Causal Heteroscedastic Structure Learning From Incomplete Temporal data

Abdellah Rahmani
LTS4, EPFL
Lausanne, Switzerland
abdellah.rahmani@epfl.ch

Pascal Frossard
LTS4, EPFL
Lausanne, Switzerland
pascal.frossard@epfl.ch

Abstract

Discovering causal structures in multivariate time series (MTS) is critical in domains such as finance and neuroscience, yet remains challenging in practice due to three distinct obstacles. First, MTS are routinely corrupted by missing values, where naive impute-then-discover pipelines distort the underlying data distribution and yield spurious causal graphs. Second, identifying instantaneous effects is difficult even on complete data, and missingness further entangles imputation with graph recovery. Third, real-world data frequently exhibits heteroscedastic noise with variance driven by both instantaneous and lagged causes, violating the core assumptions of standard discovery methods. Existing approaches fail to address these challenges concurrently, typically presuming complete data, homoscedastic noise, or strictly time-lagged relationships. Crucially, we show through ablation studies that resolving these issues sequentially, rather than jointly, leads to a substantial degradation in performance. To overcome this, we introduce Cheesefill, an optimal transport framework for causal structure learning under missingness and heteroscedastic noise. Cheesefill parameterizes causal mechanisms using conditional normalizing flows and jointly learns a stochastic correction map that refines naive imputations into trajectories consistent with the inferred causal mechanisms. Theoretically, we prove that optimizing over stochastic correction maps is equivalent to a conditional kernel reformulation of the Kantorovich problem, from which our neural parameterization yields a tractable objective. Extensive experiments on synthetic and real-world benchmarks demonstrate that Cheesefill consistently outperforms existing baselines across diverse missingness regimes.

1 Introduction

Causal structure learning from multivariate time series (MTS) is a fundamental problem with applications spanning traffic modeling [7], biology [38], climate science [37, 33], and healthcare [40]. Yet recovering causal relationships from real-world MTS confronts three intertwined challenges. First, observations are routinely corrupted by missing values [45], typically due to sensor failures or transmission loss; robustness to such missingness is essential but remains underexplored. Second, when the sampling interval is coarse relative to the underlying dynamics, multiple events collapse onto the same timestamp, giving rise to instantaneous causal links that any faithful causal model must capture. Third, real MTS exhibit heteroscedastic noise [15] whose variance depends on both instantaneous and lagged causes, a phenomenon documented in fMRI [39], EEG [18, 34], and financial data [16].

Most existing causal discovery methods for MTS assume complete observations and homoscedastic noise. On complete MTS, the field has developed in three directions: Granger-based neural approaches [24, 4] that infer dependencies through component-wise prediction; score-based formulations such as Dynotears [29], which recovers instantaneous and lagged links via the differentiable acyclicity

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

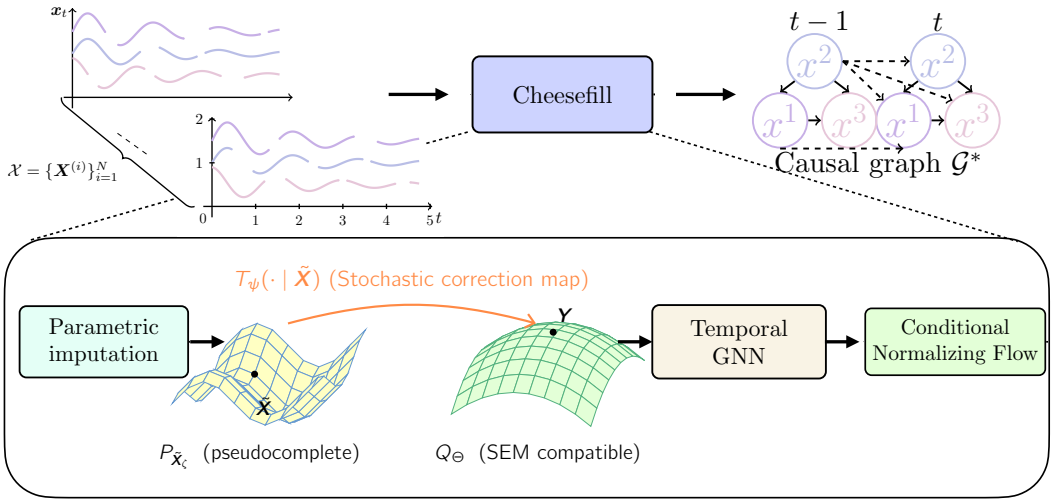


Figure 1: Cheesefill overview. The incomplete MTS dataset $\mathcal{X}$ is first processed by a parametric imputer with parameters ζ. A stochastic correction map T_ψ refines $\tilde{\boldsymbol{X}}$ into a sample $\boldsymbol{Y}$ consistent with the structural equation model marginal Q_Θ. TGNN and CNF, models causal mechanisms and heteroscedastic noise. Joint optimisation over (Θ, ψ, ζ) recovers the temporal causal graph $\mathcal{G}^*$.

constraint introduced by Zheng et al. [46]; and constraint-based methods such as PCMCI+ [36], which scale conditional independence testing to high-dimensional time series with both lagged and contemporaneous links. Closer to our setting, Gong et al. [14] consider structural equation models with historically dependent noise, and Rahmani et al. [32] handle heteroscedasticity in stationary and non-stationary regimes, but again on complete MTS. When data is missing, the prevailing approach is a two-stage pipeline: fill the gaps with zero padding, interpolation, or learned imputer [9, 5, 44], then application of a causal discovery algorithm. However, recent work [8, 6, 43] shows that lower reconstruction error does not translate into better recovered graphs: the proxy complete distribution induced by imputation encodes a different set of conditional independencies, and may therefore be compatible with a causal structure that is distinct from the actual ground truth. This calls for a joint, end-to-end treatment of imputation and structure learning. Existing methods in this direction [8, 6] however assume purely time-lagged relationships and homoscedastic noise, leaving the simultaneous treatment of instantaneous links, heteroscedasticity, and missingness as an open problem. A detailed related work appears in Appendix A.

To solve the three challenges, we propose Cheesefill (illustrated in Figure 1), an optimal transport (OT) framework [26] for causal structure learning from MTS with instantaneous and lagged effects, heteroscedastic noise [32], and missing values. To the best of our knowledge, Cheesefill is the first method to recover instantaneous and time-lagged links from incomplete MTS, with or without heteroscedastic noise. Given an MTS with missing entries, Cheesefill couples Bayesian structure learning [13, 32, 14], which filters out spurious edges, with a conditional normalizing flow [10] that models heteroscedastic noise, while simultaneously learning a stochastic correction map that refines naive imputations into samples consistent with the causal model. On the theoretical side, we show that optimizing over admissible stochastic correction maps is equivalent to a conditional kernel reformulation of the Kantorovich problem [21]; the neural parameterization used by Cheesefill then yields a tractable objective. On the experimental side, Cheesefill consistently outperforms causal discovery methods designed for incomplete MTS and naive *impute then discover* pipelines on synthetic benchmarks and two real world datasets. Our contributions are:

- We introduce Cheesefill, a framework for recovering instantaneous and time lagged causal links from incomplete MTS under homoscedastic or heteroscedastic noise.
- We cast the problem as OT-based density fitting, combining Bayesian structure learning, conditional normalizing flows for heteroscedastic mechanisms, and a stochastic correction map for aligning naive imputations with the causal model.
- We show that admissible stochastic correction maps give a conditional-kernel reformulation of the Kantorovich problem; Cheesefill's neural correction map yields a tractable objective.

- We demonstrate consistent gains over different baselines across synthetic and real datasets under multiple missingness mechanisms.

2 Preliminaries

Notation. We use calligraphic letters (e.g. $\mathcal{X}$) for sets, uppercase bold ($\boldsymbol{G}_\tau$) for matrices, lowercase bold ($\boldsymbol{x}_t$) for vectors, and lowercase letters ($x^i_{t-\tau}$) for scalars. Ground-truth quantities are marked with an asterisk, e.g. $\mathcal{G}^*$. We assume that all distributions $P(\boldsymbol{X})$ admit lower bounded densities $p(\boldsymbol{x}_t)$ with respect to the Lebesgue measure. The notation $[|0:L|]$ denotes the integer set $\{0,\dots,L\}$, and $|\cdot|$ denotes set cardinality. A MTS is written $\boldsymbol{X} = (\boldsymbol{x}_t)_{t\in\mathcal{T}} = (x^i_t)_{i\in\mathbf{V},\, t\in\mathcal{T}}$, with $|\mathbf{V}| = d$ components and length $|\mathcal{T}|$, where $\mathcal{T}$ is the time index set. We write $\boldsymbol{x}_{<t}$ for the history $\{\boldsymbol{x}_{t-L},\dots,\boldsymbol{x}_{t-1}\}$.

Optimal Transport (OT). A central object in our framework is the cost of transforming one distribution into another. Kantorovich's formulation [21] of this problem reads

$$W_c(P,Q) := \inf_{\Gamma\in\mathcal{P}(X\sim P,\, Y\sim Q)} \mathbb{E}_{(X,Y)\sim\Gamma}\big[c(X,Y)\big], \tag{1}$$

where $c : \Omega_X \times \Omega_X \to \mathbb{R}_+$ is any measurable cost function on $(\Omega_X, \mathcal{B}(\Omega_X))$, the measurable space of complete length-$|\mathcal{T}|$, d-variate MTS trajectories, which we assume to be a standard Borel space, and $\mathcal{P}(X\sim P, Y\sim Q)$ denotes the set of all couplings of (X,Y) with marginals P and Q. We will also use an equivalent conditional representation of couplings, defined as follows:

Definition 2.1 (Admissible stochastic transport kernel)**.** A stochastic kernel $T(dy \mid x)$ from Ω_X to Ω_X is admissible between P and Q if, for every measurable set $A \in \mathcal{B}_X$,

$$\int_{\Omega_X} T(A \mid x)\, dP(x) = Q(A).$$

We denote the set of admissible kernels between P and Q by $\mathcal{F}(P,Q)$.

On the finite dimensional trajectory spaces, as considered in this paper, every coupling $\Gamma \in \mathcal{P}(X\sim P,\, Y\sim Q)$ admits a conditional representation

$$\Gamma(dx,dy) = P(dx)T_\Gamma(dy \mid x).$$

We now introduce the object our method recovers. Following [1, 32, 14], we represent the causal structure of an MTS as a collection of lag-specific adjacency matrices, which separates instantaneous links ($\tau = 0$) from lagged ones ($\tau > 0$).

Definition 2.2 (Temporal Causal Graph [32, 33])**.** The temporal causal graph associated with an MTS $\boldsymbol{X} = (\boldsymbol{x}_t)_{t\in\mathcal{T}}$ is a Directed Acyclic Graph (DAG) $\mathcal{G} = (\mathbf{V}, E)$, represented by a collection of adjacency matrices $\{\boldsymbol{G}_\tau\}_{\tau\in[|0:L|]} = \{\boldsymbol{G}_0,\dots,\boldsymbol{G}_L\}$, with a fixed maximum lag L. Its vertices $\mathbf{V}$ consist of the components $x^1_{t'},\dots,x^d_{t'}$ for each $t' \in [|t-L:t|]$. For $\tau \in [|1:L|]$, the variables $x^i_{t-\tau}$ and x^j_t are connected by a lag-specific directed link $x^i_{t-\tau} \to x^j_t$ in $\mathcal{G}$ if and only if x^i at time $t-\tau$ causes x^j at time t; in that case $[G_\tau]_{ij}$ is non-zero and $x^i \in \mathbf{Pa}^{<t}_j$, the set of lagged parents of j. For instantaneous links ($\tau = 0$), self-loops are excluded ($i \neq j$): we have an edge $x^i_t \to x^j_t$ with $x^i \in \mathbf{Pa}^t_j$, the set of instantaneous parents of j at time t, if and only if x^i at time t causes x^j at time t.

Structural equation model (SEM). Let $\mathcal{X} = \big\{\boldsymbol{X}^{(1)},\cdots,\boldsymbol{X}^{(N)}\big\}$ be a collection of N MTS, where each $\boldsymbol{X}^{(j)}$ has length $|\mathcal{T}|$ and consists of $|\mathbf{V}| = d$ variables $\{x^{i,(j)}_t\}_{t\in\mathcal{T},\, i\in\mathbf{V}}$. We assume that all $\boldsymbol{X}^{(j)} \in \mathcal{X}$ share the same temporal causal graph $\mathcal{G}$ with maximum lag L as in Definition 2.2. The observation $\boldsymbol{x}^{(j)}_t$ then follows the general SEM

$$x^{i,(j)}_t = f^i\big(\mathbf{Pa}^{<t}_i,\, \mathbf{Pa}^t_i\big) + g^i\big(\mathbf{Pa}^{<t}_i,\, \mathbf{Pa}^t_i\big)\cdot \epsilon^{i,(j)}_t, \tag{2}$$

where f^i and g^i are general differentiable functions with g^i strictly positive, and $\epsilon^{i,(j)}_t$ follows an arbitrary density p_ϵ with $\mathbb{E}(\epsilon^{i,(j)}_t) = 0$ and $\mathbb{E}\big((\epsilon^{i,(j)}_t)^2\big) = 1$ (without loss of generality). This single SEM covers both regimes that interest us: when g^i is constant we recover the homoscedastic case, and when g^i depends on the parents we capture the heteroscedastic noise observed in fMRI, EEG.

SEM-induced mechanism density. Eq. (C.1) not only specifies how to sample x_t^i, it also induces a density on the value x_t^i once the parent values are fixed. For any candidate parent values $\mathbf{p}^{<t} \in \mathbb{R}^{|\mathbf{Pa}_i^{<t}|}$ and $\mathbf{p}^t \in \mathbb{R}^{|\mathbf{Pa}_i^t|}$, the random variable $f^i(\mathbf{p}^{<t}, \mathbf{p}^t) + g^i(\mathbf{p}^{<t}, \mathbf{p}^t) \cdot \epsilon$ admits a density on $\mathbb{R}$ which we denote $\rho^i\big(x\,;\, \mathbf{p}^{<t}, \mathbf{p}^t, \mathcal{G}\big)$ with $\mathbb{E}[\rho^i\big(\cdot\,;\, \mathbf{p}^{<t}, \mathbf{p}^t, \mathcal{G}\big)] = f^i(\mathbf{p}^{<t}, \mathbf{p}^t)$ and $\mathbb{E}\big[(\rho^i\big(\cdot\,;\, \mathbf{p}^{<t}, \mathbf{p}^t, \mathcal{G}\big) - f^i(\mathbf{p}^{<t}, \mathbf{p}^t))^2\big] = g^i(\mathbf{p}^{<t}, \mathbf{p}^t)^2$, with the dependence on $\mathcal{G}$ entering through the parent sets $\mathbf{Pa}_i^{<t}$, $\mathbf{Pa}_i^t$ used to read off $\mathbf{p}^{<t}, \mathbf{p}^t$.

We use ρ^i throughout the paper as a *mechanism level* density that scores observations against the SEM: it answers the question "*given that the parents took these values, how plausible is the observation x?* ". What makes ρ^i central to our setting is the fact that it is agnostic to the source of the parent values, they may come from the data, from a model roll-out, or from an imputed sample, and the density remains the same family indexed by those inputs. This locality is what makes ρ^i usable under missingness, where parent values must themselves be inferred.

Missingness. In practice, the N MTS $\mathcal{X} = \{\boldsymbol{X}^{(1)}, \cdots, \boldsymbol{X}^{(N)}\}$ we observe are rarely complete. We model missingness through a binary indicator matrix $\boldsymbol{M}^{(j)} \in \{0,1\}^{|\mathcal{T}| \times d}$, where $\boldsymbol{M}_t^{i,(j)} = 1$ marks the i^{th} feature at timestamp t of MTS (j) as missing and $\boldsymbol{M}_t^{i,(j)} = 0$ marks it as observed. Missing entries are initially set to NaN, and they arise from a range of practical causes including entry errors, deliberate non responses, and sampling drops [23]. Following Little and Rubin [23], missingness is classified into three regimes: *missing completely at random* (MCAR), where the cause is independent of the data; *missing at random* (MAR), where the probability of being missing depends only on observed values; and *missing not at random* (MNAR), where the cause itself is unobserved.

Identifiability. Throughout the paper, and following [14, 33, 32], we use the standard assumptions of temporal causal discovery: the causal Markov property, stationarity, minimality, and causal sufficiency; see Appendix J.1 for precise statements. Under these assumptions, temporal SEMs with heteroscedastic noise have been shown to be uniquely identifiable from the observed distribution [32]. For MCAR and MAR data, the missingness process is ignorable and the parameters can be consistently estimated based only on the observed data under large sample assumption[23].

MNAR data is fundamentally more challenging because the observation process may depend on unobserved values and is therefore generally non ignorable. In the fully general case, the joint distribution is not recoverable without additional information; existing works use a missingness graph to establish graphical conditions for recovery [3, 27, 25]. Such structural knowledge is difficult to obtain in practice, especially in our setting where the causal graph is also unknown. We therefore find it crucial to develop a principled approach that does not assume access to the underlying missingness mechanism. While we do not claim arbitrary MNAR identifiability without assumptions on the observation process, we provide extensive empirical evidence in Appendix G showing that Cheesefill remains effective in MNAR cases.

3 Problem formulation

Let $(\Omega_X, \mathcal{B}_X)$ denote the measurable space of complete length $|\mathcal{T}|$, d variate MTS trajectories, and let $P_{\boldsymbol{X}}$ be the unknown complete data distribution on Ω_X. We observe a finite collection of incomplete MTS $\mathcal{X} = \{\boldsymbol{X}^{(1)}, \cdots, \boldsymbol{X}^{(N)}\}$ with associated masks $\mathcal{M} = \{\boldsymbol{M}^{(1)}, \cdots, \boldsymbol{M}^{(N)}\}$. We assume that each latent complete trajectory $\boldsymbol{X}_\star^{(j)} \in \Omega_X$ is drawn i.i.d. from $P_{\boldsymbol{X}}$, generated by the SEM in Eq. (C.1) under a shared $\mathcal{G}^*$. The observed incomplete MTS $\boldsymbol{X}^{(j)}$ is obtained by corrupting $\boldsymbol{X}_\star^{(j)}$ according to the mask $\boldsymbol{M}^{(j)}$, under MCAR, MAR, or MNAR missingness. Given the finite dataset $(\mathcal{X}, \mathcal{M})$, our goal is to jointly recover $\mathcal{G}^*$ and the mechanism parameters Θ of $(f^i, g^i)_{i \in \mathbf{V}}$.

3.1 Idealised density-fitting formulation

OT motivation. We frame causal recovery as a density-fitting problem using the Kantorovich OT formulation. Unlike complete-data causal discovery, where the SEM can be scored directly on observed trajectories, the complete trajectories are unknown under missingness. Directly maximizing a score on fixed imputations would make the learned graph depend on artifacts of the imputer. The OT formulation instead compares the pseudo complete distribution and the SEM induced distribution, providing a principled way to couple imputation and graph learning through mechanism compatibility rather than treating imputation as a separate preprocessing step.

In the idealised complete data setting, we would seek $(\Theta, \mathcal{G})$ that minimise the OT cost between the model distribution Q_Θ and the data distribution $P_{\boldsymbol{X}}$,

$$\Theta^*, \mathcal{G}^* \;=\; \arg\min_{\Theta, \mathcal{G}} \; W_{c_\mathcal{G}}(P_{\boldsymbol{X}}, Q_\Theta) \quad \text{subject to } \mathbf{G}_0 \in \text{DAGs}, \tag{3}$$

where $W_{c_\mathcal{G}}$ is the Kantorovich cost in Eq. (1). The choice of transport cost is crucial. A geometric cost such as $\|X - Y\|^2$ measures pointwise proximity between trajectories, but causal structure is identified through local conditional mechanisms, not through ambient Euclidean distance. We therefore use the SEM mechanism as the transport cost: $c_\mathcal{G}(\boldsymbol{X}, \boldsymbol{Y}) = -\sum_{t=1}^{|\mathcal{T}|} \sum_{i=1}^{d} \log \rho^i\big(x\,;\, \mathbf{p}^{<t}, \mathbf{p}^t, \mathcal{G}\big)$.

For a fixed candidate parent trajectory $\boldsymbol{Y}$, this cost measures how plausible the child values in $\boldsymbol{X}$ are under the mechanisms specified by graph G. Thus, low transport cost requires not only distributional proximity, but compatibility with the causal mechanisms of the candidate SEM. Eq. (3) cannot be solved directly in our setting for two reasons: (i) we do not have access to samples from $P_{\boldsymbol{X}}$, only to the corrupted collection $(\mathcal{X}, \mathcal{M})$; and (ii) defining a model distribution Q_Θ over incomplete MTS is not straightforward. Sections 3.2 and 4 introduce Cheesefill, which resolves both through a stochastic correction map and a tractable parametric objective that can be optimized jointly with the causal graph and mechanisms.

3.2 Tractable density-fitting formulation

Cheesefill addresses the aforementioned obstacles in section 3.1 in two steps. First, a parametric initializer converts the incomplete observations into pseudo complete trajectories, giving a tractable source distribution. Second, a stochastic correction map transports these pseudo complete trajectories toward refined trajectories used by the SEM mechanisms.

Intuition. Recovering causal structure from incomplete MTS does not require necessarily reconstructing the full observational marginal $P_{\boldsymbol{X}}$. In our setting, the graph is identified through mechanism-level constraints. Thus, the role of imputation is not merely to minimize reconstruction error, but to produce completed trajectories whose conditional mechanisms are compatible with those implied by the candidate SEM. *Cheesefill implements this principle in two stages.* A parametric initializer first produces regular shaped pseudo complete trajectories. A stochastic correction map then adjusts these trajectories toward completions that both preserve the observed entries and better satisfy the local mechanism densities ρ^i_Θ.

Parametric imputer. Cheesefill first applies a parametric imputer (PI) h_ζ to fill the missing entries of each MTS, yielding pseudo complete MTS $\tilde{\mathcal{X}}_\zeta = \{\tilde{\boldsymbol{X}}^{(1)}, \ldots, \tilde{\boldsymbol{X}}^{(N)}\}, \tilde{\boldsymbol{X}} \sim P_{\tilde{\boldsymbol{X}}_\zeta}$. Thus, $P_{\tilde{\boldsymbol{X}}_\zeta}$ is the PI induced distribution on the complete MTS trajectory space Ω_X, providing a tractable source in place of the unavailable complete data distribution $P_{\boldsymbol{X}}$ in Eq. (3). Directly minimizing $W_{c_\mathcal{G}}(P_{\tilde{\boldsymbol{X}}_\zeta}, Q_\Theta)$ can inherit the failure mode of impute then discover pipelines [8, 6].

Stochastic correction map. The pseudo complete distribution $P_{\tilde{\boldsymbol{X}}_\zeta}$ gives a tractable source distribution, but its samples may contain dependencies induced by the imputer rather than by the underlying SEM. Cheesefill therefore does not treat pseudo complete trajectories as fixed causal evidence. Instead, Cheesefill introduces a *stochastic correction map* $T_\psi(d\boldsymbol{Y} \mid \tilde{\boldsymbol{X}})$, which transports each pseudo complete trajectory $\tilde{\boldsymbol{X}}$ to a refined trajectory whose values instantiate the instantaneous and lagged parents in the mechanism cost $c_\mathcal{G}(\tilde{\boldsymbol{X}}, \boldsymbol{Y})$. Thus, the following proposition shows that optimizing over such stochastic correction maps is not merely a heuristic relaxation: in the unrestricted case, it is exactly a conditional kernel reformulation of the Kantorovich problem.

Proposition 3.1. *Let $(\Omega_X, \mathcal{B}(\Omega_X))$ be the measurable space of complete length $|\mathcal{T}|$, d variate MTS trajectories. Assume that Ω_X is a standard Borel space. Let $P_{\tilde{\boldsymbol{X}}_\zeta}$ be the pseudo complete data distribution on Ω_X, and let Q_Θ be the SEM induced complete trajectory distribution on Ω_X. Let $c_\mathcal{G} : \Omega_X \times \Omega_X \to \mathbb{R}_+$ be a measurable mechanism cost satisfying $\mathbb{E}_{(X,Y)\sim\Gamma}\big[|c_\mathcal{G}(X, Y)|\big] < \infty$ for all couplings Γ considered. Then*

$$W_{c_\mathcal{G}}\big(P_{\tilde{\boldsymbol{X}}_\zeta}, Q_\Theta\big) = \inf_{T \in \mathcal{F}(P_{\tilde{\boldsymbol{X}}_\zeta}, Q_\Theta)} \mathbb{E}_{\tilde{\boldsymbol{X}} \sim P_{\tilde{\boldsymbol{X}}_\zeta}} \mathbb{E}_{\boldsymbol{Y} \sim T(\cdot \mid \tilde{\boldsymbol{X}})}\big[c_\mathcal{G}(\tilde{\boldsymbol{X}}, \boldsymbol{Y})\big], \tag{4}$$

where $\mathcal{F}(P_{\tilde{\boldsymbol{X}}_\zeta}, Q_\Theta)$ is the set of admissible stochastic kernels in Definition 2.1. Moreover, for any parametric family $\{T_\psi\}_{\psi\in\Psi}$, let $\Psi_{\mathrm{adm}} = \{\psi \in \Psi : T_\psi \in \mathcal{F}(P_{\tilde{\boldsymbol{X}}_\zeta}, Q_\Theta)\}$. Then

$$W_{c_\mathcal{G}}\big(P_{\tilde{\boldsymbol{X}}_\zeta}, Q_\Theta\big) \le \inf_{\psi\in\Psi_{\mathrm{adm}}} \mathbb{E}_{\tilde{\boldsymbol{X}}\sim P_{\tilde{\boldsymbol{X}}_\zeta}} \mathbb{E}_{\boldsymbol{Y}\sim T_\psi(\cdot|\tilde{\boldsymbol{X}})}\big[c_\mathcal{G}(\tilde{\boldsymbol{X}}, \boldsymbol{Y})\big]. \tag{5}$$

The inequality is tight whenever the parametric family contains an optimal admissible kernel. More generally, if the family contains an ε-optimal admissible kernel, then the right-hand side is at most ε above the unrestricted Kantorovich optimum.

Proposition 3.1, *proof* in Appendix J.2, justifies the correction map as a principled search over couplings rather than an auxiliary imputation heuristic. If the class of admissible kernels is unrestricted, minimizing the expected mechanism cost over T is exactly equivalent to minimizing the Kantorovich cost between $P_{\tilde{\boldsymbol{X}}_\zeta}$ and Q_Θ. In practice, Cheesefill uses a neural family T_ψ, which amortizes this coupling search: each pseudo complete trajectory is mapped directly to a distribution over refined trajectories. This restriction yields the upper bound in Eq. (5). The bound becomes tight when the neural family is expressive enough to realize an optimal admissible coupling, and it is approximately tight when it contains an ε-optimal admissible coupling. Therefore, the stochastic correction map provides a principled way to move beyond fixed imputation while still optimizing a relaxation of the original OT objective. Cheesefill optimizes the following equation:

$$\min_{\Theta,\mathcal{G},\psi} \mathbb{E}_{\tilde{\boldsymbol{X}}\sim P_{\tilde{\boldsymbol{X}}_\zeta}} \mathbb{E}_{\boldsymbol{Y}\sim T_\psi(\cdot|\tilde{\boldsymbol{X}})}\big[c_\mathcal{G}(\tilde{\boldsymbol{X}}, \boldsymbol{Y})\big], \quad \text{subject to } \mathbf{G}_0 \in \mathrm{DAGs} \quad \text{and} \quad \psi \in \Psi_{\mathrm{adm}}. \tag{6}$$

4 Causal heteroscedastic structure learning from incomplete temporal data

We now instantiate the correction-map formulation with learnable components. Cheesefill uses Bayesian structure learning to average over plausible temporal graphs, a Temporal Graph Neural Network to parametrize the conditional mean mechanisms, and conditional normalizing flows to model heteroscedastic residual densities $\rho^i_\Theta(x\,;\,\mathbf{p}^{<t}, \mathbf{p}^t, \mathcal{G})$. We then describe how the instantaneous acyclicity constraint and the correction map admissibility condition are enforced in the final optimization objective.

Bayesian structure learning. Optimizing Eq. (6) directly yields a single graph estimate. This is brittle under missingness and heteroscedastic noise: the missing entries create uncertainty about which causal relations are supported, while the noise level varies across time and samples. Hence, optimizing Eq. (6) over a deterministic single graph can therefore concentrate too early on edges that explain imputation artifacts or local fluctuations, instead of stable causal mechanisms. To reduce this sensitivity, Cheesefill follows the Bayesian structure learning strategy of [13, 14, 32]. Instead of committing to one graph during training, we maintain a variational posterior $q_\phi(\mathcal{G})$ over temporal graphs and optimize the mechanism cost averaged over graphs sampled from this posterior. Edges supported by consistent mechanism-level evidence tend to receive stable posterior mass, whereas unstable edges are averaged over competing graph samples. Since the exact graph posterior is intractable, $q_\phi(\mathcal{G})$ is parameterized as a factorized distribution over edge variables and estimated using a single Monte Carlo sample with the Gumbel Softmax reparameterization [20]. Additional details are given in Appendix E.

$$\min_{\Theta,\psi,\phi} \mathbb{E}_{\tilde{\boldsymbol{X}}\sim P_{\tilde{\boldsymbol{X}}_\zeta}} \mathbb{E}_{\boldsymbol{Y}\sim T_\psi(\cdot|\tilde{\boldsymbol{X}})}\big[\bar{c}_\phi(\tilde{\boldsymbol{X}}, \boldsymbol{Y})\big] - \beta\mathcal{H}(q_\phi), \text{subject to } \mathbf{G}_0 \in \mathrm{DAGs} \text{ and } \psi \in \Psi_{\mathrm{adm}}, \tag{7}$$

where $\bar{c}_\phi(\tilde{\boldsymbol{X}}, \boldsymbol{Y}) := \mathbb{E}_{\mathcal{G}\sim q_\phi(\mathcal{G})}\big[c_\mathcal{G}(\tilde{\boldsymbol{X}}, \boldsymbol{Y})\big]$, $\mathcal{H}(q_\phi) = -\mathbb{E}_{q_\phi}[\log q_\phi]$,and $\beta > 0$ controls the entropy regularization strength. The entropy term prevents the variational posterior from collapsing too early onto a single low cost graph, which would remove the benefit of graph averaging during the initial stages of training. As evidence accumulates, q_ϕ can concentrate on graphs whose mechanisms consistently explain the corrected trajectories. The averaged cost $\bar{c}_\phi$ evaluates how well a pair $(\tilde{\boldsymbol{X}}, \boldsymbol{Y})$ is explained across graphs considered plausible by the posterior, rather than under a single graph estimate. Finally, the equivalence and parametric upper bound results of Proposition 3.1 remain valid when $c_\mathcal{G}$ is replaced by $\bar{c}_\phi$, the formal statement is given in Appendix I.

Mechanism-level density parametrisation. Optimizing Eq. (7) requires evaluating the local mechanism density ρ^i_Θ for arbitrary pairs $(\tilde{\boldsymbol{X}}, \boldsymbol{Y})$, where $\tilde{\boldsymbol{X}}$ provides the child values being scored

and $\boldsymbol{Y}$ provides the candidate parent values. For a graph $\mathcal{G}$, let $\mathbf{p}_i^{<t}(\boldsymbol{Y}, \mathcal{G})$ and $\mathbf{p}_i^{t}(\boldsymbol{Y}, \mathcal{G})$ denote, respectively, the lagged and instantaneous parents of x_t^i read from $\boldsymbol{Y}$. We write the mechanism density through a residual model: $r_t^i = \tilde{x}_t^i - f_\Theta^i\big(\mathbf{p}_i^{<t}(\boldsymbol{Y}, \mathcal{G}), \mathbf{p}_i^{t}(\boldsymbol{Y}, \mathcal{G})\big)$, so that $\rho_\Theta^i\big(\tilde{x}_t^i; \mathbf{p}_i^{<t}, \mathbf{p}_i^t, \mathcal{G}\big) = p_\eta^i\big(r_t^i \mid \mathbf{p}_i^{<t}, \mathbf{p}_i^t, \mathcal{G}\big)$, where p_η^i is the density function of $g_\Theta^i(\mathbf{p}^{<t}, \mathbf{p}^t) \cdot \epsilon^i$. Hence the graph-averaged mechanism cost becomes

$$\bar{c}_\phi(\tilde{\boldsymbol{X}}, \boldsymbol{Y}) = \mathbb{E}_{\mathcal{G}\sim q_\phi(\mathcal{G})}\left[-\sum_{t=1}^{|\mathcal{T}|}\sum_{i=1}^{d} \log p_\eta^i\left(\tilde{x}_t^i - f_\Theta^i(\mathbf{p}_i^{<t}, \mathbf{p}_i^t) \mid \mathbf{p}_i^{<t}, \mathbf{p}_i^t, \mathcal{G}\right)\right]. \tag{8}$$

We parameterize the mean mechanism f_Θ^i with a temporal graph neural network inspired by [13, 32]. The architecture must respect the candidate graph: if $y_{t-\tau}^j$ is not a parent of x_t^i, then f_Θ^i should be insensitive to $y_{t-\tau}^j$. We enforce this by masking messages with the temporal adjacency matrices:

$$f_\Theta^i\left(\mathbf{p}^{<t}, \mathbf{p}^t\right) = \boxed{\chi}\left(\boxed{\sum_{\tau=0}^{L}\sum_{j=1}^{d} G_{\tau,ji}}\ \boxed{\vartheta}\left(y_{t-\tau}^j, \boxed{e_{\tau,j}}\right), \boxed{e_{0,i}}\right), \tag{9}$$

where χ and ϑ are neural network blocks illustrated in Figure 2 with all the other colored blocks. The trainable embeddings $e_{\tau,j}$ encode both the variable identity and the lag, allowing parameters to be shared across nodes and lags instead of using a separate neural network for each mechanism. Figure 2 illustrates this temporal graph neural architecture. Because the adjacency coefficient $G_{\tau,ji}$ masks the corresponding message, absent edges do not contribute to f_Θ^i, yielding $\partial f_\Theta^i / \partial y_{t-\tau}^j = 0$ whenever $G_{\tau,ji} = 0$.

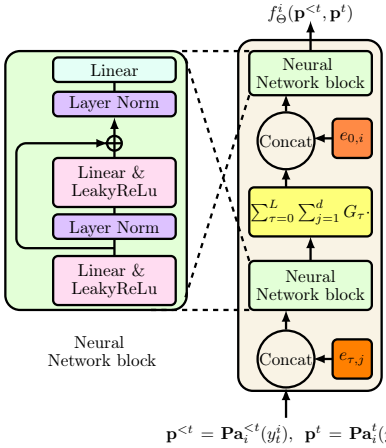


Figure 2: Temporal graph neural network (TGNN) used by Cheesefill.

To model heteroscedastic noise, we parameterize the conditional residual density p_η^i with a conditional normalizing flow. Specifically, for each variable i, we introduce an invertible map $\kappa_\Theta^i\left(\cdot; \mathbf{p}_i^{<t}, \mathbf{p}_i^t, \mathcal{G}\right) : \mathbb{R} \to \mathbb{R}$ such that

$$n_t^i = \kappa_\Theta^i\left(r_t^i; \mathbf{p}_i^{<t}, \mathbf{p}_i^t, \mathcal{G}\right), \qquad n_t^i \sim \mathcal{N}(0,1).$$

We implement κ_Θ^i as a conditional spline flow [10], whose spline parameters are predicted from the same graph masked parent representation as in Eq. (9). By the change of variables formula,

$$p_\eta^i\left(r_t^i \mid \mathbf{p}_i^{<t}, \mathbf{p}_i^t, \mathcal{G}\right) = p_{n^i}(n_t^i)\left|\frac{\partial \kappa_\Theta^i\left(r_t^i; \mathbf{p}_i^{<t}, \mathbf{p}_i^t, \mathcal{G}\right)}{\partial r_t^i}\right|, \tag{10}$$

where p_{n^i} is the standard normal density. This conditional flow lets the shape and scale of the residual distribution vary with the instantaneous and lagged parents, thereby capturing heteroscedastic noise. In the homoscedastic case, we remove the parent conditioning from the flow and learn an unconditional base residual distribution through an affine spline transformation of a standard Gaussian.

Correction map parametrization. We model the stochastic correction map T_ψ as a diagonal Gaussian kernel over the missing entries only. Given a pseudo-complete trajectory $\tilde{\boldsymbol{X}}$, the network predicts a mean $\mu_\psi(\tilde{\boldsymbol{X}})$, and we sample $\boldsymbol{Y}_{\text{mis}} \sim \mathcal{N}\big(\mu_\psi(\tilde{\boldsymbol{X}}), I\big)$. The refined trajectory is then obtained by replacing only the missing entries $\boldsymbol{Y} = \boldsymbol{M} \odot \boldsymbol{Y}_{\text{mis}} + (1 - \boldsymbol{M}) \odot \tilde{\boldsymbol{X}}$. Thus, T_ψ preserves the observed values exactly and only corrects the parts of the trajectory introduced by the initializer. The mean network μ_ψ uses the same neural blocks as in Figure 2. We fix the covariance to I because the conditional normalizing flow already models the parent dependent residual scale and shape; letting T_ψ also learn an arbitrary variance would allow the correction map to absorb heteroscedasticity that should be explained by the SEM mechanism density.

Minimizing Eq. (7) requires handling two constraints: the acyclicity of the instantaneous graph and the admissibility of the correction map. For acyclicity, we use $h(\boldsymbol{G}_0) = \operatorname{tr}\big(e^{\boldsymbol{G}_0 \odot \boldsymbol{G}_0}\big) - d$, where $\odot$ denotes the Hadamard product, and enforce $h(\boldsymbol{G}_0) = 0$ with an augmented Lagrangian penalty [46, 29, 33, 32, 13, 14]. For the admissibility condition $P_{\tilde{\boldsymbol{X}}_\zeta} T_\psi = Q_\Theta$ from Definition 2.1, we use a relaxed distributional matching penalty. Specifically, we add Maximum Mean Discrepancy regularization $\lambda_{\text{map}} \text{MMD}\big(\boldsymbol{Y}^{(\ell)}, \boldsymbol{Y}^{(j)}_{\text{SEM}}\big)$, where $\boldsymbol{Y}^{(\ell)}$ are corrected trajectories sampled from $T_\psi(\cdot \mid \tilde{\boldsymbol{X}})$, and $\boldsymbol{Y}^{(j)}_{\text{SEM}}$ are trajectories generated by the current SEM through $y^{i,(j)}_{t,\text{SEM}} = f^i_\Theta(\mathbf{p}_i^{<t}, \mathbf{p}_i^t) + g^i_\Theta(\mathbf{p}_i^{<t}, \mathbf{p}_i^t)\epsilon_t^{i,(j)}$, with $\mathbf{p}_i^{<t}$ and $\mathbf{p}_i^t$ denote, respectively, the lagged and instantaneous parents of $y^{i,(j)}_{t,\text{SEM}}$ read from $\boldsymbol{Y}^{(j)}$. Here MMD denotes the empirical multi-scale RBF MMD and the final training objective is

$$\min_{\Theta,\psi,\phi,\zeta} \quad \mathbb{E}_{\tilde{\boldsymbol{X}} \sim P_{\tilde{\boldsymbol{X}}_\zeta}} \mathbb{E}_{\boldsymbol{Y} \sim T_\psi(\cdot \mid \tilde{\boldsymbol{X}})} \Big[\bar{c}_\phi(\tilde{\boldsymbol{X}}, \boldsymbol{Y})\Big] + \mathbb{E}_{\mathcal{G} \sim q_\phi(\mathcal{G})} \left[\lambda_s \|\boldsymbol{G}_{0:L}\|_F^2 + \varrho h^2(\boldsymbol{G}_0) + \alpha h(\boldsymbol{G}_0)\right]$$
$$+ \lambda_{\text{map}} \sum_{j,\ell=1}^{N} \text{MMD}\big(\boldsymbol{Y}^{(\ell)}, \boldsymbol{Y}^{(j)}_{\text{SEM}}\big) - \beta \mathcal{H}(q_\phi).$$

5 Experiments

5.1 Synthetic data

Data generation. We conduct extensive experiments to evaluate Cheesfill's performance on synthetic datasets. For ground truth graph generation, we use the Erdos–Rényi model (degree 4) for instantaneous links and the Erdos–Rényi model (degree 2) for time-lagged relationships. For data generation process, f^i, g^i are chosen to be randomly initialized MLPs with one hidden layer and activation function $\{\texttt{Tanh}\}$. We evaluate the different models on MTS with missing values under heteroscedastic noise, details in Appendix C.1. We consider a complex settings with multiple lags $L = 2$ or 3. For each experiments, we generate 50 MTS with length 200. We test different numbers of nodes $\{10, 20, 40\}$ and different missing mechanisms (MCAR, MAR, MNAR) with varying noise level $\{0.1, 0.2, 0.3, 0.4, 0.5\}$. Each experiments is repeated three times, resulting in over 108 distinct datasets. Our code is available at Cheesefill.

Ablation studies. We investigate the contributions of the stochastic correction map and the CNF through an ablation, enabling or disabling each component. We evaluate graph recovery on a 20-node system with three time lags, MCAR missingness (0.5 noise level), and heteroscedastic noise, reporting instantaneous and lagged results separately.

Table 1: Ablation of the stochastic correction map and CNF. Inst. and Lag denote instantaneous and lagged edges, respectively. F1, precision, and recall are expressed as percentages. The final row corresponds to full Cheesefill.

		F1 (%) ↑		Ratio ↓		Precision (%) ↑		Recall (%) ↑	
Map	CNF	Inst.	Lag	Inst.	Lag	Inst.	Lag	Inst.	Lag
Off	Off	0.0	47.0	1.00	0.52	0.0	32.0	0.0	89.0
Off	On	0.0	0.0	1.00	1.00	0.0	0.0	0.0	0.0
On	Off	8.0	45.0	0.91	0.54	5.0	32.0	17.0	77.0
On	**On**	**94.2**	**94.5**	**0.05**	**0.05**	**89.0**	**91.0**	**100.0**	**98.0**

Table 1 shows that the two components are complementary in this setting. Removing both components yields zero instantaneous F1 and a lagged F1 of 47%. Introducing the stochastic correction map alone provides only limited instantaneous recovery (F1 = 8%), while lagged F1 remains low at 45%. Conversely, enabling the CNF without the correction map reduces F1 to zero for both edge types. Combining both components produces a substantial improvement: full Cheesefill achieves instantaneous and lagged F1 scores of 94.2% and 94.5%, respectively, with precision of 89%/91% and recall of 100%/98%. These results support the joint formulation: neither component in isolation

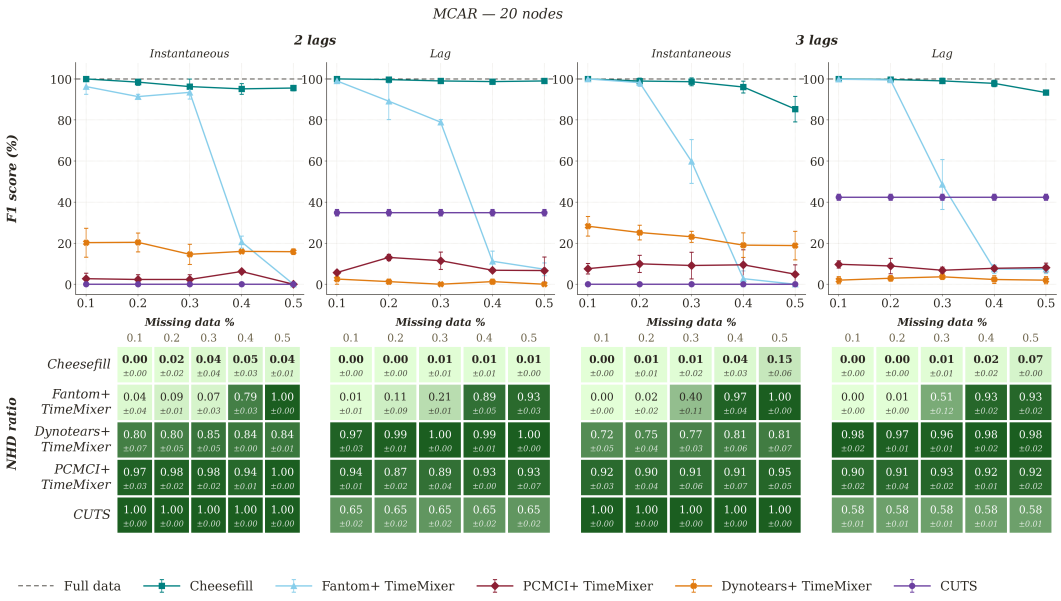


Figure 3: MCAR MTS with 20 nodes. F1 score (top, higher is better) and NHD ratio (bottom, lower is better) for instantaneous and time lagged edges, evaluated under 2 and 3 maximum lags as the missing rate increases from 10% to 50%.

reproduces the recovery performance obtained when the stochastic correction map and CNF are used together.

Benchmarks. We benchmark our model against several baselines, including causal discovery methods designed for incomplete MTS, such as CUTS [8]. Cheesefill is also compared with models that assume complete MTS, including Fantom [32], PCMCI+ [36], Dynotears [29]. *Given that these models cannot deal with incomplete MTS, we use TimeMixer imputer [44] to fill the Missing values and then training the aforementioned models on pseudo complete MTS.*

Results and discussion. Figure 3 reports results under MCAR missingness with 20 nodes, for both 2 lag and 3 lag temporal graphs. Cheesefill stays close to the full-data reference across missing rates for both instantaneous and lagged links, and keeps the NHD ratio near zero in almost all settings. In contrast, the *impute then discover* baselines degrade as the missing rate increases, showing that accurate completion alone does not necessarily preserve the causal dependencies needed for graph recovery. This effect is particularly clear for Fantom+TimeMixer: Fantom is the closest baseline to our setting because it can model instantaneous links and heteroscedastic noise, yet its performance collapses at high missingness after being fed imputed trajectories. DYNOTEARS+TimeMixer and PCMCI+TimeMixer remain far below Cheesefill throughout. CUTS does not model instantaneous links and assumes homoscedastic noise; this explains its weak performance in the instantaneous links. Overall, these results support the need for a joint correction and discovery objective rather than a pipeline that treats imputed values as fixed causal evidence.
Appendices F provides additional results under heteroscedastic noise, MAR and MNAR missing mechanisms, showing that Cheesefill remains robust across different data generating regimes and scales to graphs with 40 nodes.

5.2 Real world datasets

NetSim. We evaluate Cheesefill on NetSim, a standard benchmark for temporal causal discovery [14, 8]. Each MTS corresponds to the simulated BOLD signal of one subject across $d = 15$ brain regions. We use the `Sim-3.mat` setting (50 MTS, $T = 200$ timestamps, shared summary causal graph) and recover the effective connectivity from incomplete observations under varying missingness levels.

Figure 4 shows that Cheesefill is the most robust method as missingness increases. At low missing rates several baselines remain competitive, but their performance degrades sharply beyond 30%: FANTOM+TimeMixer and DYNOTEARS+TimeMixer collapse to near zero at high missing rates, while PCMCI+TimeMixer stays consistently low throughout. CUTS, designed for incomplete MTS, is more stable than the impute then discover baselines but trails Cheesefill at every level an expected gap, since CUTS models only time-lagged dependencies and assumes homoscedastic noise, an assumption known to be violated in fMRI data [39]. In contrast, Cheesefill stays close to the full-data reference up to 30% missingness and retains the highest F1 score even at 50%, indicating that its stochastic correction map preserves causal information that impute-then-discover pipelines lose under severe missingness.

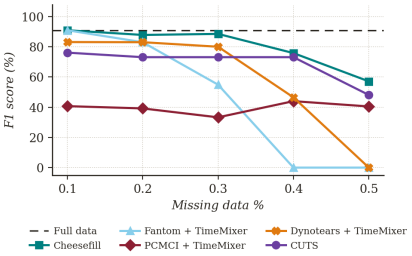


Figure 4: F1 score on NetSim across missingness levels.

Wind Tunnel. We use the wind tunnel datasets from Gamella et al. [12], featuring two controllable fans pushing air through a chamber, barometers measuring air pressure at various locations, and a hatch controlling an external opening. The dataset comprises 16 variables of 10,000 samples each (see Appendix D.1). We compare Cheesefill to the aforementioned baselines, with results in Table 2. Under 50% missingness, Cheesefill outperforms all baselines by achieving 38.1% F1 score. This benchmark is challenging because the causal graph comes from a real physical system with coupled dynamics, indirect effects, and measurement noise, making spurious associations difficult to filter out. The gain of Cheesefill over both CUTS and *impute then discover* baselines shows that its mechanism compatible correction is beneficial in a high missingness regime.

Table 2: Performance on Wind Tunnel data evaluated on summary causal graph.

Method	F1↑	Ratio↓
PCMCI+ TimeMixer	21.8	0.87
Dynotears + TimeMixer	0	1
Fantom + TimeMixer	30.1	0.69
CUTS	19.5	0.89
Cheesefill	**38.1**	**0,61**

6 Conclusion

We introduced Cheesefill, a causal structure learning framework for incomplete multivariate time series that simultaneously handles *(i)* instantaneous and time-lagged links, *(ii)* missing observations, and *(iii)* heteroscedastic noise. On the theoretical side, we showed that optimising over stochastic correction maps is equivalent to a conditional kernel reformulation of the OT problem. Extensive experiments on synthetic and real world data confirm consistent gains over strong baselines across missingness mechanisms, lags, and noise regimes.]

Limitations. Cheesefill relies on the standard assumptions of SEM based temporal causal discovery, including stationarity and causal sufficiency. These assumptions may be violated in some real world systems with regime changes, hidden confounders, or unobserved external drivers. In addition, while MCAR and MAR mechanisms are theoretically well supported under standard conditions, arbitrary MNAR missingness is not identifiable without additional assumptions on the observation process. Cheesefill is therefore best viewed as a principled and empirically robust correction framework for MNAR settings, rather than a method that guarantees recovery under all possible non-ignorable missingness mechanisms. Finally, the practical objective relaxes the exact admissibility condition of the correction map using an MMD penalty and enforces instantaneous acyclicity through an augmented Lagrangian; both relaxations can lead to suboptimal solutions when optimization is difficult or when hyperparameters are poorly chosen. Despite these limitations, the empirical results show that coupling correction with mechanism level causal learning substantially improves robustness over impute then discover pipelines across missingness levels.

Acknowledgment

We thank Ortal Senouf for helpful feedbacks and comments.

References

[1] Charles K Assaad, Emilie Devijver, and Eric Gaussier. Survey and evaluation of causal discovery methods for time series. *Journal of Artificial Intelligence Research*, 2022.

[2] Karim Assaad, Emilie Devijver, Eric Gaussier, and Ali Ait-Bachir. A mixed noise and constraint-based approach to causal inference in time series. In *ECML PKDD*, 2021.

[3] Rohit Bhattacharya, Razieh Nabi, Ilya Shpitser, and James M. Robins. Identification in missing data models represented by directed acyclic graphs. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*, volume 115 of *Proceedings of Machine Learning Research*, pages 1149–1158. PMLR, 2020.

[4] Bart Bussmann, Jannes Nys, and Steven Latré. Neural additive vector autoregression models for causal discovery in time series. In *Discovery Science*, 2021.

[5] Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. Brits: Bidirectional recurrent imputation for time series. *Advances in neural information processing systems*, 31, 2018.

[6] Yuxiao Cheng, Lianglong Li, Tingxiong Xiao, Zongren Li, Jinli Suo, Kunlun He, and Qionghai Dai. Cuts+: High-dimensional causal discovery from irregular time-series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11525–11533, 2024.

[7] Yuxiao Cheng, Ziqian Wang, Tingxiong Xiao, Qin Zhong, Jinli Suo, and Kunlun He. Causaltime: Realistically generated time-series for benchmarking of causal discovery. *arXiv preprint arXiv:2310.01753*, 2023.

[8] Yuxiao Cheng, Runzhao Yang, Tingxiong Xiao, Zongren Li, Jinli Suo, Kunlun He, and Qionghai Dai. Cuts: Neural causal discovery from irregular time-series data. *arXiv preprint arXiv:2302.07458*, 2023.

[9] Andrea Cini, Ivan Marisca, and Cesare Alippi. Filling the g_ap_s: Multivariate time series imputation by graph neural networks. *arXiv preprint arXiv:2108.00298*, 2021.

[10] Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. *Advances in neural information processing systems*, 2019.

[11] Doris Entner and Patrik O Hoyer. On causal discovery from time series data using fci. *Probabilistic graphical models*, 2010.

[12] Juan L Gamella, Jonas Peters, and Peter Bühlmann. Causal chambers as a real-world physical testbed for ai methodology. *Nature Machine Intelligence*, 7(1):107–118, 2025.

[13] Tomas Geffner, Javier Antoran, Adam Foster, Wenbo Gong, Chao Ma, Emre Kiciman, Amit Sharma, Angus Lamb, Martin Kukla, Nick Pawlowski, et al. Deep end-to-end causal inference. *arXiv preprint arXiv:2202.02195*, 2022.

[14] Wenbo Gong, Joel Jennings, Cheng Zhang, and Nick Pawlowski. Rhino: Deep causal temporal relationship learning with history-dependent noise. *Preprint arXiv:2210.14706*, 2022.

[15] Wiebke Günther, Urmi Ninad, Jonas Wahl, and Jakob Runge. Conditional independence testing with heteroskedastic data and applications to causal discovery. *Advances in Neural Information Processing Systems*, 2022.

[16] James D Hamilton and Raul Susmel. Autoregressive conditional heteroskedasticity and changes in regime. *Journal of econometrics*, 1994.

[17] Stefan Haufe, Klaus-Robert Müller, Guido Nolte, and Nicole Krämer. Sparse causal discovery in multivariate time series. In *causality: objectives and assessment*. PMLR, 2010.

[18] Hilde M Huizenga and Peter CM Molenaar. Equivalent source estimation of scalp potential fields contaminated by heteroscedastic and correlated noise. *Brain topography*, 1995.

[19] Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 2010.

[20] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[21] LK Kantorovich. On a problem of monge. *Journal of Mathematical Sciences*, 133(4), 2006.

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

[23] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2019.

[24] Sindy Löwe, David Madras, Richard Zemel, and Max Welling. Amortized causal discovery: Learning to infer causal graphs from time-series data. In *Conference on Causal Learning and Reasoning*, 2022.

[25] Chao Ma and Cheng Zhang. Identifiable generative models for missing not at random data imputation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27645–27658, 2021.

[26] Boris Muzellec, Julie Josse, Claire Boyer, and Marco Cuturi. Missing data imputation using optimal transport. In *International Conference on Machine Learning*, pages 7130–7140. PMLR, 2020.

[27] Razieh Nabi, Rohit Bhattacharya, and Ilya Shpitser. Full law identification in graphical models of missing data: Completeness results. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7153–7163. PMLR, 2020.

[28] Mark Newman. *Networks*. Oxford university press, 2018.

[29] Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. Dynotears: Structure learning from time-series data. In *International conference on artificial intelligence and statistics*, pages 1595–1605. Pmlr, 2020.

[30] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on time series using restricted structural equation models. *Advances in neural information processing systems*, 2013.

[31] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[32] Abdellah Rahmani and Pascal Frossard. Flow based approach for dynamic temporal causal models with non-gaussian or heteroscedastic noises. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

[33] Abdellah Rahmani and Pascal Frossard. Causal temporal regime structure learning. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025.

[34] Abdellah Rahmani, Arun Venkitaraman, and Pascal Frossard. A meta-gnn approach to personalized seizure detection and classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[35] Jakob Runge. Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 2018.

[36] Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Conference on Uncertainty in Artificial Intelligence*. PMLR, 2020.

[37] Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances*, 2019.

[38] Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 2005.

[39] Christof Seiler and Susan Holmes. Multivariate heteroscedasticity models for functional brain connectivity. *Frontiers in neuroscience*, 2017.

[40] Xinpeng Shen, Sisi Ma, Prashanthi Vemuri, and Gyorgy Simon. Challenges and opportunities with causal discovery algorithms: application to alzheimer's pathophysiology. *Scientific reports*, 2020.

[41] Le Song, Mladen Kolar, and Eric Xing. Time-varying dynamic bayesian networks. *In Advances in Neural Information Processing Systems*, 2009.

[42] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

[43] Vy Vo, He Zhao, Trung Le, Edwin V Bonilla, and Dinh Phung. Optimal transport for structure learning under missing data. *arXiv preprint arXiv:2402.15255*, 2024.

[44] Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y Zhang, and Jun Zhou. Timemixer: Decomposable multiscale mixing for time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2024.

[45] Ian R White, Patrick Royston, and Angela M Wood. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*, 30(4):377–399, 2011.

[46] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

Table of content

A Detailed related work **15**

B Limitations **15**

C Data generation and baselines **16**

C.1 Data generation 16

C.2 Baselines 16

C.3 Optimization parameters 17

D Real world data **17**

D.1 Causal Chambers data 17

D.2 NetSim 18

E Variational distribution Details **18**

F Additional Experiments **19**

F.1 Additional results on synthetic data 19

F.1.1 MCAR results 19

F.1.2 MAR results 20

G MNAR Experiments **22**

H Additional experiments on non linear SEM with gaussian noise **22**

H.1 MCAR and MAR results 23

H.2 MNAR results 26

I Graph average cost details **28**

J Proof of our proposition **30**

J.1 Assumptions 30

J.2 Proof 31

A Detailed related work

Causal structure learning from MTS. Causal discovery from MTS has evolved beyond the IID setting to account for temporal dependencies; Assaad et al. [1] provide a comprehensive survey. Most methods build on Dynamic Bayesian Networks, which represent discrete-time dynamics as directed graphical models with both lagged and contemporaneous edges. Three families dominate the field.

Constraint-based methods extend conditional-independence testing to the temporal setting. Early work focused on lagged relationships only [17, 41], with tsFCI [11] adapting Fast Causal Inference [42] to handle latent confounders. PCMCI [37] introduced a scalable two-stage algorithm for high-dimensional time series, initially restricted to lagged links and later extended to PCMCI+ [36] for contemporaneous edges. NBCB [2] combines noise-based and constraint-based ideas to recover summary causal graphs beyond Markov equivalence, even in the presence of instantaneous relationships.

Functional-model approaches exploit identifiability under restricted noise distributions. VAR-LiNGAM [19] integrates non-Gaussian instantaneous models with vector-autoregressive components, while TiMINo [30] studies nonlinear instantaneous effects through constrained structural equation models with independent noise.

Score-based methods cast structure learning as continuous optimisation. DYNOTEARS [29] leverages the algebraic acyclicity characterisation in [46] to recover both instantaneous and lagged dependencies via an augmented Lagrangian, without assumptions on the underlying topology. Rhino [14] extends this line to stationary MTS with history-dependent noise, modeling the noise variance as a function of lagged parents. Most recently, FANTOM [32] employs conditional normalising flows to handle heteroscedastic noise in both stationary and non-stationary regimes. Despite this variety, all of these methods assume fully complete data. None of these methods addresses heteroscedastic noise jointly with instantaneous links and missing observations, both of which Cheesefill targets.

Causal discovery from incomplete MTS. Real-world MTS routinely contain missing values, and the dominant strategy in practice is the two-stage *impute-then-discover* pipeline: complete the data with a generic imputer [9, 5, 44] and run a standard causal discovery method on top. This pipeline is known to be unreliable: the imputed distribution generally encodes a different set of conditional independencies than the true distribution, and this mismatch propagates into the recovered graph [8, 6]. A small but growing body of work targets joint imputation and structure learning under missingness. CUTS [8] alternates between a delayed-supervision imputation network and a sparse Granger graph fitting step, and CUTS+ [6] scales this idea to high-dimensional MTS through a message-passing GNN and a coarse-to-fine discovery scheme. However, both methods restrict themselves to purely time-lagged relationships and homoscedastic noise: neither can recover instantaneous links, and neither accounts for the noise heteroscedasticity observed in fMRI, EEG, and financial data.

Positioning of Cheesefill. The literature thus splits along two axes. Methods on the *complete-data* side handle instantaneous links, heteroscedastic noise, or both, but require fully complete MTS [14, 32]. Methods on the *missing data* side handle missingness, but only for time lagged, homoscedastic settings [8, 6]. Cheesefill closes both gaps simultaneously: it recovers instantaneous and time lagged links from incomplete MTS under homoscedastic and heteroscedastic noises.

B Limitations

Cheesefill relies on the standard assumptions of SEM based temporal causal discovery, including stationarity and causal sufficiency. These assumptions may be violated in some real world systems with regime changes, hidden confounders, or unobserved external drivers. In addition, while MCAR and MAR mechanisms are theoretically well supported under standard conditions, arbitrary MNAR missingness is not identifiable without additional assumptions on the observation process. Cheesefill is therefore best viewed as a principled and empirically robust correction framework for MNAR settings, rather than a method that guarantees recovery under all possible non-ignorable missingness mechanisms. Finally, the practical objective relaxes the exact admissibility condition of the correction map using an MMD penalty and enforces instantaneous acyclicity through an augmented Lagrangian; both relaxations can lead to suboptimal solutions when optimization is difficult or when hyperparameters are poorly chosen. Despite these limitations, the empirical results show that coupling correction

with mechanism level causal learning substantially improves robustness over impute then discover pipelines across missingness levels.

Spurious causality is a practical risk, so we advocate using the learned graphs as decision support rather than as ground truth. Before acting on them, especially in medical or financial settings, practitioners should adopt safeguards: (i) expert vetting of proposed edges and mechanisms; (ii) robustness checks (e.g., stability under resampling or perturbations, alternative specifications); and (iii) interventional validation when feasible.

C Data generation and baselines

C.1 Data generation

We employ the Erdos–Rényi (ER) [28] model with mean degrees of 1 or 2 to generate lagged graphs, and ER model with mean degrees 4 for instantaneous graphs. The maximum number of lags, L, is set to 2 or 3. We experiment with varying numbers of nodes $\{20, 40\}$ and different missingness mechanism MCAR, MNAR and MAR, with different level $\{10\%, 20\%, 30\%, 40\%, 50\%\}$. We generate 50 different MTS of length 200 sharing the same causal graph.

In heteroscedastic settings, noise variance shifts across both variables and observations, making the underlying DAG much harder to recover from data. Given a random set if directed acyclic graphs $\mathcal{G}$, we generate observations from the SEMs in Eq C.1 as follows:

$$x_t^{i,(j)} = f^i\big(\mathbf{Pa}_i^{<t}, \mathbf{Pa}_i^t\big) + \log(1 + \exp(\sigma(g^i\big(\mathbf{Pa}_i^{<t}, \mathbf{Pa}_i^t\big))) \cdot \epsilon_t^{i,(j)},$$

where f^i, g^i are chosen to be randomly initialized MLPs with one hidden layer of size number of nodes and `tanh` activation functions. ϵ_t^i follows either a normal distribution $\mathcal{N}(0, 1)$ or a more complex one obtained by transforming samples from a standard Gaussian with an MLP with random weights and `sin` activation function.

C.2 Baselines

Dynotears [29]. Dynotears formulates causal discovery for multivariate time series through a linear vector autoregressive (VAR) model that simultaneously captures lagged *and* instantaneous causal effects. Its key innovation is the *DAGness* penalty a smooth, continuously differentiable relaxation of the acyclicity constraint optimized via an augmented Lagrangian scheme alongside a mean squared error loss. Dynotears assumes complete MTS with linear causal relationships. Dynotears SEM emerges as the special case of CHEESEFILL SEM obtained by using linear component functions f^i, fixing the noise scaling to $g^i = 1$ and $\epsilon_t^i \sim \mathcal{N}(0, 1)$ in Eq(C.1). Moreover Cheesefill handles missingness. For comparing with this model, we use publicly available package `causalnex`[1].

PCMCI+ [36]. a scalable two-stage algorithm for time series, enabling the identification of contemporaneous causal connection. As Dynotears, PCMCI+ is a special case of CHEESEFILL, obtained by using linear or non linear component functions f^i, fixing the noise scaling to $g^i = 1$ and allowing ϵ_t^i to follow any distribution. For the comparison, we use publicly available package `Tigramite`[2].

CUTS [8]. CUTS is a neural causal discovery method designed for incomplete and irregular multivariate time series. It jointly learns an imputation model and a temporal causal graph, thereby avoiding the standard impute then discover pipeline. However, CUTS focuses on time lagged Granger style dependencies and does not model instantaneous causal effects. It also assumes a homoscedastic noise structure, whereas CHEESEFILL allows the noise scale g^i in Eq. (C.1) to depend on both instantaneous and lagged parents. Thus, CUTS can be viewed as addressing the missingness aspect of our problem, while CHEESEFILL further handles instantaneous links and heteroscedastic mechanisms through the stochastic correction map and conditional normalizing flows. For the comparison, we use publicly available package `CUTS`[3].

Fantom [32]. Fantom is a flow-based temporal causal discovery method for complete MTS with non-Gaussian and heteroscedastic noise. It models nonlinear temporal SEMs using conditional

[1] https://causalnex.readthedocs.io/en/latest/
[2] https://jakobrunge.github.io/tigramite/
[3] https://github.com/jarrycyx/UNN/tree/main/CUTS

normalizing flows, allowing the residual distribution to vary with the parents and thereby capturing richer noise mechanisms than homoscedastic methods such as Dynotears and PCMCI+. Fantom is closest to CHEESEFILL on the mechanism side: in the absence of missing values, and when the pseudo complete distribution coincides with the observed complete data distribution, CHEESEFILL reduces to a related SEM-based density fitting framework with heteroscedastic mechanisms. The key difference is that Fantom assumes complete observations, whereas CHEESEFILL explicitly learns from incomplete MTS by coupling a parametric imputer with a stochastic correction map, so that imputation and graph discovery are optimized jointly rather than sequentially. For the comparison, we use publicly available code in openreview supplementary materials `Fantom`[4].

C.3 Optimization parameters

The variational posterior $q_{\phi^r}(\mathcal{G}^r)$ is initialized to prefer sparse graphs (edge probability < 0.5). Heteroscedastic noise is modeled with conditional normalizing flows (CNFs). Every neural block is a two-layer MLP with 32 hidden units, residual connections, and layer normalization. The neural block used for stochastic map is 3 layers MLP with 16 hidden units, residual connections, and layer normalization. We use a batch size of 1024, 2 max augmented lagrangien step and 3000 inner steps.

Gradients for discrete edges are estimated via the Gumbel–Softmax trick, using a hard forward pass and a soft backward pass with temperature 0.25. All spline flows employ 16 bins, and each transformation uses an embedding dimension equal to the number of nodes.

For graphs with 20 nodes we use $\rho = 0.01$ and $\alpha = 0$, the sparsity penalty is fixed at $\lambda_s = 50$, whereas for 40 nodes we set $\rho = 0.00001$ and we use a sparsity penalty of $\lambda_s = 30$. Models are optimized with Adam [22] at a learning rate of 0.005.

D Real world data

D.1 Causal Chambers data

We use the wind tunnel datasets from Gamella et al. [12], featuring two controllable fans pushing air through a chamber, barometers measuring air pressure at various locations, and a hatch controlling an external opening. The tunnel is a chamber with two controllable fans that push air through it and barometers that measure air pressure at different locations. A hatch precisely controls the area of an additional opening to the outside (see Figure 5).

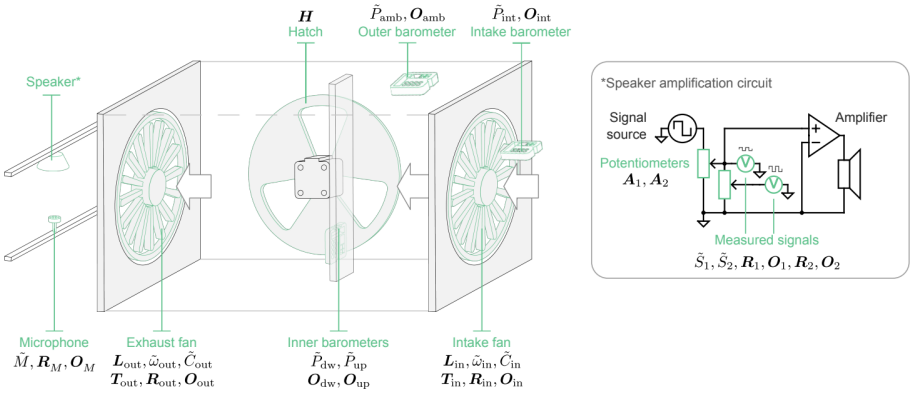


Figure 5: Figure taken from Gamella et al. [12]. Diagrams of the wind tunel causal chamber and its main components, including the amplification circuit that drives the speaker of the wind tunnel. The variables measured by the chamber are displayed in black math print. Sensor measurements are denoted by a tilde. Manipulable variables, that is, actuators and sensor parameters, are shown in bold symbols.

[4] https://openreview.net/forum?id=BgwtOwFSvY

NHD ratio

	Missing data % 0.1	0.2	0.3	0.4	0.5
Cheesefill	**0.09**	**0.12**	**0.11**	**0.24**	**0.42**
Fantom+ TimeMixer	0.09	0.16	0.45	0.94	0.94
PCMCI+ TimeMixer	0.59	0.60	0.66	0.55	0.59
Dynotears+ TimeMixer	0.16	0.16	0.20	0.53	0.94
CUTS	0.23	0.30	0.26	0.26	0.51

Figure 6: NHD ratio on NetSim across missingness levels.

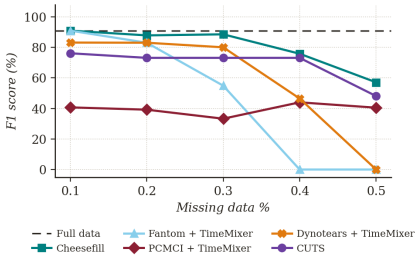


Figure 7: F1 score on NetSim across missingness levels.

The dataset comprises 16 variables: controllable load of the two fans L_{in}, L_{out}, their measurable speed $(\tilde{\omega}_{\text{in}}, \tilde{\omega}_{\text{out}})$, the current draw by the fans $\left(\tilde{C}_{\text{in}}, \tilde{C}_{\text{out}}\right)$ $(\tilde{C}_{\text{in}}, \tilde{C}_{\text{out}})$, the resulting air pressure inside the chamber $(\tilde{P}_{\text{dw}}, \tilde{P}_{\text{up}})$ or at its intake $(\tilde{P}_{\text{int}})$, and the hatch H. In the circuit that drives the speaker, we can manipulate the potentiometers (A_1, A_2) that control the amplification, monitoring the resulting signal at different points of the circuit $\left(\tilde{S}_1, \tilde{S}_2\right)$ and through the microphone output $(\tilde{M})$. $\tilde{\tilde{P}}_{\text{amb}}$ is the ambient pressure measure by the outer barometer. We evaluate all the models on MTS of 10,000 samples.

For the training procedure, we set our lag to 8 time lagged and we allow the presence of instantaneous parents. Regarding the parameters, we use a sparsity coefficient equal to 500 for time lagged links and 100 for instantaneous, spline of 128 bins and MLPs of size 32.

D.2 NetSim

We evaluate Cheesefill on NetSim, a standard benchmark for temporal causal discovery [14, 8]. Each MTS corresponds to the simulated BOLD signal of one subject across $d = 15$ brain regions. We use the `Sim-3.mat` setting (50 MTS, $T = 200$ timestamps, shared summary causal graph) and recover the effective connectivity from incomplete observations under varying missingness levels using MCAR mechanism.
Figure 4 shows that Cheesefill is the most robust method as missingness increases. At low missing rates several baselines remain competitive, but their performance degrades sharply beyond 30%: FANTOM+TimeMixer and DYNOTEARS+TimeMixer collapse to near zero at high missing rates, while PCMCI+TimeMixer stays consistently low throughout. CUTS, designed for incomplete MTS, is more stable than the impute then discover baselines but trails Cheesefill at every level an expected gap, since CUTS models only time-lagged dependencies and assumes homoscedastic noise, an assumption known to be violated in fMRI data [39]. In contrast, Cheesefill stays close to the full-data reference up to 30% missingness and retains the highest F1 score even at 50%, indicating that its stochastic correction map preserves causal information that impute-then-discover pipelines lose under severe missingness. NHD ratio metrics highlight the same trends where Cheesefill outperforms all the baseline and prove that Cheesefill learns accurate summary causal graph.

E Variational distribution Details

We provide the detailed formulations for $q_\phi(\mathcal{G})$. In order to model the temporal adjacency matrices $\boldsymbol{G}_\tau$ where $\tau \in [|1 : L|]$, we use two learnable matrices $\boldsymbol{U}_\tau, \boldsymbol{Q}_\tau \in \mathbf{R}^{d\times d}$ such that:

$$p_{k,ij} = \frac{\exp(u_{k,ij})}{\exp(u_{k,ij}) + \exp(q_{k,ij})} \tag{11}$$

For instantaneous graphs $\boldsymbol{G}_0$, we used the same trick as in [14, 32], in which we employ three lower triangular learnable matrices $\boldsymbol{U}_0, \boldsymbol{Q}_0, \boldsymbol{E}_0 \in \mathbf{R}^{d\times d}$ to characterise three scenarios: (1) $i \to j$; (2)

$j \to i$; (3) no edge between them. For node $i > j$:

$$
\begin{aligned}
p(i \to j) &= \frac{\exp(u_{ij})}{\exp(u_{ij}) + \exp(q_{ij}) + \exp(e_{ij})} \\
p(j \to i) &= \frac{\exp(q_{ij})}{\exp(u_{ij}) + \exp(q_{ij}) + \exp(e_{ij})} \\
p(\text{ no edge }) &= \frac{\exp(e_{ij})}{\exp(u_{ij}) + \exp(q_{ij}) + \exp(e_{ij})}.
\end{aligned} \tag{12}
$$

With this formulation, the instantaneous adjacency matrix is free of self-loops, eliminating any length-1 cycles.

F Additional Experiments

F.1 Additional results on synthetic data

F.1.1 MCAR results

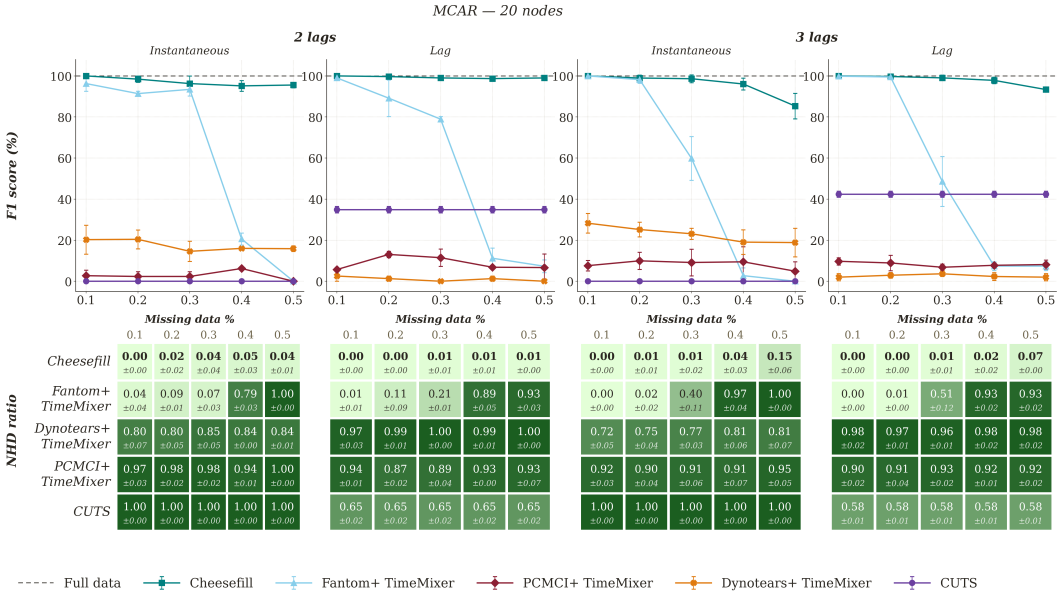


Figure 8: MCAR MTS with 20 nodes. F1 score (top, higher is better) and NHD ratio (bottom, lower is better) for instantaneous and time lagged edges, evaluated under 2 and 3 maximum lags as the missing rate increases from 10% to 50%.

Figure 8 reports results under MCAR missingness with 20 nodes, for both 2 lag and 3 lag temporal graphs. Cheesefill stays close to the full-data reference across missing rates for both instantaneous and lagged links, and keeps the NHD ratio near zero in almost all settings. In contrast, the *impute then discover* baselines degrade as the missing rate increases, showing that accurate completion alone does not necessarily preserve the causal dependencies needed for graph recovery. This effect is particularly clear for Fantom+TimeMixer: Fantom is the closest baseline to our setting because it can model instantaneous links and heteroscedastic noise, yet its performance collapses at high missingness after being fed imputed trajectories. DYNOTEARS+TimeMixer and PCMCI+TimeMixer remain far below Cheesefill throughout. Overall, these results support the need for a joint correction-and-discovery objective rather than a pipeline that treats imputed values as fixed causal evidence.

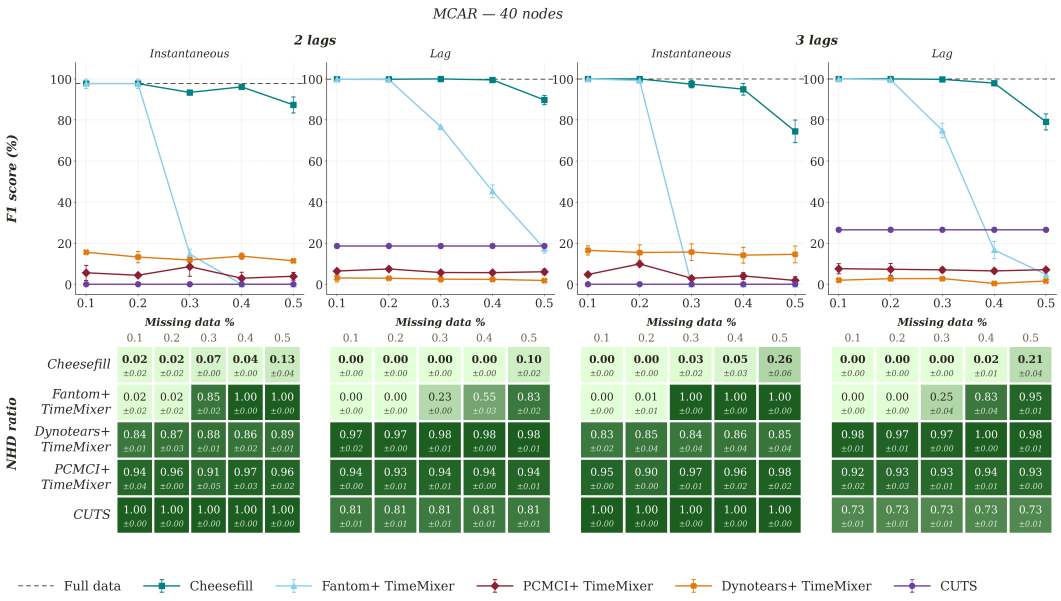


Figure 9: MCAR MTS with 40 nodes. F1 score (top, higher is better) and NHD ratio (bottom, lower is better) for instantaneous and time lagged edges, evaluated under 2 and 3 maximum lags as the missing rate increases from 10% to 50%.

MCAR, 40 nodes. Figure 9 shows the corresponding 40 node MCAR results. Although MCAR is less biased than MAR, the high dimensional setting remains difficult because missing entries reduce the effective information available for recovering both instantaneous and lagged links. Cheesefill again achieves the strongest overall performance, maintaining near perfect F1 scores for most missingness levels and keeping the NHD ratio close to zero, especially for 2 lag graphs and moderate missing rates. Performance decreases only in the hardest 3 lag setting at 50% missingness, where the graph is denser and fewer complete temporal contexts are available; nevertheless, Cheesefill still substantially outperforms all baselines. The impute then discover methods degrade rapidly as missingness increases, with Fantom+TimeMixer collapsing despite its ability to model heteroscedastic noise on complete data. These results support the central claim that causal recovery from incomplete MTS benefits from jointly correcting pseudo complete trajectories and learning SEM compatible mechanisms, rather than treating imputation as a preprocessing step.

F.1.2 MAR results

We set 10% of the variables to be fully observed.

MAR, 20 nodes. Figure 10 reports the MAR setting with 20 nodes. Cheesefill remains consistently close to the full-data reference across both 2 lag and 3 lag graphs, for instantaneous as well as lagged edges. Its NHD ratio stays near zero in almost all settings, showing that the recovered graph remains structurally close to the ground truth even as the missing rate increases. In contrast, Fantom+TimeMixer is competitive only at low missingness, then degrades sharply as the missing rate grows, especially for instantaneous links. This highlights the limitation of applying a strong complete-data method after fixed imputation: the imputer can introduce dependencies that are then interpreted as causal structure. Dynotears+TimeMixer and PCMCI+TimeMixer remain substantially below Cheesefill throughout, confirming that impute-then-discover pipelines are not robust under MAR missingness. CUTS is more stable on lagged edges than the impute then discover baselines, but it does not model instantaneous links and assumes homoscedastic mechanisms; this explains its weak performance in the instantaneous panels and its gap to Cheesefill in this heteroscedastic setting. Overall, these results show that jointly correcting pseudo-complete trajectories and learning SEM-compatible mechanisms is essential for reliable causal recovery under biased missingness.

MAR, 40 nodes. Figure 11 reports the 40 node MAR setting, which is substantially more challenging than MCAR because the missingness pattern depends on observed values and can therefore

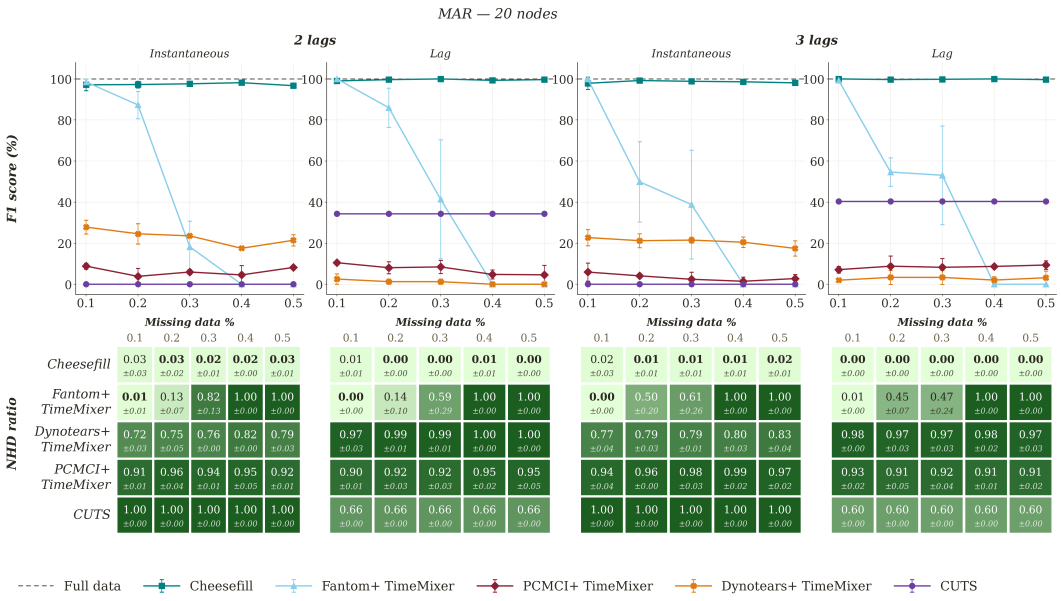


Figure 10: MAR MTS with 20 nodes. F1 score (top, higher is better) and NHD ratio (bottom, lower is better) for instantaneous and time lagged edges, evaluated under 2 and 3 maximum lags as the missing rate increases from 10% to 50%.

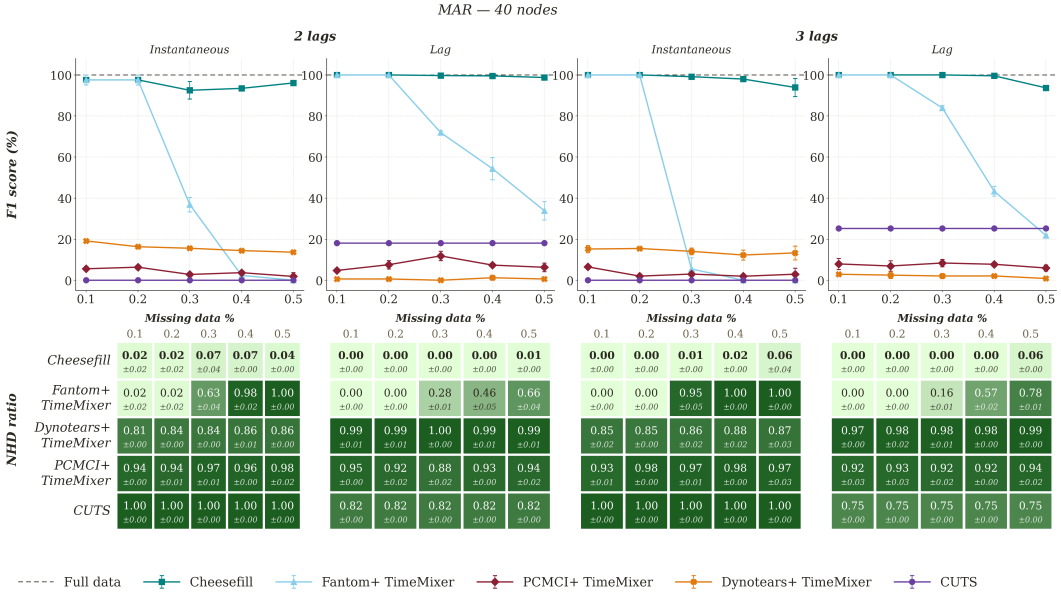


Figure 11: MAR MTS with 40 nodes. F1 score (top, higher is better) and NHD ratio (bottom, lower is better) for instantaneous and time lagged edges, evaluated under 2 and 3 maximum lags as the missing rate increases from 10% to 50%.

introduce systematic bias. Cheesefill remains close to the full-data reference across both 2 lag and 3 lag graphs, for both instantaneous and lagged edges. Its NHD ratio stays near zero in almost all settings, indicating that the recovered graph differs only marginally from the full-data reference even as the missing rate increases. In contrast, Fantom+TimeMixer performs well at low missingness but degrades sharply once the missing rate increases, especially for instantaneous edges, showing that a strong complete-data causal discovery method can still fail when it is applied after fixed imputation. Dynotears+TimeMixer, PCMCI+TimeMixer, and CUTS remain far below Cheesefill throughout,

which confirms that MAR missingness requires more than either impute then discover pipelines or incomplete MTS methods restricted to lagged homoscedastic dependencies.

G MNAR Experiments

MNAR is the most challenging missingness regime because the probability that an entry is missing may depend on the unobserved value itself. In general, this makes the complete data distribution non-identifiable without additional assumptions or knowledge of the missingness mechanism. In our experiments, we therefore do not treat MNAR as a theoretically identifiable setting; instead, we use a small observed subset to model the missingness process and test whether Cheesefill remains empirically robust under non-ignorable corruption. Specifically, we set 10% of the variables to be fully observed. Figure 12 shows that Cheesefill remains consistently close to the full data reference

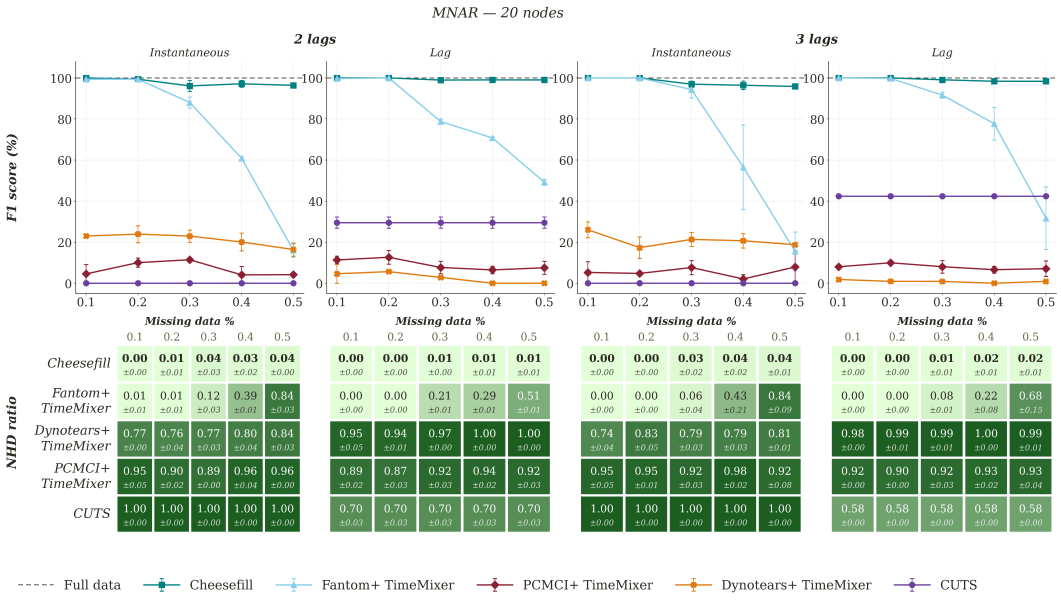


Figure 12: MNAR MTS with 20 nodes. F1 score (top, higher is better) and NHD ratio (bottom, lower is better) for instantaneous and time lagged edges, evaluated under 2 and 3 maximum lags as the missing rate increases from 10% to 50%.

across both 2 lag and 3 lag graphs. Its F1 score stays high for instantaneous and lagged edges even as the missing rate increases, and its NHD ratio remains near zero in almost all settings. This is notable because MNAR corruption can systematically bias the pseudo complete distribution, making fixed imputations especially unreliable. Fantom+TimeMixer is competitive at low missingness but degrades substantially at higher missing rates, particularly for instantaneous links, showing again that complete-data causal discovery methods are sensitive to imputation artifacts. Dynotears+TimeMixer and PCMCI+TimeMixer remain far below Cheesefill, while CUTS is stable only for lagged effects and cannot recover instantaneous links; it also does not model heteroscedastic mechanisms. Overall, these results show that Cheesefill's mechanism-compatible correction map provides strong empirical robustness even in MNAR settings, where exact recovery cannot be expected without additional assumptions on the observation process.

H Additional experiments on non linear SEM with gaussian noise

We sample causal graphs from the Erdős–Rényi (ER) model [28], with mean degree 4 for the instantaneous graph and mean degree 2 for each lagged graph, i.e., each variable has on average two instantaneous parents and two parents at every lag. The instantaneous graph is a DAG, obtained by sampling edges consistent with a random topological order, while lagged edges, including self-lags, are sampled independently. The maximum lag is $L = 3$ and the number of nodes is $d \in \{20, 40\}$.

Homoscedastic case. Here the noise variance is the same for every variable and observation. Given the sampled graph $\mathcal{G}$, we generate observations from the additive nonlinear SEM in Eq. 13:

$$x_t^{i,(j)} = \sum_{k=0}^{L} \sum_{m \in \mathbf{Pa}_i^{t-k}} w_{mi}^{(k)} \tanh\big(x_{t-k}^{m,(j)}\big) + \epsilon_t^{i,(j)}, \qquad \epsilon_t^{i,(j)} \sim \mathcal{N}(0,1), \tag{13}$$

where $\mathbf{Pa}_i^{t-k}$ is the set of parents of x_t^i at lag k, so that $\mathbf{Pa}_i^t$ $(k=0)$ are the instantaneous parents, evaluated in topological order, and $\mathbf{Pa}_i^{<t} = \bigcup_{k=1}^{L} \mathbf{Pa}_i^{t-k}$ are the lagged ones. We added also a new baseline for this setting Rhino [14].

H.1 MCAR and MAR results

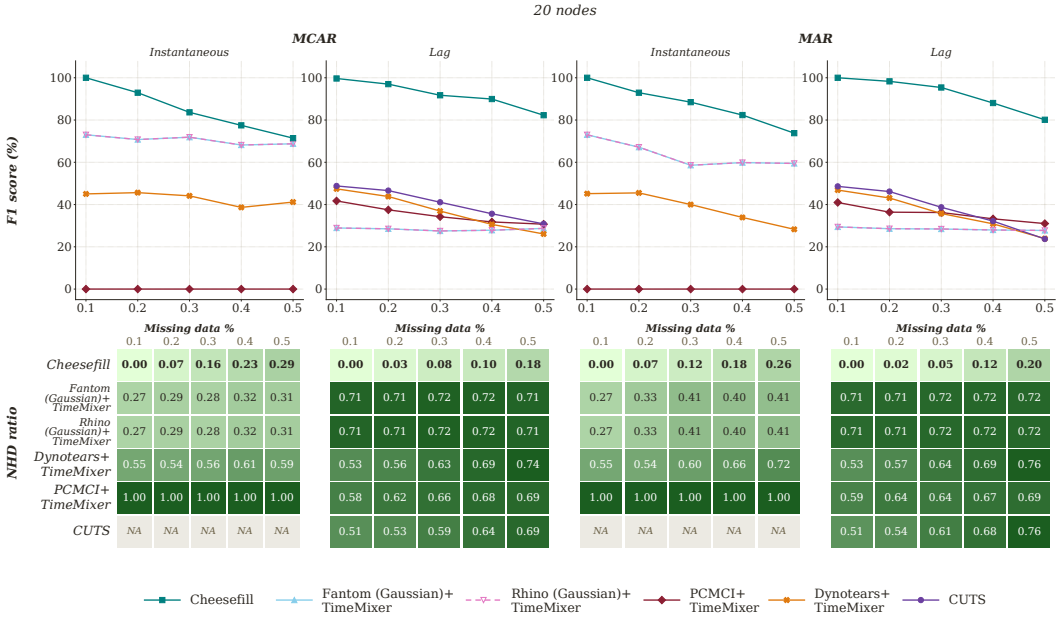


Figure 13: MCAR and MAR MTS with 20 nodes. F1 score (top, higher is better) and NHD ratio (bottom, lower is better) for instantaneous and time lagged edges, evaluated under 3 maximum lags as the missing rate increases from 10% to 50%.

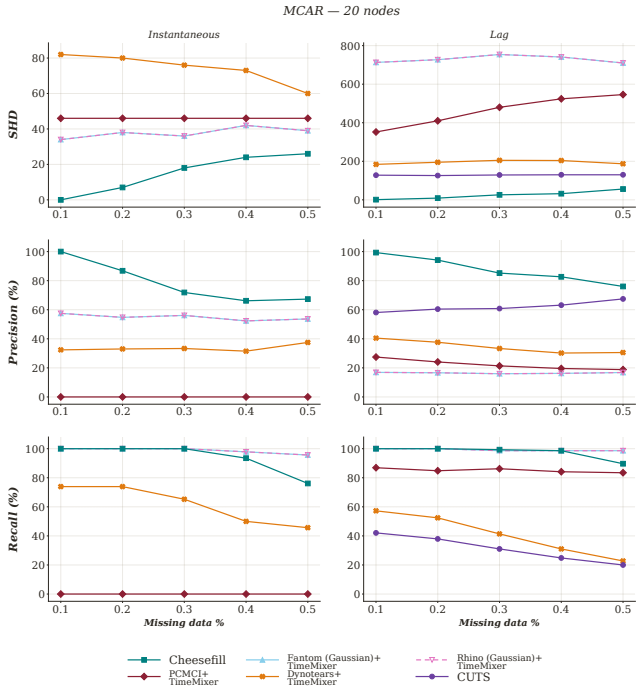


Figure 14: MCAR MTS with 20 nodes. SHD (top, lower is better), Precision (middle, higher is better) and Recall (bottom, higher is better) for instantaneous and time lagged edges, evaluated under 3 maximum lags as the missing rate increases from 10% to 50%.

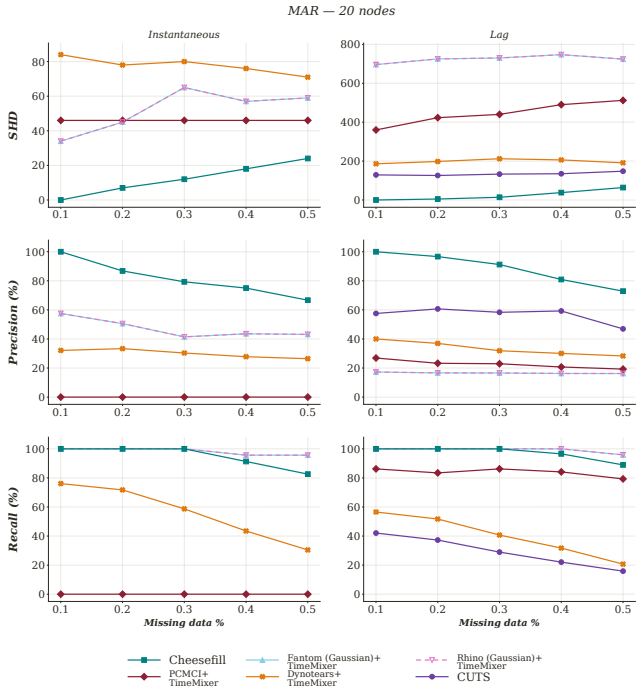


Figure 15: MAR MTS with 20 nodes. SHD (top, lower is better), Precision (middle, higher is better) and Recall (bottom, higher is better) for instantaneous and time lagged edges, evaluated under 3 maximum lags as the missing rate increases from 10% to 50%.

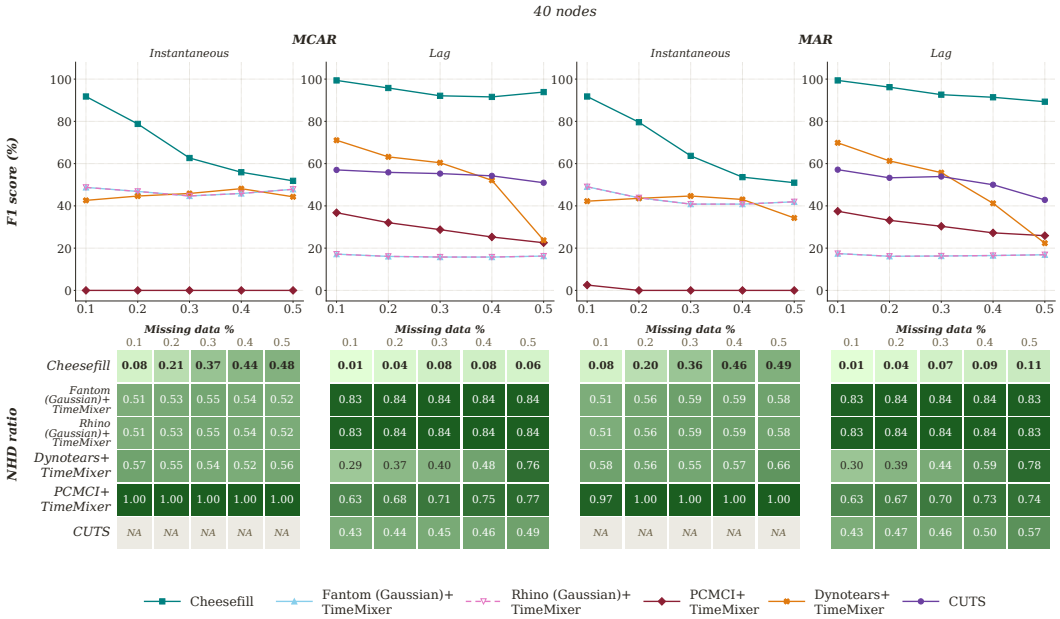


Figure 16: MCAR and MAR MTS with 40 nodes. F1 score (top, higher is better) and NHD ratio (bottom, lower is better) for instantaneous and time lagged edges, evaluated under 3 maximum lags as the missing rate increases from 10% to 50%.

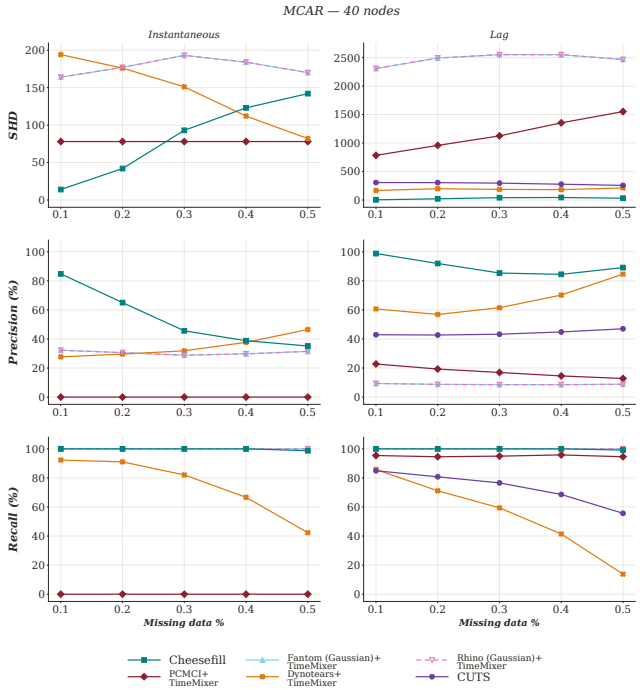


Figure 17: MCAR MTS with 40 nodes. SHD (top, lower is better), Precision (middle, higher is better) and Recall (bottom, higher is better) for instantaneous and time lagged edges, evaluated under 3 maximum lags as the missing rate increases from 10% to 50%.

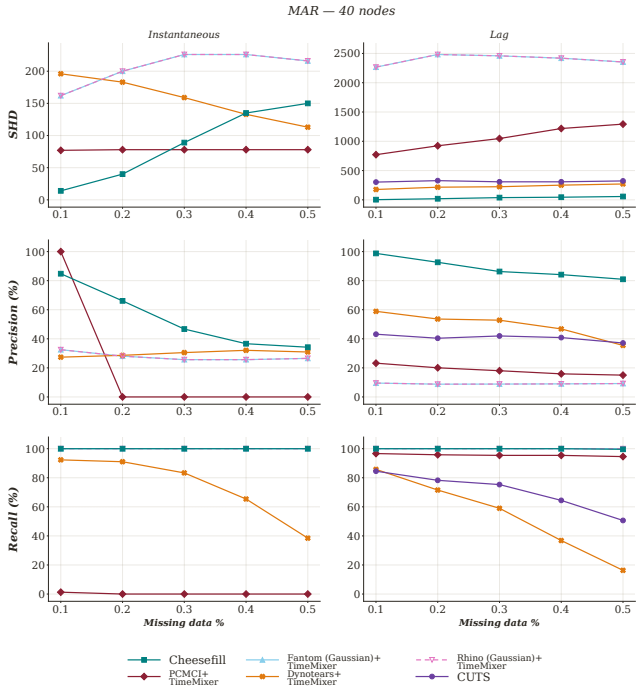
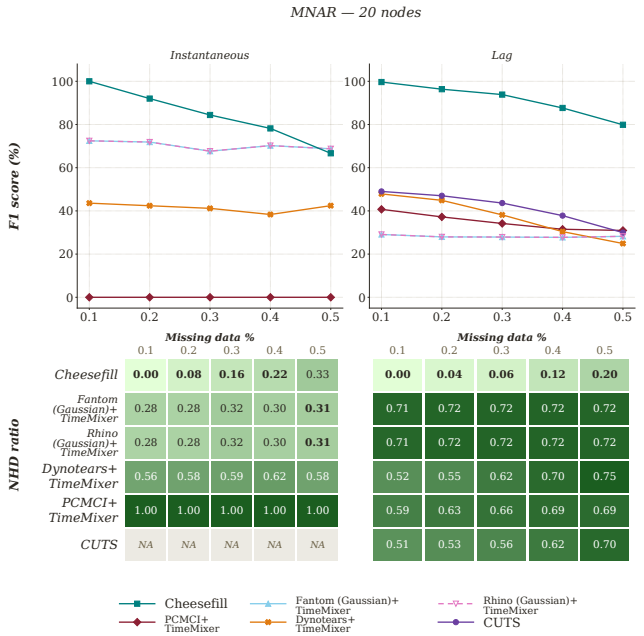


Figure 18: MAR MTS with 40 nodes. SHD (top, lower is better), Precision (middle, higher is better) and Recall (bottom, higher is better) for instantaneous and time lagged edges, evaluated under 3 maximum lags as the missing rate increases from 10% to 50%.

H.2 MNAR results



NHD ratio	0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5
Cheesefill	**0.00**	**0.08**	**0.16**	**0.22**	0.33	**0.00**	**0.04**	**0.06**	**0.12**	**0.20**
Fantom (Gaussian)+ TimeMixer	0.28	0.28	0.32	0.30	**0.31**	0.71	0.72	0.72	0.72	0.72
Rhino (Gaussian)+ TimeMixer	0.28	0.28	0.32	0.30	**0.31**	0.71	0.72	0.72	0.72	0.72
Dynotears+ TimeMixer	0.56	0.58	0.59	0.62	0.58	0.52	0.55	0.62	0.70	0.75
PCMCI+ TimeMixer	1.00	1.00	1.00	1.00	1.00	0.59	0.63	0.66	0.69	0.69
CUTS	*NA*	*NA*	*NA*	*NA*	*NA*	0.51	0.53	0.56	0.62	0.70

Figure 19: MNAR MTS with 20 nodes. F1 score (top, higher is better) and NHD ratio (bottom, lower is better) for instantaneous and time lagged edges, evaluated under 3 maximum lags as the missing rate increases from 10% to 50%.

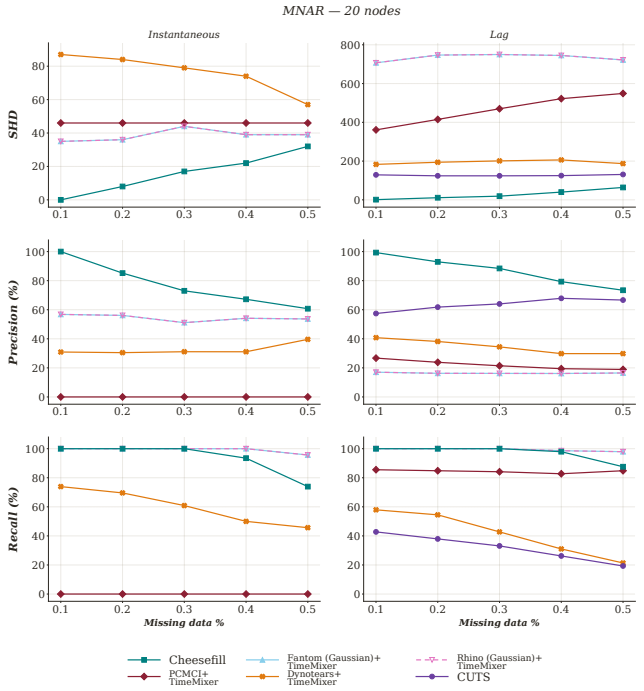


Figure 20: MNAR MTS with 20 nodes. SHD (top, lower is better), Precision (middle, higher is better) and Recall (bottom, higher is better) for instantaneous and time lagged edges, evaluated under 3 maximum lags as the missing rate increases from 10% to 50%.

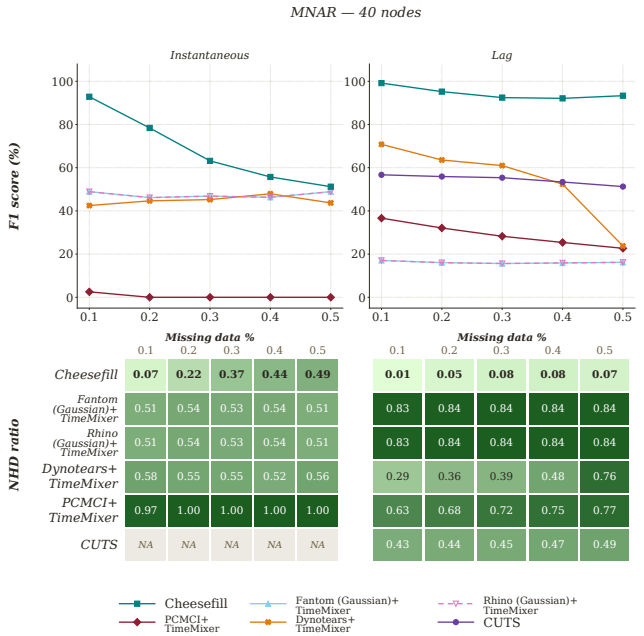


Figure 21: MNAR MTS with 40 nodes. F1 score (top, higher is better) and NHD ratio (bottom, lower is better) for instantaneous and time lagged edges, evaluated under 3 maximum lags as the missing rate increases from 10% to 50%.

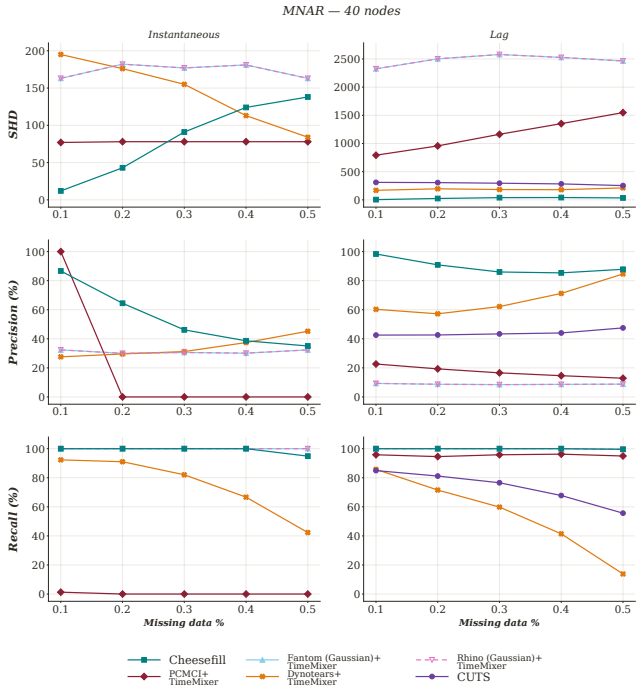


Figure 22: MNAR MTS with 40 nodes. SHD (top, lower is better), Precision (middle, higher is better) and Recall (bottom, higher is better) for instantaneous and time lagged edges, evaluated under 3 maximum lags as the missing rate increases from 10% to 50%.

I Graph average cost details

In this section, we announce the proposition for the graph averaged mechanism cost and we show that the equivalence and parametric upper bound results of Proposition 3.1 remain valid when $c_{\mathcal{G}}$ is replaced by $\bar{c}_\phi$. We present also the full proof for this proposition.

Proposition I.1 (Graph averaged correction map formulation)**.** *Let $(\Omega_X, \mathcal{B}(\Omega_X))$ be the measurable space of complete, finite length, d variate MTS trajectories, and assume that Ω_X is a standard Borel space. Let $P_{\tilde{\boldsymbol{X}}_\zeta}$ be the pseudo complete data distribution on Ω_X, and let Q_Θ be the SEM induced complete trajectory distribution on Ω_X. For a graph posterior $q_\phi(\mathcal{G})$, define the graph averaged mechanism cost $\bar{c}_\phi(\tilde{\boldsymbol{X}}, \boldsymbol{Y}) := \mathbb{E}_{\mathcal{G}\sim q_\phi(\mathcal{G})}\big[c_{\mathcal{G}}(\tilde{\boldsymbol{X}}, \boldsymbol{Y})\big]$, and assume that $\bar{c}_\phi$ is measurable and satisfies $\mathbb{E}_{(X,Y)\sim\Gamma}\big[|\bar{c}_\phi(X,Y)|\big] < \infty$ for all couplings Γ considered. Let $\mathcal{H}(q_\phi) = -\mathbb{E}_{q_\phi}[\log q_\phi]$, and assume $\mathcal{H}(q_\phi) < \infty$. Then, for any fixed ϕ and $\beta > 0$,*

$$W_{\bar{c}_\phi}\big(P_{\tilde{\boldsymbol{X}}_\zeta}, Q_\Theta\big) - \beta\mathcal{H}(q_\phi) = \inf_{T\in\mathcal{F}(P_{\tilde{\boldsymbol{X}}_\zeta}, Q_\Theta)} \left\{\mathbb{E}_{\tilde{\boldsymbol{X}}\sim P_{\tilde{\boldsymbol{X}}_\zeta}} \mathbb{E}_{\boldsymbol{Y}\sim T(\cdot|\tilde{\boldsymbol{X}})}\big[\bar{c}_\phi(\tilde{\boldsymbol{X}}, \boldsymbol{Y})\big] - \beta\mathcal{H}(q_\phi)\right\},$$

where $\mathcal{F}(P_{\tilde{\boldsymbol{X}}_\zeta}, Q_\Theta)$ is the set of admissible stochastic kernels in Definition 2.1.
Moreover, for any parametric family $\{T_\psi\}_{\psi\in\Psi}$, let

$$\Psi_{\text{adm}} = \{\psi \in \Psi : T_\psi \in \mathcal{F}(P_{\tilde{\boldsymbol{X}}_\zeta}, Q_\Theta)\}.$$

Then

$$W_{\bar{c}_\phi}\big(P_{\tilde{\boldsymbol{X}}_\zeta}, Q_\Theta\big) - \beta\mathcal{H}(q_\phi) \le \inf_{\psi\in\Psi_{\text{adm}}} \left\{\mathbb{E}_{\tilde{\boldsymbol{X}}\sim P_{\tilde{\boldsymbol{X}}_\zeta}} \mathbb{E}_{\boldsymbol{Y}\sim T_\psi(\cdot|\tilde{\boldsymbol{X}})}\big[\bar{c}_\phi(\tilde{\boldsymbol{X}}, \boldsymbol{Y})\big] - \beta\mathcal{H}(q_\phi)\right\}.$$

The inequality is tight whenever the parametric family contains an optimal admissible kernel for the graph averaged cost $\bar{c}_\phi$. More generally, if the family contains an ε-optimal admissible kernel, then the right-hand side is at most ε above the unrestricted optimum.

Proof. For readability, write

$$P := P_{\tilde{\boldsymbol{X}}_\zeta}, \qquad Q := Q_\Theta.$$

Both P and Q are probability measures on the same measurable trajectory space $(\Omega_X, \mathcal{B}(\Omega_X))$.

Fix ϕ. Since

$$\bar{c}_\phi(x, y) = \mathbb{E}_{\mathcal{G}\sim q_\phi(\mathcal{G})}[c_{\mathcal{G}}(x, y)]$$

is assumed measurable and integrable under the couplings considered, $\bar{c}_\phi$ is a valid transport cost for the Kantorovich problem

$$W_{\bar{c}_\phi}(P, Q) = \inf_{\Gamma\in\mathcal{P}(X\sim P,\, Y\sim Q)} \mathbb{E}_{(X,Y)\sim\Gamma}[\bar{c}_\phi(X, Y)].$$

We first show that this Kantorovich problem is equivalent to an optimization over admissible stochastic kernels. Take any coupling

$$\Gamma \in \mathcal{P}(X \sim P, Y \sim Q).$$

Since Ω_X is a standard Borel space, Γ admits a regular conditional distribution of the second coordinate given the first. Therefore, there exists a stochastic kernel $T_\Gamma(dy \mid x)$ such that

$$\Gamma(dx, dy) = P(dx)T_\Gamma(dy \mid x).$$

For every measurable set $A \in \mathcal{B}(\Omega_X)$,

$$\int_{\Omega_X} T_\Gamma(A \mid x)\, dP(x) = \Gamma(\Omega_X \times A).$$

Because Γ has second marginal Q,

$$\Gamma(\Omega_X \times A) = Q(A).$$

Thus,

$$\int_{\Omega_X} T_\Gamma(A \mid x)\, dP(x) = Q(A),$$

so $T_\Gamma \in \mathcal{F}(P, Q)$. Moreover, by disintegration,

$$\mathbb{E}_{(X,Y)\sim\Gamma}[\bar{c}_\phi(X, Y)] = \mathbb{E}_{X\sim P}\mathbb{E}_{Y\sim T_\Gamma(\cdot|X)}[\bar{c}_\phi(X, Y)].$$

Hence every Kantorovich coupling induces an admissible stochastic kernel with the same graph-averaged mechanism cost.

Conversely, take any admissible stochastic kernel

$$T \in \mathcal{F}(P, Q).$$

Define a probability measure Γ_T on $\Omega_X \times \Omega_X$ by

$$\Gamma_T(dx, dy) = P(dx)T(dy \mid x).$$

Equivalently, for measurable rectangles $B \times A$,

$$\Gamma_T(B \times A) = \int_B T(A \mid x)\, dP(x).$$

For any $B \in \mathcal{B}(\Omega_X)$,

$$\Gamma_T(B \times \Omega_X) = \int_B T(\Omega_X \mid x)\, dP(x) = P(B),$$

so the first marginal of Γ_T is P. For any $A \in \mathcal{B}(\Omega_X)$,

$$\Gamma_T(\Omega_X \times A) = \int_{\Omega_X} T(A \mid x)\, dP(x).$$

Since T is admissible, the right-hand side equals $Q(A)$. Therefore the second marginal of Γ_T is Q, and

$$\Gamma_T \in \mathcal{P}(X \sim P, Y \sim Q).$$

The cost of this coupling is

$$\mathbb{E}_{(X,Y)\sim\Gamma_T}[\bar{c}_\phi(X,Y)] = \mathbb{E}_{X\sim P}\mathbb{E}_{Y\sim T(\cdot|X)}[\bar{c}_\phi(X,Y)].$$

Thus, every coupling induces an admissible kernel with the same $\bar{c}_\phi$-cost, and every admissible kernel induces a coupling with the same $\bar{c}_\phi$-cost. Hence

$$W_{\bar{c}_\phi}(P,Q) = \inf_{T\in\mathcal{F}(P,Q)} \mathbb{E}_{X\sim P}\mathbb{E}_{Y\sim T(\cdot|X)}[\bar{c}_\phi(X,Y)].$$

Now consider the entropy term. For fixed ϕ, the quantity

$$-\beta\mathcal{H}(q_\phi)$$

does not depend on the coupling Γ, the admissible kernel T, or the parametric correction map T_ψ. It is therefore an additive constant with respect to the transport optimization. Subtracting this same constant from both sides gives

$$W_{\bar{c}_\phi}(P,Q) - \beta\mathcal{H}(q_\phi) = \inf_{T\in\mathcal{F}(P,Q)} \left\{\mathbb{E}_{X\sim P}\mathbb{E}_{Y\sim T(\cdot|X)}[\bar{c}_\phi(X,Y)] - \beta\mathcal{H}(q_\phi)\right\},$$

which proves the equality. Substituting back $P = P_{\tilde{\boldsymbol{X}}_\zeta}$ and $Q = Q_\Theta$ gives the stated form.

It remains to prove the parametric upper bound. By definition,

$$\{T_\psi : \psi \in \Psi_{\text{adm}}\} \subseteq \mathcal{F}(P,Q).$$

Therefore, taking the infimum over the restricted parametric family cannot give a smaller value than taking the infimum over all admissible kernels:

$$\inf_{T\in\mathcal{F}(P,Q)} \mathbb{E}_{X\sim P}\mathbb{E}_{Y\sim T(\cdot|X)}[\bar{c}_\phi(X,Y)] \le \inf_{\psi\in\Psi_{\text{adm}}} \mathbb{E}_{X\sim P}\mathbb{E}_{Y\sim T_\psi(\cdot|X)}[\bar{c}_\phi(X,Y)].$$

Again, since $-\beta\mathcal{H}(q_\phi)$ is independent of T and ψ, subtracting it from both sides preserves the inequality:

$$\begin{aligned} &W_{\bar{c}_\phi}(P,Q) - \beta\mathcal{H}(q_\phi) \\ &\quad\le \inf_{\psi\in\Psi_{\text{adm}}} \left\{\mathbb{E}_{X\sim P}\mathbb{E}_{Y\sim T_\psi(\cdot|X)}[\bar{c}_\phi(X,Y)] - \beta\mathcal{H}(q_\phi)\right\}. \end{aligned}$$

This proves the inequality.

If the parametric family contains an optimal admissible kernel for $\bar{c}_\phi$, then the restricted parametric infimum equals the unrestricted kernel infimum, so the inequality is tight.

More generally, suppose there exists $\psi_\varepsilon \in \Psi_{\text{adm}}$ such that

$$\mathbb{E}_{X\sim P}\mathbb{E}_{Y\sim T_{\psi_\varepsilon}(\cdot|X)}[\bar{c}_\phi(X,Y)] \le W_{\bar{c}_\phi}(P,Q) + \varepsilon.$$

Since the entropy term is the same constant on both sides,

$$\mathbb{E}_{X\sim P}\mathbb{E}_{Y\sim T_{\psi_\varepsilon}(\cdot|X)}[\bar{c}_\phi(X,Y)] - \beta\mathcal{H}(q_\phi) \le W_{\bar{c}_\phi}(P,Q) - \beta\mathcal{H}(q_\phi) + \varepsilon.$$

Therefore the restricted parametric objective is at most ε above the unrestricted graph-averaged Kantorovich objective with entropy regularization. □

J Proof of our proposition

J.1 Assumptions

In this section, we define and clarify the required assumptions announced in the main text for Cheesefill.

Assumption J.1 (Causal Stationarity, [35])**.** A stationary time series process $(\boldsymbol{x}_t)_{t\in\mathcal{T}}$ with graph $\mathcal{G}$ is called causally stationary over a time index set $\mathcal{T}$ if and only if for all links $\boldsymbol{x}^i_{t-\tau} \to \boldsymbol{x}^j_t$ in the graph

$$x^i_{t-\tau} \not\perp\!\!\!\perp x^j_t \mid \boldsymbol{x}_{<t} \setminus \{x^i_{t-\tau}\}.$$

This elucidates the inherent characteristics of the time-series data generation mechanism, thereby validating the choice of the auto-regressive model.

Assumption J.2. (Causal Markov Property, [31]). Given a DAG $\mathcal{G}$ and a joint distribution p, this distribution is said to satisfy causal Markov property w.r.t. the DAG $\mathcal{G}$ if each variable is independent of its non-descendants given its parents.

This is a common assumptions for the distribution induced by an SEM. With this assumption, one can deduce conditional independence between variables from the graph.

Assumption J.3 (Causal Markov Property (CMP))**.** A joint distributions $(p(\cdot|\mathcal{G}))$ satisfies the **CMP** with respect to the DAG $(\mathcal{G})$ if each variable is independent of its non-descendants given its parents.

Assumption J.4 (Causal Minimality)**.** Given DAG $(\mathcal{G})$ and a joint distribution $(p(\cdot|\mathcal{G}))$, we say that this distributions satisfies causal minimality w.r.t. the DAG $(\mathcal{G})$ if $p(\cdot|\mathcal{G})$ is Markovian w.r.t the DAG $\mathcal{G}$ but not to any proper subgraph of $\mathcal{G}$.

Assumption J.5 (Causal Sufficiency)**.** A set of observed variables $\boldsymbol{V}$ is causally sufficient for a process $\boldsymbol{x}_t$ if and only if in the process every common cause of any two or more variables in $\boldsymbol{V}$ is in $\boldsymbol{V}$ or has the same value for all units in the population.

This assumption implies there are no latent confounders present in the time-series data.

J.2 Proof

Proposition J.6. *Let $(\Omega_X, \mathcal{B}(\Omega_X))$ be the measurable space of complete length $|\mathcal{T}|$, d variate MTS trajectories. Assume that Ω_X is a standard Borel space. Let $P_{\tilde{\boldsymbol{X}}_\zeta}$ be the pseudo complete data distribution on Ω_X, and let Q_Θ be the SEM induced complete trajectory distribution on Ω_X. Let $c_\mathcal{G} : \Omega_X \times \Omega_X \to \mathbb{R}_+$ be a measurable mechanism cost satisfying $\mathbb{E}_{(X,Y)\sim\Gamma}\big[|c_\mathcal{G}(X,Y)|\big] < \infty$ for all couplings Γ considered. Then*

$$W_{c_\mathcal{G}}\big(P_{\tilde{\boldsymbol{X}}_\zeta}, Q_\Theta\big) = \inf_{T\in\mathcal{F}(P_{\tilde{\boldsymbol{X}}_\zeta}, Q_\Theta)} \mathbb{E}_{\tilde{\boldsymbol{X}}\sim P_{\tilde{\boldsymbol{X}}_\zeta}} \mathbb{E}_{\boldsymbol{Y}\sim T(\cdot|\tilde{\boldsymbol{X}})}\big[c_\mathcal{G}(\tilde{\boldsymbol{X}}, \boldsymbol{Y})\big], \tag{14}$$

where $\mathcal{F}(P_{\tilde{\boldsymbol{X}}_\zeta}, Q_\Theta)$ is the set of admissible stochastic kernels in Definition 2.1. Moreover, for any parametric family $\{T_\psi\}_{\psi\in\Psi}$, let $\Psi_{\text{adm}} = \{\psi \in \Psi : T_\psi \in \mathcal{F}(P_{\tilde{\boldsymbol{X}}_\zeta}, Q_\Theta)\}$. Then

$$W_{c_\mathcal{G}}\big(P_{\tilde{\boldsymbol{X}}_\zeta}, Q_\Theta\big) \le \inf_{\psi\in\Psi_{\text{adm}}} \mathbb{E}_{\tilde{\boldsymbol{X}}\sim P_{\tilde{\boldsymbol{X}}_\zeta}} \mathbb{E}_{\boldsymbol{Y}\sim T_\psi(\cdot|\tilde{\boldsymbol{X}})}\big[c_\mathcal{G}(\tilde{\boldsymbol{X}}, \boldsymbol{Y})\big]. \tag{15}$$

The inequality is tight whenever the parametric family contains an optimal admissible kernel. More generally, if the family contains an ε-optimal admissible kernel, then the right-hand side is at most ε above the unrestricted Kantorovich optimum.

Proof. For readability, write

$$P := P_{\tilde{\boldsymbol{X}}_\zeta}, \qquad Q := Q_\Theta.$$

Both P and Q are probability measures on the same measurable trajectory space $(\Omega_X, \mathcal{B}(\Omega_X))$. By definition, the Kantorovich cost is

$$W_{c_\mathcal{G}}(P, Q) = \inf_{\Gamma\in\mathcal{P}(X\sim P,\, Y\sim Q)} \mathbb{E}_{(X,Y)\sim\Gamma}\big[c_\mathcal{G}(X,Y)\big],$$

where $\mathcal{P}(X \sim P, Y \sim Q)$ denotes the set of couplings whose first marginal is P and whose second marginal is Q.

We prove that optimizing over couplings $\Gamma \in \mathcal{P}(X \sim P, Y \sim Q)$ is equivalent to optimizing over admissible stochastic kernels $T \in \mathcal{F}(P, Q)$.

First, take any coupling

$$\Gamma \in \mathcal{P}(X \sim P, Y \sim Q).$$

Since Ω_X is a standard Borel space, every probability measure on $\Omega_X \times \Omega_X$ admits a regular conditional distribution. Therefore, there exists a stochastic kernel $T_\Gamma(dy \mid x)$ such that

$$\Gamma(dx, dy) = P(dx)T_\Gamma(dy \mid x).$$

We now show that T_Γ is admissible. For any measurable set $A \in \mathcal{B}(\Omega_X)$,

$$\int_{\Omega_X} T_\Gamma(A \mid x)\, dP(x) = \Gamma(\Omega_X \times A).$$

Because Γ has second marginal Q, we have

$$\Gamma(\Omega_X \times A) = Q(A).$$

Hence

$$\int_{\Omega_X} T_\Gamma(A \mid x)\, dP(x) = Q(A),$$

so $T_\Gamma \in \mathcal{F}(P, Q)$. Moreover, by the disintegration identity,

$$\mathbb{E}_{(X,Y)\sim\Gamma}\big[c_{\mathcal{G}}(X, Y)\big] = \mathbb{E}_{X\sim P}\mathbb{E}_{Y\sim T_\Gamma(\cdot|X)}\big[c_{\mathcal{G}}(X, Y)\big].$$

The equality is well-defined by the assumed measurability and integrability of $c_{\mathcal{G}}$. Therefore, every Kantorovich coupling induces an admissible stochastic kernel with exactly the same cost.

Conversely, take any admissible stochastic kernel

$$T \in \mathcal{F}(P, Q).$$

Define a probability measure Γ_T on $\Omega_X \times \Omega_X$ by

$$\Gamma_T(dx, dy) = P(dx)T(dy \mid x).$$

Equivalently, for measurable rectangles $B \times A$, with $A, B \in \mathcal{B}(\Omega_X)$,

$$\Gamma_T(B \times A) = \int_B T(A \mid x)\, dP(x).$$

We verify that Γ_T is a coupling of P and Q. For any measurable set $B \in \mathcal{B}(\Omega_X)$,

$$\Gamma_T(B \times \Omega_X) = \int_B T(\Omega_X \mid x)\, dP(x) = \int_B 1\, dP(x) = P(B),$$

so the first marginal of Γ_T is P. Similarly, for any measurable set $A \in \mathcal{B}(\Omega_X)$,

$$\Gamma_T(\Omega_X \times A) = \int_{\Omega_X} T(A \mid x)\, dP(x).$$

Since T is admissible, the right-hand side equals $Q(A)$. Therefore,

$$\Gamma_T(\Omega_X \times A) = Q(A),$$

so the second marginal of Γ_T is Q. Hence

$$\Gamma_T \in \mathcal{P}(X \sim P, Y \sim Q).$$

The cost of this coupling is

$$\mathbb{E}_{(X,Y)\sim\Gamma_T}\big[c_{\mathcal{G}}(X, Y)\big] = \mathbb{E}_{X\sim P}\mathbb{E}_{Y\sim T(\cdot|X)}\big[c_{\mathcal{G}}(X, Y)\big].$$

Thus, every admissible stochastic kernel induces a Kantorovich coupling with exactly the same cost.

The two constructions show that the optimization over Kantorovich couplings and the optimization over admissible stochastic kernels have the same feasible objects up to disintegration, and the same objective value. Therefore,

$$W_{c_{\mathcal{G}}}(P, Q) = \inf_{T\in\mathcal{F}(P,Q)} \mathbb{E}_{X\sim P}\mathbb{E}_{Y\sim T(\cdot|X)}\big[c_{\mathcal{G}}(X, Y)\big].$$

Substituting back

$$P = P_{\tilde{\boldsymbol{X}}_\zeta}, \qquad Q = Q_\Theta,$$

gives Eq. (4).

It remains to prove the parametric upper bound. By definition,

$$\Psi_{\mathrm{adm}} = \{\psi \in \Psi : T_\psi \in \mathcal{F}(P, Q)\}.$$

Hence the parametric admissible family

$$\{T_\psi : \psi \in \Psi_{\mathrm{adm}}\}$$

is a subset of the unrestricted admissible kernel class $\mathcal{F}(P, Q)$. Taking an infimum over a smaller feasible set cannot produce a smaller value than taking an infimum over the full feasible set. Therefore,

$$\inf_{T \in \mathcal{F}(P,Q)} \mathbb{E}_{X\sim P}\mathbb{E}_{Y\sim T(\cdot|X)}\big[c_{\mathcal{G}}(X, Y)\big] \le \inf_{\psi\in\Psi_{\mathrm{adm}}} \mathbb{E}_{X\sim P}\mathbb{E}_{Y\sim T_\psi(\cdot|X)}\big[c_{\mathcal{G}}(X, Y)\big].$$

Using the equality already proved, the left-hand side is $W_{c_{\mathcal{G}}}(P, Q)$. Hence

$$W_{c_{\mathcal{G}}}(P, Q) \le \inf_{\psi\in\Psi_{\mathrm{adm}}} \mathbb{E}_{X\sim P}\mathbb{E}_{Y\sim T_\psi(\cdot|X)}\big[c_{\mathcal{G}}(X, Y)\big],$$

which proves Eq. (5). Substituting back $P = P_{\tilde{\boldsymbol{X}}_\zeta}$ and $Q = Q_\Theta$ gives the stated bound.

If the parametric family contains an optimal admissible kernel, then there exists $\psi^\star \in \Psi_{\mathrm{adm}}$ such that

$$T_{\psi^\star} \in \arg\min_{T\in\mathcal{F}(P,Q)} \mathbb{E}_{X\sim P}\mathbb{E}_{Y\sim T(\cdot|X)}\big[c_{\mathcal{G}}(X, Y)\big].$$

In that case, the restricted parametric infimum attains the same value as the unrestricted infimum, and the inequality in Eq. (5) is tight.

More generally, suppose the parametric family contains an ε-optimal admissible kernel. That is, suppose there exists $\psi_\varepsilon \in \Psi_{\mathrm{adm}}$ such that

$$\mathbb{E}_{X\sim P}\mathbb{E}_{Y\sim T_{\psi_\varepsilon}(\cdot|X)}\big[c_{\mathcal{G}}(X, Y)\big] \le W_{c_{\mathcal{G}}}(P, Q) + \varepsilon.$$

Then, since T_{ψ_ε} is one feasible element of the restricted parametric class,

$$\inf_{\psi\in\Psi_{\mathrm{adm}}} \mathbb{E}_{X\sim P}\mathbb{E}_{Y\sim T_\psi(\cdot|X)}\big[c_{\mathcal{G}}(X, Y)\big] \le W_{c_{\mathcal{G}}}(P, Q) + \varepsilon.$$

Combining this with the lower bound just proved gives

$$W_{c_{\mathcal{G}}}(P, Q) \le \inf_{\psi\in\Psi_{\mathrm{adm}}} \mathbb{E}_{X\sim P}\mathbb{E}_{Y\sim T_\psi(\cdot|X)}\big[c_{\mathcal{G}}(X, Y)\big] \le W_{c_{\mathcal{G}}}(P, Q) + \varepsilon.$$

Therefore the parametric objective is at most ε above the unrestricted Kantorovich optimum whenever the parametric family contains an ε-optimal admissible kernel. □

Proposition 3.1 shows that the correction map is a conditional representation of an OT coupling. A coupling $\Gamma \in \mathcal{P}(X \sim P_{\tilde{\boldsymbol{X}}_\zeta}, Y \sim Q_\Theta)$ specifies how mass from the pseudo complete distribution is matched to mass from the SEM induced distribution. Since the trajectory space is standard Borel, this coupling can be disintegrated into the source marginal $P_{\tilde{\boldsymbol{X}}_\zeta}$ and a conditional kernel $T(d\boldsymbol{Y} \mid \tilde{\boldsymbol{X}})$. Conversely, any admissible kernel defines a coupling with the required marginals. Thus, the stochastic correction map is not an auxiliary imputation heuristic; it is an amortized parameterization of the conditional component of a Kantorovich coupling.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: Yes the main claims are clear in the abstract, introduction and in the whole paper

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Yes, we have a section in appendix B in which we talk about the limitation of our work.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: All assumptions and complete proofs appear in Appendix J.2.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: Yes, we provide in the main text figures that illustrate the exact architecture used for our model. In the appendices, we provide the exact hyper-parameters needed to reproduce our results in all the presented experiments. Our code is also provided in the supplementary material.

Guidelines:

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

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

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

   Answer: [Yes]

   Justification: Our code, with all the instruction to reproduce our results, is provided in the supplementary materials

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: All the detailed needed to run our code and to reproduce our results are presented in Appendix C.3.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: We provide results with standard deviations in all our experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Our model takes 4 minutes to run on A100 with 80 Gb

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: -

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: In our introduction, we talk about paper's broader impact.

    Guidelines:

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

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: -

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: -

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: e presented all the details about our model: parameters, architecture, data. Also the detailed proofs are provided

    Guidelines:

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

14. **Crowdsourcing and research with human subjects**

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

    Answer: [N/A]

    Justification: -

    Guidelines:

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

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

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

    Answer: [N/A]

    Justification: -

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: -

    Guidelines:

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