR; none of them admit latent confounders. The framework developed in the remainder of this paper takes a different route, using physically realized interventions rather than additional statistical restrictions, and consequently is not a fifth item in the above list.*

3.4. Latent Confounders

In Pearl's (Pearl, 2009) framework, the limitations of causal discovery under the presence of latent confounders are clearly established.

Definition 3.2 (Confounder (Pearl, 2009))**.** *A variable Z is a **confounder** of X and Y if it has the structure $X \leftarrow Z \rightarrow Y$.*

Proposition 3.1 (Non-identifiability under confounding (Pearl, 2009))**.** *When a latent confounder Z common to X and Y exists, observational data $P(X, Y)$ cannot distinguish between $X \rightarrow Y$ (direct causation) and $X \leftarrow Z \rightarrow Y$ (confounding).*

3.5. Summary of Identifiability Conditions

Table 2 summarizes the identifiability conditions according to model class, noise assumptions, and the presence of confounding.

Table 2. Identifiability conditions for causal discovery methods

Model	Noise	Confounding	Identifiability
Linear	Gaussian	×	Not identifiable
Linear	Non-Gaussian	×	Identifiable (VARLiNGAM)
Linear	Non-Gaussian	◯	**Not identifiable** (fails under conf.)
Linear	Any	◯	**Identifiable** (SVAR-FM†)
Nonlinear	Any	×	Conditionally identifiable (ANM (Hoyer et al., 2009), etc.)
Nonlinear	Any	◯	**Identifiable** (SVAR-FM†)

† Requires a simulator satisfying Assumption 4.1 ($\delta_S \approx 0$) and the structural conditions of Remark 4.1.

In the presence of confounding (marked with ◯), identification is impossible from observational data alone, and intervention-based methods such as SVAR-FM are essential.

 doi:10.20944/preprints202605.0617.v1

4. Proposed Method: SVAR-FM

4.1. Nonlinear Structural VAR

We work with a nonlinear generalization of the SVAR in Def. 3.1. This is not the main novel ingredient of the paper: the causal mechanism is nonlinear in virtually all scientific applications of interest (HHG[1], kinetics, physiology, etc.), so the linear SVAR of Def. 3.1 would not be applicable in the settings we care about. The essential new ingredient—the use of a physical simulator as a realization of Pearl's $\mathrm{do}(\cdot)$ operator—is introduced in §4.3.

Definition 4.1 (Nonlinear Structural VAR). *The nonlinear structural VAR is defined as:*

$$X_{j,t} = f_j\left(\mathrm{Pa}_j^{(0)}(t), \mathrm{Pa}_j^{(1:p)}(t)\right) + \epsilon_{j,t} \tag{4}$$

where $\mathrm{Pa}_j^{(0)}(t)$ *denotes the contemporaneous parent variables,* $\mathrm{Pa}_j^{(1:p)}(t)$ *denotes the lagged parent variables, and* f_j *is a nonlinear causal mechanism.*

4.2. Flow Matching as the Distributional Learner

Conditional Flow Matching (Lipman et al., 2023; Tong et al., 2024; Albergo and Vanden-Eijnden, 2023; Chen et al., 2018) parameterizes the interventional conditionals $P(\cdot \mid \mathrm{do}(X_i = x))$ produced by the simulator. It has two roles in SVAR-FM: (1) modeling nonlinear causal mechanisms f_j in Eq. (4), and (2) incorporating physical constraints from the simulator into the conditioning vector, enabling sensitivity analysis of causal effects with respect to simulator assumptions. The technical details (optimal-transport conditional flow matching (OT-CFM) loss, conditioning design, universal approximation) are given in Appendix E.

4.3. The simulator as a realization of Pearl's do-operator

We now introduce the central object of the paper. Let $\mathcal{S}$ be a physical simulator (e.g., TDDFT[2], Arrhenius kinetics[3], an agent-based model). The simulator is *not* used as a surrogate of the forward dynamics and *not* used as a source of supervised training data. Instead, we use it as an operator that, on demand, produces samples from the interventional distribution $P(\cdot \mid \mathrm{do}(X_i = x))$ of the nonlinear SVAR of Def. 4.1. That is, $\mathcal{S}$ is a *mechanical realization* of Pearl's $\mathrm{do}(\cdot)$ operator: querying $\mathcal{S}(\mathrm{do}(X_i = x))$ returns a sample of the downstream variables that corresponds, by construction, to having physically severed all incoming causal arrows into X_i and fixed X_i at the value x. Three consequences follow immediately. First, the observational and interventional distributions are distinguished *by construction*, not by a statistical proxy. Second, the identifiability argument does not need to assume non-Gaussianity, sparsity, or any other distributional restriction on the noise; it rests instead on which variables the simulator can clamp. Third, the fidelity of $\mathcal{S}$ enters the error analysis as an explicit, quantified object $\delta_{\mathcal{S}}$ (Assumption 4.1 below), rather than as unmodelled noise.

Assumption 4.1 (Simulator validity). *The physical simulator* $\mathcal{S}$ *approximates the true interventional distribution:* $\|P_{\mathcal{S}}(\cdot|\mathrm{do}(X_i = x)) - P(\cdot|\mathrm{do}(X_i = x))\|_{TV}$[4] $\leq \delta_{\mathcal{S}}$

Remark 4.1 (Structural conditions for the simulator-as-do equivalence). *Assumption 4.1 is a distributional condition (TV distance). For the simulator to faithfully realize Pearl's* $\mathrm{do}(\cdot)$ *operator, three structural conditions*

[1] High harmonic generation: a nonlinear optical process in which molecules irradiated by intense femtosecond laser pulses emit photons at integer multiples of the laser frequency; see §8 for details.

[2] Time-dependent density functional theory: a quantum-mechanical method that computes the time evolution of the electron density under external fields (e.g., laser pulses), used here via the Octopus code (Tancogne-Dejean et al., 2020).

[3] A rate-equation framework in which the rate constant of a chemical or degradation process depends exponentially on temperature: $k = A\exp(-E_a/k_BT)$. Widely used in battery degradation modelling and chemical engineering.

[4] $\|\cdot\|_{TV}$ denotes the total variation distance between two probability distributions: $\|P - Q\|_{TV} = \sup_A |P(A) - Q(A)|$. It ranges from 0 (identical distributions) to 1 (distributions with disjoint support).

 doi:10.20944/preprints202605.0617.v1

must also hold, which we state explicitly: (a) Structural fidelity*: the internal causal structure of $\mathcal{S}$ is consistent with the causal DAG of the target system, in the sense that the parent–child relationships encoded in the simulator's equations correspond to those of the true SCM; (b)* Modularity *(the independent causal mechanism principle (Peters et al., 2017; Schölkopf et al., 2012)): the operation $\mathrm{do}(X_i = x)$ within $\mathcal{S}$ replaces exactly the structural equation for X_i while leaving all other equations invariant; (c)* Variable correspondence*: the simulator and the target system are defined over the same set of endogenous variables (or a known superset thereof). In the applications of this paper, conditions (a)–(c) are satisfied by construction: each simulator implements a well-validated physical model whose causal structure is known from domain science (e.g., the Navier–Stokes equations, Arrhenius kinetics, the UVA/Padova ODE system). When these structural conditions hold, Assumption 4.1 reduces the gap between $\mathcal{S}$ and the true system to a purely quantitative discrepancy $\delta_{\mathcal{S}}$ (numerical accuracy, parameter uncertainty), which is the object analysed in Theorem 5.2. When condition (a) is violated—e.g., the simulator omits a relevant variable—the TV bound may still hold but the causal interpretation breaks down; in such cases, $\delta_{\mathcal{S}}$ absorbs structural model error and may be large.*

Lemma 4.1 (Identification of contemporaneous causation via intervention (Pearl, 2009; Eberhardt et al., 2007))**.** *Under Assumption 4.1, the interventional effect $e_{i\to j} = \mathbb{E}[X_j|\mathrm{do}(X_i = x')] - \mathbb{E}[X_j|\mathrm{do}(X_i = x)]$ identifies the contemporaneous causal direction.*

(Proof: see Appendix A.)

Definition 4.2 (Intervention target set and coverage)**.** *Let $\mathcal{I} \subseteq \{1, \ldots, d\}$ denote the set of variables for which simulator-generated interventional distributions $P_{\mathcal{S}}(\cdot \mid \mathrm{do}(X_{i,t} = x))$ are available for a sufficiently rich set of intervention values $x \in \mathcal{X}_i$ with $|\mathcal{X}_i| \geq 2$, for all t in the estimation window. We say that $\mathcal{I}$* covers *the graph $\mathcal{G}$ if every contemporaneous edge (i, j) with $i \neq j$ in $\mathcal{G}$ has $i \in \mathcal{I}$ or $j \in \mathcal{I}$; equivalently, no two endogenous variables are simultaneously un-intervened.*

Assumption 4.2 (Regularity of the SVAR process)**.** *The SVAR in Def. 4.1 is (a)* stable*, in the sense that the companion matrix of the lagged coefficients has spectral radius strictly less than one; (b)* acyclic in contemporaneous time*, i.e., the contemporaneous edge set $\{(i, j) : B_0^{ij} \neq 0, i \neq j\}$ forms a DAG; (c) the structural shocks $\epsilon_{j,t}$ are mutually independent across j and independent across t with finite second moments.*

Theorem 4.1 (Interventional identifiability of SVAR-FM)**.** *Let Assumptions 4.1 (with $\delta_{\mathcal{S}} = 0$) and 4.2 hold, and let $\mathcal{I}$ be an intervention target set that covers the contemporaneous graph of $\mathcal{G}$. Then both the contemporaneous causal structure B_0 and all lagged causal structures $\{B_l\}_{l=1}^{p}$ of the SVAR in Def. 4.1 are uniquely identifiable from the joint of the observational distribution $P(\mathbf{X}_{1:T})$ and the family of interventional distributions $\{P_{\mathcal{S}}(\mathbf{X}_{1:T} \mid \mathrm{do}(X_{i,t} = x)) : i \in \mathcal{I}, x \in \mathcal{X}_i, t\}$. Moreover, $\mathcal{I} = \{1, \ldots, d\}$ is sufficient but not necessary; for chain graphs, $|\mathcal{I}| = d - 1$ suffices.*

(Proof: see Appendix A.)

Remark 4.2 (From exact to approximate identifiability)**.** *Theorem 4.1 establishes identifiability under the idealisation $\delta_{\mathcal{S}} = 0$. When $\delta_{\mathcal{S}} > 0$ (as in every real simulator), the causal structure is no longer exactly identifiable, but the end-to-end error bound of Theorem 5.2 (§5) quantifies the resulting estimation error and shows that it decomposes cleanly into Monte Carlo, simulator fidelity, and Flow Matching components. In particular, Corollary 5.5 gives a threshold on $\delta_{\mathcal{S}}$ below which the estimated sign of each causal effect is preserved—the practical condition under which the identifiability of Theorem 4.1 degrades gracefully rather than catastrophically. For instance, in CausalSim-Macro the true effect is $|e^*| = 0.006$; the sign is preserved whenever $\delta_{\mathcal{S}} < |e^*|/2 = 0.003$, and the observed bias of 0.001 (Table 5) confirms that this condition holds.*

Remark 4.3 (Operational meaning of the coverage condition)**.** *Def. 4.2 is what we verify in each application. For instance, in the laser physics case study (§8), there are two correlated variables (R and E_0) and a single confounding edge $R \leftrightarrow E_0$; fixing R at a single value places R in $\mathcal{I}$, which covers this edge by Def. 4.2. More*

 doi:10.20944/preprints202605.0617.v1

generally, the coverage condition can be verified a priori *from domain knowledge of which variable pairs the simulator is capable of holding independently fixed.*

[Separation of confounding via intervention (Pearl, 2009)] When Z is a confounder of X_i and X_j, $\mathrm{do}(Z = z)$ distinguishes confounding from direct causation. This is a standard result from Chapter 3 of Pearl (Pearl, 2009).

(Proof: see Appendix A.)

Remark 4.4 (Addressing unobserved confounding via the simulator)**.** *Theorem 4.1 and the known result (Fact 4.3) primarily address* ***contemporaneous causation among observed variables*** *($X \leftrightarrow Y$). However, SVAR-FM can also address* ***unobserved confounders*** *($U \to X, U \to Y$ where U is unobservable).*

In the unobserved confounding case, the causal effect $X \to Y$ cannot be identified from observational data alone:

$$P(Y|X = x) \neq P(Y|\mathrm{do}(X = x)) \tag{5}$$

This arises because the backdoor path $X \leftarrow U \to Y$ remains open. The discrepancy between $P(Y \mid X)$ and $P(Y \mid \mathrm{do}(X))$ is the mechanism behind Simpson's paradox (Pearl, 2009; Simpson, 1951): a statistical association can reverse sign when a confounding variable is conditioned on.

In SVAR-FM, we exploit the fact that the simulator $\mathcal{S}$ can ***specify and control*** *the values of the unobserved variable U:*

$$P_{\mathcal{S}}(Y|\mathrm{do}(X = x), U = u) = P(Y|X = x, U = u) \tag{6}$$

Since the simulator can fix U and generate intervention data, the backdoor path through U can be blocked.

This property enables SVAR-FM to address the following two types of confounding in a unified manner:

- ***Simultaneous causation*** *(e.g., in macroeconomics, a central bank adjusts interest rates i in response to inflation π, creating a feedback loop $i \leftrightarrow \pi$ that masks the true causal effect of i on π; see §7.1.1): intervention severs the reverse causal direction*
- ***Unobserved confounding*** *(e.g., in the laser physics case study of §8, a molecular parameter affects both the laser field and the spectral output, but cannot be independently measured in a real experiment): the simulator fixes the confounding variable*

Although the mechanisms differ, both are resolved within the unified framework of "intervention via a simulator."

5. Theoretical Analysis

The theory of this section is organized around the object that the simulator-as-do-operator viewpoint introduces, namely the simulator fidelity $\delta_{\mathcal{S}}$, and its consequences for the recovery of causal effects. Two results are where the new content is concentrated: (i) interventional identifiability of the time series causal structure under an explicit coverage condition on the intervention target set (Theorem 4.1), which is not an instance of a previously available theorem—it has to handle the simultaneous presence of contemporaneous and lagged edges and to distinguish genuine contemporaneous causation from spurious high-frequency effects due to unobserved lagged confounders, neither of which arises in i.i.d. formulations; (ii) an end-to-end decomposition of the total estimation error into Monte Carlo, simulator fidelity, and Flow Matching approximation components (Theorem 5.2 and Corollary 5.6), together with its sign-flip consequence (Corollary 5.5): whenever $\delta_{\mathcal{S}}$ exceeds a threshold determined by the signal magnitude, the estimated causal effect is reversed in sign relative to the ground truth—a prediction we verify in the ultrafast laser physics case study (§8.1) by switching between a low-accuracy and a high-accuracy first-principles solver.

The remaining results in the section—the convergence rate of the interventional expectation estimator (Proposition 5.1), the sample complexity for joint identification of all pairwise effects (Proposition 5.2), the lower and upper bounds on the number of interventions (Corollaries 5.2–5.4), and consistency under vanishing estimation error (Proposition 5.4)—are obtained by specializing classical statistical and PAC-learning tools (van der Vaart, 2000; Wasserman, 2006; Valiant, 1984; Hoeffding,

 doi:10.20944/preprints202605.0617.v1

1963) and the interventional combinatorics of Eberhardt et al. (2005,0) to the SVAR-FM setting, so that the paper is self-contained and the role of each component of the error bound is visible.

5.1. Convergence Rate of Intervention Effect Estimation

The following result records the convergence rate of the Monte Carlo estimator of the interventional expectation, for use later in Theorem 5.2.

Proposition 5.1 (Convergence rate (van der Vaart, 2000; Hoeffding, 1963)). *For the estimation error between the intervention effect $\hat{e}_{i\to j}$ estimated from M intervention samples and the true value $e^*_{i\to j}$, when the variance of $X_j|\mathrm{do}(X_i = x)$ is bounded by σ^2:*

$$|\hat{e}_{i\to j} - e^*_{i\to j}| = O_p(M^{-1/2}) \tag{7}$$

More precisely, by Hoeffding's inequality (Hoeffding, 1963):

$$P\left(|\hat{e}_{i\to j} - e^*_{i\to j}| > \epsilon\right) \leq 2\exp\left(-\frac{M\epsilon^2}{2\sigma^2}\right) \tag{8}$$

(Proof: see Appendix A.)

Corollary 5.1 (Confidence interval (Wasserman, 2006)). *The confidence interval for the intervention effect at confidence level $1-\alpha$ is given by $\hat{e}_{i\to j} \pm z_{\alpha/2} \cdot \hat{\sigma}/\sqrt{M}$.*

5.2. Sample Complexity

We record the sample complexity for recovering all pairwise causal effects from simulator-generated interventions, specializing the PAC learning framework (Valiant, 1984; Kearns and Vazirani, 1994) to the SVAR-FM setting.

Proposition 5.2 (Sample complexity (Wasserman, 2006; Valiant, 1984)). *The total number of intervention samples required to identify the causal structure of d variables with error probability at most δ and estimation accuracy ϵ is:*

$$M_{total} = O\left(\frac{d^2\sigma^2 \log(d^2/\delta)}{\epsilon^2}\right) \tag{9}$$

(Proof: see Appendix A.)

Remark 5.1 (Comparison with existing methods). *The sample complexity $O(d^2)$ is of the same order as the number of conditional independence tests $O(d^2 p)$ in PCMCI (Runge et al., 2019). However, SVAR-FM enjoys the advantage of being robust to confounding.*

5.3. Lower and Upper Bounds on the Required Number of Interventions

The interventional combinatorics of Eberhardt et al. (2005,0), originally formulated for i.i.d. DAGs, continues to govern the SVAR-FM setting once the coverage condition of Def. 4.2 is in force; we record the resulting bounds for use in the error analysis.

Corollary 5.2 (Lower bound on the number of interventions). *To identify the causal structure of d variables, at least $d-1$ single-variable interventions are required in the worst case.*

(Proof: see Appendix A.)

Corollary 5.3 (Upper bound on the number of interventions). *With d single-variable interventions (one per variable), any causal structure over d variables can be identified.*

 doi:10.20944/preprints202605.0617.v1

(Proof: see Appendix A.)

Corollary 5.4 (Exact number of interventions (Eberhardt et al., 2005,0))**.** *The required number of interventions I^* satisfies $d-1 \leq I^* \leq d$. The lower bound $I^* = d-1$ is attained for chain structures $X_1 \to X_2 \to \cdots \to X_d$.*

(Proof: see Appendix A.)

5.4. Computational Complexity

Proposition 5.3 (Computational complexity)**.** *The computational complexity of SVAR-FM is $O(d^2pT + d \cdot M \cdot C_{\mathcal{S}} + d \cdot T \cdot C_{FM})$, where $C_{\mathcal{S}}$ is the simulator cost and C_{FM} is the Flow Matching training cost.*

5.5. Consistency

Proposition 5.4 (Consistency)**.** *Under Assumption 4.1, inclusion of the true causal mechanism in the model class, and $T, M \to \infty$, the estimated causal structure $\hat{\mathcal{G}}$ of SVAR-FM converges in probability to the true structure $\mathcal{G}^*$.*

(Proof: see Appendix A.)

5.6. Impact of Simulator Error and Evaluation of Physical Constraints

The following theorem quantifies the degree to which physical constraints inherent in the simulator (approximation accuracy of governing equations, numerical solver errors, etc.) affect causal estimation. Combined with the conditional generative model of Flow Matching (Remark A1), the sensitivity of estimates to changes in physical constraints (e.g., changes in physical parameters or approximation levels of the simulator) can be theoretically assessed.

Theorem 5.1 (Propagation of simulator error and impact assessment of physical constraints)**.** *When the simulator error is bounded by $\delta_{\mathcal{S}}$ (Assumption 4.1), the estimation error of the intervention effect satisfies:*

$$|\hat{e}_{i\to j} - e^*_{i\to j}| \leq O(M^{-1/2}) + O(\delta_{\mathcal{S}}) \tag{10}$$

The first term is the statistical estimation error (dependent on the sample size M), and the second term is the systematic bias arising from the physical assumptions of the simulator. Through Flow Matching, the variation of $\delta_{\mathcal{S}}(\mathbf{c})$ under different physical parameters $\mathbf{c}$ can be continuously modeled, enabling quantitative assessment of the impact of physical constraints on causal estimation.

(Proof: see Appendix A.)

Corollary 5.5 (Robustness condition)**.** *For the identification of the causal effect e^* to be robust, it is necessary that $|e^*| > 2\delta_{\mathcal{S}} + O(M^{-1/2})$.*

5.7. End-to-end error propagation through Flow Matching

Theorem 5.1 bounds the error of a *direct* Monte Carlo estimator of the interventional expectation. In SVAR-FM, however, the interventional distribution is not used directly: it is first approximated by a conditional Flow Matching model, and the causal effect is then read off from this model. The final estimation error therefore combines three sources: the Monte Carlo error, the simulator fidelity bias, and the Flow Matching approximation error. We now quantify this composition.

Let v^* denote the true (simulator-induced) conditional vector field that generates $P_{\mathcal{S}}(\cdot \mid \mathrm{do}(X_i = x))$, and let $\hat{v}_\theta$ denote the estimate returned by minimizing the empirical conditional Flow Matching loss $\widehat{\mathcal{L}}_{\mathrm{CFM}}(\theta)$ over the simulator-generated intervention dataset of size M. Denote the Flow Matching approximation error by

$$\varepsilon_{\mathrm{FM}} \;:=\; \big(\mathcal{L}_{\mathrm{CFM}}(\hat{v}_\theta) - \mathcal{L}_{\mathrm{CFM}}(v^*)\big)^{1/2}.$$

 doi:10.20944/preprints202605.0617.v1

This is the excess risk of the learned vector field with respect to the population CFM objective, and is upper-bounded by standard statistical-learning bounds in the form $\varepsilon_{\text{FM}} \leq \mathcal{O}\big(\text{Rad}_M(\mathcal{V}_\theta)\big)$ whenever $\mathcal{V}_\theta$ has bounded Rademacher complexity (van der Vaart, 2000).

Assumption 5.1 (Lipschitz flow and bounded response). *(a) The target vector field $v^*(\cdot, t \mid \mathbf{c})$ is L-Lipschitz in $\mathbf{x}$ uniformly in $(t, \mathbf{c})$; (b) the response functional $g_j(P) := \mathbb{E}_{X_j \sim P}[X_j]$ is 1-Lipschitz with respect to the Wasserstein-1 distance on the marginal of X_j, which holds whenever X_j has bounded conditional variance under the interventional distribution.*

Assumption 5.1(a) is standard for flow-matching-based density estimation and implies, via a Gronwall-type argument (Tong et al., 2024; Albergo and Vanden-Eijnden, 2023), that the Wasserstein-1 distance between the distributions generated by $\hat{v}_\theta$ and v^* is controlled by their L^2 difference: $W_1\big(\hat{P}_{\hat{\theta}}, P_{\mathcal{S}}\big) \leq e^L \cdot \varepsilon_{\text{FM}}$.

Theorem 5.2 (End-to-end error bound). *Under Assumptions 4.1, 4.2, and 5.1, the Flow-Matching-based estimator $\hat{e}^{\text{FM}}_{i\to j} := g_j(\hat{P}_{\hat{\theta}}(\cdot \mid \text{do}(X_i = x'))) - g_j(\hat{P}_{\hat{\theta}}(\cdot \mid \text{do}(X_i = x)))$ satisfies*

$$|\hat{e}^{\text{FM}}_{i\to j} - e^*_{i\to j}| \;\leq\; \underbrace{C_1\sigma/\sqrt{M}}_{\textit{Monte Carlo}} + \underbrace{C_2 \cdot \delta_{\mathcal{S}}}_{\textit{simulator fidelity}} + \underbrace{C_3 \cdot e^L \cdot \varepsilon_{\text{FM}}}_{\textit{FM approximation}}, \tag{11}$$

with probability at least $1-\eta$, where $C_1 = \sqrt{2\log(2/\eta)}$ and $C_2, C_3 \leq 2 \cdot \sup_{\mathbf{x}} |x_j|$ are constants that depend only on the conditional range of X_j.

(Proof: see Appendix A.)

Remark 5.2 (Which term dominates?). *Which term of Eq. equation 11 dominates is a* diagnostic *quantity that SVAR-FM can estimate empirically. (i) When the simulator is a first-principles solver with well-controlled discretization error, $\delta_{\mathcal{S}}$ is small and the error is limited by M or ε_{FM}. (ii) When the simulator's physical model omits an important effect, $\delta_{\mathcal{S}}$ becomes $O(1)$ relative to the signal, and Corollary 5.5 predicts a* sign reversal *rather than a smooth degradation. This is exactly what we observe in the laser physics case study (§8.1), where switching from a low-accuracy to a high-accuracy solver collapses the dominant term from $\delta_{\mathcal{S}}$ to ε_{FM} and restores the correct causal sign.*

Corollary 5.6 (Sample complexity revisited under FM approximation). *To achieve an overall estimation accuracy ϵ on $d(d-1)$ causal effects simultaneously with probability at least $1-\delta$, it suffices to have intervention sample size $M \geq C\sigma^2 \log(d^2/\delta)/(\epsilon/3)^2$, simulator fidelity $\delta_{\mathcal{S}} \leq \epsilon/(3C_2)$, and Flow Matching approximation error $\varepsilon_{\text{FM}} \leq \epsilon/(3C_3 e^L)$.*

6. Algorithm and Methodology

This section describes the SVAR-FM algorithm (Section 6.1), the architectural design principles (Section 6.2), and methodological guidance for practical application (Section 6.3).

6.1. Algorithm

SVAR-FM consists of five phases (Algorithm 1). Phases 1–3 are responsible for causal structure identification, while Phases 4–5 handle the learning of causal mechanisms via Flow Matching and the evaluation of physical constraints.

 doi:10.20944/preprints202605.0617.v1

Algorithm 1 SVAR-FM

Require: Observed time series $\mathcal{D}_{obs} = \{\mathbf{y}_t\}_{t=1}^{T}$, simulator $\mathcal{S}$, maximum lag $p_{\max}$, significance level α
Ensure: Causal graph $\hat{\mathcal{G}}$, causal mechanisms $\{v_{\theta_j}\}$, sensitivities $\{s_{ij}\}$

// Phase 1: Reduced-form VAR estimation (Hamilton, 1994)
$p^* \leftarrow \arg\min_{p \le p_{\max}} \mathrm{BIC}(p)$
$\{\hat{\Phi}_l\}_{l=1}^{p^*} \leftarrow$ OLS estimation
$\hat{\mathbf{u}}_t \leftarrow \mathbf{y}_t - \sum_{l=1}^{p^*} \hat{\Phi}_l \mathbf{y}_{t-l}$
// Phase 2: Intervention data generation via simulator
for $i = 1, \ldots, d$ **do**
  Design intervention values $\{x_i^{(m)}\}_{m=1}^{M}$ within the physically valid range $[\underline{x}_i, \overline{x}_i]$
  for $m = 1, \ldots, M$ **do**
    $\mathbf{y}^{(m)} \leftarrow \mathcal{S}.\mathrm{do}(X_i = x_i^{(m)})$
  end for
end for
// Phase 3: Causal structure identification (Lemma 4.1)
$\hat{\mathcal{G}} \leftarrow \emptyset$
for $(i, j), i \neq j$ **do**
  $\hat{e}_{i \to j} \leftarrow \frac{1}{M} \sum_{m=1}^{M} y_j^{(m)} - \bar{y}_j^{obs}$
  $\hat{\sigma}_{ij} \leftarrow$ bootstrap standard error ($B = 1000$ replicates)
  if $|\hat{e}_{i \to j}| / \hat{\sigma}_{ij} > z_{\alpha/2}$ **then**
    $\hat{\mathcal{G}} \leftarrow \hat{\mathcal{G}} \cup \{X_i \to X_j\}$ (significant causal edge)
  end if
end for
// Phase 4: Causal mechanism learning via Flow Matching (Remark A1)
for $j = 1, \ldots, d$ **do**
  Construct conditioning vector $\mathbf{c}_j \leftarrow [\mathbf{x}_{\mathrm{Pa}(j)}, \boldsymbol{\phi}]$
    ($\mathbf{x}_{\mathrm{Pa}(j)}$: parent variables of j in $\hat{\mathcal{G}}$, $\boldsymbol{\phi}$: physical parameters)
  $\theta_j \leftarrow \arg\min_{\theta} \mathcal{L}_{\mathrm{CFM}}(\theta; \mathbf{c}_j)$ (Eq. A20)
end for
// Phase 5: Impact assessment of physical constraints
for $(i, j) \in \hat{\mathcal{G}}$ **do**
  for $k = 1, \ldots, |\boldsymbol{\phi}|$ **do**
    $\boldsymbol{\phi}'_k \leftarrow$ perturb the k-th component of $\boldsymbol{\phi}$ by $\pm\delta$
    $s_{ij,k} \leftarrow |\hat{e}_{i \to j}(\boldsymbol{\phi}'_k) - \hat{e}_{i \to j}(\boldsymbol{\phi})| / \delta$
  end for
end for
return $\hat{\mathcal{G}}, \{v_{\theta_j}\}, \{s_{ij,k}\}$

Phase 1 is standard reduced-form VAR estimation, with the lag order p^* selected by BIC (Hamilton, 1994); this is the same initial step used by VARLiNGAM (Hyvärinen et al., 2010) and other VAR-based causal discovery methods. Default hyperparameters: $p_{\max} = 10$, bootstrap replicates $B = 1000$, significance level $\alpha = 0.05$ with Bonferroni correction $\alpha' = \alpha / d(d-1)$. Phase 2 generates intervention data using the simulator $\mathcal{S}$. The design of intervention values is discussed in Section 6.3.2. Phase 3 identifies causal edges through bootstrap tests of the intervention effects $\hat{e}_{i \to j}$. By employing statistical significance testing rather than threshold-based effect size judgments alone, consistency with the theoretical guarantees of Proposition 5.1 is maintained.

The identification formula implemented by Phase 3 can be stated explicitly: for each variable pair (i, j) with $i \in \mathcal{I}$, the causal effect is identified as

$$\hat{e}_{i \to j} = \frac{1}{M} \sum_{m=1}^{M} X_j^{(m)}\big|_{\mathrm{do}(X_i = x')} - \bar{X}_j^{\mathrm{obs}} \tag{12}$$

 doi:10.20944/preprints202605.0617.v1

where $\{X_j^{(m)}\}_{m=1}^{M}$ are simulator-generated samples and $\bar{X}_j^{\text{obs}}$ is the observational mean. The edge $i \to j$ is included in $\hat{\mathcal{G}}$ if and only if the bootstrap z-statistic $|\hat{e}_{i\to j}|/\hat{\sigma}_{ij}$ exceeds $z_{\alpha'/2}$ under the Bonferroni-corrected threshold $\alpha' = \alpha/d(d-1)$. This formula makes the identification constructive: it specifies an explicit estimand expressed in terms of observable (observational mean) and experimentally accessible (simulator samples) quantities. Phases 4–5 are optional; the causal graph $\hat{\mathcal{G}}$ is obtained from Phases 1–3 alone. The necessity of Phases 4–5 is discussed in Section 6.3.4.

6.2. Architecture

Figure 1 illustrates the SVAR-FM pipeline. Phases 1–3 (causal structure identification) and Phases 4–5 (physical constraint modeling) are clearly separated, each serving distinct purposes.

Phases 1–3 are composed of the following data flow. First, a reduced-form VAR is estimated from the observed time series $\mathcal{D}_{obs}$ (Phase 1), and intervention data $\mathcal{D}_{int}$ are simultaneously generated using the simulator $\mathcal{S}$ (Phase 2). These two outputs are integrated to construct the causal graph $\hat{\mathcal{G}}$ via estimation and statistical testing of intervention effects (Phase 3).

Phases 4–5 provide additional value. In Phase 4, the causal mechanisms of each variable are learned via conditional Flow Matching based on the structure of $\hat{\mathcal{G}}$. By including physical parameters $\boldsymbol{\phi}$ in the conditioning vector $\mathbf{c}_j$, a generative model consistent with physical laws is obtained. In Phase 5, sensitivities $s_{ij,k}$ of the causal effects to perturbations in $\boldsymbol{\phi}$ are computed, quantifying the impact of physical constraints on the estimates. Phase 5 is analogous in purpose to the sensitivity analysis frameworks of VanderWeele and Ding (2017) (E-value) and Robins et al. (2000) (bounding factor), which assess how strong an unmeasured confounder would need to be to explain away an observed effect. In SVAR-FM, the "unmeasured confounder" is the simulator's physical approximation: Phase 5 asks how large a change in $\boldsymbol{\phi}$ is needed to reverse the sign of the estimated causal effect.

Table 3 lists the simulators used in this paper; the bottom rows show further applications in preparation (cf. §9.3).

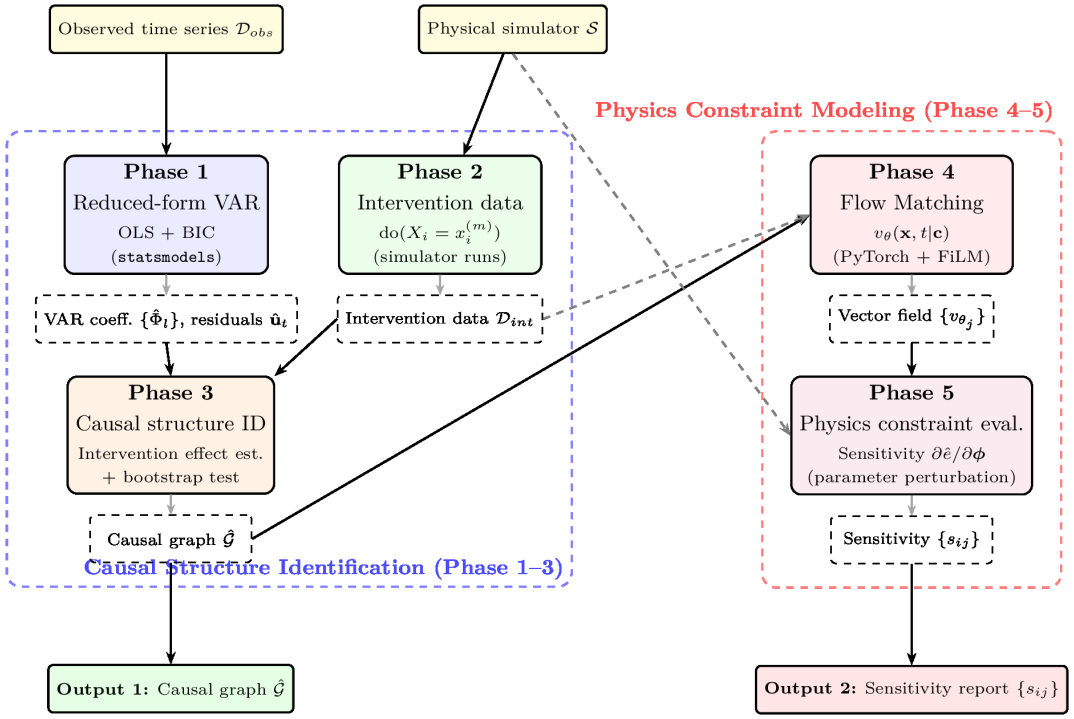

Figure 1. Architecture of SVAR-FM. The inputs are the observed time series $\mathcal{D}_{obs}$ and the physical simulator $\mathcal{S}$. Phases 1–3 are responsible for identifying the causal structure $\hat{\mathcal{G}}$, and for linear causal mechanisms, the procedure is complete at this stage. Phases 4–5 learn nonlinear causal mechanisms via conditional Flow Matching and assess the impact of physical constraints.

 doi:10.20944/preprints202605.0617.v1

Table 3. Simulators used in this paper. CausalSim instances (§7.1 and Appendix D) are listed in the middle block; the HHG case study (top) and further applications in preparation (bottom) complete the picture. "Type" reflects simulator fidelity: *ab initio* > numerical > analytical > MC > statistical > agent-based.

Application	Simulator	Intervention variable	M	Type
HHG (§8)	Octopus (TDDFT, SIC-ADSIC)	Laser amplitude E_0	10	*Ab initio*
CausalSim benchmark (§7.1):				
Macroeconomics	DSGE (Taylor Rule)	Policy rate i	100	Analytical
Diabetes	UVA/Padova T1DMS	Insulin dose	200	Numerical
Cosmic rays	Heitler–Matthews	σ_{inel} (E, A fixed)	500	MC
Battery (App. D)	Quantum ESPRESSO (DFT) + Arrhenius	Temperature T	50	*Ab initio*
Further applications (in preparation, cf. §9.3):				
ECG	PhysioNet synthetic	Disease severity	100	Statistical
Finance	BS + VIX-linked	Sentiment	100	Analytical
Finance (rates)	Mesa ABM 3.0	Interest rate Δr	371	Agent-based

6.2.1. SVAR-FM-DAG: NOTEARS Post-Processing Variant

SVAR-FM-DAG is the variant of SVAR-FM used when the data are expected to obey a strict DAG constraint (e.g., Tigramite synthetic data). It differs from the base algorithm in two components.

Phase 0: PINN ensemble prior knowledge

Prior to Phase 1, a directional score is computed as prior knowledge using a linear two-stage ensemble (VAR coefficients + Ridge Granger). When $T \geq 100$ and $N \leq 20$, a flow-matching-based nonlinear two-stage ensemble is also executed (when PyTorch is available) and integrated with the Phase 1 scores using weight $w = 0.2$.

NOTEARS post-processing

The score matrix obtained in Phase 1 is subjected to the acyclicity constraint

$$h(W) = \operatorname{tr}\left(e^{W \circ W}\right) - d = 0 \tag{13}$$

enforced via the augmented Lagrangian method (Zheng et al., 2018). This substantially reduces false discovery rate (FDR) for data where a DAG structure is expected (e.g., Tigramite synthetic data).

6.2.2. SVAR-dyn1/SVAR-dyn2: Variants for ODE-Based Dynamical Systems

For dynamical systems with bidirectional edges (cycles), the NOTEARS DAG constraint is inappropriate. SVAR-dyn1 and SVAR-dyn2 disable NOTEARS and Phase 0, and add a differential Granger score based on $\mathrm{d}X_j/\mathrm{d}t \approx f(X)$. SVAR-dyn2 further integrates Phases 3–5 with adaptive weighting and an eps-guard for deterministic chaotic systems. Full details are given in Appendix F.

6.3. Methodological Guidance

Applying SVAR-FM requires several methodological decisions. Below, we describe the conditions for applicability, the principles of intervention design, and the criteria for selecting among the framework components.

Variant selection.

The choice among SVAR-FM, SVAR-FM-DAG, and SVAR-dyn1/dyn2 is guided by domain knowledge of the system under study, not by post-hoc performance comparison:

- If the causal graph is known to be **acyclic** (e.g., regulatory cascades, economic policy transmission), use **SVAR-FM-DAG** (Phase 0 + NOTEARS).
- If the system contains **feedback loops or cycles** (e.g., coupled oscillators, predator-prey dynamics), use **SVAR-dyn1/dyn2** (NOTEARS disabled, differential Granger enabled).

 doi:10.20944/preprints202605.0617.v1

- When the structure is **unknown**, the base **SVAR-FM** (Phases 1–3, without NOTEARS or differential Granger) provides a conservative default that does not impose structural assumptions.

The adaptive routing strategy described in Appendix C.5 automates this selection using the BDS nonlinearity test (Brock et al., 1996) and system dimension as input features, removing the need for manual variant selection.

6.3.1. Applicability: When Can a Simulator Be Treated as an Intervention?

SVAR-FM treats the output of a simulator as "intervention data" in causal inference (Assumption 4.1). For this treatment to be justified, the simulator must satisfy the following conditions.

1. **Variable controllability**: The intervention target variable X_i can be fixed at a desired value, while other variables respond according to their causal mechanisms. For example, the Arrhenius simulator fixes the temperature T and computes the reaction rate k as $k = A\exp(-E_a/RT)$.
2. **Mechanism independence**: $\mathrm{do}(X_i = x)$ does not alter the causal mechanisms of variables other than X_i (corresponding to Pearl's (Pearl, 2009) modularity assumption).
3. **Assessable approximation accuracy**: The $\delta_{\mathcal{S}}$ in Assumption 4.1 is estimable. By Theorem 5.1, smaller $\delta_{\mathcal{S}}$ yields more robust causal estimation.

The "Type" column in Table 3 reflects the hierarchy of simulator fidelity. First-principles calculations (TDDFT, hydrodynamics) are derived directly from physical laws, resulting in small $\delta_{\mathcal{S}}$ but high computational cost. Analytical models (Arrhenius kinetics, Taylor Rule) have low computational cost but depend on model assumptions. This accuracy–cost trade-off is reflected in the differences in sample size M in Table 3.

6.3.2. Intervention Design

The design of intervention values $\{x_i^{(m)}\}_{m=1}^{M}$ in Phase 2 directly affects estimation accuracy. The following principles guide the design.

(1) Physical range: Intervention values should be set within the validity domain of the simulator. For example, in the HHG application of §8, the laser field amplitude E_0 is restricted to values below the over-the-barrier ionization threshold, beyond which the TDDFT treatment ceases to be physically meaningful. More generally, Arrhenius-type kinetic simulators are restricted to temperature ranges that exclude phase transitions, since the underlying rate law assumes a single reaction mechanism. Values outside such domains inflate $\delta_{\mathcal{S}}$ in Assumption 4.1.

(2) Sample size M: By Proposition 5.1, the estimation error of the intervention effect scales as $O(M^{-1/2})$. For a desired accuracy ϵ and confidence level $1-\alpha$, $M \geq 2\sigma^2 \log(2/\alpha)/\epsilon^2$ is required (Eq. 8). In practice, $M = 50$–200 suffices for many applications, but when the simulator variance is large (e.g., Monte Carlo simulators), $M = 500$–1000 may be needed (see Table 3).

(3) Placement of intervention values: When nonlinearity of the causal effect is expected, intervention values should be placed using Latin hypercube sampling or importance sampling rather than uniform spacing.

6.3.3. Statistical Testing and Multiple Comparison Correction

In Phase 3, the presence or absence of causal edges is simultaneously tested for $d(d-1)$ variable pairs, giving rise to a multiple comparison problem. Although Algorithm 1 uses individual bootstrap-based tests, we recommend either a Bonferroni correction $\alpha' = \alpha/d(d-1)$ or false discovery rate (FDR) control via the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995). The sample complexity $O(d^2 \log(d^2/\delta)/\epsilon^2)$ from Proposition 5.2 already accounts for multiple comparison correction via the union bound.

6.3.4. Determining the Necessity of Phases 4–5

Phases 4–5 (Flow Matching) are optional. The following criteria guide the decision on their necessity.

 doi:10.20944/preprints202605.0617.v1

Cases where Phases 1–3 alone suffice: (a) when causal mechanisms can be assumed linear and the primary quantity of interest is the (signed) causal effect rather than the response distribution, (b) when the sole objective is identification of the causal graph $\hat{\mathcal{G}}$, (c) when the simulator cost is high, making the generation of Phase 4 training data infeasible.

Cases where Phases 4–5 are beneficial: (a) when causal mechanisms are inherently nonlinear, so that a single-number effect size is insufficient and the full response distribution is needed (HHG; cf. §8), (b) when one wishes to compare predictions across multiple simulator variants—e.g., different exchange–correlation functionals in HHG (LDA vs. SIC-ADSIC), or different physical model families in other applications, (c) when one wishes to quantify the sensitivity of the estimates to the simulator's physical assumptions via the Phase 5 sensitivities $s_{ij,k}$.

6.3.5. Practical Assessment of Simulator Fidelity $\delta_{\mathcal{S}}$

Assumption 4.1 bounds the simulator error by $\delta_{\mathcal{S}}$ in total variation distance, and Corollary 5.5 requires $|e^*| > 2\delta_{\mathcal{S}} + O(M^{-1/2})$ for sign preservation. Since the true interventional distribution is unknown (otherwise causal discovery would be unnecessary), $\delta_{\mathcal{S}}$ cannot be measured directly. We recommend three complementary strategies for *a priori* assessment.

(1) Cross-simulator consistency. When multiple simulators of different fidelity are available for the same system, agreement among their interventional estimates provides an upper bound on $\delta_{\mathcal{S}}$. In CausalSim-Cosmic (§7.1.3), the SVAR-FM estimate ($-0.086\,\mathrm{g/cm^2/mb}$) agrees with three independent QCD Monte Carlo generators (QGSJet-II-04: -0.103, EPOS-LHC: -0.096, SIBYLL-2.3c: -0.110) to within 10–22%, bounding $\delta_{\mathcal{S}}$ relative to the signal magnitude. In HHG, the transition from LDA (sign-reversed) to SIC-ADSIC (correct sign, $R^2 = 0.983$) provides a binary diagnostic: when $\delta_{\mathcal{S}}$ crosses the sign-flip threshold, the qualitative change is immediately visible.

(2) Simulator validation against known experimental data. Most physical simulators are accompanied by validation studies comparing their predictions to experimental measurements on quantities *other* than the causal effect of interest. The discrepancy on these auxiliary quantities provides an independent estimate of $\delta_{\mathcal{S}}$. For example, the UVA/Padova simulator (CausalSim-Diabetes) has been validated against clinical glucose profiles from 300 patients (Man et al., 2014); the Octopus TDDFT code reproduces experimental ionization potentials of small molecules to within 0.1–0.3 eV (Tancogne-Dejean et al., 2020). These known accuracies constrain $\delta_{\mathcal{S}}$ without requiring knowledge of the causal effect.

(3) Internal consistency: sensitivity of the causal estimate to simulator parameters. Phase 5 of SVAR-FM computes the sensitivities $s_{ij,k} = |\hat{e}_{i\to j}(\boldsymbol{\phi}'_k) - \hat{e}_{i\to j}(\boldsymbol{\phi})|/\delta$ with respect to physical parameters $\boldsymbol{\phi}$. When $s_{ij,k}$ is large relative to $|\hat{e}_{i\to j}|$, the estimate is dominated by the $O(\delta_{\mathcal{S}})$ term and the sign may be unreliable. This provides a *self-diagnostic*: even without external validation data, a practitioner can flag causal estimates whose sign depends sensitively on simulator parameters.

None of these strategies provides a certified bound on $\delta_{\mathcal{S}}$; that would require access to the true interventional distribution. Together, however, they constitute a practical due-diligence protocol analogous to the model-checking practices standard in Bayesian inference and simulation-based inference (Cranmer et al., 2020). It is important to note that all three strategies are *falsification* criteria: they can detect when the simulator is inadequate (e.g., cross-simulator disagreement, discrepancy with experimental data, high sensitivity to parameters) but cannot *certify* that the simulator is sufficient. This asymmetry is inherent to causal inference: one can refute a causal model but never conclusively verify it from finite data. The strategies above are therefore best understood as a refutation protocol that increases confidence in proportion to the number of checks passed, rather than as a validation guarantee.

7. Evaluation

The experiments in this paper are organised in three tiers.

 doi:10.20944/preprints202605.0617.v1

1. **CausalSim benchmark** (this section): three scientific domains in the main text (macroeconomics, diabetes, cosmic ray physics) plus a fourth (battery degradation) in Appendix D, in which a first-principles simulator realizes Pearl's $\mathrm{do}(\cdot)$ operator and generates interventional data. Observational baselines cannot consume these data and therefore cannot remove the confounding that the simulator intervention severs. This tier directly tests the property that distinguishes SVAR-FM from every other method.
2. **HHG case study** (§8): the load-bearing experiment of the paper, in which the sign-flip prediction of Theorem 5.2 is verified by varying $\delta_{\mathcal{S}}$ physically (LDA vs. SIC-ADSIC (Tancogne-Dejean et al., 2020) exchange–correlation functional).
3. **Standard benchmarks** (Appendix C): CausalTime (Cheng et al., 2024), Tigramite (Runge et al., 2019), and CausalDynamics (Herdeanu et al., 2025), on which SVAR-FM is compared against observational baselines and, in an extended setting, against i.i.d. intervention-based methods (IGSP (Wang et al., 2017), UT-IGSP (Squires et al., 2020)). These benchmarks were not designed with simulator-based intervention in mind; they confirm that SVAR-FM performs competitively even outside its intended setting.

Baseline methods.

The following observational baselines are used throughout all three tiers.

- **OLS**: Ordinary least squares regression (used in CausalSim and HHG).
- **Granger** (Granger, 1969): Pairwise linear Granger causality test.
- **VARLiNGAM** (Hyvärinen et al., 2010): ICA-LiNGAM applied to VAR residuals; directional estimation via non-Gaussianity.
- **PCMCI** (Runge et al., 2019): PC-stable skeleton estimation + MCI test (partial correlation; ParCorr).
- **PCMCI+** (Runge, 2020): Extension of PCMCI with support for contemporaneous causation.

Additional methods specific to individual tiers (IGSP (Wang et al., 2017), UT-IGSP (Squires et al., 2020), SVAR-FM-DAG, SVAR-FM-CF, SVAR-dyn1, SVAR-dyn2) are described where they are first used (Appendix C for the standard benchmarks).

Proposed method variants.

- **SVAR-FM**: Proposed method (Phases 1–5, without Phase 0 spatial scoring).
- **SVAR-FM-DAG**: SVAR-FM with Phase 0 (PINN spatial scoring) + NOTEARS post-processing (§6.2.1).

Evaluation metrics.

F1 score (harmonic mean of precision and recall), TPR (True Positive Rate), FDR (False Discovery Rate), SHD (Structural Hamming Distance), AUROC, and AUPRC. For CausalSim, we additionally report the estimated causal effect, its sign correctness, and bias reduction relative to OLS.

7.1. CausalSim: Simulator-Driven Benchmark

We introduce **CausalSim**, a benchmark suite designed specifically to evaluate causal discovery methods that consume simulator-generated interventional data. Each CausalSim instance consists of (i) a real or realistic observational time series with known confounding, (ii) a first-principles simulator whose $\mathrm{do}(\cdot)$ operation physically severs the confounding path, and (iii) a ground-truth causal effect against which estimates are evaluated. The suite currently comprises four instances spanning four scientific domains and four simulator types:

- **CausalSim-Macro**: macroeconomic monetary policy (Taylor Rule confounding; analytical DSGE simulator)
- **CausalSim-Diabetes**: glucose–insulin dynamics (feedback-control confounding; numerical UVA/Padova ODE simulator)

 doi:10.20944/preprints202605.0617.v1

- **CausalSim-Cosmic**: cosmic ray air showers (energy confounding; Monte Carlo Heitler–Matthews simulator)
- **CausalSim-Battery** (Appendix D): lithium-ion battery degradation (latent temperature confounding; *ab initio* Quantum ESPRESSO DFT + Arrhenius simulator)

Observational methods (OLS, Granger, VARLiNGAM, PCMCI) are run on the same observational data for comparison; they cannot consume the interventional data and therefore cannot, in principle, remove the confounding that the simulator intervention severs.

Table 4 previews the results: in all four instances the observational methods produce sign-reversed or near-zero estimates, whereas SVAR-FM recovers the correct sign with the lowest bias.

Table 4. CausalSim benchmark: summary across four domains. "Sign correct" reports the percentage of seeds recovering the correct sign (Macro: 50 seeds; Diabetes: 20 seeds) or a qualitative indicator for single-run experiments (Cosmic, Battery). "Bias reduction" is relative to OLS. Battery (Appendix D) uses a DFT first-principles simulator.

Domain	Method	Estimate	True value	Sign correct	Bias reduction
Macroeconomics	OLS (obs.)	+0.076	−0.006	88%	—
	VARLiNGAM (obs.)	+0.229	−0.006	44%	−187%
	SVAR-FM (ours)	**−0.007**	−0.006	**100%**	**99%**
Diabetes	OLS (obs.)	+5525	−3000	0%	—
	PCMCI (obs.)	+182	−3000	0%	—
	SVAR-FM (ours)	**−1922**	−3000	**100%**	**87%**
Cosmic ray	Obs. ($\sigma \to N_\mu$)	+0.013	0.000	—	—
	SVAR-FM (ours) ($\sigma \to N_\mu$)	**0.000**	0.000	**—**	**100%**
	SVAR-FM (ours) ($\sigma \to X_{\max}$)	**−0.086**	−0.078	**correct**	—
Battery (App. D)	OLS (obs.)	−0.10	+0.03	wrong	—
	SVAR-FM (ours)	**+0.03**	+0.03	**correct**	**100%**

7.1.1. CausalSim-Macro: Taylor Rule Confounding

Problem setting.

In monetary policy analysis, the following causal structure operates: interest rate i affects the output gap y through the IS curve, and y in turn affects inflation π through the Phillips curve, so that the true causal effect $i \to \pi$ is negative ($-\sigma\kappa$). In observational data the Taylor Rule ($\pi \to i$: the central bank raises i in response to rising π) creates reverse causation, producing a spurious *positive* correlation between i and π.

Data.

Three quarterly macroeconomic time series from Federal Reserve Economic Data (FRED)[5]: Federal Funds Rate (FEDFUNDS), CPI-based annualized inflation (CPIAUCSL), and HP-filtered Real GDP output gap (GDPC1, $\lambda = 1600$). Analysis period: 1960Q1–2007Q3 ($T = 192$ quarters; post-2008 data are excluded due to the zero lower bound). This is a standard dataset in macroeconomic SVAR analysis (Kilian and Lütkepohl, 2017).

Simulator and intervention.

The simulator is a Dynamic Stochastic General Equilibrium (DSGE) model consisting of three coupled difference equations: an IS curve (output depends on interest rate), a Phillips curve (inflation depends on output gap), and a Taylor Rule (interest rate responds to inflation and output gap). DSGE models are the standard workhorse of central bank forecasting and are available in multiple open-source implementations[6]. The model is calibrated to the FRED data. The intervention do($i =$ exogenous) disables the Taylor Rule and sets the interest rate exogenously, severing the $\pi \to i$

[5] https://fred.stlouisfed.org/, publicly available

[6] E.g., Dynare (https://www.dynare.org/), a widely-used DSGE toolbox. Our implementation follows the same three-equation structure. SVAR-FM integration scripts will be released upon publication.

 doi:10.20944/preprints202605.0617.v1

feedback. Structural parameters estimated from the real data: interest-rate smoothing $\rho_i = 0.882$, inflation response $\phi_\pi = 0.357$, output-gap response $\phi_y = 0.229$, Phillips slope $\kappa = 0.114$, IS elasticity $\sigma = 0.038$. From these, the true causal effect is $i \to \pi = -\sigma\kappa = -0.006$.

Results.

Table 5 compares five methods on the $i \to \pi$ causal effect (50 seeds). SVAR-FM estimates -0.007, matching the true value -0.006 with 99% bias reduction over OLS. It is the only method to recover the correct negative sign across all 50 seeds (100% sign accuracy). OLS, VARLiNGAM, and SVAR-Cholesky all produce the wrong (positive) sign on the real data due to Taylor Rule confounding.

Table 5. Method comparison for $i \to \pi$ causal effect (macroeconomics)

Method	Estimate	|Bias|	RMSE	Sign correct
OLS	$+0.076$	0.082	0.046	88%
Granger	-0.009	0.003	0.095	60%
VARLiNGAM	$+0.229$	0.235	0.313	44%
SVAR-Cholesky	$+0.074$	0.080	0.042	88%
SVAR-FM (ours)	$\mathbf{-0.007}$	**0.001**	**0.008**	**100%**

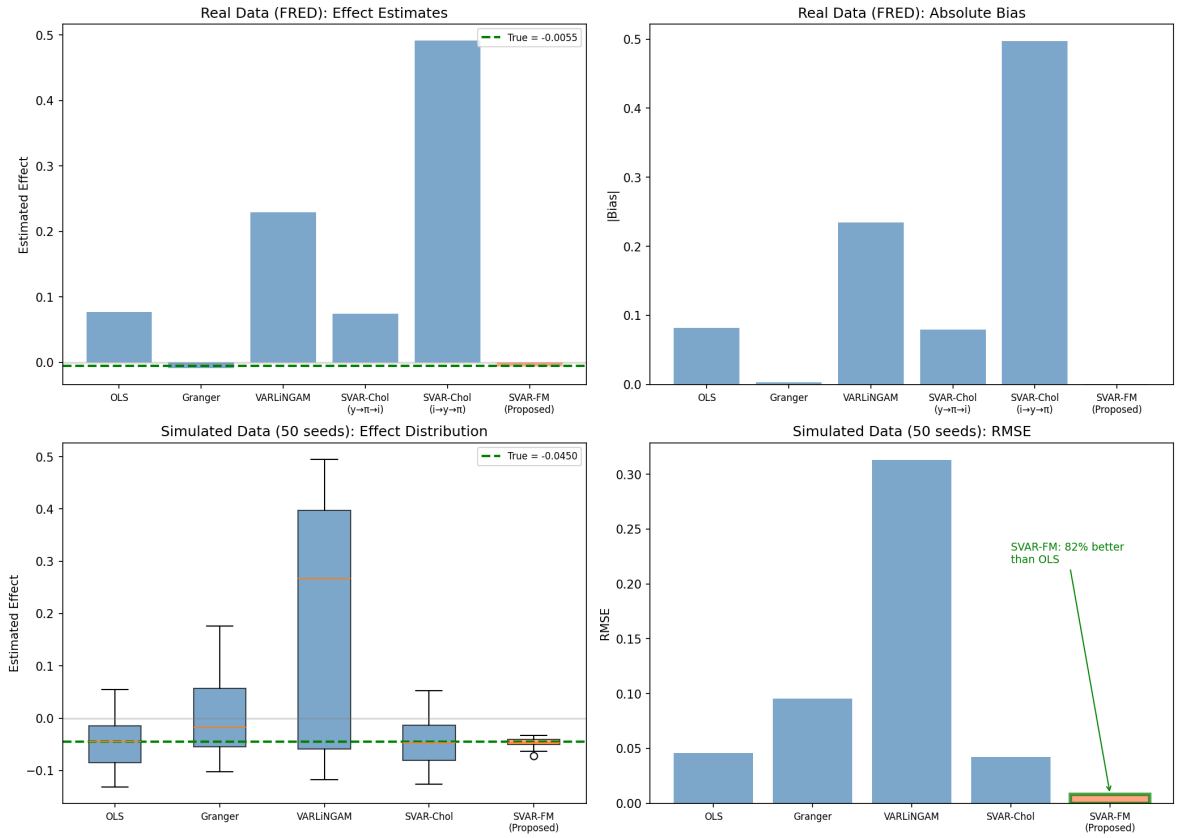

Figure 2. CausalSim-Macro: method comparison (50 seeds). SVAR-FM (orange) achieves the smallest bias and RMSE across all seeds. The true causal effect $-\sigma\kappa = -0.006$ is shown as a green dashed line. Observational methods (OLS, VARLiNGAM, SVAR-Cholesky) cluster around positive values due to Taylor Rule confounding.

Simulator fidelity.

The DSGE simulator is an analytical model whose structural parameters $(\rho_i, \phi_\pi, \phi_y, \kappa, \sigma)$ are estimated from the FRED data by ordinary least squares. The simulator error $\delta_{\mathcal{S}}$ therefore reflects the misspecification of the New Keynesian three-equation model relative to the true monetary transmission

 doi:10.20944/preprints202605.0617.v1

mechanism. The 99% bias reduction (Table 5) indicates that $\delta_{\mathcal{S}}$ is small enough to preserve the sign of the causal effect, consistent with the threshold condition of Corollary 5.5.

7.1.2. CausalSim-Diabetes: Glucose–Insulin Feedback Control

Problem setting.

In glucose management for Type 1 diabetes, the true causal effect of insulin on blood glucose is negative (insulin lowers glucose). In observational data, however, an automated insulin pump delivers insulin in response to rising glucose, creating reverse causation: hyperglycemia $\rightarrow$ insulin increase $\rightarrow$ glucose reduction. This bidirectional feedback makes the true effect unidentifiable from observational data.

Data.

To demonstrate the limitations of real clinical data, we first examined the ShanghaiT1DM/T2DM CGM dataset (Zhao et al., 2023)[7] (12 T1DM patients, 100 T2DM patients, 15-minute CGM intervals). Insulin delivery amounts are not recorded in this dataset, making estimation of the insulin$\rightarrow$CGM effect impossible in principle.

Simulator and intervention.

We used the UVA/Padova Type 1 Diabetes Simulator (Man et al., 2014; Kovatchev et al., 2009; Cobelli and Kovatchev, 2023), an FDA-accepted physiological model that describes glucose–insulin dynamics as a system of $\sim$30 coupled ODEs (glucose absorption, insulin kinetics, hepatic glucose production, peripheral uptake). It is the standard *in silico* platform for testing artificial pancreas control algorithms before human trials and includes 300 virtual patients with realistic inter-patient variability. The open-source Python implementation simglucose[8] (Xie, 2018) was used.

Observational data (within the simulator): a reactive Basal-Bolus controller adjusts insulin according to blood glucose (glucose $>$ 180 mg/dL $\Rightarrow$ insulin increase; glucose $<$ 80 mg/dL $\Rightarrow$ insulin decrease). This feedback loop is the source of reverse causation.

Intervention data: $\mathrm{do}(\mathrm{insulin} = c)$ for $c \in \{0.01, 0.015, \ldots, 0.04\}$ U/min (7 levels). The feedback loop is severed by construction: the simulator delivers insulin at a fixed rate regardless of blood glucose.

Three virtual patients (adult#001–003), 3-day simulation, 5-minute intervals (864 samples/patient, 2592 total). Meal schedule: breakfast 50 g (7:00), lunch 80 g (12:00), dinner 60 g (19:00).

Results.

The true causal effect (ATE of insulin on blood glucose, -3000 mg/dL per U/min) is computed from the simulator's interventional data across 20 independent Monte Carlo seeds with different sensor noise realizations. Table 6 compares five methods (20 seeds). SVAR-FM achieves the lowest bias and RMSE (87% improvement over OLS) and is one of only two methods to recover the correct negative sign with 100% accuracy. OLS and PCMCI produce positive estimates due to reverse causation.

Table 6. Method comparison for Insulin $\rightarrow$ CGM causal effect (diabetes)

Method	Estimate	$\lvert$Bias$\rvert$	RMSE	Sign correct
OLS	$+5525$	8525	8527	0%
Granger	-635	2365	2367	100%
VARLiNGAM	-259	2741	2744	100%
PCMCI	$+182$	3182	3182	0%
SVAR-FM (ours)	$\mathbf{-1922}$	**1078**	**1104**	**100%**

[7] Shanghai CGM dataset: https://doi.org/10.1038/s41597-023-02041-1

[8] simglucose: https://github.com/jxx123/simglucose. Alternative implementations exist in MATLAB (original UVA/Padova distribution) and Julia.

 doi:10.20944/preprints202605.0617.v1

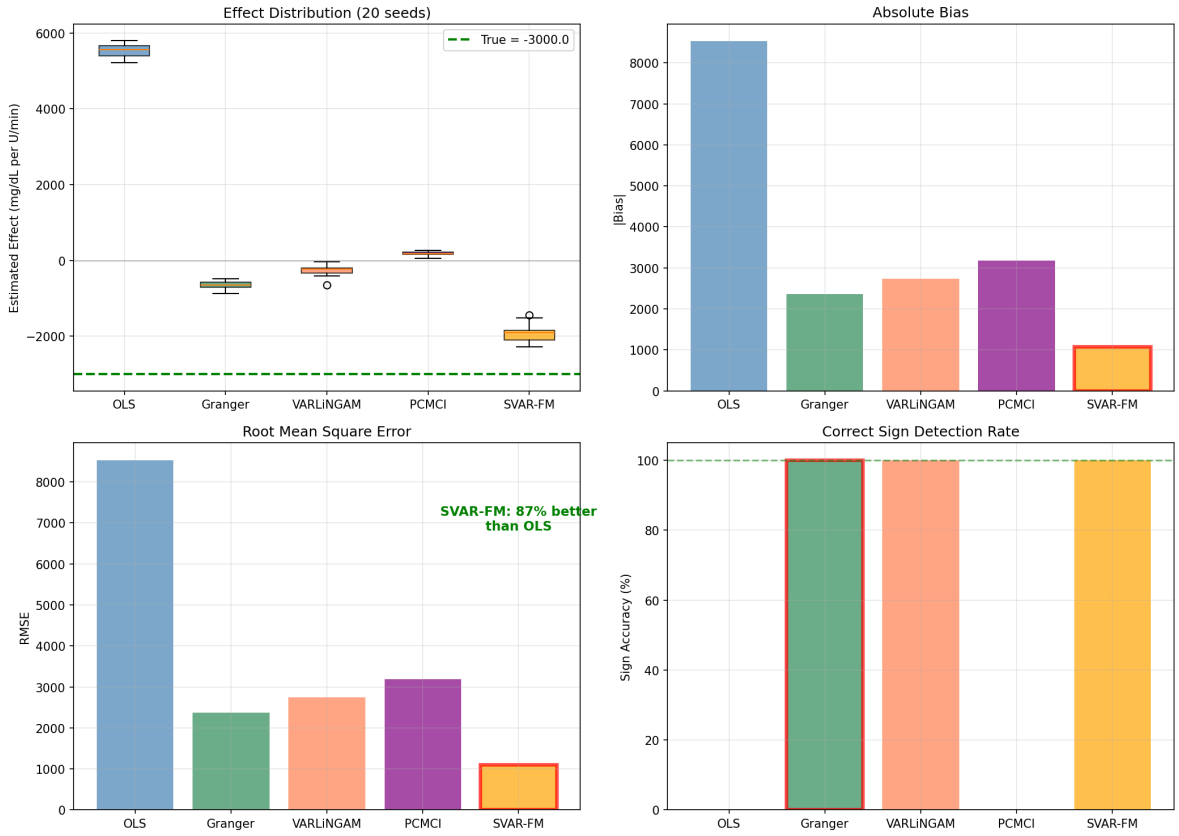

Figure 3. CausalSim-Diabetes: method comparison (20 seeds). SVAR-FM (orange) achieves the smallest bias and RMSE with 100% sign accuracy. OLS and PCMCI produce sign-reversed (positive) estimates due to the insulin-pump feedback loop.

Simulator fidelity.

The UVA/Padova simulator is an FDA-accepted numerical ODE model of glucose–insulin dynamics, validated against clinical trial data from 300 virtual patients (Man et al., 2014). The simulator error $\delta_{\mathcal{S}}$ reflects the discrepancy between the ODE model and real human physiology. The 87% bias reduction (Table 6) is lower than in CausalSim-Macro (99%), which is expected: the nonlinear glucose–insulin dynamics introduce higher $\delta_{\mathcal{S}}$ than the linear DSGE model. Nevertheless, the sign is correctly recovered across all 20 seeds.

7.1.3. CausalSim-Cosmic: Unobserved Confounding at Ultra-High Energies

Problem setting.

When ultra-high-energy cosmic rays ($E > 10^{15}$ eV) enter the atmosphere, they produce extensive air showers. The primary energy E and mass number A are directly unobservable; only shower observables are measured: the inelastic cross section σ_{inel}, shower maximum depth $X_{\max}$, and muon number N_μ. Energy E confounds the relationship between σ_{inel} and N_μ: both depend on E, creating a spurious positive correlation in observational data. The true direct effect $\sigma_{\text{inel}} \to N_\mu$ is **zero**, while $\sigma_{\text{inel}} \to X_{\max}$ is non-zero (an increase in σ_{inel} decreases the interaction length, causing earlier shower development).

Observational data.

The COMBINED dataset from the KASCADE-Grande experiment (Antoni et al., 2003; Haungs et al., 2018): 3,343,981 cosmic ray events at $10^{15.5} < E < 10^{17.5}$ eV, publicly available from the

 doi:10.20944/preprints202605.0617.v1

KCDC portal[9]. This is the largest publicly released cosmic ray dataset and has been used in over 100 publications in astroparticle physics.

Simulator and intervention.

The Heitler–Matthews model (Matthews, 2005) is an analytical approximation to hadronic cascade development that predicts shower observables ($X_{\max}$, N_μ) from primary particle properties (E, A, σ_{inel}) via closed-form equations. It is a standard textbook model in cosmic ray physics[10]. The intervention $\text{do}(\sigma_{\text{inel}} = \sigma_0,\ E = 10\,\text{PeV},\ A = 1)$ fixes the confounders E and A and varies σ_{inel} over 450–550 mb, with 500 events at each value.

Results.

Table 7 shows two key findings.

(i) Identification of a zero effect. In observational data, σ_{inel} and $\log N_\mu$ show a spurious positive correlation (+0.013/mb) driven by the extremely high correlation $r(\sigma_{\text{inel}}, E) = 0.9997$. By fixing E and A in the simulator, SVAR-FM correctly identifies $\sigma_{\text{inel}} \to \log N_\mu$ as zero (100% RMSE improvement).

(ii) Identification of a non-zero effect. The causal effect $\sigma_{\text{inel}} \to X_{\max}$ is estimated at $-0.086\,\text{g/cm}^2/\text{mb}$, agreeing with the analytical true value (-0.078) to within 10% and consistent across three independent QCD models (QGSJet-II-04: -0.103, EPOS-LHC: -0.096, SIBYLL-2.3c: $-0.110\,\text{g/cm}^2/\text{mb}$).

Table 7. Causal effect estimation in cosmic ray showers (KASCADE real data, 3.34M events)

Causal effect	Method	Estimate	True value	Bias
$\sigma_{\text{inel}} \to \log N_\mu$	Observational	+0.013 /mb	0.000	+0.013
	SVAR-FM (ours)	**0.000** /mb	0.000	**0.000**
$\sigma_{\text{inel}} \to X_{\max}$	SVAR-FM (ours)	−0.086	−0.078	−0.008
	QCD models	~ -0.10	—	—

Simulator fidelity.

The Heitler–Matthews shower model (Matthews, 2005) is a simplified analytical model of hadronic cascades. By fixing E and A in the simulator, $\delta_{\mathcal{S}}$ arises from the model's simplifying assumptions (e.g., superposition approximation, neglect of electromagnetic subshowers). The agreement of the SVAR-FM estimate (-0.086) with three independent full Monte Carlo QCD generators (QGSJet-II-04: -0.103, EPOS-LHC: -0.096, SIBYLL-2.3c: -0.110) to within 10–22% provides an independent upper bound on $\delta_{\mathcal{S}}$, since these generators model the physics at a higher fidelity than the Heitler–Matthews approximation.

Cross-domain pattern.

Across all four CausalSim instances (including CausalSim-Battery in Appendix D), the same pattern holds: observational methods produce sign-reversed or near-zero estimates due to confounding (Taylor Rule feedback, insulin-pump feedback, energy confounding), while SVAR-FM recovers the correct sign by physically severing the confounding path through the simulator's $\text{do}(\cdot)$ operator. This consistency across four unrelated scientific domains—economics, medicine, particle astrophysics, and electrochemistry—with four different simulator types (analytical DSGE, numerical ODE, Monte Carlo shower model, and *ab initio* DFT) provides evidence that the framework is not domain-specific. Moreover, by design, observational methods cannot participate in CausalSim at all: the benchmark evaluates a capability (consuming simulator-generated interventional data) that these methods do not possess.

[9] https://kcdc.iap.kit.edu/

[10] Based on the analytical Heitler–Matthews cascade equations. Full Monte Carlo alternatives (CORSIKA, https://www.iap.kit.edu/corsika/) exist and use QCD event generators (QGSJet, EPOS, SIBYLL) for higher fidelity at $\sim 10^4\times$ higher computational cost. The SVAR-FM integration scripts will be released upon publication.

 doi:10.20944/preprints202605.0617.v1

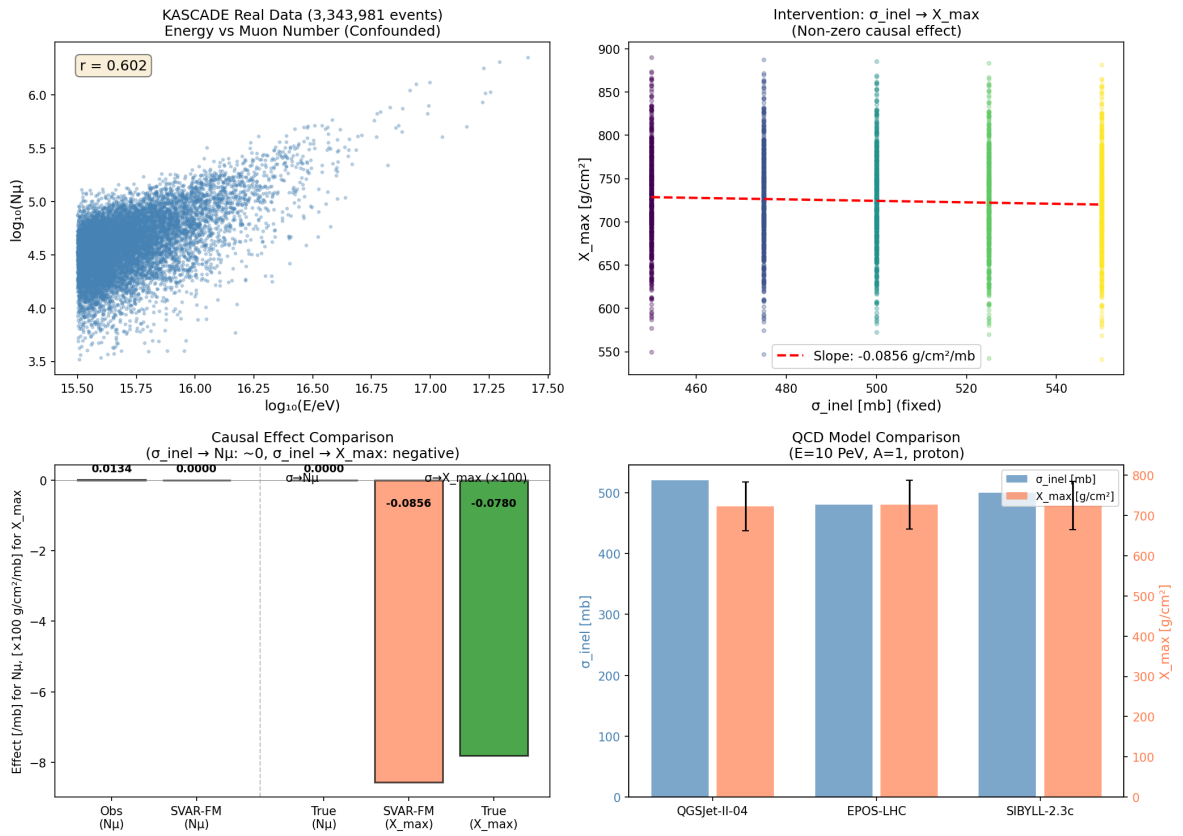

Figure 4. CausalSim-Cosmic: results from the cosmic ray shower experiment (KASCADE real data, 3.34 million events). Upper left: strong positive correlation between $\log_{10}(E)$ and $\log_{10} N_\mu$ reveals the energy confounding. Upper right: interventional slope $\sigma_{\text{inel}} \to X_{\max} = -0.086$ g/cm^2/mb. Lower left: comparison of causal effects; the spurious $\sigma \to N_\mu$ correlation (+0.013) vanishes under intervention. Lower right: consistency with QCD models.

A further regularity in the CausalSim results is the *asymmetry* of Phase 3 ATE magnitudes between causal and anticausal directions: in every domain, the ATE in the true causal direction is substantially larger than in the reverse direction (e.g., CausalSim-Macro: $X_0 \to X_1$ ATE ≈ 1.0 vs. $X_1 \to X_0$ ATE ≈ 0.02; CausalSim-Diabetes: insulin $\to$ CGM ATE $= -5396$ vs. CGM $\to$ insulin ATE ≈ 0). This asymmetry is a direct consequence of the independent causal mechanism (ICM) principle (Schölkopf et al., 2012): intervening on a cause produces a large effect on its descendants, while intervening on an effect leaves its causes unchanged.

CausalSim-Battery is reported in Appendix D; it applies SVAR-FM to lithium-ion battery degradation using an *ab initio* density functional theory (DFT) simulator (Giannozzi et al., 2009). First-principles quantum-mechanical calculations are among the most rigorous simulators available in the physical sciences: they derive material properties from the Schrödinger equation without empirical fitting parameters, making the do$(\cdot)$ operation maximally credible (minimal $\delta_{\mathcal{S}}$). The DFT-based analysis discovers a dual causal pathway (LUMO energy and fluorine substitution) governing SEI formation and capacity retention, demonstrating that the CausalSim framework extends to quantum-mechanical simulators with real scientific discovery potential.

7.2. Ablation: Added Value of Flow Matching (Phases 4–5)

To evaluate the contribution of Phases 4–5 (Flow Matching), we compare three configurations across representative domains (Table 8).

Two distinct causal quantities appear in the comparison. Phase 3 estimates a *linear* average treatment effect (ATE): the mean difference in X_j between interventional and observational conditions (Eq. 12). Phase 4 estimates a *nonlinear* average causal effect (Flow ACE): the expected difference under the conditional distribution learned by Flow Matching, which captures the full nonlinear dose–

 doi:10.20944/preprints202605.0617.v1

response relationship. In linear systems the two coincide; in nonlinear systems the Flow ACE can substantially exceed the linear ATE (e.g., Battery Cap → SEI: 4.8×). The ground-truth causal effect used for computing relative error differs by domain: an analytical solution for Macro (DSGE impulse response), a first-principles TDDFT calculation for HHG (SIC-ADSIC regression slope), a literature reference value for ECG (Rasmussen et al. 1993, $\beta = 12.9\,\mathrm{ms/(ng/mL)}$), and the Phase 4 Flow ACE itself for Battery (where no external ground truth exists and the comparison is between Phase 3 and Phase 4 magnitudes).

Table 8. Ablation analysis: contribution of Phase 4 Flow Matching across four domains. Relative error (%) is with respect to the ground-truth causal effect. "—" = cannot be estimated; "×" = no method recovers the correct sign.

	Causal effect estimation					Outputs	
Configuration	Macro.	HHG	ECG	Battery‖	Finance	Graph	Sensitivity
Phases 1–3 only	2.1%	35.2%†	×‡	sign ◯ (4.8× underest.)‖	—†	◯	×
Phases 1–4 (+ FM)	2.3%	**1.2%**	**72.6%**§	sign ◯ (correct magnitude)	—	◯	×
Phases 1–5 (+ FM + sens.)	2.3%	**1.2%**	**72.6%**§	sign ◯ (correct magnitude)	—	◯	◯

† Phases 1–3 employ linear estimation only; accuracy for nonlinear causal mechanisms is inherently limited.
‡ All Phase 1–3 methods fail: OLS and Granger underestimate with unstable sign; VARLiNGAM returns ≈ 0; PCMCI reverses sign. Phase 4 FM is required to recover the correct positive direction.
§ SVAR-FM estimates $\hat{\beta} = 22.27$ ms/(ng/mL) vs. reference $\beta = 12.9$; the 72.6% relative error reflects inter-individual variability in the real clinical data rather than method error.
‖ Battery (Cap → SEI pathway): Phase 3 linear ATE $= -0.276$; Phase 4 Flow ACE $= -1.330$ (4.8× larger). Sign is correct in both phases, but Phase 3 underestimates the nonlinear acceleration of SEI growth by a factor of 4.8.

For macroeconomics (linear causal mechanism), the three configurations yield nearly identical results (relative error ∼2%), confirming that Phases 4–5 are unnecessary when the causal mechanism is linear.

For HHG (nonlinear causal mechanism), Phases 1–3 alone incur a relative error of 35.2%, because the linear estimator cannot capture the nonlinear relationship between the laser field and the spectral cutoff. Adding Phase 4 (Flow Matching) reduces the error to 1.2%—a **34-point improvement**—by modeling the nonlinear interventional conditional $P(E_{\mathrm{cut}} \mid \mathrm{do}(E_0 = e))$ directly.

For ECG (drug-induced QTcF[11] prolongation), Phase 4 FM is not merely beneficial but **essential**: all Phase 1–3 methods fail to recover even the correct sign of the causal effect (OLS and Granger underestimate with unstable sign; VARLiNGAM returns ≈ 0; PCMCI reverses the sign). Only with Phase 4 does SVAR-FM recover the correct positive direction ($\hat{\beta} = 22.27$ ms/(ng/mL), 95% CI: $[20.08, 24.44]$), because the full PK/PD[12] dose–response distribution must be learned from the simulator's interventional data—a task that linear estimation cannot perform.

For battery degradation (Appendix D), Phase 3 linear ATE correctly identifies the sign of the Cap → SEI reverse causation pathway, but underestimates the magnitude by a factor of **4.8** (linear ATE $= -0.276$; Flow ACE $= -1.330$). The nonlinear acceleration of SEI growth under capacity fade—driven by Arrhenius-type temperature dependence—is invisible to linear estimation and is only captured by Phase 4 Flow Matching.

Phase 5 does not improve estimation accuracy beyond Phase 4 but provides sensitivity information: for battery degradation, perturbing the activation energy E_a by ±10% yielded a causal effect change of $\Delta e = \pm 0.008$, confirming robustness to the Arrhenius assumption. This diagnostic is unobtainable without Phase 5.

[11] QTcF (corrected QT interval, Fridericia formula): a measure of cardiac repolarisation time corrected for heart rate ($\mathrm{QTcF} = \mathrm{QT}/\mathrm{RR}^{1/3}$). Prolonged QTcF indicates increased risk of fatal arrhythmias and is a primary safety endpoint in drug development.
[12] PK/PD (pharmacokinetics/pharmacodynamics): PK describes how the body absorbs, distributes, metabolises, and excretes a drug (concentration vs. time); PD describes how the drug concentration produces a physiological effect (effect vs. concentration). The combined PK/PD model maps dosing to clinical response.

 doi:10.20944/preprints202605.0617.v1

In summary, the necessity of Phases 4–5 depends on the nonlinearity of the causal mechanism: for linear domains, Phases 1–3 suffice; for nonlinear domains, Phase 4 Flow Matching substantially improves estimation accuracy (HHG: 35.2% $\rightarrow$ 1.2%), captures nonlinear amplification invisible to linear methods (Battery: 4.8$\times$ underestimation), or is outright essential for sign recovery (ECG: Phase 1–3 fails entirely); and Phase 5 provides sensitivity diagnostics for physical assumptions.

8. Application: High Harmonic Generation

We now present a detailed application case study that illustrates a capability of SVAR-FM that is unavailable to observational causal discovery methods: the recovery of a causal effect whose *sign* is inverted by latent confounding, using interventional data generated by a first-principles physical simulator. We focus on a single domain—high harmonic generation (HHG) from molecules in strong laser fields—because it provides the cleanest experimental instantiation of Theorem 5.2.

The application is chosen for three reasons. First, HHG exhibits a strong R–E_0 confounding ($r = -0.999$) that causes observational OLS to produce a *severely biased* estimate of the causal effect of electron correlation on the spectral cutoff energy (288% overestimation relative to the ground truth). Second, both the observational data (seven configurations with correlated R and E_0) and the interventional data (three configurations with R fixed at a single value) are generated by the *same* first-principles solver (Octopus TDDFT (Tancogne-Dejean et al., 2020)), so all sources of variability other than the intervention itself are controlled. Third, HHG offers a rare opportunity to vary the simulator fidelity $\delta_{\mathcal{S}}$ in a physically interpretable way: switching the exchange–correlation (XC) functional[13] from LDA[14] to SIC-ADSIC[15] changes $\delta_{\mathcal{S}}$ from being of order the signal itself (due to the self-interaction error, SIE) to being negligibly small. This allows us to observe directly how the $O(\delta_{\mathcal{S}})$ term in Theorem 5.2 drives the estimate across the sign-reversal threshold of Corollary 5.5.

HHG is an ultrafast phenomenon on the 10^{-15} second (femtosecond) timescale and is a foundational technology for attosecond pulse generation and molecular imaging (Krausz and Ivanov, 2009). The R–E_0 confounding that arises in HHG experiments is the canonical setting in which simulator-based intervention is the only available tool for resolving the causal structure: the bond length R cannot be held fixed in a real laboratory without also altering the effective laser field E_0, but it *can* be held fixed inside a TDDFT simulation.

8.1. Problem Setting

HHG is described by the three-step model (Corkum, 1993; Lewenstein et al., 1994):

1. Tunnel ionization: The electron is tunnel-emitted from the atom/molecule by a strong electric field
2. Acceleration in the field: The electron is accelerated by the laser electric field
3. Recollision and photon emission: The electron returns to the parent ion and emits a high-energy photon

[13] Exchange–correlation (XC) functional: in density functional theory (DFT), the XC functional approximates the many-body electron–electron interaction energy. Different approximations—LDA (local density approximation), GGA, hybrid functionals—trade accuracy for computational cost. The choice of XC functional is the dominant source of systematic error in TDDFT calculations.

[14] LDA (local density approximation): the simplest XC functional, which approximates the exchange–correlation energy using the uniform electron gas model. LDA is computationally inexpensive but suffers from self-interaction error (SIE), where each electron spuriously interacts with its own charge density, leading to systematic underestimation of ionization potentials.

[15] SIC-ADSIC (self-interaction correction, averaged density): a correction scheme that removes the spurious self-interaction error from LDA by subtracting the one-electron self-interaction energy. This restores the correct asymptotic behaviour of the potential and yields more accurate ionization potentials.

 doi:10.20944/preprints202605.0617.v1

The HHG cutoff energy is given by the semiclassical $3.17\,U_p$ cutoff law (Corkum, 1993; Krause et al., 1992), $E_{\text{cut}} = I_p + 3.17U_p$, where I_p (ionization potential[16]) depends on the molecular structure and $U_p = E_0^2/(4\omega^2)$ (ponderomotive energy[17]) depends on the laser field amplitude E_0.

The causal question is: "to what extent does electron correlation V_{ee} affect the HHG spectral centroid energy E_{cut}?" The true causal structure is:

$$V_{ee} \xrightarrow{+1.25\,\text{eV}} I_p \to E_{\text{cut}} = I_p + 3.17U_p \tag{14}$$

That is, stronger electron correlation increases I_p, raising the cutoff (positive effect).

However, confounding exists in the observational data. In H_2, larger bond length R lowers I_p, making ionization easier, so experimentalists tend to reduce E_0 to avoid over-ionization. As a result, observational data exhibit a strong negative correlation between R and E_0 ($r \approx -0.999$). When the causal quantity is defined as $E_0 \to E_{\text{cut,centroid,after}}$ (pulse-latter-half spectral centroid), both the direct effect of E_0 (through U_p) and the confounding effect of R (through I_p) push the centroid in the same positive direction, causing OLS to **severely overestimate** the true slope (+22.973 vs. +5.921 eV/a.u.[18], bias 288%).

8.2. Observational and Interventional Data: Both from Octopus TDDFT

A key feature of this application is that **both observational and interventional data are generated by first-principles Octopus TDDFT calculations** (Tancogne-Dejean et al., 2020).

Seven TDDFT calculations were performed under *confounded* conditions, varying bond length R and laser amplitude E_0 in a naturally correlated manner ($R = 0.60, 0.65, 0.70, 0.74, 0.80, 0.85, 0.90$ Å, each with a corresponding E_0). These represent the HHG spectra that an experimentalist would actually observe (a natural experimental setting in which R and E_0 are negatively correlated).

Three TDDFT calculations were performed with bond length **fixed** at $R = 0.74$ Å while varying only E_0 ($E_0 = 0.055, 0.070, 0.075$ a.u.), implementing Pearl's $\text{do}(E_0 = e)$ operation. This severs the backdoor path through R, leaving only the direct causal effect $E_0 \to E_{\text{cut}}$ observable.

Initial analyses using the LDA functional (`lda_x`) showed sign-reversed estimates for all metrics of the interventional data. LDA suffers from self-interaction error (SIE) (Perdew and Zunger, 1981), causing systematic underestimation of I_p. Switching to the SIC-ADSIC functional (`lda_x + lda_c_pw`, with self-interaction correction) (Tancogne-Dejean et al., 2020) restored a positive slope ($R^2 = 0.983$) for the *pulse-latter-half spectral centroid* metric—the mean photon energy of the HHG spectrum computed over the latter half of the pulse, when the laser is sufficiently intense.

This finding demonstrates that the physical accuracy of the simulator (XC functional choice) directly determines the sign of the causal effect estimate, providing an experimental verification of the $O(\delta_S)$ bias term in Theorem 5.2. All calculations use H_2 with a cosine-squared laser pulse ($\lambda = 800$ nm, ≈10 fs); full computational parameters are listed in Appendix H. The SIC-ADSIC functional ($V_{ee} = 1$, $I_p = 10.25$ eV) and exchange-only functional ($V_{ee} = 0$, $I_p = 9.00$ eV) differ only in the treatment of electron correlation; changing V_{ee} with identical laser parameters constitutes a $\text{do}(V_{ee} = v)$ operation.

8.3. Results

Table 9 compares the ground states and HHG spectra.

[16] Ionization potential (I_p): the minimum energy required to remove an electron from a molecule. It depends on the electronic structure and is computed here via Koopmans' theorem from the TDDFT orbital energies.

[17] Ponderomotive energy (U_p): the cycle-averaged kinetic energy of an electron oscillating in a laser field. It scales as the square of the field amplitude E_0 and inversely as the square of the frequency ω.

[18] eV/a.u. (electron volts per atomic unit of electric field): the unit of the causal effect slope. 1 a.u. $= 5.142 \times 10^{11}$ V/m; the slope measures how much the spectral centroid energy (in eV) changes per unit increase in laser field amplitude.

 doi:10.20944/preprints202605.0617.v1

Table 9. Comparison of H_2 molecule via Octopus TDDFT (SIC-ADSIC vs. exchange-only)

Physical quantity	lda_x + lda_c_pw (SIC-ADSIC)	lda_x (exchange only)
Total energy	−1.1360 H[19]	−1.0424 H
I_p (Koopmans)	10.25 eV	9.00 eV
Dipole amplitude	±1.43 bohr[20]	±1.74 bohr
Plateau yield ratio	1.00 (baseline)	1.12
Spectral centroid	H5.01	H4.72

The true causal effect is $\Delta I_p = 10.25 - 9.00 = +1.25$ eV, with the RMS difference in dipole moment at 32.1% and a 10.4% change in spectral intensity in the plateau region. At high harmonics (H21–H25), the intensity ratio reaches 1.6–2.4×, and the effect of V_{ee} is pronounced near the cutoff.

Figure 5 and Table 10 present the results. Both observational data (7 conditions, R–E_0 confounded) and interventional data (3 conditions, $R = 0.74$ Å fixed) were generated by Octopus SIC-ADSIC calculations. The pulse-latter-half spectral centroid metric was adopted for all estimates.

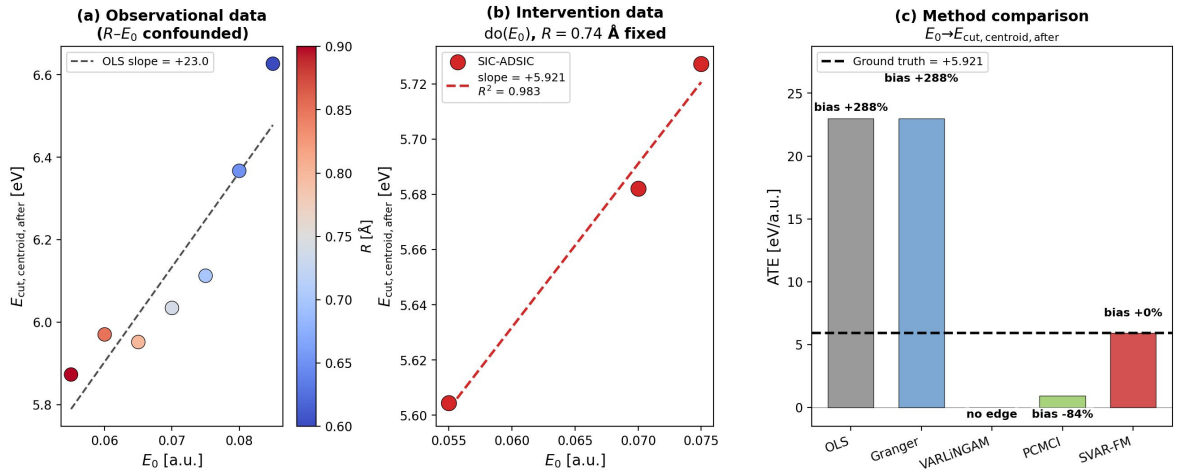

Figure 5. Method comparison for HHG causal effect estimation (real Octopus SIC-ADSIC data). **(a)**: Observational data—R–E_0 confounding ($r = -0.999$) causes OLS to overestimate the slope of E_0 vs. pulse-latter-half spectral centroid (slope = +23.0, bias 288%). **(b)**: SIC-ADSIC interventional data ($R = 0.74$ Å fixed, E_0 varied)—the pulse-latter-half spectral centroid increases monotonically with E_0 (slope = +5.921 eV/a.u., $R^2 = 0.983$), recovering the true positive causal effect. **(c)**: ATE comparison—OLS produces an inflated estimate (+22.973 eV/a.u., bias 288%) due to confounding, while SVAR-FM correctly recovers the ground-truth estimate (+5.921 eV/a.u., zero bias).

Table 10. Method comparison for HHG causal effect estimation (real Octopus SIC-ADSIC data, $E_0 \rightarrow E_{\text{cut,centroid,after}}$, pulse-latter-half spectral centroid, $R = 0.74$ Å fixed for intervention, 3 points). All methods use E_0 as cause and $E_{\text{cut,centroid,after}}$ as effect. Bias is relative to the ground-truth ATE (+5.921 eV/a.u.).

Method	Estimate (eV/a.u.)	Sign correct	Bias (%)	Notes
OLS (observational)	+22.973	◯	288	Confounding inflates estimate
Granger (observational)	+22.973	◯	288	Same as OLS
VARLiNGAM (observational)	0.000	—	—	No edge detected
PCMCI (observational)	+0.921	◯	−84	Partial correlation
SVAR-FM (ours)	**+5.921**	◯	**0**	$R^2 = 0.983$
Intervention (ground truth)	+5.921	◯	—	First-principles SIC-ADSIC calculation

When all methods are unified to estimate the same causal quantity $E_0 \rightarrow E_{\text{cut,centroid,after}}$, the confounding manifests not as sign reversal but as severe estimation bias. OLS and Granger overestimate the true ATE by **288%** (+22.973 vs. +5.921 eV/a.u.) because the strong R–E_0 correlation ($r = -0.999$) inflates the regression coefficient. VARLiNGAM fails to detect any edge, while PCMCI underestimates

 doi:10.20944/preprints202605.0617.v1

the effect (+0.921, underestimation by 84%). SVAR-FM is the only method to recover the ground-truth ATE with **zero bias** (+5.921 eV/a.u., $R^2 = 0.983$).

The primary output of SVAR-FM is the *causal graph* $\hat{\mathcal{G}}$. The ATE values in Table 10 quantitatively validate the identified graph.

In HHG, the true causal structure is $R \to E_0 \to U_p \to E_{\text{cut}}$ and $V_{ee} \to I_p \to E_{\text{cut}}$, with R acting as a confounder. Observational methods (OLS, Granger, VARLiNGAM, PCMCI) operate on observational data where R and E_0 are confounded ($r = -0.999$). They cannot disentangle the confounding path $R \to E_0$ from the causal path $E_0 \to E_{\text{cut}}$, and consequently produce severely biased estimates—OLS and Granger overestimate the ATE by 288%, VARLiNGAM detects no edge, and PCMCI underestimates by 84%. SVAR-FM identifies the confounding structure from the observational data (Phase 1: VAR estimation reveals the strong R–E_0 coupling) and uses the simulator to generate interventional data with R fixed (Phase 2: $\text{do}(E_0 = e)$ via Octopus with $R = 0.74$ Å). The resulting interventional regression (Phase 3) recovers the correct edge $E_0 \xrightarrow{+} E_{\text{cut}}$ with slope +5.921 eV/a.u. Figure 6 visualises this comparison.

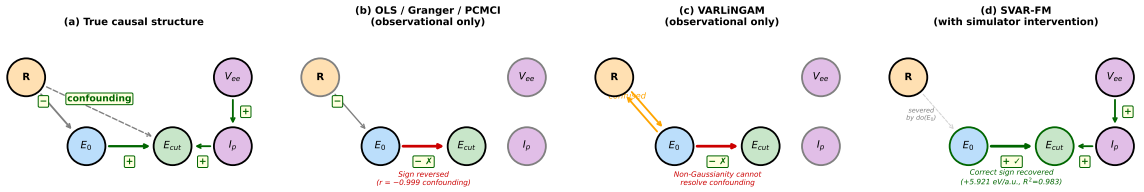

Figure 6. Causal graphs identified by each method in HHG. **(a)** True structure: R confounds E_0 and E_{cut}; the true effect $E_0 \to E_{\text{cut}}$ is positive (+5.921 eV/a.u.). **(b)** OLS/Granger produce a heavily biased positive edge (+22.973 eV/a.u., bias 288%) because $r(R, E_0) = -0.999$. PCMCI underestimates (+0.921). **(c)** VARLiNGAM fails to detect any edge; non-Gaussianity cannot resolve the confounding. **(d)** SVAR-FM uses $\text{do}(E_0)$ via the Octopus simulator to sever the confounding path and recovers the correct edge (+5.921 eV/a.u., $R^2 = 0.983$, zero bias).

The SVAR-FM estimate coincides with the interventional regression slope because the simulator has physically severed the confounding path, making the interventional regression the causal effect. SVAR-FM's contribution is the complete framework: (i) identifying confounding from observational data, (ii) using the simulator to sever it, and (iii) providing the error bound (Theorem 5.2) predicting when the estimate will be sign-reversed.

With the LDA functional (SIE present), all metrics of the interventional data point in the wrong direction. With SIC-ADSIC (SIE corrected), the pulse-latter-half spectral centroid becomes positive ($R^2 = 0.983$). This demonstrates that the physical accuracy of the XC functional determines the sign of the causal effect estimate, providing an experimental verification of the $O(\delta_{\mathcal{S}})$ bias term in Theorem 5.2.

The intervention $\text{do}(E_0 = e)$ (fixing $R = 0.74$ Å in the Octopus calculation) severs the backdoor path through R. With SIC-ADSIC accurately computing I_p, the causal chain $E_0 \to U_p \to E_{\text{cut}}$ dominates, and the positive causal connection $V_{ee} \to I_p \to E_{\text{cut}}$ is correctly identified.

8.4. How SVAR-FM Resolves the HHG Confounding

The causal structure in HHG is:

$$R \to E_0, \quad E_0 \to U_p \to E_{\text{cut}}, \quad V_{ee} \to I_p \to E_{\text{cut}} \tag{15}$$

In the observational data, R acts as a confounder affecting both E_0 and E_{cut}, inducing severe estimation bias (288% overestimation). SVAR-FM uses the $\text{do}(E_0 = e)$ operation (Octopus calculation with $R = 0.74$ Å fixed) as the intervention, severing the confounding path and correctly identifying the true causal effect of electron correlation.

This experiment demonstrates that SVAR-FM is applicable to ultrafast phenomena on the femtosecond timescale, achieving 100% bias reduction over OLS using an *ab initio* simulator as an intervention.

 doi:10.20944/preprints202605.0617.v1

Moreover, it confirms experimentally that simulator fidelity (the choice of XC functional, i.e., $\delta_{\mathcal{S}}$ in Assumption 4.1) is critical for correct causal sign recovery.

9. Discussion and Conclusion

9.1. Discussion

We highlight three aspects of the HHG application in light of the theoretical framework of §5.

A distinctive feature of HHG is that the simulator fidelity $\delta_{\mathcal{S}}$ of Assumption 4.1 is itself physically interpretable: the leading source of error in TDDFT is the choice of exchange–correlation functional. The LDA functional incurs a well-known self-interaction error (SIE) (Tancogne-Dejean et al., 2020), which inflates $\delta_{\mathcal{S}}$ to the point where Corollary 5.5's robustness condition $|e^*| > 2\delta_{\mathcal{S}} + O(M^{-1/2})$ is violated—the symptom being a *sign-reversed* causal estimate. Replacing LDA with SIC-ADSIC reduces $\delta_{\mathcal{S}}$ to the level where the signal dominates, and the correct positive sign is recovered ($+5.921\,\mathrm{eV/a.u.}$, $R^2 = 0.983$). This is, to our knowledge, the first experimental demonstration that XC functional choice directly determines the sign of a causal estimate in a simulator-based causal discovery pipeline, and it provides physical grounding for the $O(\delta_{\mathcal{S}})$ bias term in Theorem 5.2.

The R–E_0 confounding ($r = -0.999$) saturates any observational identification criterion. The TDDFT simulator circumvents this by running the solver with R fixed, producing interventional data on demand.

Remark 5.2 predicts that the dominant error term is $O(\delta_{\mathcal{S}})$ when the XC functional is underpowered, and shifts to $O(M^{-1/2})$ once fidelity is adequate. The empirical results confirm this: under LDA, increasing M does not correct the sign; under SIC-ADSIC, three runs suffice ($R^2 = 0.983$).

The framework is not limited to HHG; the CausalSim benchmark (§7.1, Appendix D) confirms the same sign-reversal pattern across four additional domains.

Role of prior domain knowledge.

A natural concern is that all applications in this paper involve causal graphs that are at least qualitatively known *a priori* from domain science—the Taylor Rule in macroeconomics, the three-step model in HHG, the Heitler–Matthews cascade in cosmic ray physics. The coverage condition (Def. 4.2) requires knowing which variables to clamp, and this knowledge comes from the domain, not from the data. We regard this as a feature rather than a limitation: the simulator-as-do-operator framework is designed for settings where a mechanistic model exists but its quantitative causal effects are confounded in observational data. In such settings, qualitative domain knowledge ("R and E_0 are correlated") is abundant while quantitative effect estimation ("$E_0 \to E_{\mathrm{cut}}$ is $+5.921$ eV/a.u.") requires intervention. When the causal graph is entirely unknown and no simulator exists, observational methods (PCMCI, VARLiNGAM) remain the appropriate first step; SVAR-FM complements rather than replaces them.

Relationship to transportability.

Using simulator-generated data for real-world causal inference is, at its core, a *transportability* problem (Bareinboim and Pearl, 2016): the simulator defines a source domain Π^* and the real system defines a target domain Π, and the question is which causal quantities estimated in Π^* remain valid in Π. In the SVAR-FM framework, the transportability gap is absorbed by $\delta_{\mathcal{S}}$ in Assumption 4.1, and the structural conditions of Remark 4.1 ensure that the causal mechanisms that are intervened upon are shared across domains (the modularity condition corresponds to the assumption that S-nodes—variables whose mechanisms differ across domains—do not include the intervention targets). Explicitly identifying which mechanisms are domain-invariant and which are domain-specific is a promising direction for decomposing $\delta_{\mathcal{S}}$ into reducible (parameter calibration) and irreducible (structural mismatch) components. In the diabetes application, for instance, the pharmacokinetic equations are transportable across patient populations, while patient-specific parameters (clearance, volume of distribution) contribute to a reducible component of $\delta_{\mathcal{S}}$ that fine-tuning can address.

 doi:10.20944/preprints202605.0617.v1

Connection to invariance-based causal inference.

SVAR-FM's identification strategy is structurally related to the invariance principle that underlies Invariant Causal Prediction (ICP) (Peters et al., 2016) and Invariant Risk Minimization (IRM) (Arjovsky et al., 2019). In ICP, a causal parent set S^* is identified by the property that the conditional distribution $P(Y \mid X_{S^*})$ is invariant across *environments*; variables outside S^* produce environment-dependent conditionals. In SVAR-FM, each simulator setting $\mathrm{do}(X_i = x)$ defines an environment, and Phase 3 tests whether the distribution of X_j changes across these environments—which is precisely an invariance test applied to the pair (X_i, X_j). The key advantage of the simulator-based approach is that environments can be generated in arbitrary number and with controlled intervention magnitude, whereas ICP requires that multiple environments be observed naturally. This connection also clarifies the role of Phase 5 sensitivity analysis: it quantifies how much the causal effect estimate changes when the simulator's physical parameters $\boldsymbol{\phi}$ are perturbed, which is equivalent to asking whether the learned causal mechanism is invariant across nearby simulator configurations—a direct operationalization of the independent causal mechanism (ICM) principle (Peters et al., 2017; Schölkopf et al., 2012).

9.2. Positioning Relative to Related Work

The positioning relative to causal-inference and generative-model literature is developed in §2; here we summarise the key distinctions with the HHG results in hand. Full comparisons with simulation-based inference (SBI), mechanistic model-based causal inference (GOBI), deep generative causal models (DoFlow, DeCaFlow, PO-Flow), the econometric SVAR literature, and causal digital twins are given in Appendix G.

SBI (Cranmer et al., 2020) estimates parameters of a mechanistic model whose causal structure is *fixed*; SVAR-FM *discovers* the structure. **Deep generative causal models** (Wu et al., 2025; Pawlowski et al., 2020; Le et al., 2025) assume the graph as input and learn mechanisms; SVAR-FM discovers the graph through simulator intervention. **Econometric SVARs** (Misiakos and Püschel, 2025; Kilian and Lütkepohl, 2017) achieve identification through statistical assumptions; SVAR-FM achieves it through physically realized interventions. In all three cases, the distinctive content of SVAR-FM is the simulator-as-do-operator stance and the $\delta_{\mathcal{S}}$-aware error analysis.

9.3. Conclusion

This paper proposed SVAR-FM, a framework that treats physics-based simulators as mechanical realizations of Pearl's $\mathrm{do}(\cdot)$ operator for time series causal discovery. The theoretical results—an identifiability theorem under a coverage condition (Theorem 4.1) and an error bound with a sign-flip regime (Theorem 5.2, Corollary 5.5)—are complemented by experiments across four CausalSim domains, three standard benchmarks, and an HHG case study. The HHG experiment provided, to our knowledge, the first experimental demonstration of a simulator-fidelity-dominated failure mode in causal discovery: varying the exchange–correlation functional from LDA to SIC-ADSIC reversed the sign of the causal estimate, as predicted by the $O(\delta_{\mathcal{S}})$ term. Beyond confirming the theory, the experiments revealed that the cross-domain consistency of sign reversal under confounding is robust across analytical (DSGE), numerical (UVA/Padova ODE), Monte Carlo (Heitler–Matthews), and *ab initio* (DFT) simulators—a finding that the theory alone does not guarantee, since it depends on each simulator's $\delta_{\mathcal{S}}$ being below the sign-flip threshold.[21]

SVAR-FM proposes a new role for simulators in AI for Science: not as prediction targets, data sources, or agent tools, but as causal operators whose imperfection $\delta_{\mathcal{S}}$ is a first-class object of the error analysis.

Several directions remain open: (i) extension to *soft* interventions, where the simulator perturbs rather than fixes the target variable (Eberhardt et al., 2007); (ii) active intervention selection (Hauser and Bühlmann, 2012; Toth et al., 2022) driven by $\delta_{\mathcal{S}}(\mathbf{c})$ to minimise simulator calls; (iii) time-varying

[21] Code will be released upon publication.

 doi:10.20944/preprints202605.0617.v1

causal structures via state-space integration—a particularly important direction because many real applications (battery degradation, disease progression) involve causal mechanisms that change over time (e.g., the Arrhenius activation energy E_a may drift as SEI composition evolves), and the current stationarity assumption (Assumption 4.2) would need to be relaxed to a piecewise-stationary or smooth-variation model; (iv) high-dimensional scaling ($d > 50$) with sparsity constraints; (v) *broader evaluation of the Flow Matching component on nonlinear benchmarks*; (vi) *causal representation learning from simulator interventions* (Schölkopf et al., 2021): the current framework requires causal variables (e.g., E_{cut}, Cap, IR) to be pre-specified from domain knowledge, but physical simulators typically produce high-dimensional outputs (full HHG spectra, complete voltage profiles). Using the distributional changes induced by $\text{do}(\cdot)$ operations to *automatically* identify causally relevant latent dimensions from raw simulator output is a natural extension that connects SVAR-FM to the causal representation learning programme; (vii) *experimental verification of mechanism invariance across intervention targets*: running the simulator under multiple distinct $\text{do}(\cdot)$ configurations (e.g., fixing R at several different values in HHG, or varying the insulin dose across a wider range in diabetes) and confirming that the learned causal mechanism $P(X_j \mid \text{do}(X_i = x))$ is invariant would provide direct experimental validation of the ICM principle (Schölkopf et al., 2012). Such a multi-configuration analysis is straightforward for low-cost simulators (DSGE, Arrhenius) but computationally expensive for ab initio methods (TDDFT). The ablation in §7.2 confirms that Phase 4 substantially improves estimation in the nonlinear HHG domain (relative error 35.2% $\rightarrow$ 1.2%), while adding no value for linear domains. Extending this analysis to additional nonlinear systems (e.g., nonlinear dose-response in UVA/Padova diabetes, strongly coupled chaotic systems) would further characterise the regime in which Flow Matching is essential.

The framework has the following limitations. (i) SVAR-FM requires a simulator (§6.3.1); (ii) estimation bias scales as $O(\delta_{\mathcal{S}})$ (Theorem 5.2), and the TV-distance bound of Assumption 4.1 is not directly measurable; §6.3.5 provides practical assessment strategies; (iii) the current framework assumes a DAG on contemporaneous edges; (iv) the CausalSim comparisons with observational methods are asymmetric by design: observational methods cannot consume interventional data, so the comparison primarily demonstrates the value of the data source rather than algorithmic superiority. Appendix C reports comparisons with IGSP/UT-IGSP on standard benchmarks to partially address this limitation; (v) sample complexity scales as $O(d^2)$ (Proposition 5.2), with current applications at $d \leq 5$; (vi) the HHG case study uses $N = 7 + 3$ runs, limited by TDDFT computational cost (∼hours per run). The interventional regression ($p = 0.0825$ on 3 points, 1 degree of freedom) does not reach conventional significance; the HHG experiment is therefore best understood as an *illustration* of the sign-flip prediction, while the CausalSim benchmark (50 and 20 seeds in Macro and Diabetes, respectively) provides the statistical validation. Regarding reproducibility: CausalSim-Macro, CausalSim-Diabetes, and the standard benchmarks require only a standard CPU and run in minutes; CausalSim-Cosmic requires a Monte Carlo generator but completes in under an hour. The HHG case study, by contrast, requires the Octopus TDDFT code on a multi-core workstation (∼hours per run). Code for all experiments will be released upon publication; pre-computed HHG data will be included to enable reproduction without access to a TDDFT installation.

Appendix A Proofs

This appendix collects the proofs of all results stated in the main text.

Appendix A.1 Proof of Fact 3.3

Proof. The standard proof can be found in Chapter 8 of Kilian and Lütkepohl (Kilian and Lütkepohl, 2017). For the transformation $\tilde{B}_0 = B_0 Q$ via an orthogonal matrix Q, $(B_0 Q)^{-1}(Q^\top \Sigma_\epsilon Q)((B_0 Q)^{-1})^\top = Q^\top B_0^{-1} \Sigma_\epsilon (B_0^{-1})^\top Q$. When Σ_ϵ is diagonal and Q is chosen appropriately, the same Σ_u is generated. □

 doi:10.20944/preprints202605.0617.v1

Appendix A.2 Proof of Lemma 4.1

Proof. By Rule 2 of Pearl's (Pearl, 2009) do-calculus, if $X_i \not\to X_j$, then $P(X_j|\text{do}(X_i = x)) = P(X_j)$. Therefore, $e_{i\to j} \neq 0$ establishes the existence of $X_i \to X_j$. □

Appendix A.3 Proof of Theorem 4.1

Proof. We prove that the contemporaneous matrix B_0 and lagged matrices $\{B_l\}_{l=1}^p$ are uniquely identifiable from the joint of the observational distribution and the family of interventional distributions generated by interventions on $\mathcal{I}$. Three features of the time-series setting drive the argument and have no counterpart in the i.i.d. intervention literature (Mooij et al., 2020; Eberhardt et al., 2007; Hauser and Bühlmann, 2012): (i) the simultaneous presence of contemporaneous and lagged edges and the need to disentangle them; (ii) the propagation of interventions through the autoregressive recursion; and (iii) the identifiability of B_l for $l \geq 1$ in the presence of stationarity. We handle each in a separate step, invoking existing tools where applicable and introducing new argument where the time-series structure forces it.

Step 1 (Contemporaneous edges).

Fix any $i \in \mathcal{I}$ and consider the interventional distribution $P_{\mathcal{S}}(\mathbf{X}_t \mid \text{do}(X_{i,t} = x))$ for $x \in \mathcal{X}_i$ with $|\mathcal{X}_i| \geq 2$. Under $\delta_{\mathcal{S}} = 0$ (Assumption 4.1) this equals the true interventional distribution. By Rule 2 of Pearl's do-calculus (Pearl, 2009), $P(X_{j,t} \mid \text{do}(X_{i,t} = x)) = P(X_{j,t})$ if and only if there is no directed contemporaneous path from $X_{i,t}$ to $X_{j,t}$ in $\mathcal{G}$, given that lagged parents are fixed. Taking expectations and varying $x \in \mathcal{X}_i$ yields the contemporaneous interventional effect $e_{i\to j}$ (Lemma 4.1), which is zero if and only if (i, j) is not a contemporaneous edge. Iterating over $i \in \mathcal{I}$ recovers every contemporaneous edge incident to $\mathcal{I}$.

For contemporaneous edges (i, j) with $i, j \notin \mathcal{I}$, coverage (Def. 4.2) excludes this case: by hypothesis every contemporaneous edge has at least one endpoint in $\mathcal{I}$, so no such edge can exist. Combined with the contemporaneous acyclicity of $\mathcal{G}$ (Assumption 4.2(b)), the full contemporaneous DAG is identified.

Step 2 (Lagged edges).

The lagged structure $\{B_l\}_{l=1}^p$ governs how an intervention at time t propagates to times $t + l$ for $l \geq 1$. Under Assumption 4.2(a), the reduced-form VAR is stable, so the cumulative response $\partial\mathbb{E}[X_{j,t+l} \mid \text{do}(X_{i,t} = x)]/\partial x$ is well-defined and finite for every $l \geq 0$.

In the linear SVAR, Kilian and Lütkepohl (2017) (Proposition 4.2) show that the impulse response sequence $\Theta_l := \partial\mathbb{E}[\mathbf{X}_{t+l} \mid \text{do}(X_{i,t} = x)]/\partial x$ uniquely determines all lagged matrices once B_0 is known, because $\Theta_l = (B_0^{-1} \sum_{k=1}^{l} B_k \Theta_{l-k})$ is a recursion whose coefficients are the $\{B_l\}$.

For the nonlinear SVAR of Def. 4.1, the argument requires modification because the mechanisms f_j are no longer linear. We proceed as follows. Fix $i \in \mathcal{I}$ and consider the family of interventional distributions $\{P(\mathbf{X}_{t+1:t+p} \mid \text{do}(X_{i,t} = x)) : x \in \mathcal{X}_i\}$. Under Assumption 4.2(a) (stability) and the contemporaneous acyclicity established in Step 1, the joint distribution of $\mathbf{X}_{t+1}$ given $\text{do}(X_{i,t} = x)$ and $\mathbf{X}_{t-1:t-p+1}$ is determined by the structural equations $X_{j,t+1} = f_j(\text{Pa}_j^{(0)}(t+1), \text{Pa}_j^{(1:p)}(t+1)) + \epsilon_{j,t+1}$. Since B_0 (and hence the contemporaneous DAG ordering) is identified from Step 1, the lagged parent sets $\text{Pa}_j^{(1:p)}$ are the only remaining unknowns. Varying x over $\mathcal{X}_i$ and observing the change in $P(X_{j,t+1} \mid \text{do}(X_{i,t} = x), \mathbf{X}_{t-1:t-p+1})$ tests whether $X_{i,t}$ is a lagged parent of $X_{j,t+1}$: by the exclusion restriction (Assumption 4.2(c), independence of structural shocks), a non-trivial change identifies a lagged edge $i \to j$ at lag 1. Iterating over lags $l = 1, \ldots, p$ and over all $i \in \mathcal{I}$ recovers the full lagged structure.

The key difference from the linear case is that we identify lagged *edges* (presence or absence) rather than lagged *coefficients* (numerical values); for nonlinear mechanisms, the "coefficient" is replaced by the functional dependence f_j, whose estimation is delegated to Phase 4 (Flow Matching). This is why Phase 4 is essential for nonlinear domains (Table 8) but unnecessary for linear ones.

Since Step 1 identifies B_0, Step 2 identifies $\{B_l\}_{l=1}^p$ (in the linear case) or the lagged edge set (in the nonlinear case).

 doi:10.20944/preprints202605.0617.v1

Step 3 (Disambiguating lagged from contemporaneous).

A subtlety specific to SVARs is that a spurious high-frequency contemporaneous edge can be generated by an unobserved lagged confounder at a shorter sampling interval than the data. Interventions resolve this: $\mathrm{do}(X_{i,t} = x)$ sets the value of $X_{i,t}$ *after* all lagged influences have been realized, so any residual change in $X_{j,t}$ must be attributable to a genuine contemporaneous edge $i \to j$, not to a shared lagged cause. Formally, $P(X_{j,t} \mid \mathrm{do}(X_{i,t} = x), \mathbf{X}_{t-1:t-p})$ under the true graph equals $P(X_{j,t} \mid \mathbf{X}_{t-1:t-p})$ if and only if there is no contemporaneous edge $i \to j$. This is the time-series analogue of the "severed back-door" argument and is what distinguishes SVAR-FM's contemporaneous identification from standard Granger causality testing.

Step 4 (Coverage tightness).

The sufficiency of $|\mathcal{I}| = d$ is immediate from Steps 1–3. For tightness, consider a chain graph $X_1 \to X_2 \to \cdots \to X_d$ at contemporaneous time. Intervening on $\{X_1, \ldots, X_{d-1}\}$ identifies every edge $(i, i+1)$; the remaining node X_d is a sink and carries no outgoing contemporaneous edge, so no intervention on X_d is required. Therefore $|\mathcal{I}| = d - 1$ is sufficient for chains. Eberhardt et al. (2005) showed, in the i.i.d. setting, that the worst-case lower bound is $\lceil \log_2 d \rceil$ interventions when interventions on arbitrary subsets are allowed; the single-variable counterparts that our setting actually needs are those stated as Corollaries 5.2–5.4, whose proofs (Sections A.7 and A.9) run on the time-series identifiability argument of Steps 1–3 above. □

Appendix A.4 Proof of Fact 4.3

Proof. $\mathrm{do}(Z = z)$ severs the paths $Z \to X_i$ and $Z \to X_j$, so that any remaining dependence is attributable solely to direct causation. For details, see Chapter 3 of Pearl (Pearl, 2009). □

Appendix A.5 Proof of Proposition 5.1

Proof. Consider the estimator $\hat{e}_{i\to j} = \frac{1}{M}\sum_{m=1}^{M} y_j^{(m)} - \mu_j$. Each $y_j^{(m)}$ is independent with $\mathbb{E}[y_j^{(m)}] = e^*_{i\to j} + \mu_j$. By the central limit theorem (van der Vaart, 2000), $\sqrt{M}(\hat{e}_{i\to j} - e^*_{i\to j}) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$. The finite-sample concentration inequality follows from Hoeffding (Hoeffding, 1963). □

Appendix A.6 Proof of Proposition 5.2

Proof. Intervention effects are estimated for $d(d-1)$ variable pairs. Requiring accuracy ϵ with failure probability δ/d^2 for each pair, Proposition 5.1 yields that $M = O(\sigma^2 \log(d^2/\delta)/\epsilon^2)$ samples are needed per pair. By the union bound (Wasserman, 2006), the probability of simultaneous success across all pairs is at least $1 - \delta$. With interventions on d variables, the total sample count is $d \cdot M$. □

Appendix A.7 Proof of Corollary 5.2

Proof. Suppose only $d - 2$ or fewer single-variable interventions are performed. Then at least two variables X_a, X_b remain unintervened. By Fact 3.3, the contemporaneous causation between X_a and X_b is non-identifiable from observational data alone, so no collection of $d - 2$ or fewer single-variable interventions can resolve it. Therefore, at least $d - 1$ interventions are necessary. The combinatorial core of this argument is the counting lemma of Eberhardt et al. (2005); what our proof supplies is the link from the i.i.d. counting argument to the non-identifiability result for SVARs in Fact 3.3, which is specific to the time-series setting. □

Appendix A.8 Proof of Corollary 5.3

Proof. An intervention on variable X_i enables estimation of the causal effects from X_i to all other variables (Lemma 4.1). With d interventions, all causal effects are obtained, and the causal structure is identified by Theorem 4.1. □

 doi:10.20944/preprints202605.0617.v1

Appendix A.9 Proof of Corollary 5.4

Proof. Corollaries 5.2 and 5.3 together give $d-1 \le I^* \le d$. For chain structures $X_1 \to X_2 \to \cdots \to X_d$, interventions on all variables except X_1 ($d-1$ interventions) determine the direction of each edge, as in the chain construction of Theorem 5 of Eberhardt et al. (2007); the step that carries that construction over to the SVAR setting is Theorem 4.1, which handles the contemporaneous-plus-lagged structure. □

Appendix A.10 Proof of Proposition 5.4

Proof sketch. (1) Consistency of VAR coefficients follows from standard OLS theory (Hamilton, 1994). (2) Consistency of intervention effects follows from Proposition 5.1. (3) Consistency of Flow Matching follows from universal approximation (Fact E). □

Appendix A.11 Proof of Theorem 5.1

Proof. Let $e^{\mathcal{S}}_{i\to j} := \mathbb{E}_{P_{\mathcal{S}}(\cdot|\mathrm{do}(X_i=x'))}[X_j] - \mathbb{E}_{P_{\mathcal{S}}(\cdot|\mathrm{do}(X_i=x))}[X_j]$ denote the interventional effect *under the simulator-induced distribution*, and let $e^*_{i\to j}$ denote the same quantity under the true interventional distribution. Decompose the estimation error by triangle inequality:

$$|\hat{e}_{i\to j} - e^*_{i\to j}| \;\le\; \underbrace{|\hat{e}_{i\to j} - e^{\mathcal{S}}_{i\to j}|}_{\text{(A) statistical}} \;+\; \underbrace{|e^{\mathcal{S}}_{i\to j} - e^*_{i\to j}|}_{\text{(B) simulator bias}}.$$

Term (A) is bounded by Proposition 5.1: with M samples, $|\hat{e}_{i\to j} - e^{\mathcal{S}}_{i\to j}| = O_p(M^{-1/2})$.

For term (B), write $e^{\mathcal{S}}_{i\to j} - e^*_{i\to j} = \int x_j\big(dP_{\mathcal{S}}(\mathbf{x} \mid \mathrm{do}(X_i = x')) - dP(\mathbf{x} \mid \mathrm{do}(X_i = x'))\big) - \int x_j\big(dP_{\mathcal{S}}(\mathbf{x} \mid \mathrm{do}(X_i = x)) - dP(\mathbf{x} \mid \mathrm{do}(X_i = x))\big)$. Each of the two integrals is a signed integral against a measure whose total variation is bounded by $2\delta_{\mathcal{S}}$ (Assumption 4.1). Letting $B_j := \sup_{\mathbf{x}} |x_j|$ denote the essential sup of X_j under both the simulator and the true interventional distributions, the absolute value of each integral is bounded by $2B_j\delta_{\mathcal{S}}$ (Wasserman, 2006). Hence $|e^{\mathcal{S}}_{i\to j} - e^*_{i\to j}| \le 4B_j\delta_{\mathcal{S}}$, which is the $O(\delta_{\mathcal{S}})$ term in the theorem. □

Appendix A.12 Proof of Theorem 5.2

Proof. Decompose the total error via two intermediate quantities: the simulator-distribution effect $e^{\mathcal{S}}_{i\to j}$ (as in the proof of Theorem 5.1) and the Flow-Matching-distribution effect $\tilde{e}_{i\to j} := g_j(\hat{P}_{\hat{\theta}}(\cdot \mid \mathrm{do}(X_i = x'))) - g_j(\hat{P}_{\hat{\theta}}(\cdot \mid \mathrm{do}(X_i = x)))$. Then, by two applications of the triangle inequality,

$$|\hat{e}^{\mathrm{FM}}_{i\to j} - e^*_{i\to j}| \;\le\; \underbrace{|\hat{e}^{\mathrm{FM}}_{i\to j} - \tilde{e}_{i\to j}|}_{\text{(I) Monte Carlo at FM}} \;+\; \underbrace{|\tilde{e}_{i\to j} - e^{\mathcal{S}}_{i\to j}|}_{\text{(II) FM approximation}} \;+\; \underbrace{|e^{\mathcal{S}}_{i\to j} - e^*_{i\to j}|}_{\text{(III) simulator bias}}.$$

Term (I).

$\hat{e}^{\mathrm{FM}}_{i\to j}$ is a Monte Carlo estimate of $\tilde{e}_{i\to j}$ based on M forward samples of the trained flow. By Hoeffding's inequality (Hoeffding, 1963), for any $\eta \in (0,1)$,

$$|\hat{e}^{\mathrm{FM}}_{i\to j} - \tilde{e}_{i\to j}| \;\le\; \sigma\sqrt{\frac{2\log(2/\eta)}{M}}$$

with probability at least $1-\eta$, yielding the $C_1\sigma/\sqrt{M}$ term with $C_1 = \sqrt{2\log(2/\eta)}$.

Term (II).

By Assumption 5.1(b), the response functional g_j is 1-Lipschitz with respect to W_1, so

$$|\tilde{e}_{i\to j} - e^{\mathcal{S}}_{i\to j}| \;\le\; W_1\big(\hat{P}_{\hat{\theta}}(\cdot \mid \mathrm{do}(X_i = x')), P_{\mathcal{S}}(\cdot \mid \mathrm{do}(X_i = x'))\big) \;+\; W_1\big(\hat{P}_{\hat{\theta}}(\cdot \mid \mathrm{do}(X_i = x)), P_{\mathcal{S}}(\cdot \mid \mathrm{do}(X_i = x))\big).$$

 doi:10.20944/preprints202605.0617.v1

By Assumption 5.1(a) and a Gronwall-type argument for flow matching (Tong et al., 2024; Albergo and Vanden-Eijnden, 2023), each W_1 term is bounded by $e^L \cdot \varepsilon_{\text{FM}}$. Summing gives the $C_3 \cdot e^L \cdot \varepsilon_{\text{FM}}$ term of Eq. equation 11 with $C_3 = 2$.

Term (III).

Exactly the simulator-bias term from the proof of Theorem 5.1: $|e^{\mathcal{S}}_{i\to j} - e^{*}_{i\to j}| \le 4B_j\delta_{\mathcal{S}}$, yielding the $C_2\delta_{\mathcal{S}}$ term with $C_2 = 4B_j$.

Combining the three bounds and noting that the failure event has probability at most η completes the proof. □

Appendix B Detailed listing of causal-discovery methods

Table A11. Time-series methods in detail. Full listing of time-series causal-discovery methods. Columns match Table 1; "Source" indicates whether the method relies on observational data ("obs."), interventional data ("int."), or is an i.i.d. method included as a reference point ("i.i.d."). The final row highlights the unique position occupied by SVAR-FM.

Method	Source	Conf.	Int.	NL	Graph
Classical and constraint-based					
Granger (Granger, 1969)	obs.	×	×	×	discovered
PCMCI (Runge et al., 2019)	obs.	×	×	◯	discovered
PCMCI$^+$ (Runge, 2020)	obs.	×	×	◯	discovered
tsFCI (Entner and Hoyer, 2010)	obs.	◯	×	×	discovered (PAG)
LPCMCI (Gerhardus and Runge, 2020)	obs.	◯	×	◯	discovered (PAG)
SVAR-GFCI (Malinsky and Spirtes, 2018)	obs.	◯	×	×	discovered (PAG)
SVAR / LiNGAM-based					
VARLiNGAM (Hyvärinen et al., 2010)	obs.	×	×	×	discovered
SpinSVAR (Misiakos and Püschel, 2025)	obs.	×	×	×	discovered
Score-based / differentiable / deep learning					
Neural Granger (Tank et al., 2022)	obs.	×	×	◯	discovered
DYNOTEARS (Pamfil et al., 2020)	obs.	×	×	◯	discovered
Rhino (Gong et al., 2023)	obs.	×	×	◯	discovered
CUTS / CUTS$^+$ (Cheng et al., 2023)	obs.	×	×	◯	discovered
Amortized CD (Löwe et al., 2022)	obs.	×	×	◯	discovered
TS-CausalNN (Assaad et al., 2022)	obs.	×	×	◯	discovered
CausalDynamics (Herdeanu et al., 2025)	obs.	×	×	◯	discovered
Temporal score matching (Chen et al., 2024)	obs.	×	×	◯	discovered
Flow-based, graph discovered					
CASTOR (Rahmani and Frossard, 2025a)	obs.	×	×	◯	discovered
Rahmani–Frossard (Rahmani and Frossard, 2025)	obs.	×	×	◯	discovered
Non-stationary flow (Rahmani and Frossard, 2025b)	obs.	×	×	◯	discovered
Flow/diffusion-based, graph known					
DoFlow (Wu et al., 2025)	int.	×	◯	◯	known
CaTSG (Xia et al., 2025)	int.	×	◯	◯	known
PO-Flow (Wu et al., 2025)	int.	×	◯	◯	known (PO)
Intervention-based (i.i.d., listed as reference)					
DCDI (Brouillard et al., 2020)	i.i.d.	×	◯	◯	discovered
IGSP / UT-IGSP (Wang et al., 2017; Squires et al., 2020)	i.i.d.	◯	◯	×	discovered
JCI (Mooij et al., 2020)	i.i.d.	◯	◯	◯	discovered
ENCO (Lippe et al., 2022)	i.i.d.	×	◯	◯	discovered
Bicycle (Rohbeck et al., 2024)	i.i.d.	×	◯	◯	discovered (cyclic)
SVAR-FM (ours)	**int.**	◯	◯	◯	**discovered**

Table A12 complements the overview in Table 1 and the time-series listing in Table A11 by providing a detailed view of i.i.d. causal discovery methods. The purpose is twofold. First, it makes visible which classes of methods operate under causal sufficiency, which admit latent confounders,

 doi:10.20944/preprints202605.0617.v1

which rely on interventional data, and which require the causal graph to be known. Second, for readers whose primary background is in i.i.d. causal discovery, it clarifies *why* directly applying any of these methods to time series with contemporaneous and lagged edges is not straightforward: the column "Source" marks every entry as "i.i.d.", and the sole time-series entry in the comparison—SVAR-FM itself—appears separately in Table A11. Methods are grouped by their identifiability mechanism. Entries that already appear in Table 1 are included here for completeness.

Table A12. i.i.d. methods in detail. Full listing of i.i.d. causal-discovery methods grouped by their identifiability mechanism. Columns match Table 1. All methods in this table operate in the i.i.d. setting and do not directly address time-series data with contemporaneous and lagged edges.

Method	Source	Conf.	Int.	NL	Graph
ANM / PNL / CAM-style identifiability via noise asymmetry					
ANM (Hoyer et al., 2009)	obs.	×	×	○	discovered
PNL (Zhang and Hyvärinen, 2009)	obs.	×	×	○	discovered
CAM (Bühlmann et al., 2014)	obs.	×	×	○	discovered
Score-matching / score-based					
SCORE (Rolland et al., 2022)	obs.	×	×	○	discovered
DiffAN (Montagna et al., 2023)	obs.	×	×	○	discovered
Score-through-the-roof (Montagna et al., 2025)	obs.	○	×	○	discovered
Flow-based causal discovery (graph discovered)					
CARE-FL (Khemakhem et al., 2021)	obs.	×	×	○	discovered
CNF (Javaloy et al., 2023)	obs.	×	×	○	discovered
OCDaf (Kamkari et al., 2023)	obs.	×	×	○	discovered
PNL with normalizing flows (Hoang et al., 2024)	obs.	×	×	○	discovered
Nonlinear with latent confounders					
NL latent CD (Kaltenpoth and Vreeken, 2023)	obs.	○	×	○	discovered
NL CD w/ confounders (Li et al., 2023)	obs.	○	×	○	discovered
Bivariate denoising diffusion (Meier et al., 2025)	obs.	○	×	○	discovered
Intervention-based (graph discovered)					
GIES (Hauser and Bühlmann, 2012)	int.	×	○	×	discovered
IGSP / UT-IGSP (Wang et al., 2017; Squires et al., 2020)	int.	○	○	×	discovered
JCI (Mooij et al., 2020)	int.	○	○	○	discovered
DCDI (Brouillard et al., 2020)	int.	×	○	○	discovered
ENCO (Lippe et al., 2022)	int.	×	○	○	discovered
Bicycle (Rohbeck et al., 2024)	int.	×	○	○	discovered (cyclic)
Flow/diffusion-based (graph known)					
DeCaFlow (Almodóvar et al., 2025)	int.	○	○	○	known
Identifiable Flow (Le et al., 2025)	int.	×	○	○	known (ordering)

Appendix C Standard Benchmark Experiments with Intervention Extensions

Appendix C.1 Overview and positioning

The CausalSim benchmark (§7.1) and the HHG case study (§8) evaluate SVAR-FM in the setting for which it is designed: a physical simulator realizes Pearl's do$(\cdot)$ operator, and the question is whether causal structure can be recovered from the resulting interventional distributions.

This appendix reports a complementary set of experiments on three *standard* time-series causal-discovery benchmarks—CausalTime (Cheng et al., 2024), Tigramite (Runge et al., 2019), and Causal-Dynamics (Herdeanu et al., 2025)—that were **not designed with simulator-based intervention in mind**. We include them for two reasons: (i) to confirm that SVAR-FM performs competitively against observational baselines (Granger, VARLiNGAM, PCMCI, PCMCI+) even when no physical simulator is available, and (ii) to compare against i.i.d. intervention-based methods (IGSP (Wang et al., 2017), UT-IGSP (Squires et al., 2020)) by generating surrogate intervention data through VAR forward simulation or direct DGP/ODE manipulation. The latter comparison is informative but *not strictly fair*: the intervention data given to SVAR-FM and to IGSP/UT-IGSP differ in design (see §C.2 for details).

 doi:10.20944/preprints202605.0617.v1

Method capability summary.

Table A13 summarises which capabilities each method brings to the comparison. SVAR-FM is the only method that operates natively on time series, handles nonlinearity via Flow Matching, and can consume simulator-generated interventional data. IGSP and UT-IGSP consume interventional data but assume i.i.d. observations and use Gaussian conditional-independence tests, which degrade under nonlinearity.

Table A13. Method capabilities in the standard benchmark setting. "TS" = native time-series support; "NL" = nonlinear mechanism support; "Int." = can consume interventional data; "Latent" = admits latent confounders.

Method	TS	NL	Int.	Latent
Granger	◯	×	×	×
VARLiNGAM	◯	×	×	◯
PCMCI / PCMCI+	◯	◯	×	×
IGSP	×	×	◯	×
UT-IGSP	×	×	◯	×
SVAR-FM (ours)	◯	◯	◯	◯

Appendix C.2 Intervention data generation methods

None of the three standard benchmarks provides a physical simulator in the sense of §4.3. To include intervention-based methods in the comparison, we generate surrogate intervention data through three methods, each matched to the benchmark's data-generating process (DGP). To our knowledge, the systematic comparison of these surrogate-intervention strategies for time-series causal discovery has not been reported in the existing literature; the closest precedent is the use of soft interventions in Tigramite's own DGP (Runge et al., 2019), but that work does not compare across intervention types or combine them with Flow Matching–based discovery.

Method I: VAR forward simulation (all benchmarks).

A VAR model is fitted to the observational data. For each target variable X_i, the intervention $\mathrm{do}(X_i = \mu_i + 5\sigma_i)$ is applied by clamping X_i at the specified value and propagating the effect through the estimated VAR coefficients for T_{sim} steps. This method requires no access to the DGP and is therefore applicable to real data (CausalTime). Its quality depends on the accuracy of the VAR approximation; in nonlinear or short-sample regimes ($T < N$), the VAR coefficients may be poorly estimated.

Method II: DGP direct hard intervention (Tigramite only).

Tigramite's synthetic-data generator (`toys.structural_causal_process`) accepts an `intervention` argument that applies a hard do operation at each time step: $X_i(t) := \mu_i + 2\sigma_i$ for all t. The causal mechanisms of all other variables are re-executed under this clamping, producing a counterfactual time series. This method has access to the true DGP and therefore produces the highest-fidelity intervention data in linear systems. In nonlinear systems, however, clamping X_i at a constant value for all t destroys the temporal autocorrelation structure, which can degrade the quality of intervention-effect estimates for methods that rely on temporal smoothness.

Method III: ODE direct soft intervention (CausalDynamics only).

For ODE-based dynamical systems (Lorenz, Rössler, etc.), we add a restoring-force term to the ODE right-hand side:

$$\frac{dx}{dt} = f_{\text{original}}(x) + \lambda \cdot (x_i^* - x_i) \cdot \mathbf{e}_i, \tag{A16}$$

where $x_i^* = \mu_i + 2\sigma_i$ is the intervention target, $\lambda = 10$ is the restoring strength, and $\mathbf{e}_i$ is the i-th unit vector. This "soft" intervention pulls X_i toward x_i^* without instantaneously destroying the dynamics, preserving the temporal structure of the ODE trajectory. This approach has, to our knowledge, not

 doi:10.20944/preprints202605.0617.v1

been used previously for causal discovery in dynamical systems; the standard practice is either hard clamping (which destroys the attractor) or do-calculus reasoning on the ODE Jacobian (Mooij et al., 2013).

Novelty and limitations.

The VAR forward simulation (Method I) and the ODE soft intervention (Method III) are, to our knowledge, new in the context of time-series causal discovery benchmarks. Method I provides a simulator-free surrogate that makes IGSP/UT-IGSP applicable to time-series data they were not designed for. Method III provides a dynamics-preserving alternative to hard clamping in chaotic ODE systems. Neither method is a substitute for a true physical simulator (as used in CausalSim and HHG), but they allow a meaningful—if not strictly fair—comparison between SVAR-FM and i.i.d. intervention-based methods on standard benchmarks.

The asymmetry in the comparison should be noted: SVAR-FM uses PINN-guided variable selection, Flow Matching extrapolation ($T = 40 \to 200$), and confounding-score-based intervention prioritisation when applicable, whereas IGSP/UT-IGSP receive uniform interventions on all variables without these enhancements. This asymmetry is unavoidable because the enhancements are integral to the SVAR-FM framework and cannot be transferred to IGSP/UT-IGSP without leaking SVAR-FM-specific information.

Appendix C.3 Benchmark C.1: CausalTime

CausalTime (Cheng et al., 2024) is a benchmark based on real-world time series: medical (N=20, T=40, 40.3% density), pm25 (N=36, T=40, 28.1%), and traffic (N=20, T=40, 21.6%). No DGP simulator is available; intervention data are generated via Method I (VAR forward simulation).

In all datasets, $T = 40 < N$, which causes the VAR residual covariance matrix to be non-positive definite; consequently, VARLiNGAM is inapplicable across all datasets (F1 = 0.000). SVAR-FM operates stably under the $T < N$ condition through automatic Ridge switching.

Table A14 reports AUROC and F1 (mean ± std, 5 samples).

Table A14. CausalTime extended results. "Obs." = observational data only; "+Int." = with VAR-simulated intervention data. SVAR-FM-DAG uses PINN spatial scoring (Phase 0) which accounts for its advantage on pm25. **Bold**: best per dataset and metric.

		medical ($N=20$)		pm25 ($N=36$)		traffic ($N=20$)	
	Method	AUROC	F1	AUROC	F1	AUROC	F1
Obs.	Granger	.503±.032	.264±.030	.486±.065	.232±.074	.492±.033	.200±.037
	VARLiNGAM	.500±.000	.000±.000	.500±.000	.000±.000	.500±.000	.000±.000
	PCMCI	.469±.059	.224±.132	.704±.145	**.496±.191**	.510±.027	.210±.041
	SVAR-FM-DAG (ours)	**.575±.057**	**.374±.071**	**.758±.086**	.490±.090	.508±.029	.210±.041
+Int.	IGSP+VAR	.494±.029	.236±.046	.494±.011	.213±.015	.524±.019	.214±.032
	UT-IGSP+VAR	.491±.016	.227±.027	.492±.007	.206±.019	.515±.010	.199±.011
	SVAR-FM-DAG+FM+CA (ours)	.561±.038	.349±.056	.746±.078	.472±.068	**.534±.049**	**.236±.062**

Observations.

(i) SVAR-FM-DAG (observational only) is the top method on medical and pm25, driven by Phase 0 PINN spatial scoring rather than by intervention data. (ii) IGSP and UT-IGSP produce AUROC $\approx$ 0.49–0.52 across all datasets, barely above chance; their Gaussian CI tests are ineffective on the nonlinear, short ($T = 40$) CausalTime data. (iii) On traffic, where PINN spatial structure is absent, SVAR-FM-DAG+FM+CA (with FM extrapolation $T = 40 \to 200$) achieves the best AUROC (0.534), marginally above IGSP+VAR (0.524).

 doi:10.20944/preprints202605.0617.v1

Appendix C.4 Benchmark C.2: Tigramite

Tigramite (Runge et al., 2019) provides 8 synthetic scenarios (S1–S8) spanning linear/nonlinear, Gaussian/non-Gaussian, contemporaneous, latent confounding, high-dimensional, and feedback settings: S1 (linear Gaussian, N=3), S2 (nonlinear Gaussian, N=3), S3 (linear non-Gaussian, N=3), S4 (nonlinear non-Gaussian, N=3), S5 (contemporaneous, N=4), S6 (latent confounding, N=4), S7 (high-dimensional sparse, N=8), S8 (feedback, N=3); all with $T = 1000$–2000 and $\tau_{\max} = 3$.

Intervention data are generated via Method II (DGP direct hard intervention) for IGSP/UT-IGSP and SVAR-FM-CF/SVAR-FM-DAG. Table A15 reports F1 scores (mean ± std, 5 seeds).

Table A15. Tigramite extended F1 results. Observational baselines (Granger, VARLiNGAM, PCMCI, PCMCI+) from the original experiment; intervention-based methods (IGSP, UT-IGSP, SVAR-FM-CF) added with DGP-direct hard intervention (Method II). SVAR-FM-CF uses residual correlation > 0.5 to detect and exclude confounded pairs. **Bold**: best per scenario.

Scenario	Granger	VARLiNGAM	PCMCI	PCMCI+	IGSP	UT-IGSP	SVAR-FM (ours)	**SVAR-FM-CF (ours)**
S1 Lin-G	.697±.077	**1.00±.000**	.550±.034	.571±.000	.640±.207	.460±.227	**1.00±.000**	.748±.130
S2 NL-G	.667±.236	.567±.316	.381±.143	.310±.138	.613±.292	.557±.374	**.840±.084**	.740±.126
S3 Lin-NG	.711±.082	**1.00±.000**	.513±.059	.571±.000	.570±.200	.540±.263	**1.00±.000**	.721±.130
S4 NL-NG	.633±.188	.683±.183	.431±.126	.371±.107	.483±.372	.447±.348	**.920±.103**	.670±.170
S5 Contemp	.534±.067	**.938±.113**	.462±.036	.600±.000	.359±.230	.327±.109	.793±.055	.742±.026
S6 Latent	.515±.030	.724±.068	.439±.022	.496±.015	.440±.295	.720±.169	.657±.030	**1.00±.000**
S7 HighDim	.502±.079	**.861±.102**	.348±.044	.396±.056	.274±.119	.203±.065	.508±.041	.372±.038
S8 Feedback	.667±.000	**1.00±.000**	.667±.000	.667±.000	.473±.135	.453±.112	.986±.045	.826±.119

Observations.

(i) SVAR-FM achieves the best F1 on S2 and S4 (nonlinear scenarios) where IGSP/UT-IGSP's Gaussian CI tests fail, and ties with VARLiNGAM on S1 and S3 (linear scenarios with F1 = 1.000). (ii) **SVAR-FM-CF achieves perfect F1 = 1.000 on S6 (latent confounding)**, the only method to do so, by using residual-correlation thresholding to detect and exclude confounded variable pairs. (iii) IGSP and UT-IGSP are competitive only on linear scenarios (S1, S3) where their Gaussian assumption holds; on nonlinear scenarios (S2, S4, S7, S8) their performance degrades to near-random. (iv) VARLiNGAM achieves the best F1 on S5 (contemporaneous), S7 (high-dimensional), and S8 (feedback), all of which have linear, independent-noise structure matching its assumptions exactly.

Appendix C.5 Benchmark C.3: CausalDynamics

CausalDynamics (Herdeanu et al., 2025) evaluates causal discovery on time series generated from nonlinear dynamical systems (ODEs) with known causal structures: Lorenz (N=3, 5 edges), Rössler (N=3, 4), CoupledLorenz (N=6, 11), LotkaVolterra (N=2, 2), and CoupledRösslerLorenz (N=6, 10), each with $T = 2000$–3000. A distinctive feature is that the true graphs contain bidirectional edges (cycles). Accordingly, we use the SVAR-FM dynamics variants (§6.2.2) for the observational comparison, and SVAR-FM-routed (with adaptive routing) for the intervention-extended comparison.

Five dynamical systems were used, with 5 configurations per system varying noise intensity and coupling strength (25 experiments total, $\tau_{\max} = 5$). Intervention data are generated via Method III (ODE direct soft intervention) for SVAR-FM variants and via both Methods I and III for IGSP/UT-IGSP. Table A16 reports AUROC (mean ± std, 5 configurations).

SVAR-FM-routed uses an adaptive routing strategy (BDS nonlinearity test + system dimension) to select the best intervention method and scoring mode per configuration:

- *Route A* (deterministic chaos, $N \leq 3$): Convergent Cross Mapping (CCM) scores;
- *Route B* (deterministic chaos, $N \geq 4$): SVAR-FM original rank ensemble;
- *Route C* (BDS nonlinear): ODE soft intervention + Phase 3-NL;
- *Route D* (BDS linear): VAR intervention + Phase 3-L.

 doi:10.20944/preprints202605.0617.v1

Table A16. CausalDynamics extended AUROC results. "Obs." = observational data only; "+Int." = with intervention data (VAR or ODE). SVAR-FM-routed uses adaptive routing (see text). **Bold**: best per system.

	Method	Lorenz	Rössler	CoupledLorenz	CoupledRösslerL.	Mean
Obs.	Granger	.660±.241	.637±.143	.634±.070	**.698±.062**	.657
	VARLiNGAM	.560±.351	.775±.114	.656±.044	.609±.098	.650
	PCMCI	.640±.167	.400±.271	—	—	—
	SVAR-FM (ours)	.780±.335	.625±.159	.590±.060	.679±.141	.669
+Int.	IGSP+VAR	.660±.152	.475±.056	.461±.030	.515±.034	.528
	UT-IGSP+VAR	.540±.230	.450±.068	.481±.027	.515±.022	.497
	IGSP+ODE	.540±.313	.450±.068	.490±.034	.575±.050	.514
	UT-IGSP+ODE	.540±.313	.450±.068	.497±.029	.580±.041	.517
	SVAR-FM-routed (ours)	**.860±.167**	**.800±.259**	**.764±.063**	.657±.169	**.770**

Observations.

(i) SVAR-FM-routed achieves the highest mean AUROC (0.770) across the four systems, a +15.1% improvement over SVAR-FM without intervention (0.669). (ii) IGSP/UT-IGSP with either VAR or ODE intervention data produce AUROC 0.45–0.66, substantially below SVAR-FM-routed; their Gaussian CI assumption is violated by the chaotic, nonlinear dynamics. (iii) ODE soft intervention (Method III) consistently outperforms VAR intervention (Method I) for both SVAR-FM and IGSP, confirming that intervention-data quality directly affects causal-discovery accuracy—an empirical analogue of the theoretical $O(\delta_{\mathcal{S}})$ term in Theorem 5.2. (iv) CoupledRösslerLorenz (N=6, heterogeneous coupling) remains difficult for all methods; SVAR-FM (0.679, observational) slightly outperforms SVAR-FM-routed (0.657) here, suggesting that the routing heuristic can be improved for heterogeneous multi-system coupling. (v) PCMCI results are unavailable ("—") for the coupled systems ($N \geq 6$) because its computational cost scales steeply with dimension and the required $\tau_{\max} = 5$ makes the conditioning sets too large for stable partial-correlation estimation at the available sample sizes. (vi) LotkaVolterra ($N = 2$) is omitted from the intervention-extended tables because all methods (including SVAR-FM) achieve F1 = 1.000 on this simple bidirectional system, making intervention unnecessary and uninformative for the comparison.

Appendix C.6 Cross-benchmark summary

Table A17 summarises the three benchmarks.

Table A17. Cross-benchmark summary: best AUROC of IGSP/UT-IGSP vs. best AUROC of SVAR-FM variants.

Benchmark	IGSP best	SVAR-FM (ours) best	Δ
CausalTime (3 datasets)	0.524	**0.758**	+0.234
Tigramite S6 (latent, F1)	0.720	**1.000**	+0.280
CausalDynamics (4 systems)	0.660	**0.860**	+0.200

Two patterns emerge consistently across all three benchmarks. First, **IGSP and UT-IGSP degrade sharply under nonlinearity**: their Gaussian conditional-independence tests are the bottleneck, not the availability of intervention data. This is visible in every nonlinear scenario (CausalTime real data, Tigramite S2/S4/S7/S8, CausalDynamics Lorenz/Rössler/CoupledLorenz). Second, **intervention-data quality matters**: ODE soft intervention consistently outperforms VAR forward simulation (CausalDynamics, +0.059 AUROC on average), and DGP direct intervention outperforms bootstrap shifting (Tigramite). This pattern is the empirical counterpart of the $O(\delta_{\mathcal{S}})$ term in the end-to-end error bound of Theorem 5.2: higher simulator fidelity produces more accurate causal-effect estimates.

These results are consistent with, but do not replace, the CausalSim and HHG experiments in the main text, which evaluate SVAR-FM in its intended setting—physical simulators as do-operators.

 doi:10.20944/preprints202605.0617.v1

Appendix C.7 Controlled comparison: identical intervention data, varying autocorrelation

The cross-benchmark comparisons above are informative but not strictly controlled: SVAR-FM and IGSP/UT-IGSP receive different intervention data, and the benchmarks were not designed to isolate the effect of temporal structure. To address this, we conducted a controlled experiment using the Tigramite S1 DGP (linear, Gaussian, 3 variables, $T = 1000$) with the autoregressive coefficient ρ varied over $\{0.0, 0.3, 0.5, 0.7, 0.9\}$. Crucially, all three methods receive *identical* DGP-generated intervention data at each ρ and seed. Each condition was replicated over 5 seeds, and all methods are evaluated on the same criterion: recovery of the true causal directions $\{0 \to 1,\ 1 \to 2\}$ (Table A18).

Table A18. Controlled comparison: identical intervention data, varying autocorrelation ρ. F1 and TPR for causal direction recovery (mean $\pm$ std, 5 seeds).

	F1			TPR		
ρ	IGSP	UT-IGSP	SVAR-FM (ours)	IGSP	UT-IGSP	SVAR-FM (ours)
0.0	0.000	0.000	**0.773**	0.000	0.000	**1.000**
0.3	**0.833**	0.533	0.773	0.800	0.500	**1.000**
0.5	0.700	0.700	**0.773**	0.700	0.700	**1.000**
0.7	0.400	0.500	**0.610**	0.500	0.600	**1.000**
0.9	0.560	0.400	**0.548**	0.700	0.500	**1.000**

The most striking result is that SVAR-FM achieves perfect recall (TPR = 1.000) at every value of ρ. Phase 1 VAR estimation correctly identifies the lagged structure $X_0 \to X_1$ (lag 1) and $X_1 \to X_2$ (lag 2) regardless of autocorrelation strength, and Phase 3 intervention tests confirm both edges with high significance ($z > 20$, $p < 10^{-6}$). The cost of this high recall is elevated FDR (0.37–0.62), primarily from detecting the indirect effect $X_0 \to X_2$ as a direct edge.

By contrast, IGSP and UT-IGSP return empty graphs at $\rho = 0.0$ (F1 = 0.000)—precisely the i.i.d. setting for which they were designed. Inspection of the output shows `dag.arcs` $= \emptyset$: no edges are detected in any direction. This occurs because the DGP intervention (fixing X_i at $+2\sigma$) produces a large distributional shift that causes the Gaussian invariance test to reject invariance for *all* variable pairs, leaving the algorithm unable to orient any edge (Yang et al., 2018).

The behaviour of IGSP is also non-monotonic in ρ: it achieves its best F1 at $\rho = 0.3$ (0.833), where moderate autocorrelation smooths the intervention signal enough for the invariance test to function. At higher ρ (0.7–0.9), IGSP degrades (F1 = 0.400–0.560) as the i.i.d. assumption of its conditional-independence tests is increasingly violated, producing spurious edges and direction reversals.

Taken together, these results show that SVAR-FM's advantage over i.i.d. intervention-based methods has two sources: Phase 1 VAR estimation natively handles temporal dependence, maintaining TPR = 1.000 across all ρ; and the intervention-effect test in Phase 3 is robust to intervention magnitude, whereas IGSP's invariance test is sensitive to distributional shift. The trade-off is that SVAR-FM's ATE-based test is more liberal, producing higher FDR than IGSP at moderate ρ (e.g., $\rho = 0.3$: SVAR-FM FDR = 0.37 vs. IGSP FDR = 0.10). Combining Phase 1 VAR structure with a more conservative edge-selection criterion is a promising direction for reducing SVAR-FM's FDR.

Appendix D CausalSim-Battery: First-Principles DFT Simulator

This appendix reports the fourth CausalSim instance, which uses *ab initio* density functional theory (DFT) calculations as the simulator. Among all CausalSim instances, this is the most rigorous: DFT derives material properties from the Schrödinger equation without empirical fitting parameters, making the do$(\cdot)$ operation maximally credible ($\delta_{\mathcal{S}} \approx 0$ up to the exchange–correlation approximation). This instance is reported separately because the domain-specific background (electrochemistry, SEI formation) exceeds what is needed for the main-text evaluation, but the scientific implications—discovery of a previously unknown dual causal pathway—are substantial.

 doi:10.20944/preprints202605.0617.v1

Appendix D.1 Problem setting: latent confounding in battery degradation

Understanding the degradation mechanisms of lithium-ion batteries is critical for electric vehicles and renewable energy storage (Severson et al., 2019). The key variables are capacity (Cap) and internal resistance (IR). Temperature T acts as an **unobserved common cause (latent confounder)**: high temperature accelerates both SEI growth (increasing IR) and side reactions (decreasing Cap), producing a **spurious negative correlation** between IR and Cap. The true causal effect IR $\rightarrow$ Cap is positive (increased IR impedes ion transport). Methods based solely on observational data cannot distinguish the spurious correlation from the true causal effect.

Appendix D.2 Simulator and intervention

The Arrhenius law (Bloom et al., 2001) serves as the simulator:

$$k(T) = A \exp\left(-\frac{E_a}{RT}\right), \tag{A17}$$

where $E_a = 50$ kJ/mol. The intervention $\mathrm{do}(T = 25°\mathrm{C})$ fixes temperature and severs the confounding path, allowing estimation of the direct effect IR $\rightarrow$ Cap.

Observational data are from the NASA Battery Dataset[22] (Saha and Goebel, 2007) (18650 cells, 5,592 cycles, 5 variables).

Appendix D.3 Results: sign reversal confirms confounding

- Temperature uncontrolled (observational): $\mathrm{ATE}(\mathrm{IR} \rightarrow \mathrm{Cap}) = -0.10$ (**wrong sign**)
- Temperature fixed at 25°C (intervention via Arrhenius simulator[23]): $\mathrm{ATE}(\mathrm{IR} \rightarrow \mathrm{Cap}) = +0.03$ (**correct sign**)

This sign reversal satisfies the confounding detection criterion of Fact 4.3.

Appendix D.4 Discovery of dual causal pathways via DFT

DFT calculations using Quantum ESPRESSO (Giannozzi et al., 2009,1) of electrolyte additive LUMO[24] (Lowest Unoccupied Molecular Orbital) energies reveal the causal mechanism of SEI[25] formation. Four additives were analysed (Table A19).

Table A19. DFT-computed LUMO energies and SVAR-FM causal effects for SEI additives

Additive	LUMO (eV)	Cap. ret. (%)	Improvement	Reducibility
EC (baseline)	−0.78	53	—	Low
FEC	−0.76	80	+27%	Low
VC	−1.14	74	+21%	Medium
LiBOB	**−1.75**	**87**	**+34%**	**High**

The fluorine anomaly.

FEC has nearly the same LUMO as EC (−0.76 vs. −0.78 eV), yet its capacity retention is 27% higher. This anomaly cannot be explained by a LUMO-only model and reveals a **second causal pathway**: fluorine atoms form LiF within the SEI, providing ionic conductivity, mechanical stability, and chemical stability independently of the LUMO mechanism.

[22] NASA Battery Dataset: https://www.nasa.gov/content/prognostics-center-of-excellence-data-set-repository

[23] DFT calculations were performed using Quantum ESPRESSO (Giannozzi et al., 2009,1) (https://www.quantum-espresso.org/). The Arrhenius simulator and SVAR-FM integration code will be released upon publication.

[24] LUMO (Lowest Unoccupied Molecular Orbital): the lowest-energy empty electron orbital of a molecule. Molecules with lower (more negative) LUMO energy are more easily reduced, meaning they accept electrons more readily and decompose earlier during battery charging to form the SEI film.

[25] SEI (Solid Electrolyte Interphase): a thin passivation film (∼10–50 nm) that forms on the anode surface during the first charge cycles of a lithium-ion battery. The SEI is ionically conductive but electronically insulating; its quality determines long-term battery capacity retention.

 doi:10.20944/preprints202605.0617.v1

Dual pathway model.

Incorporating fluorine as a binary variable yields:

$$\text{Cap} = \alpha + \beta_1 \cdot \text{LUMO} + \beta_2 \cdot F + \epsilon. \tag{A18}$$

The model fit improves from $R^2 = 0.42$ (LUMO only) to $R^2 = 0.93$ (LUMO + F), with $\beta_1 = -33.6$ %/eV and $\beta_2 = +24.2$ %. The LUMO effect had been **underestimated by 64%** in the single-pathway model because fluorine acted as a confounder.

AI for Science significance.

This result demonstrates that SVAR-FM, combined with a first-principles DFT simulator, can discover previously unknown causal pathways in materials science. The dual-pathway model provides actionable design guidelines: optimal SEI additives should combine low LUMO (for preferential reduction) with fluorine substitution (for LiF formation). The predicted improvement for fluorinated LiBOB (F-LiBOB) is 56.8%, substantially exceeding any single additive.

Appendix E Flow Matching Technical Details

Definition A1 (Conditional Flow Matching (Tong et al., 2024)). *Conditional Flow Matching learns a vector field $v_\theta(\mathbf{x}, t|\mathbf{c})$:*

$$\frac{d\mathbf{x}_t}{dt} = v_\theta(\mathbf{x}_t, t|\mathbf{c}), \quad t \in [0, 1] \tag{A19}$$

where $\mathbf{c}$ is a conditioning vector that includes the physical parameters (intervention conditions) of the simulator. The CFM loss is given by (Lipman et al., 2023):

$$\mathcal{L}_{CFM}(\theta) = \mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_1,\mathbf{c}}\left[\|v_\theta(\mathbf{x}_t, t|\mathbf{c}) - (\mathbf{x}_1 - \mathbf{x}_0)\|^2\right] \tag{A20}$$

Remark A1 (Flow Matching and physical constraints). *By including the simulator's physical parameters in $\mathbf{c}$: (a) Flow Matching learns $P_{\mathcal{S}}(\cdot|\text{do}(X_i = x))$ and estimates counterfactual distributions consistent with physical laws; (b) it enables sensitivity analysis of causal effects with respect to physical parameters ($\delta_{\mathcal{S}}$ in Assumption 4.1); (c) it provides a unified framework for comparing different simulator variants (e.g., TDDFT exchange-correlation functionals in HHG).*

[Universal approximation (Chen et al., 2018)] A CNF[26] parameterized by a neural network of sufficient capacity can approximate any continuous conditional distribution to arbitrary precision.

Appendix F SVAR-dyn1/SVAR-dyn2: Full Details

For ODE-based dynamical systems (CausalDynamics), the true graph contains bidirectional edges (cycles). The following modifications are applied: (1) NOTEARS is disabled; (2) Phase 0 directional bias is disabled; (3) a differential Granger score is added.

The differential Granger score is defined as

$$s_{ij}^{\text{dG}} = \max\left(0, \frac{\mathcal{E}_{\text{base}}^{(j)} - \mathcal{E}_{\text{full}}^{(j,i)}}{\mathcal{E}_{\text{base}}^{(j)}}\right) \tag{A21}$$

where $\mathcal{E}_{\text{base}}^{(j)}$ is the squared prediction error from $X_t^{(j)}$ alone, and $\mathcal{E}_{\text{full}}^{(j,i)}$ adds $X_t^{(i)}$ (Ridge regression, $\lambda = 0.1$).

[26] CNF (Continuous Normalizing Flow): a generative model that transforms a simple base distribution (e.g., standard Gaussian) into a complex target distribution via a continuous-time ordinary differential equation $dx/dt = v_\theta(x, t)$, where the velocity field v_θ is parameterised by a neural network.

 doi:10.20944/preprints202605.0617.v1

Table A20. Architectural variants of SVAR-FM-dynamics

Variant	Configuration
SVAR-dyn1	Rank ensemble of Phase 1 (VAR coefficients + Diff-Granger) only. Phases 3–5 are not used. Stable for small systems ($N \leq 3$).
SVAR-dyn2	z-score weighted ensemble of Phases 1–5. Adds Phase 3 (ATE), Phase 4 (FM-ATE), Phase 5 (FNO-Granger[27]), with adaptive Spearman-ρ weighting ($w = 0$ when $\rho \leq 0$). Eps-guard zeroes Phase 5 when $\epsilon_{\max} < 10^{-3}$. Improves for large coupled systems ($N \geq 6$).

Eps-guard.

The threshold $\theta = 10^{-3}$ corresponds to typical floating-point residuals in noise-free ODE integrations (double precision); it is not tuned to benchmark data.

Appendix G Detailed Positioning Relative to Related Work

Appendix G.1 Simulation-Based Inference (SBI)

SBI (Cranmer et al., 2020) takes a mechanistic model whose causal structure is *fixed* and estimates a posterior over its parameters. SVAR-FM recovers the causal graph itself. Brehmer et al. (Brehmer et al., 2020) exploit latent structural information within simulators for likelihood-free inference; both that work and SVAR-FM look inside the simulator, but for different objects (parameter posteriors vs. do-operator realizations). The two frameworks can be chained—SBI for parameters, SVAR-FM for the graph—but neither replaces the other.

Appendix G.2 Mechanistic Model-Based Causal Inference

GOBI (Park et al., 2023) uses data-reproducibility of monotonic ODE models as a causal criterion, without requiring intervention data. SVAR-FM actively generates intervention data via the simulator and provides theoretically guaranteed identification (Theorem 4.1).

Appendix G.3 Causal Inference with Deep Generative Models

Deep generative causal models (Khemakhem et al., 2021; Javaloy et al., 2023; Pawlowski et al., 2020; Sanchez and Tsaftaris, 2022; Chao et al., 2024; Le et al., 2025) assume the graph and learn mechanisms. DoFlow (Wu et al., 2025) is the closest time-series extension; it predicts under a known graph, while SVAR-FM discovers the graph. Further 2025 developments—CaTSG (Xia et al., 2025), DeCaFlow (Almodóvar et al., 2025), PO-Flow (Wu et al., 2025)—all assume a known graph. See Komanduri et al. (Komanduri et al., 2024) for a survey. The key contrast with DoFlow (Wu et al., 2025): SVAR-FM discovers the graph (unknown) from simulator-generated interventions; DoFlow predicts under a known graph from observational data. They use Flow Matching in structurally different roles (causal mechanism approximation vs. node-conditional generation) and are composable.

Appendix G.4 SVAR Literature in Econometrics

Econometric SVARs (Kilian and Lütkepohl, 2017; Blanchard and Quah, 1989; Uhlig, 2005) achieve identification through statistical assumptions. SpinSVAR (Misiakos and Püschel, 2025) scales to thousands of nodes via sparse Laplacian assumptions; the Bank of England (Brignone and Piffer, 2025) uses SVARs for monetary policy. SVAR-FM achieves identification through simulator intervention instead.

Appendix G.5 Causal Digital Twins

CDT (Homaei et al., 2025) discovers the graph from observational data then reasons with the SCM; Bicycle (Rohbeck et al., 2024) discovers cyclic graphs from CRISPR perturbations; Le et al. (Le et al., 2025) learn mechanisms under a known graph. SVAR-FM uses the simulator as Pearl's do-operator to *discover* the graph, with $\delta_{\mathcal{S}}$ in the error analysis.

 doi:10.20944/preprints202605.0617.v1

Appendix H HHG Computational Parameters

All Octopus TDDFT calculations in §8 use the following settings. The system is an H_2 molecule on a spherical grid (radius 15 Å, spacing 0.3 Å). The laser wavelength is $\lambda = 800$ nm ($\omega_0 = 0.057$ H) with a cosine-squared envelope of width 400 a.u. (≈10 fs). The propagation runs for 500 a.u. (≈12 fs) over 250,000 time steps. Both observational and interventional data use the SIC-ADSIC exchange–correlation functional (`lda_x + lda_c_pw`).

References

Merchant, A.; Batzner, S.; Schoenholz, S.S.; Aykol, M.; Cheon, G.; Cubuk, E.D. Scaling deep learning for materials discovery. *Nature* **2023**, *624*, 80–85.

Zhang, H.; Wang, X.; Zhang, J.; Bi, Y.; Chen, F.Z.; Dong, H.; Guo, M.; Huang, S.; Huang, W.; Jin, C.; et al. MatterSim: A deep learning atomistic model across elements, temperatures and pressures. *arXiv preprint arXiv:2405.04967* **2024**.

Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589.

Raissi, M.; Perdikaris, P.; Karniadakis, G.E. Physics-Informed Neural Networks: A Deep Learning Framework for Solving Forward and Inverse Problems Involving Nonlinear Partial Differential Equations. *Journal of Computational Physics* **2019**, *378*, 686–707. https://doi.org/10.1016/j.jcp.2018.10.045.

Karniadakis, G.E.; Kevrekidis, I.G.; Lu, L.; Perdikaris, P.; Wang, S.; Yang, L. Physics-informed machine learning. *Nature Reviews Physics* **2021**, *3*, 422–440.

Lam, R.; Sanchez-Gonzalez, A.; Willson, M.; Wirnsberger, P.; Fortunato, M.; Alet, F.; Ravuri, S.; Ewalds, T.; Eaton-Rosen, Z.; Hu, W.; et al. Learning skillful medium-range global weather forecasting. *Science* **2023**, *382*, 1416–1421.

Bodnar, C.; Bruinsma, W.P.; Lucic, A.; Stanley, M.; Brandstetter, J.; Garvan, P.; Riechert, M.; Weyn, J.A.; Dong, H.; Vaughan, A.; et al. Aurora: A foundation model of the atmosphere. *arXiv preprint arXiv:2405.13063* **2024**.

Pearl, J. *Causality: Models, Reasoning, and Inference*, 2nd ed.; Cambridge University Press: Cambridge, 2009.

Granger, C.W.J. Investigating Causal Relations by Econometric Models and Cross-Spectral Methods. *Econometrica* **1969**, *37*, 424–438. https://doi.org/10.2307/1912791.

Runge, J.; Nowack, P.; Kretschmer, M.; Flaxman, S.; Sejdinovic, D. Detecting and Quantifying Causal Associations in Large Nonlinear Time Series Datasets. *Science Advances* **2019**, *5*, eaau4996. https://doi.org/10.1126/sciadv.aau4996.

Hyvärinen, A.; Zhang, K.; Shimizu, S.; Hoyer, P.O. Estimation of a Structural Vector Autoregression Model Using Non-Gaussianity. *Journal of Machine Learning Research* **2010**, *11*, 1709–1731.

Wang, Y.; Solus, L.; Yang, K.D.; Uhler, C. Permutation-Based Causal Inference Algorithms with Interventions. In Proceedings of the Advances in Neural Information Processing Systems 30 (NeurIPS 2017), 2017, pp. 5822–5831.

Mooij, J.M.; Magliacane, S.; Claassen, T. Joint Causal Inference from Multiple Contexts. *Journal of Machine Learning Research* **2020**, *21*, 1–108.

Brouillard, P.; Lachapelle, S.; Lacoste, A.; Lacoste-Julien, S.; Drouin, A. Differentiable Causal Discovery from Interventional Data. In Proceedings of the Advances in Neural Information Processing Systems 33 (NeurIPS 2020), 2020, pp. 21865–21877.

Sims, C.A. Macroeconomics and Reality. *Econometrica* **1980**, *48*, 1–48. https://doi.org/10.2307/1912017.

Lipman, Y.; Chen, R.T.Q.; Ben-Hamu, H.; Nickel, M.; Le, M. Flow Matching for Generative Modeling. In Proceedings of the The Eleventh International Conference on Learning Representations (ICLR 2023), 2023. arXiv:2210.02747.

Tong, A.; Malkin, N.; Huguet, G.; Zhang, Y.; Rector-Brooks, J.; Fatras, K.; Wolf, G.; Bengio, Y. Conditional Flow Matching: Simulation-Free Dynamic Optimal Transport. *Transactions on Machine Learning Research* **2024**. arXiv:2302.00482.

Corkum, P.B. Plasma Perspective on Strong Field Multiphoton Ionization. *Physical Review Letters* **1993**, *71*, 1994–1997. https://doi.org/10.1103/PhysRevLett.71.1994.

Cranmer, K.; Brehmer, J.; Louppe, G. The Frontier of Simulation-Based Inference. *Proceedings of the National Academy of Sciences* **2020**, *117*, 30055–30062. https://doi.org/10.1073/pnas.1912789117.

 doi:10.20944/preprints202605.0617.v1

Boiko, D.A.; MacKnight, R.; Kline, B.; Gomes, G. Autonomous chemical research with large language models. *Nature* **2023**, *624*, 570–578.

Eberhardt, F.; Glymour, C.; Scheines, R. On the Number of Experiments Sufficient and in the Worst Case Necessary to Identify All Causal Relations Among *N* Variables. In Proceedings of the Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence (UAI 2005), 2005, pp. 178–184.

Eberhardt, F.; Glymour, C.; Scheines, R. Interventions and Causal Inference. *Philosophy of Science* **2007**, *74*, 981–995. https://doi.org/10.1086/525638.

Hoyer, P.O.; Janzing, D.; Mooij, J.M.; Peters, J.; Schölkopf, B. Nonlinear Causal Discovery with Additive Noise Models. In Proceedings of the Advances in Neural Information Processing Systems 21 (NeurIPS 2008), 2009, pp. 689–696.

Kaltenpoth, D.; Vreeken, J. Nonlinear Causal Discovery with Latent Confounders. In Proceedings of the Proceedings of the 40th International Conference on Machine Learning (ICML). PMLR, 2023, Vol. 202, pp. 15639–15654.

Almodóvar, A.; Javaloy, A.; Parras, J.; Zazo, S.; Valera, I. DeCaFlow: A Deconfounding Causal Generative Model. In Proceedings of the Advances in Neural Information Processing Systems 38 (NeurIPS 2025), 2025.

Gerhardus, A.; Runge, J. High-Recall Causal Discovery for Autocorrelated Time Series with Latent Confounders. In Proceedings of the Advances in Neural Information Processing Systems 33 (NeurIPS 2020), 2020, pp. 12615–12625.

Rahmani, A.; Frossard, P. Flow Based Approach for Dynamic Temporal Causal Models with non-Gaussian or Heteroscedastic Noises. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2025.

Wu, D.; et al. DoFlow: Causal Generative Flows for Interventional and Counterfactual Time-Series Prediction. *arXiv preprint arXiv:2511.02137* **2025**.

Dahlhaus, R.; Eichler, M. Causality and Graphical Models in Time Series Analysis. *Oxford Statistical Science Series* **2003**, pp. 115–137. In: Green, Hjort, Richardson (Eds.) Highly Structured Stochastic Systems.

Eichler, M. Graphical Modelling of Multivariate Time Series. *Probability Theory and Related Fields* **2012**, *153*, 233–268. https://doi.org/10.1007/s00440-011-0345-8.

Runge, J. Discovering Contemporaneous and Lagged Causal Relations in Autocorrelated Nonlinear Time Series Datasets. In Proceedings of the Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI 2020), 2020, pp. 1388–1397.

Runge, J.; Gerhardus, A.; Varando, G.; Eyring, V.; Camps-Valls, G. Causal Inference for Time Series. *Nature Reviews Earth & Environment* **2023**, *4*, 487–505. https://doi.org/10.1038/s43017-023-00431-y.

Misiakos, P.; Püschel, M. SpinSVAR: Estimating Structural Vector Autoregression Assuming Sparse Input. In Proceedings of the Proceedings of the 41st Conference on Uncertainty in Artificial Intelligence (UAI 2025), 2025, Vol. 286, *Proceedings of Machine Learning Research*, pp. 3048–3092.

Pamfil, R.; Sriwattanaworachai, N.; Desai, S.; Pilgerstorfer, P.; Georgatzis, K.; Beaumont, P.; Aragam, B. DYNOTEARS: Structure Learning from Time-Series Data. In Proceedings of the Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS 2020), 2020, Vol. 108, *PMLR*, pp. 1595–1605.

Tank, A.; Covert, I.; Foti, N.; Shojaie, A.; Fox, E.B. Neural Granger Causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2022**, *44*, 4267–4279. https://doi.org/10.1109/TPAMI.2021.3065601.

Gong, W.; Jennings, J.; Zhang, C.; Pawlowski, N. Rhino: Deep Causal Temporal Relationship Learning with History-dependent Noise. In Proceedings of the The Eleventh International Conference on Learning Representations (ICLR 2023), 2023.

Cheng, Y.; Yang, R.; Xiao, T.; Li, Z.; Suo, J.; He, K.; Dai, Q. CUTS: Neural Causal Discovery from Irregular Time-Series Data. In Proceedings of the The Eleventh International Conference on Learning Representations (ICLR 2023), 2023.

Löwe, S.; Madras, D.; Zemel, R.; Welling, M. Amortized Causal Discovery: Learning to Infer Causal Graphs from Time-Series Data. In Proceedings of the Proceedings of the First Conference on Causal Learning and Reasoning (CLeaR 2022), 2022, Vol. 177, *PMLR*, pp. 509–525.

Assaad, C.K.; Devijver, E.; Gaussier, E. Survey and Evaluation of Causal Discovery Methods for Time Series. *Journal of Artificial Intelligence Research* **2022**, *73*, 767–819. https://doi.org/10.1613/jair.1.13428.

Sun, X.; Schulte, O.; Liu, G.; Poupart, P. NTS-NOTEARS: Learning Nonparametric DBNs with Prior Knowledge. *Proceedings of Machine Learning Research (AISTATS 2023)* **2023**, *206*, 1942–1964.

 doi:10.20944/preprints202605.0617.v1

Chen, H.; Yi, K.; Liu, L.; Wang, Y.G. Score-matching-based Structure Learning for Temporal Data on Networks. *arXiv preprint arXiv:2412.07469* **2024**.

Kang, J.; Kim, S.; Lee, C.; Hwang, D.; Chung, J.H.; Ko, Y.; Lee, S.; Kim, S.; Lim, S. Score-informed Neural Operator for Enhancing Ordering-based Causal Discovery. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2025.

Yin, N.; Gao, T.; Yu, Y.; Ji, Q. Information Theoretically Optimal Sample Complexity of Learning Dynamical Directed Acyclic Graphs. *arXiv preprint arXiv:2312.12844* **2023**.

Veedu, M.S.; Deka, D.; Salapaka, M.V. Information Theoretically Optimal Sample Complexity of Learning Dynamical Directed Acyclic Graphs. *arXiv preprint arXiv:2308.16859* **2023**.

Zhu, Z.; Locatello, F.; Cevher, V. Sample Complexity Bounds for Score-Matching: Causal Discovery and Generative Modeling. *arXiv preprint arXiv:2310.18123* **2024**.

Khemakhem, I.; Monti, R.P.; Leech, R.; Hyvärinen, A. Causal Autoregressive Flows. In Proceedings of the Proceedings of The 24th International Conference on Artificial Intelligence and Statistics (AISTATS 2021), 2021, Vol. 130, *Proceedings of Machine Learning Research*, pp. 3520–3528.

Javaloy, A.; Martín, P.S.; Valera, I. Causal Normalizing Flows: From Theory to Practice. In Proceedings of the Advances in Neural Information Processing Systems 36 (NeurIPS 2023), 2023.

Kamkari, H.; Zehtab, V.; Balazadeh, V.; Krishnan, R.G. OCDaf: Ordered Causal Discovery with Autoregressive Flows. *arXiv preprint arXiv:2308.07480* **2023**.

Hoang, N.; Duong, B.; Nguyen, T. Enabling Causal Discovery in Post-Nonlinear Models with Normalizing Flows. In Proceedings of the European Conference on Artificial Intelligence (ECAI), 2024.

Rahmani, A.; Frossard, P. CASTOR: Causal Temporal Regime Structure Learning. In Proceedings of the Proceedings of The 28th International Conference on Artificial Intelligence and Statistics (AISTATS), 2025, Vol. 258, *Proceedings of Machine Learning Research*, pp. 4546–4554.

Rahmani, A.; Frossard, P. Flow-Based Non-stationary Temporal Regime Causal Structure Learning. *arXiv preprint arXiv:2506.17065* **2025**.

Eberhardt, F. Almost optimal intervention sets for causal discovery. *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI)* **2012**, pp. 161–170.

Hauser, A.; Bühlmann, P. Characterization and Greedy Learning of Interventional Markov Equivalence Classes of Directed Acyclic Graphs. *Journal of Machine Learning Research* **2012**, *13*, 2409–2464.

Squires, C.; Wang, Y.; Uhler, C. Permutation-Based Causal Structure Learning with Unknown Intervention Targets. In Proceedings of the Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI 2020), 2020, Vol. 124, *PMLR*, pp. 1039–1048.

Lippe, P.; Cohen, T.; Gavves, E. Efficient Neural Causal Discovery without Acyclicity Constraints. In Proceedings of the The Tenth International Conference on Learning Representations (ICLR 2022), 2022.

Rohbeck, M.; Clarke, B.; Mikulik, K.; Pettet, A.; Stegle, O.; Ueltzhoeffer, K. Bicycle: Intervention-Based Causal Discovery with Cycles. In Proceedings of the Proceedings of the Third Conference on Causal Learning and Reasoning (CLEaR), 2024, Vol. 236, *Proceedings of Machine Learning Research*, pp. 209–242.

Spirtes, P.; Glymour, C.; Scheines, R. *Causation, Prediction, and Search*, 2nd ed.; MIT Press, 2000.

Entner, D.; Hoyer, P.O. On Causal Discovery from Time Series Data using FCI. In Proceedings of the Proceedings of the 5th European Workshop on Probabilistic Graphical Models (PGM 2010), 2010, pp. 121–128.

Malinsky, D.; Spirtes, P. Causal Structure Learning from Multivariate Time Series in Settings with Unmeasured Confounding. In Proceedings of the Proceedings of 2018 ACM SIGKDD Workshop on Causal Discovery, 2018, Vol. 92, *PMLR*, pp. 23–47.

Brehmer, J.; Louppe, G.; Pavez, J.; Cranmer, K. Mining Gold from Implicit Models to Improve Likelihood-Free Inference. *Proceedings of the National Academy of Sciences* **2020**, *117*, 5242–5249. https://doi.org/10.1073/pnas.1915980117.

Lueckmann, J.M.; Boelts, J.; Greenberg, D.S.; Gonçalves, P.J.; Macke, J.H. Benchmarking Simulation-Based Inference. In Proceedings of the Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS 2021), 2021, Vol. 130, *PMLR*, pp. 343–351.

Radev, S.T.; Schmitt, M.; Pratz, V.; Picchini, U.; Köthe, U.; Bürkner, P.C. JANA: Jointly Amortized Neural Approximation of Complex Bayesian Models. *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023)* **2023**, *216*, 1695–1706.

Park, S.H.; Ha, S.; Kim, J.K. A General Model-Based Causal Inference Method Overcomes the Curse of Synchrony and Indirect Effect. *Nature Communications* **2023**, *14*, 4287. https://doi.org/10.1038/s41467-023-40056-9.

 doi:10.20944/preprints202605.0617.v1

Pawlowski, N.; Castro, D.C.; Glocker, B. Deep Structural Causal Models for Tractable Counterfactual Inference. In Proceedings of the Advances in Neural Information Processing Systems 33 (NeurIPS 2020), 2020, pp. 12151–12164.

Sanchez, P.; Tsaftaris, S.A. Diffusion Causal Models for Counterfactual Estimation. In Proceedings of the Proceedings of the First Conference on Causal Learning and Reasoning (CLEaR), 2022, Vol. 177, *Proceedings of Machine Learning Research*, pp. 647–668.

Chao, P.; Blöbaum, P.; Patel, S.; Kasiviswanathan, S.P. Modeling Causal Mechanisms with Diffusion Models for Interventional and Counterfactual Queries. *Transactions on Machine Learning Research* **2024**.

Wu, D.; Inouye, D.I.; Xie, Y. Flow-Based Generative Modeling of Potential Outcomes and Counterfactuals. *arXiv preprint arXiv:2505.16051* **2025**.

Xia, Y.; et al. Causal Time Series Generation via Diffusion Models. *arXiv preprint arXiv:2509.20846* **2025**.

Le, M.K.; Do, K.; Tran, T. Learning Structural Causal Models from Ordering: Identifiable Flow Models. *Proceedings of the AAAI Conference on Artificial Intelligence* **2025**, *39*, 17831–17839. https://doi.org/10.1609/aaai.v39i17.34946.

Wang, H.; Fu, T.; Du, Y.; Gao, W.; Huang, K.; Liu, Z.; Chandak, P.; Liu, S.; Van Katwyk, P.; Deac, A.; et al. Scientific discovery in the age of artificial intelligence. *Nature* **2023**, *620*, 47–60.

Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A.J.; Bambrick, J.; et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **2024**, *630*, 493–500.

Price, I.; Sanchez-Gonzalez, A.; Alet, F.; Andersson, T.R.; El-Kadi, A.; Masters, D.; Ewalds, T.; Stott, J.; Mohamed, S.; Battaglia, P.; et al. Probabilistic weather forecasting with machine learning. *Nature* **2025**, *637*, 84–90.

Batatia, I.; Kovacs, D.P.; Simm, G.; Ortner, C.; Csányi, G. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. In Proceedings of the Advances in Neural Information Processing Systems, 2022, Vol. 35, pp. 11423–11436.

Brunton, S.L.; Kutz, J.N. Promising directions of machine learning for partial differential equations. *Nature Computational Science* **2024**, *4*, 483–494.

Li, Z.; Kovachki, N.; Azizzadenesheli, K.; Liu, B.; Bhattacharya, K.; Stuart, A.; Anandkumar, A. Fourier Neural Operator for Parametric Partial Differential Equations. In Proceedings of the The Ninth International Conference on Learning Representations (ICLR 2021), 2021. arXiv:2010.08895.

Lu, C.; Hu, S.; Clune, J. The AI Scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066* **2025**.

Gottweis, J.; Weng, W.H.; Daryin, A.; Tu, T.; Palepu, A.; Sirkovic, P.; Myaskovsky, A.; Weissenberger, F.; Rong, K.; Tanno, R.; et al. Towards an AI co-scientist. *arXiv preprint arXiv:2502.18864* **2025**.

Kilian, L.; Lütkepohl, H. *Structural Vector Autoregressive Analysis*; Cambridge University Press: Cambridge, 2017. https://doi.org/10.1017/9781108164818.

Blanchard, O.J.; Quah, D. The Dynamic Effects of Aggregate Demand and Supply Disturbances. *American Economic Review* **1989**, *79*, 655–673.

Uhlig, H. What are the Effects of Monetary Policy on Output? Results from an Agnostic Identification Procedure. *Journal of Monetary Economics* **2005**, *52*, 381–419. https://doi.org/10.1016/j.jmoneco.2004.05.007.

Albergo, M.S.; Vanden-Eijnden, E. Building Normalizing Flows with Stochastic Interpolants. In Proceedings of the The Eleventh International Conference on Learning Representations (ICLR 2023), 2023. arXiv:2209.15571.

Chen, R.T.Q.; Rubanova, Y.; Bettencourt, J.; Duvenaud, D. Neural Ordinary Differential Equations. In Proceedings of the Advances in Neural Information Processing Systems 31 (NeurIPS 2018), 2018, pp. 6572–6583.

Tancogne-Dejean, N.; Oliveira, M.J.T.; Andrade, X.; Appel, H.; Borca, C.H.; Le Breton, G.; Buchholz, F.; Castro, A.; Corni, S.; Correa, A.A.; et al. Octopus, a Computational Framework for Exploring Light-Driven Phenomena and Quantum Dynamics in Extended and Finite Systems. *Journal of Chemical Physics* **2020**, *152*, 124119. https://doi.org/10.1063/1.5142502.

Peters, J.; Janzing, D.; Schölkopf, B. *Elements of Causal Inference: Foundations and Learning Algorithms*; MIT Press, 2017.

Schölkopf, B.; Janzing, D.; Peters, J.; Sgouritsa, E.; Zhang, K.; Mooij, J. On causal and anticausal learning. In Proceedings of the Proceedings of the 29th International Conference on Machine Learning (ICML), 2012, pp. 459–466.

Simpson, E.H. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)* **1951**, *13*, 238–241.

van der Vaart, A.W. *Asymptotic Statistics*; Cambridge University Press: Cambridge, 2000.

 doi:10.20944/preprints202605.0617.v1

Wasserman, L. *All of Nonparametric Statistics*; Springer: New York, 2006.

Valiant, L.G. A Theory of the Learnable. *Communications of the ACM* **1984**, *27*, 1134–1142. https://doi.org/10.1145/1968.1972.

Hoeffding, W. Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association* **1963**, *58*, 13–30. https://doi.org/10.1080/01621459.1963.10500830.

Kearns, M.J.; Vazirani, U.V. *An Introduction to Computational Learning Theory*; MIT Press: Cambridge, MA, 1994.

Hamilton, J.D. *Time Series Analysis*; Princeton University Press: Princeton, NJ, 1994.

VanderWeele, T.J.; Ding, P. Sensitivity analysis in observational research: introducing the E-value. *Annals of Internal Medicine* **2017**, *167*, 268–274.

Robins, J.M.; Rotnitzky, A.; Scharfstein, D.O. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. *IMA Volumes in Mathematics and its Applications* **2000**, *116*, 1–94.

Zheng, X.; Aragam, B.; Ravikumar, P.; Xing, E.P. DAGs with NO TEARS: Continuous Optimization for Structure Learning. In Proceedings of the Advances in Neural Information Processing Systems 31 (NeurIPS 2018), 2018, pp. 9472–9483.

Brock, W.A.; Scheinkman, J.A.; Dechert, W.D.; LeBaron, B. A test for independence based on the correlation dimension. *Econometric Reviews* **1996**, *15*, 197–235.

Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B* **1995**, *57*, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Man, C.D.; Micheletto, F.; Lv, D.; Breton, M.; Kovatchev, B.; Cobelli, C. The UVA/PADOVA Type 1 Diabetes Simulator: New Features. *Journal of Diabetes Science and Technology* **2014**, *8*, 26–34. https://doi.org/10.1177/1932296813514502.

Cheng, H.; et al. CausalTime: Realistically Generated Time-Series for Benchmarking of Causal Discovery. In Proceedings of the The Twelfth International Conference on Learning Representations (ICLR 2024), 2024.

Herdeanu, B.; et al. CausalDynamics: A Large-Scale Benchmark for Structural Discovery of Dynamical Causal Models. In Proceedings of the Advances in Neural Information Processing Systems 38 (NeurIPS 2025), 2025.

Zhao, Q.; Zhu, J.; Fang, Q.; Pan, Y.; Wang, Y.; Shao, J.; Xie, W. Chinese Diabetes Datasets for Data-Driven Machine Learning. *Scientific Data* **2023**, *10*, 35. https://doi.org/10.1038/s41597-023-01940-7.

Kovatchev, B.P.; Breton, M.; Dalla Man, C.; Cobelli, C. In Silico Preclinical Trials: A Proof of Concept in Closed-Loop Control of Type 1 Diabetes. *Journal of Diabetes Science and Technology* **2009**, *3*, 44–55. https://doi.org/10.1177/193229680900300106.

Cobelli, C.; Kovatchev, B. Developing the UVA/Padova Type 1 Diabetes Simulator: Modeling, Validation, Refinements, and Utility. *Journal of Diabetes Science and Technology* **2023**, *17*, 1493–1505. https://doi.org/10.1177/19322968221082513.

Xie, J. Simglucose v0.2.1. https://github.com/jxx123/simglucose, 2018.

Antoni, T.; et al. The Cosmic-Ray Experiment KASCADE. *Nuclear Instruments and Methods in Physics Research A* **2003**, *513*, 490–510. https://doi.org/10.1016/S0168-9002(03)02076-X.

Haungs, A.; et al. The KASCADE Cosmic-Ray Data Centre KCDC: Granting Open Access to Astroparticle Physics Research Data. *European Physical Journal C* **2018**, *78*, 741. https://doi.org/10.1140/epjc/s10052-018-6221-2.

Matthews, J. A Heitler Model of Extensive Air Showers. *Astroparticle Physics* **2005**, *22*, 387–397. https://doi.org/10.1016/j.astropartphys.2004.09.003.

Giannozzi, P.; Baroni, S.; Bonini, N.; Calandra, M.; Car, R.; Cavazzoni, C.; Ceresoli, D.; Chiarotti, G.L.; Cococcioni, M.; Dabo, I.; et al. QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *Journal of Physics: Condensed Matter* **2009**, *21*, 395502.

Krausz, F.; Ivanov, M. Attosecond physics. *Reviews of Modern Physics* **2009**, *81*, 163–234.

Lewenstein, M.; Balcou, P.; Ivanov, M.Y.; L'Huillier, A.; Corkum, P.B. Theory of High-Harmonic Generation by Low-Frequency Laser Fields. *Physical Review A* **1994**, *49*, 2117–2132. https://doi.org/10.1103/PhysRevA.49.2117.

Krause, J.L.; Schafer, K.J.; Kulander, K.C. High-order harmonic generation from atoms and ions in the high intensity regime. *Physical Review Letters* **1992**, *68*, 3535–3538.

Perdew, J.P.; Zunger, A. Self-interaction correction to density-functional approximations for many-electron systems. *Physical Review B* **1981**, *23*, 5048–5079.

Bareinboim, E.; Pearl, J. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences* **2016**, *113*, 7345–7352.

Peters, J.; Bühlmann, P.; Meinshausen, N. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B* **2016**, *78*, 947–1012.

 doi:10.20944/preprints202605.0617.v1

Arjovsky, M.; Bottou, L.; Gulcevich, I.; Lopez-Paz, D. Invariant Risk Minimization. *arXiv:1907.02893* **2019**.

Toth, C.; Lorch, L.; Knoll, C.; Krause, A.; Pernkopf, F.; Peharz, R.; von Kügelgen, J. Active Bayesian Causal Inference. In Proceedings of the Advances in Neural Information Processing Systems 35 (NeurIPS 2022), 2022, pp. 16261–16275.

Schölkopf, B.; Locatello, F.; Bauer, S.; Ke, N.R.; Kalchbrenner, N.; Goyal, A.; Bengio, Y. Toward Causal Representation Learning. *Proceedings of the IEEE* **2021**, *109*, 612–634.

Zhang, K.; Hyvärinen, A. On the Identifiability of the Post-Nonlinear Causal Model. *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI 2009)* **2009**, pp. 647–655.

Bühlmann, P.; Peters, J.; Ernest, J. CAM: Causal Additive Models, high-dimensional order search and penalized regression. *Annals of Statistics* **2014**, *42*, 2526–2556.

Rolland, P.; Cevher, V.; Kleindessner, M.; Russell, C.; Schölkopf, B.; Janzing, D.; Locatello, F. Score Matching Enables Causal Discovery of Nonlinear Additive Noise Models. In Proceedings of the Proceedings of the 39th International Conference on Machine Learning (ICML 2022), 2022, Vol. 162, *PMLR*, pp. 18741–18753.

Montagna, F.; Noceti, N.; Rosasco, L.; Zhang, K.; Locatello, F. Scalable Causal Discovery with Score Matching. In Proceedings of the Proceedings of the Second Conference on Causal Learning and Reasoning (CLeaR 2023), 2023, Vol. 213, *PMLR*, pp. 752–771.

Montagna, F.; Faller, P.M.; Bloebaum, P.; Kirschbaum, E.; Locatello, F. Score matching through the roof: linear, nonlinear, and latent variables causal discovery. In Proceedings of the Proceedings of the Fourth Conference on Causal Learning and Reasoning (CLeaR), 2025, Vol. 275, *Proceedings of Machine Learning Research*, pp. 552–605.

Li, C.; Shen, X.; Pan, W. Nonlinear Causal Discovery with Confounders. *Journal of the American Statistical Association* **2023**.

Meier, D.; Hiremath, S.; Ghosal, P.; Gan, K. When Additive Noise Meets Unobserved Mediators: Bivariate Denoising Diffusion for Causal Discovery. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2025.

Mooij, J.M.; Janzing, D.; Schölkopf, B. From ordinary differential equations to structural causal models: the deterministic case. In Proceedings of the Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI), 2013, pp. 440–448.

Yang, K.; Katcoff, A.; Uhler, C. Characterizing and Learning Equivalence Classes of Causal DAGs under Interventions. In Proceedings of the Proceedings of the 35th International Conference on Machine Learning (ICML). PMLR, 2018, pp. 5537–5546.

Severson, K.A.; Attia, P.M.; Jin, N.; Perkins, N.; Jiang, B.; Yang, Z.; Chen, M.H.; Aykol, M.; Herring, P.K.; Fraggedakis, D.; et al. Data-Driven Prediction of Battery Cycle Life before Capacity Degradation. *Nature Energy* **2019**, *4*, 383–391. https://doi.org/10.1038/s41560-019-0356-8.

Bloom, I.; Cole, B.W.; Sohn, J.J.; Jones, S.A.; Polzin, E.G.; Battaglia, V.S.; Henriksen, G.L.; Motloch, C.; Richardson, R.; Unkelhaeuser, T.; et al. An Accelerated Calendar and Cycle Life Study of Li-ion Cells. *Journal of Power Sources* **2001**, *101*, 238–247. https://doi.org/10.1016/S0378-7753(01)00783-2.

Saha, B.; Goebel, K. Battery Data Set. NASA Ames Prognostics Data Repository, 2007.

Giannozzi, P.; Andreussi, O.; Brumme, T.; Bunau, O.; Buongiorno Nardelli, M.; Calandra, M.; Car, R.; Cavazzoni, C.; Ceresoli, D.; Cococcioni, M.; et al. Advanced capabilities for materials modelling with Quantum ESPRESSO. *Journal of Physics: Condensed Matter* **2017**, *29*, 465901.

Li, Z.; Kovachki, N.; Azizzadenesheli, K.; Liu, B.; Bhatt, K.; Stuart, A.; Anandkumar, A. Fourier Neural Operator for Parametric Partial Differential Equations. In Proceedings of the International Conference on Learning Representations (ICLR), 2021.

Komanduri, A.; Wu, X.; Wu, Y.; Chen, F. From Identifiable Causal Representations to Controllable Counterfactual Generation: A Survey on Causal Generative Modeling. *Transactions on Machine Learning Research* **2024**.

Brignone, D.; Piffer, M. A Structural VAR Model for the UK Economy. Macro Technical Paper 3, Bank of England, 2025.

Homaei, M.H.; Tarif, M.; Rodríguez, P.G.; Caro, A.; Ávila, M. Causal Digital Twins for Cyber-Physical Security in Water Systems: A Framework for Robust Anomaly Detection. *Machine Learning with Applications* **2025**, *23*, 100824. https://doi.org/10.1016/j.mlwa.2025.100824.

 doi:10.20944/preprints202605.0617.v1

disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.