DIFFERENTIABLE LATENT-TREE CAUSAL DISCOVERY VIA TRIPLET BRANCH-POINT OBJECTIVES AND THREE-TERMINAL ROOT IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Discovering directed causal structure in the presence of latent variables remains challenging, particularly when identifiability principles must be translated into practical learning objectives. We study latent-tree structural causal models, in which observed variables are leaves and latent variables form the internal structure of a directed tree. We introduce a differentiable framework for recovering both the latent-tree skeleton and its causal orientation from observational data. For skeleton learning, we develop a differentiable triplet branch-point (TBP) formulation for latent-tree completion: using the branch-additivity form of the established median property, simplex-valued triplet-center assignments jointly learn unknown branch-point identities and latent-related distances. For causal orientation, we establish a tree-conditioned three-terminal independence (TTI) characterization of the latent root under linear non-Gaussian LT-SCM assumptions; orienting all edges away from the identified root then recovers the directed latent-tree causal graph. Synthetic experiments across diverse latent-tree structures demonstrate strong structure-recovery performance, while experiments on the White Wine Quality dataset and a VAE-based generative model illustrate the applicability of the proposed framework to real data and end-to-end representation learning.

1 INTRODUCTION

Latent-variable causal discovery seeks to recover causal structure among interacting observed and unobserved factors, and is fundamental to causal reasoning in complex systems (Pearl, 2009; Spirtes et al., 2000; Imbens & Rubin, 2015; Peters et al., 2017). Hidden variables are ubiquitous in scientific and real-world data, making latent-variable causal models an important interface between observational data and mechanistic interpretation (Spirtes et al., 2000; Glymour et al., 2019; Schölkopf et al., 2021).

Among latent-variable models, latent tree models (Zhang, 2004; Mourad et al., 2013) provide a fundamental structured class in which observed and latent variables interact through a tree. Although the tree constraint removes cycles and restricts the combinatorial structure, causal discovery remains nontrivial: the internal variables are unobserved, the latent topology itself must be reconstructed from observed dependence, and an undirected tree does not determine the causal directions. Existing statistically consistent methods primarily target recovery of the undirected latent-tree topology (Choi et al., 2011; Wang et al., 2017; Zhou et al., 2020; Wang et al., 2024), often through discrete reconstruction procedures such as neighbor joining or sibling grouping. Consequently, they do not directly provide a differentiable route to full directed causal recovery or integration into end-to-end learning systems.

Recent work has established identifiability for broader classes of latent-variable causal models under a variety of structural and distributional assumptions (Kong et al., 2023; Xie et al., 2024; Prashant et al., 2025; Li et al., 2025). However, identifiability alone does not directly yield an explicit learning objective that reliably reconstructs latent structure and supports gradient-based optimization.

This gap is already substantial in latent-tree causal models: different methods rely on different structural conditions and may therefore apply to different subsets of latent-tree topologies. Figure 1 illustrates the structural diversity considered in our study, comprising 190 latent-tree topologies with

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

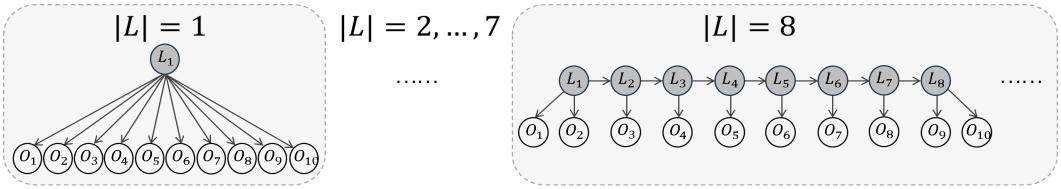


Figure 1: Latent-tree topology collection considered in our comprehensive synthetic benchmark. With 10 observed leaves, the collection contains 190 valid latent-tree structures with varying branching patterns and between 1 and 8 latent internal nodes.

varying branching patterns and numbers of latent internal nodes. We evaluate structural recovery across this complete collection; comprehensive results are reported in Section 6.2.

Motivated by the challenge of converting identifiability insights into explicit and optimizable criteria for latent causal structure learning, we focus on the latent-tree setting and develop a differentiable framework for learning the latent-tree skeleton, followed by a root-identification and edge-orientation stage. Our key idea is to turn established triplet branch-point geometry into a learning principle for incomplete latent-tree distances. For a tree metric, requiring the median (branch point) of every triplet to belong to the represented vertex set is an established fully labeled-tree property (Hayamizu et al., 2017); in our setting, however, the branch-point identities and all latent-related distances are unknown. We therefore introduce simplex-valued candidate-center assignments and branch-wise TBP residuals, and jointly optimize these assignments with latent-related distances while anchoring the observed-leaf information distances. This lifts a geometric existence condition into a differentiable latent-variable objective that can also be integrated into end-to-end learning pipelines. Finally, building on Triad/GIN-style non-Gaussian linear-combination independence (Cai et al., 2019; Xie et al., 2020; 2024), we use the recovered tree geometry to define TTI, an iff characterization of the unique latent root; orienting away from that root yields the directed LT-SCM. Overall, our framework connects identifiability to actionable learning through latent-tree completion and root characterization. This work makes the following contributions:

Lifted latent-tree completion. Building on established branch-point geometry, we jointly optimize unknown triplet branch-point assignments and latent-related distances: simplex-valued assignments lift the existential center choice, branch-wise TBP residuals enforce compatibility, and observed-leaf distances anchor completion.

Tree-conditioned root characterization. Given the recovered skeleton, TTI organizes observed terminal triplets by their tree branch point and gives a necessary-and-sufficient characterization of the unique latent root under our LT-SCM assumptions; orienting away from that root yields the full directed graph.

Unified directed recovery and differentiable integration. The resulting pipeline combines skeleton completion, root identification, and orientation, while the continuous TBP module supports gradient-based integration, as illustrated by the VAE proof-of-concept experiment.

2 Related Work

Causal discovery with latent variables. Causal discovery with latent variables is extensively studied via constraint-based frameworks that represent Markov equivalence classes under latent confounding and selection bias, such as ancestral graphs and PAGs, together with sound orientation rules and fast variants (Spirtes et al., 2000; Richardson & Spirtes, 2002; Zhang, 2008; Glymour et al., 2019; Spirtes & Zhang, 2016; Colombo et al., 2012; Ogarrio et al., 2016). Complementary to equivalence-class characterizations, functional-model approaches seek identifiability by leveraging stronger distributional/structural assumptions. Representative examples include Triad/GIN independent-noise criteria (Cai et al., 2019; Xie et al., 2020; 2024), nonlinear latent-hierarchy identification (Kong et al., 2023), rank-based discovery (Dong et al., 2024), and homologous surrogates (Li et al., 2025). Triad uses a three-variable pseudo-residual and non-Gaussian independence; GIN generalizes this to variable sets and describes Triad as a restrictive special case. TTI instead applies this principle after skeleton recovery, conditioning on tree-defined terminal triplets to characterize

the unique root. These methods operate under structural assumptions different from the latent-tree setting.

Structural discovery for latent tree models. Learning *undirected* latent-tree graphical models has a long history, spanning tree-metric characterizations and phylogenetic-style reconstruction methods (Buneman, 1971; 1974; Semple & Steel, 2003; Saitou & Nei, 1987; Atteson, 1999). The classical four-point condition characterizes tree metrics while allowing additional unlabeled internal vertices; the fourth-point condition requires the median of every triplet to lie in the metric space itself, and together these conditions give the established fully labeled-tree characterization (Hayamizu et al., 2017). In machine learning and statistics, consistent and scalable procedures have been developed based on score maximization, neighbor-joining variants, and sibling-grouping techniques (Zhang, 2004; Mourad et al., 2013; Choi et al., 2011; Wang et al., 2017; Zhou et al., 2020; Wang et al., 2024). These methods primarily target the recovery of the undirected latent-tree topology, while causal edge orientations typically require additional assumptions or interventional information. Our use of branch-point geometry differs at the learning level: the latent vertices, latent-related distances, and triplet branch-point identities are unknown and are optimized jointly.

Differentiable structural learning. Continuous optimization has substantially improved scalability for observed-variable DAG learning (Zheng et al., 2018; Yu et al., 2019; Ng et al., 2020; Lachapelle et al., 2020; Wei et al., 2020; Bello et al., 2022; Ng et al., 2022). Recent work has started to incorporate differentiable objectives into latent hierarchical causal models and to connect causal discovery with representation learning pipelines (Prashant et al., 2025; Schölkopf et al., 2021). A key open challenge is to connect identifiability with actionable learning guarantees: while existing differentiable formulations can achieve identifiability under suitable structural assumptions, their optimization objectives are often detached from identifiability in latent-variable models, and they lack reliable criteria for ensuring correct structural recovery.

3 Latent Tree Structural Causal Model

In this paper, we study a class of *Latent Tree Structural Causal Models (LT-SCMs)* that combines (i) a latent-tree structural constraint over observed and latent variables and (ii) a linear non-Gaussian acyclic structural equation model.

Latent tree structure. An LT-SCM is associated with an undirected tree $T = (V, E)$, with respect to which the joint distribution is Markov. The vertex set V can be partitioned as $O \cup L$, where O denotes the observed variables and L the latent variables. Following standard assumptions for latent tree models (Wang et al., 2017), we further impose: (A1) each latent node has degree at least three, (A2) for every edge $(u, v) \in E$, the two endpoints are neither perfectly dependent nor independent, and (A3) observed variables correspond exactly to the leaf nodes of T, whereas latent variables occupy the internal nodes. Let $\pi_T(u, v)$ denote the unique simple path between u and v in T.

Structural equation model. We further assume that the designated root is a latent internal node, $r \in L$. We orient every undirected edge of T away from r, yielding a directed tree (and hence a DAG) $G = (V, E_r)$ (so no additional acyclicity constraint is needed). For each node $v \in V \setminus \{r\}$, let $\mathrm{pa}(v)$ denote its parent in G. Assume that the data-generating process follows a linear structural equation model according to this directed tree G:

$$X_v \;=\; \sum_{u \in \mathrm{pa}(v)} b_{vu} X_u \;+\; \varepsilon_v, \qquad v \in V, \tag{1}$$

where $b_{vu} \in \mathbb{R}$ is the edge coefficient (each node has at most one parent) and $\{\varepsilon_v\}_{v \in V}$ are mutually independent non-Gaussian noises with finite positive variances. We observe i.i.d. samples of $X_O = \{X_v : v \in O\}$, while $X_L = \{X_v : v \in L\}$ remains latent.

4 Triplet Branch-Point Geometry for Tree Structures

This section gives the geometric basis for our learning objective. We first establish the LT-SCM information distance, then state TBP as the branch-additivity form of the established triplet median property. For metrics, TBP is equivalent to the fourth-point condition; together with the tree-metric premise, it gives the established fully labeled-tree property used below.

4.1 Additive Distances for LT-SCMs

We equip the vertex set V with a correlation-based distance as follows. For any $u, v \in V$, let ρ_{uv} denote the correlation coefficient of two variables X_u and X_v. We can define a distance

$$d_{uv} \;=\; -\log|\rho_{uv}| \tag{2}$$

between u and v. We define $d_{uu} = 0$ for any u, and assume d_{uv} is well-defined for any $u, v \in V$.

We next establish the key identity: in an LT-SCM with mutually independent noises, correlations factorize multiplicatively through intermediate vertices along collider-free paths (Spirtes et al., 2000), and therefore the distance in Equation 2 is additive.

Lemma 4.1 (Correlation factorization along collider-free paths)**.** *Consider the LT-SCM in equation 1 with mutually independent noises and finite second moments. Let $u, v, w \in V$ be distinct vertices such that w lies on the unique path $\pi_T(u, v)$. Then*

$$\mathrm{Cov}(X_u, X_w)\mathrm{Cov}(X_w, X_v) = \mathrm{Cov}(X_u, X_v)\mathrm{Var}(X_w), \qquad \rho_{uv} = \rho_{uw}\rho_{wv}.$$

Lemma 4.1 shows that correlations factorize multiplicatively through intermediate vertices along collider-free paths. By applying the negative logarithm transformation in equation 2, this multiplicative composition is converted into an additive decomposition, yielding an additive metric property that will be used in our structural characterization. (Proofs of Lemma 4.1 and Theorem 4.2 are given in Appendix A.1 and A.2)

Theorem 4.2 (Additivity of correlation distance)**.** *Under the conditions of Lemma 4.1, if $d_{uv} = -\log|\rho_{uv}|$ is well-defined, then $d_{uv} = d_{uw} + d_{wv}$.*

Population tree-metric property. Because the LT-SCM is collider-free, the unique path between any two vertices contains no collider. Applying Lemma 4.1 recursively along the successive vertices of this path gives $|\rho_{uv}| = \prod_{(i,j)\in\pi_T(u,v)} |\rho_{ij}|$, and hence $d_{uv} = \sum_{(i,j)\in\pi_T(u,v)} d_{ij}$. Therefore, at the population level, the information distance $d_{uv} = -\log|\rho_{uv}|$ is a tree metric realized by the ground-truth tree T. Under assumption (A2), every edge distance is finite and strictly positive.

4.2 Triplet Branch-Point Condition

TBP asserts that every triple of distinct vertices has a *branch point* at which all three pairwise distances decompose additively. This is the branch-additivity form of the established median/fourth-point property; we retain the TBP name for the learning formulation below.

Definition 4.3 (Triplet Branch-Point Condition)**.** Let $d : V \times V \to \mathbb{R}_{\geq 0}$ be a distance on a vertex set V. We say the distance matrix $D = (d_{uv})_{u,v\in V}$ satisfies the *Triplet Branch-Point (TBP)* condition if for any three distinct nodes $u, v, w \in V$, there exists a node $m \in V$ such that

$$\begin{cases} d_{uv} = d_{um} + d_{mv}, \\ d_{uw} = d_{um} + d_{mw}, \\ d_{vw} = d_{vm} + d_{mw}. \end{cases} \tag{3}$$

4.3 TBP and Fully Labeled Tree Realizations

A metric on V is a *tree metric* if it is realized on a weighted tree that may contain extra vertices; equivalently, it satisfies the four-point condition (Buneman, 1971; 1974). The fourth-point condition of Hayamizu et al. (2017) additionally requires, for every $u, v, w \in V$, some $m \in V$ with $d_{um} + d_{vm} + d_{wm} = \frac{1}{2}(d_{uv} + d_{uw} + d_{vw})$. For a metric this is equivalent to the three TBP equalities, since the three triangle slacks are nonnegative and sum to zero. Hence the next theorem restates the established fully labeled-tree characterization in TBP notation.

Theorem 4.4 (Fully labeled tree realization in TBP form)**.** *Let V be a finite set and let $D = (d_{uv})_{u,v\in V}$ be a tree metric on V. Then the following are equivalent:*

(i) D satisfies the Triplet Branch-Point (TBP) condition;

(ii) D admits an additive realization on a weighted tree $T = (V, E)$ whose vertex set is exactly V, with nonnegative edge lengths $\{\ell_e\}_{e\in E}$, such that, for all $u, v \in V$,

$$d_{uv} = \sum_{e\in\pi_T(u,v)} \ell_e. \tag{4}$$

A direct proof is given in Appendix A.3. On $T = (V, E)$, every triplet u, v, w has a unique median on all three pairwise paths; we denote it by $m = t(u, v, w)$. Path additivity gives the TBP equalities in equation 3.

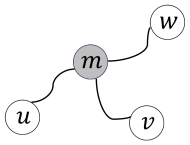


Figure 2: For any tree T and three nodes u, v, w, there exists a unique triplet branch-point m lying on all three paths between each pair of nodes.

Relation to tree-metric conditions. Four-point permits an additive realization with extra unlabeled branch vertices; fourth-point, equivalently TBP for a metric, requires triplet medians to lie in V (Hayamizu et al., 2017). This equivalence is established in the metric-tree literature; here it provides the geometric basis for our lifted latent-tree learning objective. Our contribution is the branch-wise optimization over unknown branch-point assignments and latent-related distances.

5 Latent-Tree Causal Discovery

We now turn the geometric existence statement "for each triplet, there exists a branch point" into a latent-variable objective. The branch-point identity and latent-related distances d_{OL}, d_{LL} are unknown, so we optimize them jointly via simplex-valued candidate-center assignments and branch-wise TBP residuals. Zero TBP loss certifies TBP; the tree-metric premise remains separate.

5.1 Lifted TBP Objective

For a latent tree model, observed variables are leaves, while latent variables form the internal branching structure. Consequently, under (A3), for any observed triplet $(u, v, w) \in O^3$ with distinct u, v, w, the triplet branch-point $t(u, v, w)$ is necessarily an internal (latent) node. We go further with this fact by introducing, for each triplet in V^3, a simplex-valued assignment that places weights on candidate branch points in V. Let $\mathcal{S} := \{(u, v, w) \in V^3 : u < v < w\}$ be the set of unordered triplets, and let $\alpha^{uvw} = (\alpha^{uvw}_m)_{m\in V}$ be a simplex weight vector over candidate centers $m \in V$.

Loss function for differentiable learning. We optimize the following objective, which combines the TBP loss, a data-fitting term, and an observed-center penalty:

$$\begin{aligned} \min_{d,\alpha} \quad & \mathcal{L}_{\mathrm{TBP}}(d, \alpha) + \lambda_{\mathrm{data}}\mathcal{L}_{\mathrm{data}}(d) + \lambda_O\mathcal{L}_O(\alpha) \\ \text{s.t.} \quad & \alpha^{uvw}_m \geq 0, \quad \forall (u, v, w) \in \mathcal{S},\ \forall m \in V, \\ & \sum_{m\in V} \alpha^{uvw}_m = 1, \quad \forall (u, v, w) \in \mathcal{S}, \\ & d_{uv} = d_{vu}, \quad d_{uu} = 0, \quad d_{uv} \geq 0, \quad \forall u, v \in V. \end{aligned} \tag{5}$$

Each α^{uvw} is a simplex-valued relaxation of the otherwise existential branch-point choice. These basic distance constraints do not enforce tree-metricity, which remains a separate premise.

(1) TBP loss. For each triplet $(u, v, w) \in \mathcal{S}$ and candidate center $m \in V$, define the TBP additivity residual

$$r^{uvw}_m(d) = \big(d_{uv} - d_{um} - d_{mv}\big)^2 + \big(d_{uw} - d_{um} - d_{mw}\big)^2 + \big(d_{vw} - d_{vm} - d_{mw}\big)^2. \tag{6}$$

When d is a tree distance and $m = t(u, v, w)$, the residual is 0 (up to estimation noise). Although TBP and the fourth-point condition have the same zero set on metrics, an optimization iterate need not be metric. The scalar equality underlying the fourth-point condition constrains only the sum of the three path-additivity deviations, so positive and negative deviations may cancel, whereas the branch-wise squared residual penalizes each deviation separately. We aggregate residuals using the simplex assignment weights:

$$\mathcal{L}_{\mathrm{TBP}}(d, \alpha) = \sum_{(u,v,w)\in\mathcal{S}} \sum_{m\in V} \alpha^{uvw}_m\, r^{uvw}_m(d). \tag{7}$$

The TBP loss provides a differentiable way to enforce the triplet branch-point condition. Here, d denotes the learned distance matrix, while the simplex variables α assign weights to candidate branch points for each triplet. For a fixed triplet (u, v, w), the residual $r_m^{uvw}(d)$ measures whether the candidate node m is compatible with the branch-point role under the current distance matrix. Thus, minimizing the simplex-weighted TBP loss encourages the triplet-wise additive decompositions required by TBP. In particular, $\mathcal{L}_{\mathrm{TBP}}(d, \alpha) = 0$ certifies the TBP condition. Tree-metric membership is a separate premise in the identification results below.

The following theorem states the exact zero-loss implication within the tree-metric class (Proof in Appendix A.4). Zero TBP loss certifies the TBP condition; when d is additionally a tree metric, Theorem 4.4 guarantees an additive realization whose vertex set is exactly V.

Theorem 5.1 (Zero TBP loss under the tree-metric premise)**.** *Suppose that d is a tree metric on V. If*

$$\mathcal{L}_{\mathrm{TBP}}(d, \alpha) = 0,$$

then d satisfies the TBP condition on V and, by Theorem 4.4, admits an additive realization on a weighted tree $T = (V, E)$ whose vertex set is exactly V.

Moreover, if all edge lengths are strictly positive, then this fully labeled weighted-tree realization is uniquely determined by d.

(2) Data fitting loss. Let $\widehat{d}_{uv}$ be the empirical observed–observed distance estimated from samples (e.g., $\widehat{d}_{uv} = -\log|\widehat{\rho}_{uv}|$). Since only observed variables are measured, we fit the observed–observed block of d to $\widehat{d}$:

$$\mathcal{L}_{\mathrm{data}}(d) = \sum_{\substack{u<v \\ u,v\in O}} \left(d_{uv} - \widehat{d}_{uv}\right)^2. \tag{8}$$

This term softly anchors the observed–observed block of d to its empirical estimate. Alternatively, when this block is held fixed, we impose $d_{uv} = \widehat{d}_{uv}$ for all $u, v \in O$ and omit $\mathcal{L}_{\mathrm{data}}$; distances involving latent nodes are then inferred through the TBP constraints.

(3) Observed-center penalty. Under (A3), observed variables are leaves and cannot be branch points for triplets of distinct vertices. We therefore penalize observed centers via

$$\mathcal{L}_O(\alpha) = \sum_{(u,v,w)\in\mathcal{S}} \sum_{m\in O} \alpha_m^{uvw}. \tag{9}$$

This term discourages observed nodes from being selected as triplet branch points.

5.2 Characterization for the Root in LT-SCMs

Given the recovered undirected latent tree, we next identify its root and orient all edges accordingly. Triad/GIN already use non-Gaussian linear-combination independence for latent-structure discovery, with Triad as a three-variable special case (Cai et al., 2019; Xie et al., 2020; 2024). TTI differs by conditioning this principle on recovered tree geometry: candidate m is tested only through terminal triplets with $t(u, v, w) = m$. We formalize this tree-conditioned property as the *three-terminal independence condition* (TTI), which will serve as the criterion for root identification.

Definition 5.2 (Three-terminal independence condition)**.** Let $T = (V, E)$ be a tree over $V = O \cup L$ and let $m \in V$. We say that m satisfies the *three-terminal independence condition* (TTI) if for every observed triplet $(u, v, w) \in O^3$ such that $m = t(u, v, w)$, and for each choice of $c \in \{u, v, w\}$ with $\{a, b\} = \{u, v, w\} \setminus \{c\}$, there exist coefficients $(\lambda_a, \lambda_b) \neq (0, 0)$ such that $\lambda_a X_a + \lambda_b X_b \perp\!\!\!\perp X_c$.

The TTI condition captures a root-specific independence signature: for every observed triplet branching at the root, a nontrivial linear combination of two terminals can be made independent of the third. Under our LT-SCM assumptions, the following theorem shows that TTI is necessary and sufficient for root identification.

Theorem 5.3 (Three-terminal independences for root characterization)**.** *Let $T = (V, E)$ be a collider-free directed tree over $V = O \cup L$ with a unique latent root $r \in L$. Assume the variables follow an LT-SCM on T with mutually independent non-Gaussian noises and nonzero edge coefficients. Then an internal node $m \in L$ is the root if and only if it satisfies the TTI condition.*

On root identification. Theorem 5.3 provides an operational criterion for root recovery via the three-terminal independence condition (TTI) (The proof can be found in Appendix A.5). In practice, this makes root identification straightforward: for each candidate internal node $m \in L$, we collect observed triplets $(u, v, w) \in O^3$ with $m = t(u, v, w)$ and test whether the corresponding three-terminal independences hold (e.g., by estimating the coefficients (λ_a, λ_b) from second-order moments and applying a kernel-based independence test such as HSIC). The true root is the (unique) internal node that satisfies TTI, after which orienting all edges away from it deterministically yields the directed latent-tree causal graph.

5.3 Learning the LT-SCMs

A key ingredient for our overall recovery result is that, under the standard non-degeneracy assumption (in particular, ruling out degree-two internal nodes), complete pairwise information among observed variables suffices to recover the undirected latent tree. Specifically, Wang et al. (2017, Theorem 3.2) shows that, under their stated latent-tree model assumptions, the underlying latent tree can be recovered exactly from complete pairwise information among observed variables; equivalently, in the corresponding distance representation, the observed pairwise information distances are consistent with an additive tree metric induced by the recovered tree. Accordingly, suppose that d is a tree metric on V whose observed–observed block matches the population pairwise information distances, and that there exists a feasible α such that $\mathcal{L}_{\mathrm{TBP}}(d, \alpha) = 0$. By Theorem 5.1, d then admits an additive realization on a fully labeled tree. Under the standard nondegeneracy and reducedness conditions for latent-tree models, the population observed pairwise distances uniquely determine the reduced undirected latent-tree skeleton, up to relabeling of latent variables.

Theorem 5.4 (Population recovery of the latent-tree causal structure)**.** *Consider an LT-SCM on a collider-free directed tree over $V = O \cup L$ with a unique latent root $r \in L$. Assume that the exogenous noises are mutually independent, non-Gaussian, and have finite, strictly positive variances, and that all edge coefficients are nonzero. Let $d^\star_{uv}$ denote the population information distances for $u, v \in O$.*

Suppose there exists a distance matrix d on V such that (i) *d is a tree metric on V,* (ii) *$d_{uv} = d^\star_{uv}$ for all $u, v \in O$, and* (iii) *there exists a feasible α with $\mathcal{L}_{\mathrm{TBP}}(d, \alpha) = 0$. Assume further that the fully labeled additive realization of d is compatible with the LT-SCM model: O is exactly the leaf set, every latent vertex has degree at least three, and all edge lengths are strictly positive.*

Then the reduced undirected latent-tree skeleton coincides with the ground-truth skeleton up to relabeling of latent variables. Moreover, the TTI condition of Theorem 5.3 identifies the latent root, and orienting all edges away from it recovers the directed latent-tree causal graph. Consequently, under the stated assumptions, the population observational distribution uniquely determines the directed latent-tree causal structure up to relabeling of latent variables.

On minimizing the loss function. The observed-center penalty $\mathcal{L}_O$ encodes the leaf constraint on observed variables and is not required for the population identification argument. The relevant identification conditions concern the other two terms: zero data-fitting loss enforces $d_{uv} = d^\star_{uv}$ for all $u, v \in O$, while zero TBP loss certifies that d satisfies the TBP condition.

Importantly, zero TBP loss alone does not establish that d is a tree metric. Theorems 4.4 and 5.1 therefore apply to population solutions that additionally lie in the tree-metric class. In the population LT-SCM, the ground-truth information distance has this property because it is induced by the underlying latent tree.

6 Experiments

6.1 Setup

We evaluate our method on synthetic latent-tree SCMs, the White Wine Quality dataset, and an end-to-end VAE-based generative modeling task. Synthetic data are generated from the SCM in Equation 1. Section 6.2 first benchmarks structural recovery across all 190 topologies in Figure 1, and then examines representative 6-leaf and 30-leaf LT-SCMs in greater detail. Additional results across structures, noise distributions, and sensitivity settings are provided in Appendices C.2 and E.

Baselines: We compare with LaHiCSL (Xie et al., 2024), CDHS (Li et al., 2025), and DCD (Prashant et al., 2025). Because these methods impose different structural assumptions, their applicable topology sets may differ.

Metrics: LCER measures exact recovery of the number of latent variables, SER measures exact recovery of the undirected skeleton, and SHD measures directed structural errors after minimizing over latent-node permutations. For synthetic experiments, we hard-anchor the empirical observed–observed distances, whereas real-data experiments use the soft data-fitting term in equation 8. Formal definitions and complete experimental details are provided in Appendix C.1.

6.2 Synthetic Results

Comprehensive benchmark across 190 topologies. We first evaluate structural recovery across the complete collection of 190 latent-tree topologies shown in Figure 1. All structures contain 10 observed leaves and between 1 and 8 latent internal nodes. Data are generated under exponential noise with $n = 5000$. Because the competing methods impose different structural assumptions, their applicable topology sets differ. LaHiCSL applies to all 190 structures, CDHS to 106, and DCD to 8, while our method is evaluated on the full collection. We therefore report, for each method, average performance over the structures to which that method is applicable.

Table 1: Structural recovery across the 190 latent-tree topologies in Figure 1 under exponential noise with $n = 5000$. Results for each method are averaged over its applicable topology set, whose size is shown in parentheses.

LaHiCSL (190/190)		**CDHS (106/190)**		**DCD (8/190)**		**Ours (190/190)**	
SER↑	**SHD↓**	**SER↑**	**SHD↓**	**SER↑**	**SHD↓**	**SER↑**	**SHD↓**
0.02	8.60 ± 4.81	0.14	15.52 ± 9.30	0.00	–	**0.53**	$\mathbf{5.40 \pm 4.49}$

Among the methods applicable to all 190 structures, our method substantially improves exact skeleton recovery over LaHiCSL (SER 0.53 versus 0.02) and also achieves lower SHD. CDHS and DCD are applicable only to restricted subsets of the benchmark, reflecting their additional structural requirements. Thus, this experiment evaluates not only recovery accuracy but also robustness across substantial variation in latent-tree topology.

Representative 6-leaf setting. We next examine the 6-leaf LT-SCM in Figure 3 in more detail across sample sizes. Additional synthetic results are provided in Appendix C.2.

Table 2 reports results under exponential noise for $n \in \{1\text{k}, 3\text{k}, 5\text{k}\}$. Our method achieves perfect skeleton recovery at all three sample sizes and reaches SHD $= 0$ for $n \geq 3\text{k}$, indicating accurate recovery of both the skeleton and its orientation. LaHiCSL and DCD retain structural errors, while CDHS is not applicable to this topology.

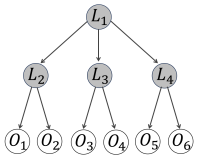


Figure 3: The 6-leaf LT-SCM used in Table 2.

Table 2: Causal discovery performance on the LT-SCM in Figure 3 under exponential noise.

Sample	**LaHiCSL**			**DCD**			**Ours**		
	LCER↑	**SER↑**	**SHD↓**	**LCER↑**	**SER↑**	**SHD↓**	**LCER↑**	**SER↑**	**SHD↓**
1k	0.08	0.08	5.08 ± 1.87	**1.00**	0.52	2.08 ± 2.25	**1.00**	**1.00**	$\mathbf{0.32 \pm 0.74}$
3k	0.50	0.46	2.60 ± 2.49	**1.00**	0.74	0.98 ± 1.72	**1.00**	**1.00**	**0**
5k	0.78	0.76	1.12 ± 2.09	**1.00**	0.72	1.12 ± 1.89	**1.00**	**1.00**	**0**

Larger 30-leaf setting. We further evaluate a 30-leaf LT-SCM to assess performance at a larger scale; the structure is shown in Appendix C.2, Figure 7. For computational efficiency, we use 30% triplet subsampling. Under exponential noise with $n = 5\text{k}$, our method achieves LCER $= 0.76$, SER $= 0.52$, and SHD $= 2.82 \pm 4.01$. LaHiCSL and DCD obtain SER $= 0$ in this setting, while CDHS

is not applicable. These results indicate that the TBP objective remains effective as the latent-tree scale and structural complexity increase.

6.3 Real-Data Analysis

We apply our method to the White Wine Quality dataset. As shown in Figure 4, the estimated hierarchy groups fixed acidity, citric acid, and pH under one latent branch, while chlorides, sulfur-dioxide measures, density, and alcohol form another major branch connected with residual sugar. These patterns admit natural acidity- and fermentation/flavor-related interpretations, which we treat as exploratory. DCD (Prashant et al., 2025) produces two disconnected trees, whereas LaHiCSL (Xie et al., 2024) and CDHS (Li et al., 2025) do not produce a comparable latent hierarchy under their respective settings. Detailed interpretations and the DCD result are provided in Appendix C.3.

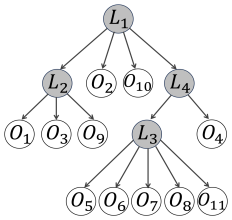


Figure 4: Estimated directed latent tree on White Wine.

6.4 End-to-End LT-SCM Learning with a VAE Backbone

We further examine whether the proposed TBP-based structural module can be integrated into an end-to-end generative model. We use a VAE backbone and compare with CausalVAE (Yang et al., 2021). Both methods use the same encoder and decoder architectures, but ours also uses the LT-SCM non-Gaussian noise specification whereas CausalVAE uses Gaussian noise. This is therefore a proof-of-concept of end-to-end integration, not an isolated structural-regularizer ablation or a claim that MNIST's true latent structure is a tree.

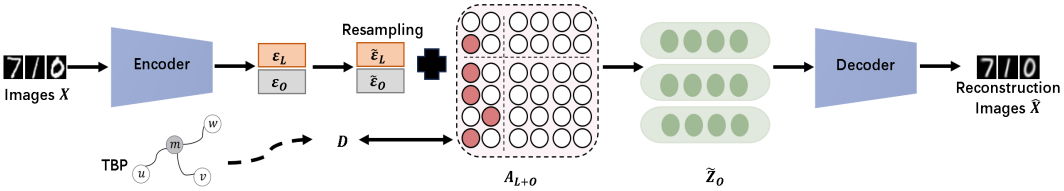


Figure 5: Overview of our end-to-end LT-SCM learning pipeline with a VAE backbone.

We augment the standard VAE objective with the TBP structure loss: $\mathcal{L} = \mathcal{L}_{\text{Rec}} + \beta\,\mathcal{L}_{\text{KL}} + \lambda_{\text{TBP}}\,\mathcal{L}_{\text{TBP}}(D(A), \alpha)$, where A is the learned latent structural matrix and $D(A)$ denotes the information-distance surrogate computed from the induced latent representation. At 50 epochs, our reconstruction loss is comparable to that of CausalVAE, while the learned structure is sparser and more tree-consistent. Additional training details, loss curves, reconstructed images, and extracted latent graphs are provided in Appendix D.

7 Discussion and Conclusion

Our method combines lifted TBP completion based on established branch-point geometry with a tree-conditioned TTI root characterization. TTI builds on Triad/GIN-style non-Gaussian linear-combination independence (Cai et al., 2019; Xie et al., 2020; 2024), but uses recovered-tree terminal triplets to give an iff characterization of the unique latent root under our assumptions; orientation away from that root completes directed recovery. Our guarantees rely on the latent-tree assumption and LT-SCM conditions. When these assumptions are violated (e.g., non-tree latent structure or weak non-Gaussianity), orientation may degrade and the learned metric may only approximate the underlying dependency geometry. Computationally, the TBP objective scales as $T|V|^4$, where T is the number of iterations, motivating triplet subsampling/minibatching and constraint screening for larger problems. Future work will explore extensions beyond trees (e.g., forests or mixtures) and tighter integration of the differentiable TBP module into end-to-end representation learning pipelines.

AI USE STATEMENT

In this work, we used generative AI tools, including ChatGPT, to assist with three types of tasks: language editing and polishing of the manuscript; literature retrieval and discovery of potentially relevant prior work; and partial mathematical derivation, proof development, and proof checking.

All AI-assisted work was reviewed and verified by the authors. Mathematical arguments and theoretical claims were independently checked by the authors, and bibliographic information and citation-related claims were checked against the relevant source materials. AI-assisted language was reviewed and revised by the authors before inclusion in the manuscript. Synthetic datasets used in the experiments were generated by the specified data-generating processes and computational code, rather than by generative AI.

Generative AI tools did not make final scientific decisions or replace author verification of mathematical claims, experimental results, or cited sources. We take responsibility for the final content of this work.

ETHICS STATEMENT

The theoretical results of this work do not require collection of new human-subject data. The dataset used is publicly available.

REPRODUCIBILITY STATEMENT

Implementation details are provided in the appendix. The code will be made publicly available upon cleanup.

REFERENCES

Kevin Atteson. The performance of neighbor-joining methods of phylogenetic reconstruction. *Algorithmica*, 25(2–3):251–278, 1999. doi: 10.1007/PL00008277.

Kevin Bello, Bryon Aragam, and Pradeep K. Ravikumar. DAGMA: Learning DAGs via M-matrices and a log-determinant acyclicity characterization. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-0598.

Peter Buneman. The recovery of trees from measures of dissimilarity. In *Mathematics in the Archaeological and Historical Sciences*, pp. 387–395. Edinburgh University Press, Edinburgh, 1971.

Peter Buneman. A note on the metric properties of trees. *Journal of Combinatorial Theory, Series B*, 17(1):48–50, 1974. doi: 10.1016/0095-8956(74)90047-1.

Ruichu Cai, Feng Xie, Clark Glymour, Zhifeng Hao, and Kun Zhang. Triad constraints for learning causal structure of latent variables. In *Advances in Neural Information Processing Systems*, volume 32, pp. 12863–12872, 2019.

Myung Jin Choi, Vincent Y. F. Tan, Animashree Anandkumar, and Alan S. Willsky. Learning latent tree graphical models. *Journal of Machine Learning Research*, 12:1771–1812, 2011.

Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, 40(1):294–321, 2012. doi: 10.1214/11-AOS940.

Xinshuai Dong, Biwei Huang, Ignavier Ng, Xiangchen Song, Yujia Zheng, Songyao Jin, Roberto Legaspi, Peter Spirtes, and Kun Zhang. A versatile causal discovery framework to allow causally-related hidden variables. In *International Conference on Learning Representations*, 2024.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10:524, 2019. doi: 10.3389/fgene.2019.00524.

Arthur Gretton, Olivier Bousquet, Alexander J. Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert–Schmidt norms. In *Algorithmic Learning Theory*, pp. 63–77. Springer, 2005. doi: 10.1007/11564089_7.

Momoko Hayamizu, Hiroshi Endo, and Kenji Fukumizu. A characterization of minimum spanning tree-like metric spaces. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 14(2):468–471, 2017. doi: 10.1109/TCBB.2016.2550431.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

A. M. Kagan, Yu. V. Linnik, and C. R. Rao. *Characterization Problems in Mathematical Statistics*. Wiley, 1973.

Lingjing Kong, Biwei Huang, Feng Xie, Eric P. Xing, Yuejie Chi, and Kun Zhang. Identification of nonlinear latent hierarchical models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural DAG learning. In *International Conference on Learning Representations*, 2020.

Xiu-Chuan Li, Jun Wang, and Tongliang Liu. Recovery of causal graph involving latent variables via homologous surrogates. In *International Conference on Learning Representations*. OpenReview.net, 2025.

Raphaël Mourad, Christine Sinoquet, Nevin L. Zhang, Tengfei Liu, and Philippe Leray. A survey on latent tree models and applications. *Journal of Artificial Intelligence Research*, 47:157–203, 2013. doi: 10.1613/jair.3879.

Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and DAG constraints for learning linear DAGs. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Ignavier Ng, Sébastien Lachapelle, Nan Rosemary Ke, Simon Lacoste-Julien, and Kun Zhang. On the convergence of continuous constrained optimization for structure learning. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 8176–8198. PMLR, 2022.

Juan Miguel Ogarrio, Peter Spirtes, and Joe Ramsey. A hybrid causal search algorithm for latent variable models. In *Proceedings of the Eighth International Conference on Probabilistic Graphical Models*, volume 52 of *Proceedings of Machine Learning Research*, pp. 368–379. PMLR, 2016.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2 edition, 2009.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017.

Parjanya Prajakta Prashant, Ignavier Ng, Kun Zhang, and Biwei Huang. Differentiable causal discovery for latent hierarchical causal models. In *International Conference on Learning Representations*. OpenReview.net, 2025.

Thomas S. Richardson and Peter Spirtes. Ancestral graph markov models. *The Annals of Statistics*, 30(4):962–1030, 2002. doi: 10.1214/aos/1031689015.

Naruya Saitou and Masatoshi Nei. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, 1987. doi: 10.1093/oxfordjournals.molbev.a040454.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. doi: 10.1109/JPROC.2021.3058954.

Charles Semple and Mike Steel. *Phylogenetics*, volume 24 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, 2003.

Peter Spirtes and Kun Zhang. Causal discovery and inference: Concepts and recent methodological advances. *Applied Informatics*, 3:3, 2016. doi: 10.1186/s40535-016-0018-x.

Peter Spirtes, Clark N. Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2 edition, 2000.

Shuai Wang, Lizhu Hao, Xiaofei Wang, and Jianhua Guo. Paralinear distance and its algorithm for hierarchical clustering of high-dimensional discrete variables. *International Journal of Approximate Reasoning*, 167:109133, 2024. doi: 10.1016/j.ijar.2024.109133.

Xiaofei Wang, Jianhua Guo, Lizhu Hao, and Nevin L. Zhang. Spectral methods for learning discrete latent tree models. *Statistics and Its Interface*, 10(4):677–698, 2017. doi: 10.4310/SII.2017.v10.n4.a11.

Dennis Wei, Tian Gao, and Yue Yu. DAGs with no fears: A closer look at continuous optimization for learning bayesian networks. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Feng Xie, Ruichu Cai, Biwei Huang, Clark Glymour, Zhifeng Hao, and Kun Zhang. Generalized independent noise condition for estimating latent variable causal graphs. In *Advances in Neural Information Processing Systems*, volume 33, pp. 14891–14902, 2020.

Feng Xie, Biwei Huang, Zhengming Chen, Ruichu Cai, Clark Glymour, Zhi Geng, and Kun Zhang. Generalized independent noise condition for estimating causal structure with latent variables. *Journal of Machine Learning Research*, 25(191):1–61, 2024.

Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. CausalVAE: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9593–9602, 2021. doi: 10.1109/CVPR46437.2021.00947.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. DAG-GNN: DAG structure learning with graph neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7154–7163. PMLR, 2019.

Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16–17):1873–1896, 2008. doi: 10.1016/j.artint.2008.08.001.

Nevin L. Zhang. Hierarchical latent class models for cluster analysis. *Journal of Machine Learning Research*, 5:697–723, 2004.

Xun Zheng, Bryon Aragam, Pradeep K. Ravikumar, and Eric P. Xing. Dags with no tears: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Can Zhou, Xiaofei Wang, and Jianhua Guo. Learning mixed latent tree models. *Journal of Machine Learning Research*, 21(249):1–35, 2020.

A Proofs

A.1 Proof of Lemma 4.1

Lemma 4.1 [Correlation factorization along collider-free paths] Consider the LT-SCM in equation 1 with mutually independent noises and finite second moments. Let $u, v, w \in V$ be distinct vertices such that w lies on the unique path $\pi_T(u, v)$. Then

$$\mathrm{Cov}(X_u, X_w)\mathrm{Cov}(X_w, X_v) = \mathrm{Cov}(X_u, X_v)\mathrm{Var}(X_w), \qquad \rho_{uv} = \rho_{uw}\rho_{wv}.$$

Proof. Because all edges are oriented away from the unique root, every undirected path in the LT-SCM is collider-free. Let w lie on the unique path $\pi_T(u, v)$. We distinguish two cases.

Case I: w is an ancestor of both u and v. Since w lies on $\pi_T(u, v)$, the nodes u and v lie in two different descendant branches of w. By recursively substituting the linear structural equations along the two directed paths from w, we can write

$$X_u = a_u X_w + \eta_u, \qquad X_v = a_v X_w + \eta_v,$$

where a_u and a_v are the products of the edge coefficients along the corresponding directed paths. The random variables η_u and η_v depend only on exogenous noises strictly below w in the two disjoint branches. Hence η_u, η_v, and X_w are mutually independent.

Therefore,

$$\mathrm{Cov}(X_u, X_w) = a_u \mathrm{Var}(X_w),$$
$$\mathrm{Cov}(X_w, X_v) = a_v \mathrm{Var}(X_w),$$

and

$$\mathrm{Cov}(X_u, X_v) = a_u a_v \mathrm{Var}(X_w).$$

Consequently,

$$\mathrm{Cov}(X_u, X_w)\mathrm{Cov}(X_w, X_v) = \mathrm{Cov}(X_u, X_v)\mathrm{Var}(X_w).$$

Case II: w is an ancestor of exactly one of u and v. Without loss of generality, suppose that v is a descendant of w. Then

$$X_v = a_v X_w + \eta_v,$$

where a_v is the product of the edge coefficients along the directed path from w to v, and η_v depends only on exogenous noises strictly below w on the v-side of the tree. Because u lies in the other component of $T \setminus \{w\}$, η_v is independent of both X_u and X_w. Hence

$$\mathrm{Cov}(X_w, X_v) = a_v \mathrm{Var}(X_w),$$

and

$$\mathrm{Cov}(X_u, X_v) = a_v \mathrm{Cov}(X_u, X_w).$$

Multiplying the latter identity by $\mathrm{Var}(X_w)$ gives

$$\mathrm{Cov}(X_u, X_w)\mathrm{Cov}(X_w, X_v) = \mathrm{Cov}(X_u, X_v)\mathrm{Var}(X_w).$$

The case in which u is a descendant of w is symmetric.

Finally, assuming the relevant variances are positive, dividing the covariance identity by

$$\mathrm{Var}(X_w)\sqrt{\mathrm{Var}(X_u)\mathrm{Var}(X_v)}$$

yields

$$\rho_{uv} = \rho_{uw}\rho_{wv}.$$

□

A.2 Proof of Theorem 4.2

Theorem 4.2 [Additivity of correlation distance] Under the conditions of Lemma 4.1, if $d_{uv} = -\log|\rho_{uv}|$ is well-defined, then $d_{uv} = d_{uw} + d_{wv}$.

Proof. By Lemma 4.1, whenever w lies on the unique path $\pi_T(u, v)$,

$$\rho_{uv} = \rho_{uw}\rho_{wv}.$$

Assume the correlation-based distances are well-defined, i.e., $\rho_{uv} \neq 0$, $\rho_{uw} \neq 0$, and $\rho_{wv} \neq 0$ so that

$$d_{uv} = -\log|\rho_{uv}| = -\log\big(|\rho_{uw}| \cdot |\rho_{wv}|\big) = -\log|\rho_{uw}| - \log|\rho_{wv}| = d_{uw} + d_{wv}.$$

□

A.3 Proof of Theorem 4.4

Theorem 4.4 is the branch-additivity form, specialized to our notation, of the established fully labeled-tree characterization in Hayamizu et al. (2017): the tree-metric premise supplies the four-point condition, while TBP is equivalent to their fourth-point condition for a metric. We include a short direct proof for completeness.

Four-point condition. For completeness, a finite metric δ on a set X satisfies the four-point condition if, for every $a, b, c, e \in X$, the two largest values among

$$\delta_{ab} + \delta_{ce}, \qquad \delta_{ac} + \delta_{be}, \qquad \delta_{ae} + \delta_{bc}$$

are equal. This condition is equivalent to δ being a tree metric (Buneman, 1971; 1974).

Theorem 4.4 [Fully labeled tree realization in TBP form] Let V be a finite set and let $D = (d_{uv})_{u,v\in V}$ be a tree metric on V. Then D satisfies the TBP condition if and only if it admits an additive realization on a weighted tree $T = (V, E)$ whose vertex set is exactly V.

Proof. We prove both directions.

(ii)⇒(i). Suppose that D is realized by a weighted tree $T = (V, E)$. For any three distinct vertices $u, v, w \in V$, let m be their unique tree median, i.e.,

$$m \in \pi_T(u, v) \cap \pi_T(u, w) \cap \pi_T(v, w).$$

Because the vertex set of T is exactly V, we have $m \in V$. Path additivity then gives

$$d_{uv} = d_{um} + d_{mv}, \qquad d_{uw} = d_{um} + d_{mw}, \qquad d_{vw} = d_{vm} + d_{mw},$$

so the TBP condition holds.

(i)⇒(ii). Since D is a tree metric, it admits a weighted-tree realization $\widetilde{T} = (\widetilde{V}, \widetilde{E})$ with $V \subseteq \widetilde{V}$. Choose a reduced realization by removing unlabeled leaves, suppressing unlabeled degree-two vertices, and contracting zero-length edges whenever necessary. Thus every remaining edge has strictly positive length, every remaining unlabeled vertex has degree at least three, and the realization preserves all pairwise distances among vertices in V.

Suppose, for contradiction, that there exists an unlabeled vertex $q \in \widetilde{V} \setminus V$. Since the realization is reduced, $\deg(q) \geq 3$, and each component of $\widetilde{T} \setminus \{q\}$ contains at least one labeled vertex from V. Choose $u, v, w \in V$ from three distinct such components. Then q is the unique common intersection of the three pairwise paths:

$$\pi_{\widetilde{T}}(u, v) \cap \pi_{\widetilde{T}}(u, w) \cap \pi_{\widetilde{T}}(v, w) = \{q\}.$$

By the TBP condition, there exists $m \in V$ such that

$$d_{uv} = d_{um} + d_{mv}, \qquad d_{uw} = d_{um} + d_{mw}, \qquad d_{vw} = d_{vm} + d_{mw}.$$

Because all remaining edge lengths are strictly positive, in this reduced realization the equality $d_{ab} = d_{am} + d_{mb}$ holds if and only if m lies on the unique path between a and b. Hence m lies on all three pairwise paths above, so necessarily $m = q$. This contradicts $m \in V$ and $q \notin V$.

Therefore no unlabeled vertices remain in the reduced realization, so its vertex set is exactly V. Hence D admits an additive realization on a weighted tree $T = (V, E)$. □

A.4 Proof of Theorem 5.1

Theorem 5.1 [Zero TBP loss under the tree-metric premise] Suppose that d is a tree metric on V. If

$$\mathcal{L}_{\mathrm{TBP}}(d, \alpha) = 0,$$

then d satisfies the TBP condition on V and, by Theorem 4.4, admits an additive realization on a weighted tree $T = (V, E)$ whose vertex set is exactly V.

Moreover, if all edge lengths are strictly positive, then this fully labeled weighted-tree realization is uniquely determined by d.

Proof. Recall that, for each triplet $(u, v, w) \in \mathcal{S}$ and each candidate center $m \in V$, the TBP residual is defined as

$$r_m^{uvw}(d) = \big(d_{uv} - d_{um} - d_{mv}\big)^2 + \big(d_{uw} - d_{um} - d_{mw}\big)^2 + \big(d_{vw} - d_{vm} - d_{mw}\big)^2.$$

Each residual is nonnegative:

$$r_m^{uvw}(d) \geq 0, \qquad \forall (u, v, w) \in \mathcal{S},\ \forall m \in V.$$

Moreover, by the simplex constraints on the assignment variables,

$$\alpha_m^{uvw} \geq 0, \qquad \sum_{m \in V} \alpha_m^{uvw} = 1, \qquad \forall (u, v, w) \in \mathcal{S}.$$

Suppose that

$$\mathcal{L}_{\mathrm{TBP}}(d, \alpha) = 0.$$

By the definition of the TBP loss,

$$0 = \mathcal{L}_{\mathrm{TBP}}(d, \alpha) = \sum_{(u,v,w) \in \mathcal{S}} \sum_{m \in V} \alpha_m^{uvw} r_m^{uvw}(d).$$

Since every summand is nonnegative, the equality above implies

$$\alpha_m^{uvw} r_m^{uvw}(d) = 0, \qquad \forall (u, v, w) \in \mathcal{S},\ \forall m \in V.$$

Now fix an arbitrary triplet $(u, v, w) \in \mathcal{S}$. Since α^{uvw} is a probability vector over V, there exists at least one node

$$m^\star \in V$$

such that

$$\alpha_{m^\star}^{uvw} > 0.$$

For this node, the identity

$$\alpha_{m^\star}^{uvw} r_{m^\star}^{uvw}(d) = 0$$

implies

$$r_{m^\star}^{uvw}(d) = 0.$$

Therefore,

$$\big(d_{uv} - d_{um^\star} - d_{m^\star v}\big)^2 + \big(d_{uw} - d_{um^\star} - d_{m^\star w}\big)^2 + \big(d_{vw} - d_{vm^\star} - d_{m^\star w}\big)^2 = 0.$$

Since each term in the sum is a square and hence nonnegative, all three terms must be zero. Hence,

$$d_{uv} = d_{um^\star} + d_{m^\star v},$$

$$d_{uw} = d_{um^\star} + d_{m^\star w},$$

and

$$d_{vw} = d_{vm^\star} + d_{m^\star w}.$$

Thus, for the triplet (u, v, w), there exists a node $m^\star \in V$ through which the three pairwise distances decompose additively.

Because $(u, v, w) \in \mathcal{S}$ was arbitrary, the same conclusion holds for every triplet in $\mathcal{S}$. Therefore, d satisfies the TBP condition on V.

Since d is a tree metric by hypothesis and satisfies the TBP condition, Theorem 4.4 implies that d admits an additive realization on a weighted tree $T = (V, E)$ whose vertex set is exactly V. Hence, there exists a weighted tree

$$T = (V, E)$$

with edge-length function

$$\ell : E \to \mathbb{R}_{\geq 0}$$

such that, for every pair of nodes $a, b \in V$,

$$d_{ab} = \sum_{e \in \mathrm{path}_T(a,b)} \ell(e),$$

where $\mathrm{path}_T(a, b)$ denotes the unique path between a and b in T. Therefore, d is the additive path metric induced by the weighted tree T.

It remains to establish uniqueness when all edge lengths are strictly positive. For distinct $u, v \in V$, observe that

$$\{u, v\} \in E \quad \Longleftrightarrow \quad \text{there is no } w \in V \setminus \{u, v\} \text{ such that } d_{uv} = d_{uw} + d_{wv}.$$

Indeed, if u and v are not adjacent, their unique path in $T = (V, E)$ contains an intermediate vertex $w \in V \setminus \{u, v\}$, and path additivity gives $d_{uv} = d_{uw} + d_{wv}$. Conversely, if u and v are adjacent, their path contains no third vertex; since all edge lengths are strictly positive, no $w \in V \setminus \{u, v\}$ can satisfy this equality.

Hence the edge set E is determined uniquely by the full distance matrix d. For every recovered edge $\{u, v\} \in E$, its weight is necessarily $\ell_{uv} = d_{uv}$. Therefore the fully labeled weighted tree realizing d is unique. □

A.5 Proof of Theorem 5.3

Theorem A.1 (Darmois–Skitovich (DS) theorem (Kagan et al., 1973)). *Let $\{\varepsilon_s\}_{s\in\mathcal{X}}$ be mutually independent random variables, and consider two linear forms*

$$Y = \sum_{s\in\mathcal{X}} a_s \varepsilon_s, \qquad Z = \sum_{s\in\mathcal{X}} b_s \varepsilon_s,$$

with real coefficients $\{a_s\}_{s\in\mathcal{X}}$ and $\{b_s\}_{s\in\mathcal{X}}$. If $Y \perp\!\!\!\perp Z$, then for every $s \in \mathcal{X}$ such that $a_s b_s \neq 0$, the random variable ε_s must be Gaussian.

Theorem 5.3 [Three-terminal independences for root characterization] Let $T = (V, E)$ be a collider-free directed tree over $V = O \cup L$ with a unique latent root $r \in L$. Assume the variables follow an LT-SCM on T with mutually independent non-Gaussian noises and nonzero edge coefficients. Then an internal node $m \in L$ is the root if and only if it satisfies the TTI condition.

Proof. We first introduce several notations used only in this proof.

Notation. For a node $x \in V$, define its ancestor set (including itself)

$$\mathrm{Anc}(x) := \{\, s \in V : \; s = x \text{ or there exists a directed path } s \to \cdots \to x \,\}.$$

Under the LT-SCM on a directed tree, each variable admits a unique source expansion

$$X_x = \sum_{s\in\mathrm{Anc}(x)} \pi_{s\to x}\, \varepsilon_s,$$

where ε_s is the exogenous noise at node s (mutually independent and non-Gaussian by assumption), and $\pi_{s\to x}$ denotes the product of edge coefficients along the unique directed path from s to x (with the convention $\pi_{x\to x} = 1$). Since all edge coefficients are assumed nonzero, whenever $s \in \mathrm{Anc}(x)$ we have $\pi_{s\to x} \neq 0$.

We prove the two directions.

($\Rightarrow$) Root implies the three-terminal condition. Assume that for the observed triplet $(u, v, w) \in \mathcal{O}^3$, its TBP center equals the root, i.e., $m = t(u, v, w) = r$. In an undirected tree, $t(u, v, w) = r$ means that the three undirected paths from r to u, v, w are internally disjoint. Because the directed tree is collider-free and all edges are oriented away from r, r is the unique common ancestor of any two terminals among $\{u, v, w\}$, hence

$$\mathrm{Anc}(u) \cap \mathrm{Anc}(v) = \{r\}, \quad \mathrm{Anc}(u) \cap \mathrm{Anc}(w) = \{r\}, \quad \mathrm{Anc}(v) \cap \mathrm{Anc}(w) = \{r\}.$$

Fix any $c \in \{u, v, w\}$ and let $\{a, b\} = \{u, v, w\} \setminus \{c\}$. Using the source expansions, we can write

$$X_a = \pi_{r\to a}\varepsilon_r + \sum_{s\in\mathrm{Anc}(a)\setminus\{r\}} \pi_{s\to a}\varepsilon_s, \qquad X_b = \pi_{r\to b}\varepsilon_r + \sum_{t\in\mathrm{Anc}(b)\setminus\{r\}} \pi_{t\to b}\varepsilon_t.$$

By the intersection relations above, the noise sets

$$\{\varepsilon_s : s \in \mathrm{Anc}(a) \setminus \{r\}\}, \qquad \{\varepsilon_t : t \in \mathrm{Anc}(b) \setminus \{r\}\}, \qquad \{\varepsilon_q : q \in \mathrm{Anc}(c)\}$$

are pairwise disjoint. Choose

$$(\lambda_a, \lambda_b) := (\pi_{r\to b}, -\pi_{r\to a}) \neq (0,0),$$

and define $X_{ab} := \lambda_a X_a + \lambda_b X_b$. With this choice, the ε_r term cancels:

$$X_{ab} = \lambda_a \sum_{s\in \mathrm{Anc}(a)\setminus\{r\}} \pi_{s\to a}\varepsilon_s + \lambda_b \sum_{t\in \mathrm{Anc}(b)\setminus\{r\}} \pi_{t\to b}\varepsilon_t.$$

Therefore X_{ab} depends only on noises in $(\mathrm{Anc}(a) \setminus \{r\}) \cup (\mathrm{Anc}(b) \setminus \{r\})$, while X_c depends only on noises in $\mathrm{Anc}(c)$. These two noise sets are disjoint and all exogenous noises are mutually independent, hence $X_{ab} \perp\!\!\!\perp X_c$. Since c was arbitrary, the stated condition holds for every choice of $c \in \{u, v, w\}$.

($\Leftarrow$) Non-root TBP center violates the condition for some terminal choice. Assume $m \neq r$, and let $p := \mathrm{pa}(m)$ be the unique parent of m in the out-arborescence rooted at r. Since $m \in L$ has degree at least three and has one parent, it has at least two child-subtrees. Choose observed terminals $v, w \in O$ from two distinct child-subtrees of m.

Moreover, since the root is latent and observed vertices are leaves, p is also a latent internal node. Choose an observed terminal $u \in O$ from a child-subtree of p different from the child-subtree containing m. Then the three undirected paths $\pi(u, v)$, $\pi(u, w)$, and $\pi(v, w)$ intersect exactly at m, so

$$t(u, v, w) = m.$$

We claim that for the triple (u, v, w), the condition in the theorem fails by taking $c = w$ (and $a = u, b = v$): for all $(\lambda_a, \lambda_b) \neq (0, 0)$,

$$\lambda_a X_a + \lambda_b X_b \not\perp\!\!\!\perp X_c.$$

Let $X_{ab} := \lambda_a X_a + \lambda_b X_b$. We separate the source expansions into (i) noises in $\mathrm{Anc}(p)$, (ii) the noise ε_m, and (iii) noises strictly below the branching point on the corresponding branches:

$$X_a = \sum_{s\in \mathrm{Anc}(p)} \pi_{s\to a}\varepsilon_s \; + \sum_{t\in \mathrm{Desc}(p\rightsquigarrow u)} \pi_{t\to a}\varepsilon_t,$$

$$X_b = \sum_{s\in \mathrm{Anc}(p)} \pi_{s\to b}\varepsilon_s \; + \; \pi_{m\to b}\varepsilon_m \; + \sum_{t\in \mathrm{Desc}(m\rightsquigarrow v)} \pi_{t\to b}\varepsilon_t,$$

$$X_c = \sum_{s\in \mathrm{Anc}(p)} \pi_{s\to c}\varepsilon_s \; + \; \pi_{m\to c}\varepsilon_m \; + \sum_{t\in \mathrm{Desc}(m\rightsquigarrow w)} \pi_{t\to c}\varepsilon_t.$$

Here $\mathrm{Desc}(p \rightsquigarrow a)$ denotes the set of nodes strictly below p on the unique directed path from p to a, while $\mathrm{Desc}(m \rightsquigarrow b)$ and $\mathrm{Desc}(m \rightsquigarrow c)$ denote the sets of nodes strictly below m on the corresponding directed paths to b and c. By the nonzero edge-coefficient assumption,

$$\pi_{p\to a} \neq 0, \qquad \pi_{m\to b} \neq 0, \qquad \pi_{m\to c} \neq 0.$$

Consider first the coefficients of $\{\varepsilon_s : s \in \mathrm{Anc}(p)\}$ in X_{ab}. For each $s \in \mathrm{Anc}(p)$, the coefficient of ε_s in X_{ab} equals

$$\lambda_a \pi_{s\to a} + \lambda_b \pi_{s\to b}.$$

If there exists some $s \in \mathrm{Anc}(p)$ such that this coefficient is nonzero, then X_{ab} and X_c share the same non-Gaussian source ε_s with nonzero coefficients (since $\pi_{s\to c} \neq 0$), which implies $X_{ab} \not\perp\!\!\!\perp X_c$. This step invokes Theorem A.1, namely the standard Darmois–Skitovich type fact that two linear forms of independent non-Gaussian sources cannot be independent if they share at least one common source with nonzero coefficients.

Otherwise, suppose

$$\lambda_a \pi_{s\to a} + \lambda_b \pi_{s\to b} = 0 \qquad \text{for all } s \in \mathrm{Anc}(p).$$

Then necessarily $\lambda_b \neq 0$ (otherwise $\lambda_a = 0$ as well, contradicting $(\lambda_a, \lambda_b) \neq (0, 0)$). In this case, the coefficient of ε_m in X_{ab} equals $\lambda_b \pi_{m\to b} \neq 0$, while the coefficient of ε_m in X_c equals $\pi_{m\to c} \neq 0$. Hence X_{ab} and X_c share the non-Gaussian source ε_m with nonzero coefficients, and again $X_{ab} \not\perp\!\!\!\perp X_c$ by Theorem A.1.

In either case, there is no nonzero (λ_a, λ_b) such that $\lambda_a X_a + \lambda_b X_b \perp\!\!\!\perp X_c$. Therefore, when $m \neq r$, the condition in the theorem fails for the constructed triple (u, v, w) (taking $c = w$).

Combining the two directions concludes the proof. □

B Latent-Tree Distance Learning and Rooted Orientation Algorithm

In this section, we present our main algorithm 1. We introduce a lifted TBP formulation and optimize it with an alternating scheme that updates the distance parameters and auxiliary triplet weights in turn, exploiting the induced biconvex structure. Given an explicit upper bound on the number of latent variables, we further apply a structure-collapsing step to remove redundant degree-two internal nodes, enforcing the standard non-degeneracy requirement that each latent node has at least three neighbors. Finally, based on the recovered undirected latent tree, we leverage a three-terminal root characterization to select a root and orient edges, yielding a rooted, directed latent tree consistent with the underlying causal model.

Optimization viewpoint. The objective in equation 5 remains amenable to alternating minimization. For fixed d, both the TBP term and the observed-center penalty are linear in each α^{uvw}, so the problem decomposes across triplets into linear programs over simplices. For fixed α, $\mathcal{L}_O(\alpha)$ is constant, while the TBP and data-fitting terms are convex in d; hence the d-step reduces to a constrained least-squares problem.

Theorem B.1 (Biconvexity of the TBP loss)**.** *Let V be a finite node set, let $d = \{d_{ab}\}_{a,b\in V}$ denote the distance variables, and let $\mathcal{S} = \{(u, v, w) \in V^3 : u < v < w\}$ be the set of unordered node triplets. Define $r_m^{uvw}(d)$ by equation 6 and $\mathcal{L}_{\mathrm{TBP}}(d, \alpha)$ by equation 7, where each weight vector $\alpha^{uvw} = (\alpha_m^{uvw})_{m\in V}$ lies in the simplex*

$$\Delta_V := \Big\{\alpha \in \mathbb{R}_{\geq 0}^{|V|} : \sum_{m\in V} \alpha_m = 1\Big\}.$$

Then $\mathcal{L}_{\mathrm{TBP}}(d, \alpha)$ is biconvex in (d, α), i.e.,

- *for any fixed feasible α, the mapping $d \mapsto \mathcal{L}_{\mathrm{TBP}}(d, \alpha)$ is convex;*
- *for any fixed d, the mapping $\alpha \mapsto \mathcal{L}_{\mathrm{TBP}}(d, \alpha)$ is convex over $\prod_{(u,v,w)\in\mathcal{S}} \Delta_V$ (equivalently, it is convex in each block $\alpha^{uvw} \in \Delta_V$).*

Proof. **Convexity in d for fixed α.** Fix any feasible α, so $\alpha_m^{uvw} \geq 0$ for all $(u, v, w) \in \mathcal{S}$ and $m \in V$. For each (u, v, w, m), the three expressions

$$d_{uv} - d_{um} - d_{mv}, \qquad d_{uw} - d_{um} - d_{mw}, \qquad d_{vw} - d_{vm} - d_{mw}$$

are affine functions of d. Hence each squared term is convex in d, and their sum $r_m^{uvw}(d)$ is convex in d. Since $\mathcal{L}_{\mathrm{TBP}}(d, \alpha)$ is a nonnegative weighted sum of the convex functions $\{r_m^{uvw}(d)\}$, it follows that $d \mapsto \mathcal{L}_{\mathrm{TBP}}(d, \alpha)$ is convex.

Convexity in α for fixed d. Fix any d. For each triplet $(u, v, w) \in \mathcal{S}$, define the vector

$$r^{uvw}(d) := \big(r_m^{uvw}(d)\big)_{m\in V} \in \mathbb{R}^{|V|}.$$

Then the contribution of (u, v, w) to the loss can be written as

$$\sum_{m\in V} \alpha_m^{uvw}\, r_m^{uvw}(d) = \big\langle \alpha^{uvw},\, r^{uvw}(d)\big\rangle,$$

which is affine (indeed linear) in α^{uvw}. Therefore it is convex over the simplex Δ_V. Summing over all triplets preserves convexity, hence $\alpha \mapsto \mathcal{L}_{\mathrm{TBP}}(d, \alpha)$ is convex on $\prod_{(u,v,w)\in\mathcal{S}} \Delta_V$. This establishes biconvexity. □

Algorithm 1 Latent-Tree Causal Structural Learning with TBP Loss and Rooted Orientation

Require: Observed samples $X_O = \{x_O^{(i)}\}_{i=1}^n$; hyperparameters $\lambda_{\text{data}}, \lambda_O$; max iterations T; tolerance ε.

Ensure: Estimated distance matrix $\widehat{D} = (\widehat{d}_{uv})_{u,v\in V}$ and directed tree $\widehat{\mathcal{T}} = (V, \widehat{\vec{E}})$.

1: Compute empirical distances $\widehat{d}_{uv}$ for all $u, v \in O$ from X_O (e.g., $\widehat{d}_{uv} = -\log|\widehat{\rho}_{uv}|$).

2: Initialize $d^{(0)}$ by setting $d_{uv}^{(0)} = \widehat{d}_{uv}$ for $u, v \in O$ and random values for entries involving L within $(0, \max \widehat{d}_{uv})$, $u, v \in O$; enforce $d_{uv}^{(0)} = d_{vu}^{(0)}$ and $d_{uu}^{(0)} = 0$ for all $u, v \in V$.

3: **for** $t = 0 \to T-1$ **do**

4: $\quad$ **α-step (triplet-wise simplex update).**

5: $\quad$ **for** each $(u, v, w) \in \mathcal{S} = \{(u, v, w) \in V^3 : \ u < v < w\}$ **do**

6: $\quad\quad$ Update $\alpha^{uvw,(t+1)} \in \Delta_V$ by solving

$$\alpha^{uvw,(t+1)} \in \arg\min_{\alpha^{uvw}\in\Delta_V} \sum_{m\in V} \alpha_m^{uvw}\, r_m^{uvw}\big(d^{(t)}\big) + \lambda_O \sum_{m\in O} \alpha_m^{uvw},$$

where $r_m^{uvw}(d)$ is defined in equation 6 and $\Delta_V = \{\alpha \in \mathbb{R}_{\geq 0}^{|V|} : \sum_{m\in V} \alpha_m = 1\}$.

7: $\quad$ **end for**

8: $\quad$ **d-step (least-squares update).**

9: $\quad$ Update $d^{(t+1)}$ by solving

$$d^{(t+1)} \in \arg\min_{d} \ \mathcal{L}_{\text{TBP}}\big(d, \alpha^{(t+1)}\big) + \lambda_{\text{data}} \mathcal{L}_{\text{data}}(d),$$

subject to $d_{uv} = d_{vu}$, $d_{uu} = 0$, and $d_{uv} \geq 0$ for all $u, v \in V$. When the observed distance block is hard-anchored, additionally impose $d_{uv} = \widehat{d}_{uv}$ for all $u, v \in O$ and set $\lambda_{\text{data}} = 0$.

10: $\quad$ **if** $\mathcal{L}_{\text{TBP}} < \varepsilon$ **then**

11: $\quad\quad$ **break**

12: $\quad$ **end if**

13: **end for**

14: Set $\widehat{D} \leftarrow d^{(t^\star)}$ where $t^\star$ is the iterate that minimizes $\mathcal{L}_{\text{TBP}}$.

15: **Undirected topology recovery.** Construct $\widehat{\mathcal{T}}_{\text{und}} = (V, \widehat{E})$ as the MST of the complete graph on V weighted by $\widehat{D}$.

16: **Two-stage Redundant Latent Node Collapse.** collapse edges with excessively small d and L of degree 1 or 2; then update the $\widehat{D}'$ and $\widehat{\mathcal{T}}'_{\text{und}}$.

17: **Root selection.** For each candidate $s \in L$, remove s from $\widehat{\mathcal{T}}'_{\text{und}}$ to form terminal triples $(u, v, w) \in O^3$ from distinct components. For each such triple and each choice of $c \in \{u, v, w\}$ with $\{a, b\} = \{u, v, w\} \setminus \{c\}$, estimate (λ_a, λ_b) by enforcing $\text{Cov}(\lambda_a X_a + \lambda_b X_b,\ X_c) = 0$ and $\|(\lambda_a, \lambda_b)\|_2 = 1$, and compute an HSIC statistic between $\lambda_a X_a + \lambda_b X_b$ and X_c. Aggregate these statistics, e.g., by averaging them over the tested triplets and terminal choices, to obtain $\text{Score}(s)$, and set

$$\widehat{r} \leftarrow \arg\min_{s\in L} \text{Score}(s),$$

since smaller HSIC values indicate stronger evidence of independence.

18: **Orientation.** Orient all edges of $\widehat{\mathcal{T}}'_{\text{und}}$ away from $\widehat{r}$ via a BFS traversal to obtain $\widehat{\mathcal{T}} = (V, \widehat{\vec{E}})$.

19: **return** $\widehat{D}'$ and $\widehat{\mathcal{T}} = (V, \widehat{\vec{E}})$.

Choosing the number of latent nodes. Since the latent (internal) nodes are unobserved, their number is typically unknown. We therefore initialize the model with an over-parameterized latent budget $|L|_0 = |O| - 2$. This is justified by the universal upper bound below: when all observed nodes are leaves and all latent nodes are branching vertices with degree at least 3, the number of latent nodes cannot exceed $|O| - 2$. In subsequent optimization, redundant latent nodes can be naturally eliminated by edge contraction/merging, yielding an effective latent structure consistent with the data.

Theorem B.2. *Let $T = (V, E)$ be a tree whose vertex set is partitioned into the set of observed (leaf) vertices O and the set of latent (internal) vertices L. Assume that every observed vertex is a*

leaf and every latent vertex has degree at least 3 *. Then*

$$|L| \leq |O| - 2.$$

Proof. By the handshaking lemma,

$$\sum_{v \in V} \deg(v) = 2|E|.$$

Since T is a tree, $|E| = |V| - 1 = |O| + |L| - 1$, hence

$$\sum_{v \in V} \deg(v) = 2(|O| + |L| - 1).$$

On the other hand, using the degree assumptions,

$$\sum_{v \in V} \deg(v) = \sum_{o \in O} \deg(o) + \sum_{\ell \in L} \deg(\ell) \;\geq\; |O| \cdot 1 + |L| \cdot 3 \;=\; |O| + 3|L|.$$

Combining the two displays yields

$$2(|O| + |L| - 1) \;\geq\; |O| + 3|L|.$$

Rearranging gives $|O| - 2 \geq |L|$, i.e., $|L| \leq |O| - 2$. □

Root selection in practice. Given $\widehat{\mathcal{T}}_{\text{und}}$, removing a candidate node $s \in L$ splits the tree into connected components. We evaluate s using triples $(u, v, w) \in O^3$. For each such triple and each choice of $c \in \{u, v, w\}$, we estimate coefficients (λ_a, λ_b) by $\text{Cov}(\lambda_a X_a + \lambda_b X_b,\ X_c) = 0$ and normalizing $\|\lambda\|_2 = 1$. In the orientation stage, our method employs the Hilbert–Schmidt Independence Criterion (HSIC) (Gretton et al., 2005) for independence testing. Specifically, we compute an empirical HSIC statistic between $\lambda_a X_a + \lambda_b X_b$ and X_c. We aggregate these statistics over the tested triplets and terminal choices to obtain $\text{Score}(s)$, where smaller values indicate stronger evidence for the TTI condition. We then set

$$\widehat{r} := \arg\min_{s \in L} \text{Score}(s).$$

Finally, we orient $\widehat{\mathcal{T}}_{\text{und}}$ away from $\widehat{r}$ via a BFS traversal, producing the directed estimate $\widehat{\mathcal{T}} = (V, \widehat{\vec{E}})$.

C Synthetic Data and Wine Data

C.1 Experimental Details

Data Generation. For the latent tree structures shown in Figures 1, 3, 7, 6a, and 6b, we construct the corresponding weighted adjacency matrix $B \in \mathbb{R}^{d \times d}$ for each structure, where each element B_{ij} is independently sampled from a uniform distribution, and d denotes the total number of variables, including both observed and latent variables. Based on B, we generate data $\mathbf{X}$ for each latent tree model using the structural equation model $\mathbf{X} = B^T\mathbf{X} + \boldsymbol{\varepsilon}$, where the noise term $\boldsymbol{\varepsilon}$ is sampled from the Laplace distribution $Laplace(0, 1)$ and exponential distribution $Exp(1)$, respectively. For Figures 1 and 7, we generate datasets using an exponential noise model with a sample size of $n = 5000$. For the remaining three structures, we generate datasets under both noise types, each with sample sizes $n \in \{1000, 3000, 5000\}$.

Baseline details. LaHiCSL (Xie et al., 2024): the method introduces a Generalized Independent Noise (GIN) condition for linear non-Gaussian acyclic causal models with latent variables and develops an efficient algorithm that leverages this condition to estimate Linear, Non-Gaussian Latent Hierarchical Models (LiNGLaHs). CDHS (Li et al., 2025): this method introduces the concept of homologous surrogates to relax the pure children assumption in latent-variable causal discovery and develops an algorithm that leverages these surrogates to partially and, under stronger conditions, fully recover latent causal graphs from observational data. DCD (Prashant et al., 2025): this method proposes a differentiable causal discovery framework for nonlinear latent hierarchical causal models

that establishes new identifiability results under relaxed assumptions and learns causal structures via continuous optimization.

These baseline methods are developed under different structural assumptions and therefore do not necessarily apply to all latent-tree topologies considered in our simulations. For instance, DCD requires an equal-depth latent hierarchical structure, CDHS relies on the homologous surrogate assumption, and LaHiCSL targets LiNGLaH-type hierarchical latent structures satisfying the corresponding GIN-based conditions. As a result, each baseline is evaluated on the subset of structures for which its assumptions are satisfied. The applicable subsets are explicitly reported in the simulation results below.

Evaluation metrics. The *latent count exact recovery (LCER)* measures whether a method recovers the correct number of latent variables. It is defined as the proportion of experimental runs in which the estimated number of latent variables, denoted by N_{est}, exactly matches the ground-truth number of latent variables, denoted by N_{true}:

$$\text{LCER} = \frac{\#\{N_{\text{est}} = N_{\text{true}}\}}{N_{\text{exp}}}.$$

Here, $\#\{\cdot\}$ denotes the number of experimental runs satisfying the condition, and N_{exp} denotes the total number of experimental runs.

To evaluate skeleton recovery, we use the *exact skeleton recovery (SER)*, defined as the proportion of experimental runs in which the estimated undirected skeleton exactly matches the ground-truth skeleton:

$$\text{SER} = \frac{\#\{\widehat{G}_{\text{skel}} = G_{\text{skel}}\}}{N_{\text{exp}}}.$$

Here, $\widehat{G}_{\text{skel}}$ and G_{skel} denote the estimated and ground-truth undirected skeletons, respectively.

We further report the *structural Hamming distance (SHD)*, which counts discrepancies in edge presence, absence, or direction between the estimated graph and the ground truth. Since latent variables are inherently unlabeled, we compute SHD over all possible permutations of latent variables and take the minimum value, denoted by minSHD. We report the mean and standard deviation of minSHD over all experimental runs.

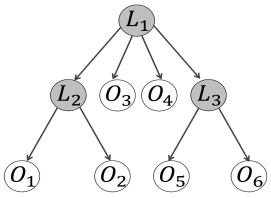
Hyperparameters. For synthetic experiments, we hard-anchor the observed–observed distance block by imposing $d_{uv} = \widehat{d}_{uv}$ for all $u, v \in O$ throughout optimization. Accordingly, we omit the soft data-fitting term ($\lambda_{\text{data}} = 0$). Because the synthetic data are generated exactly from LT-SCMs, this design isolates the latent-distance completion and structural-recovery stages: the empirical observed distances are treated as the input, while only distances involving latent nodes are optimized. For all structures except those in Figure 1, we generate 50 datasets using random seeds from 0 to 49; for the 190 structures shown in Figure 1, we generate 5 datasets per structure. For the DCD (Prashant et al., 2025) method, we set the number of epochs to 500 and the batch size to 128. For the Wine dataset, where exact LT-SCM specification is not assumed, we instead use the soft data-fitting formulation with $\lambda_{\text{data}} = 0.3$, allowing the observed distance block to adjust to the TBP structural constraints while remaining close to its empirical estimate. For all synthetic and real-data experiments, we set $\lambda_O = 1.0$. We use $\mathcal{L}_{\text{TBP}}$ as the early-stopping criterion with a mildly relaxed threshold $\tau_{\text{TBP}} = 0.05$: if $\mathcal{L}_{\text{TBP}} < \tau_{\text{TBP}}$, we stop early; otherwise, we select the information-distance matrix $\hat{D}$ that achieves the smallest $\mathcal{L}_{\text{TBP}}$. We perform contraction in two stages. **(i) Distance-based contraction.** During pruning, Assumption A2 implies that the correlation coefficient is bounded away from 1, and hence the corresponding information distance is bounded away from 0. Therefore, an extremely small latent–latent information distance suggests that two latent vertices may correspond to the same node rather than two distinct latent factors. We introduce a cutoff threshold τ_{cut} and contract two latent vertices u and v into a single vertex whenever $d_{uv} < \tau_{\text{cut}}$. This motivates the use of a small cutoff value in practice, such as 0.05. **(ii) Degree-based contraction.** We further contract redundant latent vertices based on degree information: in a tree, vertices of degree 1 or 2 are treated as redundant latent nodes and are contracted accordingly, with the relevant information distances updated after each contraction.

In the directional phase, we adopt an adaptive design and do not require hyperparameter tuning; when calculating the HSIC statistic, we employ the Gaussian kernel, a choice consistent with that used in the LaHiCSL codebase.

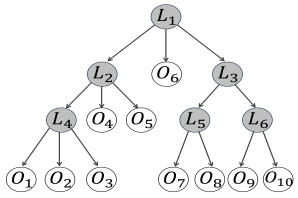
Hardware and environment. Computational Setup: All experiments in this paper, including those on synthetic simulation data, the White Wine Quality dataset and the MNIST dataset, were conducted on Linux workstations equipped with an NVIDIA RTX A6000 GPU (48 GB) and an Intel® Xeon® W-3465X processor (256 GB RAM).

C.2 Additional Experiments

We evaluate the proposed hierarchical latent variable modeling approach on three different latent-tree SCMs with exponential and Laplace noise. The performance of our method is compared to baseline models including LaHiCSL, CDHS, and DCD. Across all experiments, our method demonstrates consistent improvements in key causal discovery metrics.



(a) A 6-leaf latent-tree causal structure with 3 latent internal nodes.



(b) A 10-leaf latent-tree causal structure with 6 latent internal nodes.

Figure 6: Additional representative latent-tree structures used in the synthetic experiments. Observed variables are leaves, and latent variables are internal nodes.

Table 3: Causal discovery performance on a latent-tree SCM with exponential noise (Figure 6a).

Metric	Sample	LaHiCSL	CDHS	DCD	Ours
LCER↑	1k	0.54	0.14	-	**0.92**
	3k	0.88	0.66	-	**1.00**
	5k	0.86	0.84	-	**1.00**
SER↑	1k	0.54	0.14	-	**0.92**
	3k	0.88	0.66	-	**1.00**
	5k	0.90	0.84	-	**1.00**
SHD↓	1k	2.96 ± 3.38	7.40 ± 4.13	-	$\mathbf{0.40 \pm 1.37}$
	3k	0.70 ± 1.92	2.34 ± 3.58	-	**0**
	5k	0.92 ± 2.46	0.94 ± 2.18	-	**0**

Table 4: Causal discovery performance on a latent-tree SCM with exponential noise (Figure 6b).

Metric	Sample	LaHiCSL	CDHS	DCD	Ours
LCER↑	1k	0.10	-	-	**0.60**
	3k	0.12	-	-	**0.64**
	5k	0.06	-	-	**0.70**
SER↑	1k	0.04	-	-	**0.54**
	3k	0.08	-	-	**0.64**
	5k	0	-	-	**0.62**
SHD↓	1k	7.56 ± 2.44	-	-	$\mathbf{2.50 \pm 2.82}$
	3k	6.40 ± 2.14	-	-	$\mathbf{1.80 \pm 2.45}$
	5k	7.04 ± 0.83	-	-	$\mathbf{2.14 \pm 2.95}$

Table 3 reports causal discovery performance on the exponential-noise latent-tree SCM in Figure 6a. Across all sample sizes, our method consistently outperforms the applicable baselines (LaHiCSL

Table 5: Causal discovery performance on a latent-tree SCM with laplace noise (Figure 3).

Metric	Sample	LaHiCSL	CDHS	DCD	Ours
LCER↑	1k	0	-	**1.00**	0.94
	3k	0	-	**1.00**	**1.00**
	5k	0.02	-	**1.00**	**1.00**
SER↑	1k	0	-	0.50	**0.94**
	3k	0	-	0.82	**1.00**
	5k	0	-	0.90	**1.00**
SHD↓	1k	7.82 ± 2.46	-	2.20 ± 2.43	$\mathbf{1.60 \pm 1.21}$
	3k	6.70 ± 2.34	-	$\mathbf{0.72 \pm 1.55}$	0.88 ± 1.00
	5k	5.78 ± 1.53	-	$\mathbf{0.40 \pm 1.21}$	0.68 ± 0.96

Table 6: Causal discovery performance on a latent-tree SCM with laplace noise (Figure 6a).

Metric	Sample	LaHiCSL	CDHS	DCD	Ours
LCER↑	1k	0.28	0	-	**0.60**
	3k	0.44	0	-	**0.68**
	5k	0.34	0.02	-	**0.74**
SER↑	1k	0.28	0	-	**0.60**
	3k	0.40	0	-	**0.68**
	5k	0.30	0.02	-	**0.74**
SHD↓	1k	5.34 ± 3.88	14.06 ± 2.67	-	$\mathbf{2.12 \pm 2.42}$
	3k	4.24 ± 3.83	10.80 ± 2.21	-	$\mathbf{1.64 \pm 2.35}$
	5k	4.58 ± 3.41	10.04 ± 2.15	-	$\mathbf{1.30 \pm 2.22}$

Table 7: Causal discovery performance on a latent-tree SCM with laplace noise (Figure 6b).

Metric	Sample	LaHiCSL	CDHS	DCD	Ours
LCER↑	1k	0	-	-	**0.74**
	3k	0	-	-	**0.68**
	5k	0.06	-	-	**0.62**
SER↑	1k	0	-	-	**0.62**
	3k	0	-	-	**0.60**
	5k	0	-	-	**0.60**
SHD↓	1k	10.72 ± 2.89	-	-	$\mathbf{3.18 \pm 3.29}$
	3k	10.24 ± 2.63	-	-	$\mathbf{2.32 \pm 2.74}$
	5k	8.32 ± 2.60	-	-	$\mathbf{2.02 \pm 2.54}$

and CDHS) in terms of both exact skeleton recovery (LCER/SER) and SHD. In particular, our method achieves near-perfect recovery at $n = 1\text{k}$ (SER$= 0.92$) and reaches perfect skeleton recovery at $n \geq 3\text{k}$ (SER$= 1.00$, SHD$= 0$). We note that DCD (Prashant et al., 2025) is not applicable to this structure, and hence is omitted "-".

Table 4 reports causal discovery performance on the exponential-noise latent-tree SCM in Figure 6b. CDHS (Li et al., 2025) and DCD (Prashant et al., 2025) are not applicable to this structure (hence "-"). Across all sample sizes, our method consistently outperforms the applicable baseline (LaHiCSL) on all metrics, achieving substantially higher LCER/SER and markedly lower SHD. For example, LCER improves from 0.60 at $n = 1\text{k}$ to 0.70 at $n = 5\text{k}$ (vs. 0.10 to 0.06 for LaHiCSL), and SER increases from 0.54 to 0.62 (vs. 0.04 to 0). In addition, our SHD remains around 2 and is markedly below LaHiCSL across sample sizes (with the lowest SHD at $n = 3\text{k}$).

Table 5 reports causal discovery performance on the latent-tree SCM in Figure 3 under Laplace noise. Our method achieves substantially higher exact skeleton recovery than the baselines, with SER increasing from 0.94 at $n = 1\text{k}$ to 1.00 at $n \geq 3\text{k}$. Its SHD also decreases from 1.60 at $n = 1\text{k}$ to 0.68 at $n = 5\text{k}$. DCD attains slightly lower SHD at $n = 3\text{k}$ and $n = 5\text{k}$, but has lower SER across all three sample sizes, while LaHiCSL fails to recover the exact skeleton in these settings.

Table 6 presents results for another latent-tree SCM with Laplace noise (Figure 6a). Our method again shows superior performance, particularly as the sample size increases. For 1000 samples, our method achieves an LCER of 0.60, compared to LaHiCSL's 0.28. At 5000 samples, our method improves the LCER to 0.74, while LaHiCSL's performance remains relatively static at 0.34. The SHD for our method decreases from 2.12 at 1000 samples to 1.30 at 5000 samples, while LaHiCSL's SHD remains considerably higher.

In Table 7, which reports on causal discovery performance on a latent-tree SCM with Laplace noise (Figure 6b), our method excels once again. At 1000 samples, our method achieves an LCER of 0.74, outperforming LaHiCSL, which has an LCER of 0. At 5000 samples, our method maintains an LCER of 0.62, compared to LaHiCSL's 0.06. The SHD is markedly reduced, from 3.18 at 1000 samples to 2.02 at 5000 samples.

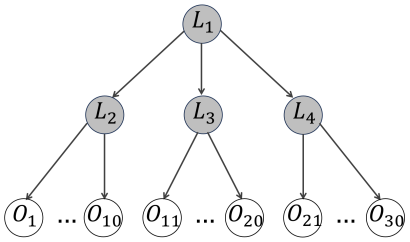


Figure 7: The 30-leaf LT-SCM used in the larger-scale synthetic experiment.

C.3 Additional Details on the White Wine Experiment

Interpretation of the estimated hierarchy. Figure 4 shows the directed latent tree estimated from the White Wine Quality dataset[1]. The branch centered at L_2 groups fixed acidity (O_1), citric acid (O_3), and pH (O_9), suggesting an acidity-related latent factor. The branch associated with L_3 groups chlorides (O_5), sulfur-dioxide measures (O_6, O_7), density (O_8), and alcohol (O_{11}), and may be interpreted as capturing a broader fermentation/flavor-related pattern. Since L_4 connects this branch with residual sugar (O_4), it may reflect variation linking residual sugar to the broader fermentation-related profile. The top-level node L_1 connects these major branches and can be viewed as summarizing higher-level physicochemical dependence. These semantic descriptions are exploratory interpretations of the estimated hierarchy rather than independently validated meanings of the latent variables.

Comparison with DCD. For comparison, DCD (Prashant et al., 2025) produces two disconnected trees on this dataset, as shown in Figure 8. LaHiCSL (Xie et al., 2024) and CDHS (Li et al., 2025) do not produce a comparable latent hierarchy under their respective model assumptions/settings.

[1] `https://archive.ics.uci.edu/dataset/186/wine+quality`

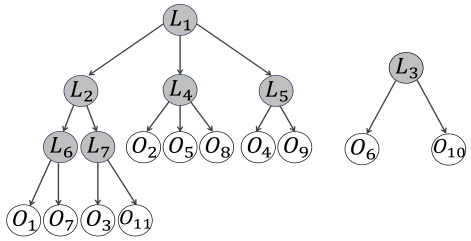


Figure 8: Structure estimated by DCD (Prashant et al., 2025) on the White Wine Quality dataset. The estimated graph consists of two disconnected trees.

D Additional Details on the MNIST Experiment

This section provides implementation details and additional results for the end-to-end VAE experiment in Section 6.4. The MNIST [2] experiment is used as a proof-of-concept benchmark to examine whether the proposed TBP-based structural module can be integrated into a neural generative model.

Architecture. We adopt a CausalVAE-style backbone (Yang et al., 2021). The model consists of three components: (i) a convolutional encoder $q_\phi(\epsilon \mid x)$ that outputs elementwise location and scale parameters, (ii) a linear SEM solver

$$z = (I - A)^{-1}\epsilon,$$

and (iii) a convolutional decoder $p_\theta(x \mid z_O)$ that takes only the observed block z_O as input. Here, A denotes the learned latent structural matrix.

Let $N = |O| + |L|$ denote the number of SEM nodes in the structural module, where $|O|$ is the dimension of the decoder input block and $|L|$ is the number of latent internal nodes introduced by the latent-tree module. Following the CausalVAE setting, which uses a structural module over 10 one-dimensional latent variables, we set $|O| = 10$ in our structural module. Only the $|O|$-dimensional block z_O is passed to the decoder $p_\theta(x \mid z_O)$. For consistency with the latent-tree notation, we refer to these coordinates as "observed nodes" within the structural module, although they are learned latent representations rather than directly observed image variables. The additional $|L|$ variables correspond to latent internal nodes used only for the tree-structured dependency module.

We use reparameterizable non-Gaussian noise for ϵ, implemented with a Laplace distribution, while CausalVAE uses Gaussian noise.

Given

$$z = (I - A)^{-1}\epsilon$$

and independent noise variables ϵ with diagonal covariance, the implied covariance matrix of z is

$$\Sigma_z(A) = (I - A)^{-1}\Sigma_\epsilon(I - A)^{-T}.$$

We obtain the corresponding correlation matrix $\rho(A)$ by standard normalization and then construct the information-distance surrogate $D(A)$ using Equation 2.

For each sampled observed triplet (u, v, w), we compute the TBP residuals $\{r_m^{uvw}\}_{m \in V}$ and define the soft branch-point assignment by

$$\alpha_m^{uvw} = \frac{\exp(-r_m^{uvw}/\tau)}{\sum_{m' \in V} \exp(-r_{m'}^{uvw}/\tau)}.$$

This makes $\mathcal{L}_{\mathrm{TBP}}(D(A), \alpha)$ differentiable with respect to the structural matrix A.

Training stability. We use several numerical stabilization techniques, including lower and upper bounds on the encoder scale parameter, clipping of encoder outputs, gradient clipping, and skip-and-retry for non-finite batches. We also use KL warm-up during the early training stage, while the TBP term is not warmed up.

[2] https://tensorflow.google.cn/datasets/catalog/mnist

Hyperparameters. We train for 50 epochs with batch size 128 on a single GPU using 4 dataloader workers. The loss weights are set to

$$\beta_{\mathrm{KL}} = 0.001, \qquad \lambda_{\mathrm{TBP}} = 0.01.$$

For numerical stability, the Laplace scale is clipped to

$$b \in [2 \times 10^{-3}, 3.0],$$

the encoder location and sampled noise are clipped with thresholds

$$\texttt{mu_clip} = 10, \qquad \texttt{eps_clip} = 20,$$

and gradient clipping is applied with max-norm 5.

Extracted structures. For visualization, we extract a directed graph from the learned structural matrix by retaining the strongest incoming edge for each node when applicable, and then simplify the resulting graph by contracting degree-one or degree-two internal nodes. The same extraction procedure is applied to both our method and CausalVAE.

Figure 9 shows the directed latent tree extracted from our method after 50 epochs. The learned structure provides a compact hierarchical summary of the latent dependencies encouraged by the TBP regularization. Before edge selection, the structural matrix learned by CausalVAE is relatively dense, with many entries having similar magnitudes. Under the same visualization heuristic, this dense pattern is reduced to a chain-like graph, as shown in Figure 10. This provides limited hierarchical information compared with the tree-consistent structure encouraged by our method.

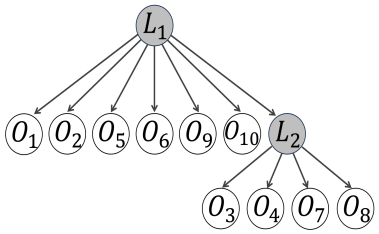


Figure 9: MNIST: directed latent tree extracted from the learned structure of our method after 50 epochs.

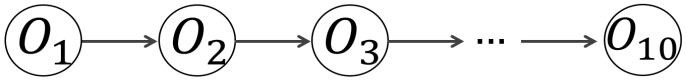
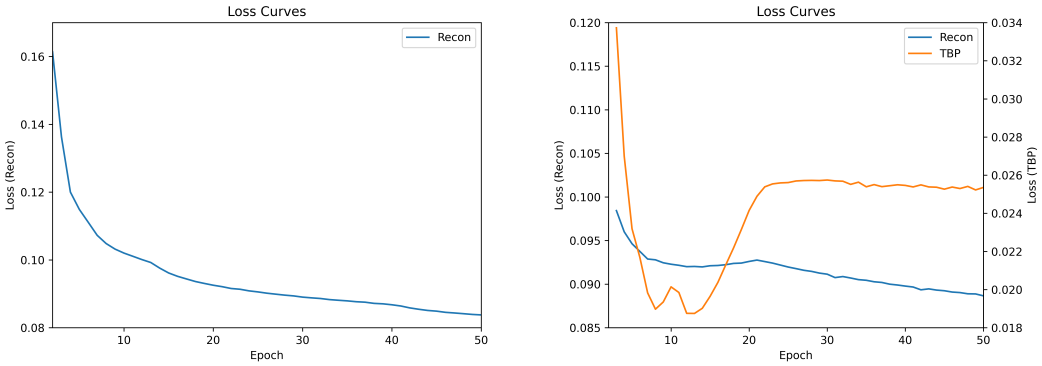


Figure 10: MNIST: graph structure extracted from CausalVAE using the same edge-selection procedure as in our visualization.

Loss curves. Figure 11 presents the training curves of CausalVAE and our method. At epoch 50, the reconstruction loss of CausalVAE is 0.08378, while that of our method is 0.08866. This small gap indicates that adding the TBP structural regularizer preserves comparable reconstruction quality while encouraging a more structured latent dependency pattern.

Reconstructions. Figure 12 shows the original images and the reconstructed images obtained by CausalVAE and our method at epoch 50. The reconstructions indicate that the proposed TBP-based structural regularization can be integrated into the VAE backbone while maintaining comparable reconstruction quality.

Remark D.1 (Interpretation of the MNIST experiment)**.** The MNIST experiment should be interpreted as an end-to-end integration and proof-of-concept demonstration rather than as a claim that the true latent generative structure of MNIST necessarily follows a tree. The purpose of this experiment is to show that the proposed TBP-based structural module can be incorporated into VAE- or CausalVAE-style generative models and can serve as a structured regularizer that encourages sparse,



(a) CausalVAE (b) Ours

Figure 11: MNIST: training curves for reconstruction and TBP losses.



(a) Original (b) CausalVAE (c) Ours

Figure 12: MNIST: original and reconstructed images at epoch 50.

hierarchical, and tree-consistent latent dependencies. From this perspective, the experiment demonstrates the feasibility of using the proposed differentiable structure-learning objective within a neural generative framework.

If the goal were only to impose an undirected structural bias, then an undirected tree or a generic sparsity regularizer might already be useful in some representation-learning settings. However, for intervention-aware or counterfactual generation, the orientation and identifiability of the latent structure become important, because different causal structures can imply different responses to interventions. Thus, the MNIST experiment is intended as a preliminary step toward integrating identifiable latent causal structure learning with generative models, rather than as a definitive validation of a tree-shaped causal mechanism in image data.

E Sensitivity to Near-Gaussian Noise

We further conduct sensitivity experiments to evaluate the behavior of the methods near the boundary of the non-Gaussianity assumption used for causal orientation. The near-Gaussian setting mainly tests the robustness of the TTI-based orientation step, because the skeleton recovery stage of our method relies on information distances and the TBP-based latent-tree completion objective rather than on non-Gaussianity. For the near-Gaussian noise setting, we use the structure shown in Figure 3 of the main text as an example. We generate the noise variables from Student's t-distributions with degrees of freedom $v = 5$ and $v = 20$, which approximate Gaussian noise to different extents. The results are reported in Tables 8 and 9.

Table 8: Causal discovery performance under near-Gaussian noise with t-distributed errors and degrees of freedom $v = 5$.

Metric	Sample	LaHiCSL	DCD	Ours
SER↑	1k	0	0.42	**1.00**
	3k	0	0.76	**1.00**
	5k	0	0.90	**1.00**
SHD↓	1k	6.30 ± 1.98	1.66 ± 1.65	$\mathbf{1.24 \pm 0.98}$
	3k	5.74 ± 2.43	0.98 ± 1.83	$\mathbf{0.36 \pm 0.78}$
	5k	6.02 ± 1.74	0.38 ± 1.16	$\mathbf{0.20 \pm 0.61}$

Table 9: Causal discovery performance under near-Gaussian noise with t-distributed errors and degrees of freedom $v = 20$.

Metric	Sample	LaHiCSL	DCD	Ours
SER↑	1k	0	0.24	**0.94**
	3k	0	0.76	**1.00**
	5k	0	0.74	**1.00**
SHD↓	1k	5.84 ± 1.45	2.42 ± 1.74	$\mathbf{1.70 \pm 1.34}$
	3k	5.52 ± 1.13	$\mathbf{1.00 \pm 1.82}$	1.44 ± 0.91
	5k	5.00 ± 0.70	$\mathbf{1.12 \pm 1.94}$	1.28 ± 0.97

Tables 8 and 9 show that our method remains effective under near-Gaussian noise. This is consistent with the design of our method: skeleton recovery relies on information distances and the TBP-based latent-tree completion objective, while non-Gaussianity is mainly used for TTI-based root identification and edge orientation. Consequently, as the noise distribution becomes closer to Gaussian, the main degradation is expected to appear in SHD rather than in SER. Indeed, even when $v = 20$, our method maintains high SER, while SHD increases compared with the $v = 5$ setting.