Robust Causal Discovery Under Imperfect Structural Constraints

Zidong Wang[1], **Xi Lin**[1], **Chuchao He**[2,*], **Xiaoguang Gao**[3]

[1] Department of Computer Science, City University of Hong Kong, Hong Kong, China
[2] School of Electronic Information Engineering, Xi'an Technological University, Xi'an, China
[3] School of Electronics And Information, Northwestern Polytechnical University, Xi'an, China
{zidowang, xilin4}@cityu.edu.hk, hechuchao@xatu.edu.cn, cxg2012@nwpu.edu.cn

Abstract

Robust causal discovery from observational data under imperfect prior knowledge remains a significant and largely unresolved challenge. Existing methods typically presuppose perfect priors or can only handle specific, pre-identified error types. And their performance degrades substantially when confronted with flawed constraints of unknown location and type. This decline arises because most of them rely on inflexible and biased thresholding strategies that may conflict with the data distribution. To overcome these limitations, we propose to harmonizes knowledge and data through prior alignment and conflict resolution. First, we assess the credibility of imperfect structural constraints through a surrogate model, which then guides a sparse penalization term measuring the loss between the learned and constrained adjacency matrices. We theoretically prove that, under ideal assumption, the knowledge-driven objective aligns with the data-driven objective. Furthermore, to resolve conflicts when this assumption is violated, we introduce a multi-task learning framework optimized via multi-gradient descent, jointly minimizing both objectives. Our proposed method is robust to both linear and nonlinear settings. Extensive experiments, conducted under diverse noise conditions and structural equation model types, demonstrate the effectiveness and efficiency of our method under imperfect structural constraints.

Code — https://github.com/wzd2502/RoaDs
Extended version — https://arxiv.org/abs/2511.06790

Introduction

Causal discovery from observational data is a cornerstone of artificial intelligence and scientific inquiry (Spirtes, Glymour, and Scheines 2000; Pearl 2009). By revealing the underlying causal mechanism and representing as a directed acyclic graph (DAG), it provides the fundamental structure required for downstream tasks such as causal inference (Hernán and Robins 2010; Peters, Janzing, and Schölkopf 2017), and causal representation learning (Schölkopf et al. 2021; Brehmer et al. 2022). A central topic in causal discovery is identifiability (Vowels, Camgöz, and Bowden 2023). Under the causal sufficiency and faithfulness assumptions (Koller and Friedman 2009), traditional combinatorial optimization methods can identify the structure up to its Markov Equivalence Class (MEC), which is also known as Bayesian network structure learning (Glymour, Zhang, and Spirtes 2019; Kitson et al. 2023). This full DAG-level identifiability can be achieved either by using interventional data or by imposing stricter assumptions on the data-generating process, such as non-Gaussian noise or nonlinear structural equation models (SEMs) (Vowels, Camgöz, and Bowden 2023). These stronger assumptions often enable the problem to be cast as a continuous optimization problem, making it solvable by zero-order (Shimizu et al. 2011), first-order (Zheng et al. 2018; Ng, Ghassami, and Zhang 2020), or second-order optimization methods (Rolland et al. 2022).

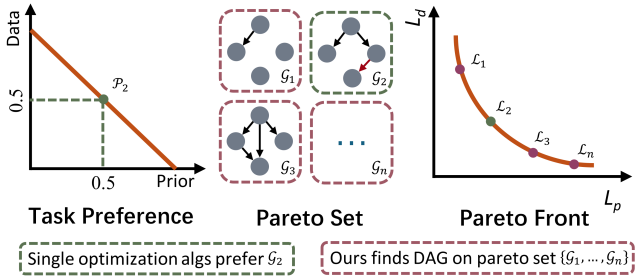

Figure 1: Robustness to imperfect constraints. A single-objective baseline assigns equal weight and is misled by the flawed prior (red arrow in $\mathcal{G}_2$), whereas ours identifies the conflict and discovers DAG on the Pareto set.

However, in numerous real-world applications, such as rare disease diagnosis or industrial fault analysis, high-quality observational data are often scarce and difficult to obtain. These domains typically possess a wealth of expert prior knowledge (e.g., positive or negative edge constraints) (Constantinou, Guo, and Kitson 2023; Brouillard et al. 2024). Consequently, how to effectively integrate such prior knowledge with data-driven methods has become an important yet challenging research direction.

Most existing methods are designed for perfect priors (no errors in constraints): combinatorial-based approaches typically treat priors as hard constraints, such as initializing the search or populating a tabu list (de Campos and

*Corresponding author.

Castellano 2007; Chen et al. 2025b; Wang, Gao, and Zhang 2025); continuous-based approaches incorporate priors as soft penalty terms or as hard optimization goals (Sun et al. 2023; Chen et al. 2025a). In practice, however, expert knowledge is often imperfect, potentially containing overlooked true causal edges or erroneously introduced spurious ones. When faced with such imperfect priors, the performance of existing methods degrades sharply. We explicate this issue from a multi-objective optimization perspective in Figure 1. Previous works typically use weighted sum scalarization to combine the data-driven and knowledge-driven objectives, which restricts the solution to a single, predetermined point on the Pareto front. Furthermore, these methods can neither adaptively correct erroneous priors nor adjust the weight of the knowledge-based objective to reflect its credibility. When priors are unreliable, a fixed, high weight inevitably forces the model to overfit to this incorrect DAG, such as $\mathcal{G}_2$ in Pareto set.

To tackle this dilemma, we build upon continuous optimization methods to develop a robust framework capable of handling imperfect structural constraints. Our approach achieves this through two components: **Prior Alignment**, which employs a surrogate model dynamically modulating the weights of imperfect constraints based on the observational data; **Conflict Resolution**, which leverages multi-task learning (MTL) to explicitly manage the trade-off between the data-driven and knowledge-driven objectives. We named it as Robust Causal Discovery under Imperfect structural constraints (**RoaDs**). Our main contributions are as follows:

- We introduce a consistent constraint assumption and use a surrogate model to learn continuous weights for priors.
- We design the knowledge-based optimization goal based on consistent constraints, and theoretically prove the asymptotic consistency of it.
- We employ Multi Gradient Descent Algorithm (MGDA), enhanced with gradient normalization, to efficiently find a balanced Pareto stationary point for MTL problem.
- In experimental evaluation, we demonstrate the superior robustness and effectiveness of RoaDs against SOTA methods across diverse and challenging settings.

Related Works

Causal discovery under structural constraints For combinatorial-based methods, integrating edge constraints is relatively straightforward, typically by restricting the search space (de Campos and Castellano 2007; Colombo and Maathuis 2014; Constantinou, Guo, and Kitson 2023). However, path constraints, which are weaker and non-decomposable, need the graphical search space or specialized data structures to entail (Chen et al. 2016; Wang et al. 2021, 2025). A key limitation of these approaches is their reliance on the assumption that all provided constraints are perfect and error-free. For continuous-based approaches, perfect edge constraints are often handled in two ways: either enforced as hard constraints that are optimized simultaneously with the acyclicity constraint (Hasan and Gani 2022; Sun et al. 2023; Wang et al. 2024), or by directly modifying the gradients of the adjacency matrix to steer the search (Bello, Aragam, and Ravikumar 2022). Imperfect priors are typically handled via soft penalties, where constraints are formulated as differentiable terms, such as a cross-entropy loss measuring constraint violation (Li et al. 2024; Chen et al. 2025a). To handle path constraints, this paradigm involves employing partial order-based optimization strategies (Ban et al. 2025c).

More recently, a nascent line of work has explored using Large Language Models (LLMs) as a proxy for domain experts (Kiciman et al. 2024). LLMs have been used to generate initial graphs (Ban et al. 2025b), suggest post-hoc adjustments (Khatibi et al. 2024), or fuse structural priors from text (Zhou et al. 2024; Ban et al. 2025a). For a broader survey of general causal discovery, we refer the reader to Appendix A.

Multi-task Learning MTL is quite a hot topic in the machine learning community (Zhang and Yang 2022). It improves the generalization and reduce the cost of learned models (Zhao and Gordon 2022). Key research in MTL involves designing shared architectures and managing conflicting task objectives (Lin and Zhang 2023). Our work concentrates on the latter, employing multi-objective optimization (MOO) to mitigate the conflict between data-driven and knowledge-driven objectives for causal discovery.

MOO solvers can be broadly categorized into two families (Zhang et al. 2024). The first, aggregation-based methods, transforms the multi-objective problem into a single-objective one by aggregating individual loss functions, such as Linear scalarization (Miettinen 1998), the Tchebycheff method (Zhang and Li 2007), Smooth TCH (Lin et al. 2024). The second family, gradient-manipulation-based methods, operates directly on the gradients of each task to find a descent direction that improves all objectives. Prominent examples include the MGDA (Sener and Koltun 2018), its preference-based extensions (Lin et al. 2019), and normalization version (Chen et al. 2018).

Preliminary

Causal discovery

A causal structure can be represented by a DAG $\mathcal{G} = (\boldsymbol{V}, \boldsymbol{E})$, where $\boldsymbol{V} = \{X_1, \dots, X_{n_v}\}$ is a set of variables and $\boldsymbol{E}$ is the set of edges. An edge $X_i \to X_j$ implies that X_i is a direct cause (parent) of X_j (Koller and Friedman 2009), denoted as $X_i \in \Pi_j^{\mathcal{G}}$. We consider the Additive Noise Model (ANM) (Hoyer et al. 2008), where each variable is generated by a function of its parents plus an independent noise term $X_j = f_j(\Pi_j^{\mathcal{G}}) + \epsilon_j$. Here, f_j is a causal mechanism, and the noise terms $\boldsymbol{\epsilon} = \{\epsilon_1, \dots, \epsilon_{n_v}\}$ are assumed to be mutually independent with zero mean ($\mathbb{E}[\epsilon_j] = 0$) and covariance matrix $\mathrm{diag}(\sigma_1, \dots, \sigma_{n_v})$. Given an i.i.d. dataset $\mathbf{X} = [\mathbf{x}_1 | \dots | \mathbf{x}_{n_v}] \in \mathbb{R}^{n_d \times n_v}$, the goal of causal discovery is to find the optimal DAG $\mathcal{G}$ by solving a continuous optimization problem:

$$\begin{aligned} &\min_{\mathbf{f}} \textstyle\sum_{j=1}^{n_v} \mathcal{L}(\mathbf{x}_j, f_j(\mathbf{X})) \\ &s.t.\ \mathcal{G}(\mathbf{f}) \text{ is acyclic,} \end{aligned} \tag{1}$$

where $\mathcal{L}(\cdot)$ is a least squares loss or negative log-likelihood loss, and $\mathcal{G}(\mathbf{f})$ is the DAG induced by the functional depen-

dencies in $\mathbf{f} = \{f_1, \ldots, f_{n_v}\}$. Each f_j can be parameterized using a Multilayer Perceptron: $f_j(\mathbf{X}) = \text{MLP}(\mathbf{X}; \theta_j)$, where $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_{n_v}\}$. $\theta_j = \{A_j^{(k)}\}_{k=1}^{n_h}$ are the parameters for the j-th MLP, and $A_j^{(k)} \in \mathbb{R}^{d_{k-1} \times d_k}$ denotes the weights of the k-th layer (Lachapelle et al. 2020; Zheng et al. 2020). Under such condition, the weighted adjacency matrix can be approximately expressed as $W(\boldsymbol{\theta}) \in \mathbb{R}^{n_v \times n_v}$. The entry $[W(\boldsymbol{\theta})]_{ij}$ quantifies the causal influence from X_i to X_j and is defined as $[W(\boldsymbol{\theta})]_{ij} = \|[A_j^{(1)}]_{:,i}\|_2$, Consequently, the optimization problem from Eq. (1) is reformulated as:

$$\begin{aligned} &\min_{\boldsymbol{\theta}} \ \frac{1}{n_d} \sum_{j=1}^{n_v} \|\mathbf{x}_j - \text{MLP}(\mathbf{X}; \theta_j)\|_F^2 + \lambda_1 \|W(\boldsymbol{\theta})\|_1 \\ &s.t.\ h(W(\boldsymbol{\theta})) = \text{tr}(e^{W(\boldsymbol{\theta}) \circ W(\boldsymbol{\theta})}) - n_v = 0. \end{aligned} \tag{2}$$

Problem (2) can be transformed into unconstrained optimization form using Augmented Lagrangian Method (ALM) (Zheng et al. 2018). For brevity, we will henceforth denote the original objective function as $\mathcal{F}_{\mathbf{X}}(\boldsymbol{\theta})$, respectively, and the constrained objective function as $\mathcal{H}(W(\boldsymbol{\theta}))$

$$\mathcal{H}(W(\boldsymbol{\theta})) = \varphi h(W(\boldsymbol{\theta})) + \frac{\rho}{2} |h(W(\boldsymbol{\theta}))|^2, \tag{3}$$

where φ and ρ are parameters in ALM.

Multi-task learning

A MTL problem can be formulated as a multi-objective optimization problem, where the goal is to simultaneously minimize a vector of loss functions corresponding to different tasks (Caruana 1993; Miettinen 1998):

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathbf{L}(\boldsymbol{\theta}) = (\mathcal{L}_1(\boldsymbol{\theta}), \ldots, \mathcal{L}_{n_p}(\boldsymbol{\theta}))^T, \tag{4}$$

A solution $\boldsymbol{\theta}_a$ is said to dominate $\boldsymbol{\theta}_b$, denoted as $\mathbf{L}(\boldsymbol{\theta}_a) \prec \mathbf{L}(\boldsymbol{\theta}_b)$, if $\mathcal{L}_k(\boldsymbol{\theta}_a) \leq \mathcal{L}_k(\boldsymbol{\theta}_b)$ holds $\forall k \in \{1, \ldots, n_p\}$, and there exists at least one index j for which $\mathcal{L}_j(\boldsymbol{\theta}_a) < \mathcal{L}_j(\boldsymbol{\theta}_b)$.

Definition 1. *(Pareto Optimality) A solution $\boldsymbol{\theta}^* \in \Theta$ is Pareto optimal if no other solution $\boldsymbol{\theta} \in \Theta$ dominates it, i.e., there is no $\boldsymbol{\theta}$ such that $\mathbf{L}(\boldsymbol{\theta}) \prec \mathbf{L}(\boldsymbol{\theta}^*)$.*

For MTL with conflicting objectives, there not exists a single solution that minimizes all task losses simultaneously. Instead, a set of trade-off solutions exists. The set of all Pareto optimal solutions is called the *Pareto set*, and its image in the objective space is the *Pareto front*.

Framework

This paper focuses on causal discovery where the available prior knowledge may conflict with the ground-truth graph. And such knowledge can be formally defined as follow.

Definition 2. *(Imperfect constraints.) Let the constraints be encoded in a matrix $\mathbf{B}^c \in \{0, 1, -1\}^{n_v \times n_v}$, where $\mathbf{B}^c_{ij} = 1, -1, 0$ signifies positive constraint ($X_i \to X_j$), negative constraint ($X_i \not\to X_j$), and no constraint. Let $\mathbf{B}^*$ be the adjacency matrix of the ground-truth graph. $\mathbf{B}^c$ is considered **imperfect** if there exist entries (i, j) such that $\mathbf{B}^c_{ij} = 1$ but $\mathbf{B}^*_{ij} = 0$, or $\mathbf{B}^c_{ij} = -1$ but $\mathbf{B}^*_{ij} = 1$.*

The propose RoaDs refines the imperfect constraints by aligning them with the observational data to against the terrible influence from flawed priors, and resolves the remain conflict between the data-driven and knowledge-driven objectives using a MOO solver, as illustrated in Figure 2.

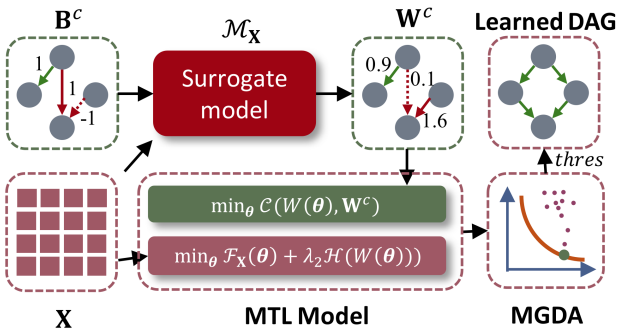

Figure 2: Pipeline. RoaDs constructs a data-driven objective from a continuous score and a knowledge-driven objective using a surrogate model to align imperfect constraints (Red arrows in figure). These are formulated as a MTL problem, which is then solved via the MGDA to recover the final causal graph.

Prior alignment

The reliability of the prior alignment is fundamentally compromised by the highly non-convex optimization landscape of continuous-based methods. We therefore lay the foundation for RoaDs by first defining a theoretical criterion that acts as the *tool* for overriding flawed priors and simultaneously establishes the *bounds* of its valid application. Note that the subsequent analysis still holds under causal faithfulness and sufficiency assumption.

Tool for alignment. To enable a uniform re-evaluation of all priors, the negative constraints are firstly converted into positive constraints. Then, a surrogate model $\boldsymbol{\mathcal{M}} = \{\mathcal{M}^{(1)}, \ldots, \mathcal{M}^{(n_v)}\}$ is employed to test the credibility of different constraints under the dataset $\mathbf{X}$. For arbitrary X_j, $\mathcal{M}^{(j)}$ solves $\mathbb{E}[X_j | \Pi_j^{\mathbf{B}}]$ to find the weights of edges point to X_j, which is defined as

$$\mathbf{W}_{:,j} = \mathcal{M}_{\mathbf{X}}^{(j)}([\mathbf{B}_{:,j}]_{\neq 0}). \tag{5}$$

Thus, when $n_d \to \infty$, the ground-truth DAG satisfies

$$\mathbf{W}^*_{:,j} = \mathcal{M}_{\mathbf{X}}^{(j)}([\mathbf{B}^*_{:,j}]_{\neq 0}) = \mathcal{M}_{\mathbf{X}}^{(j)}(\mathbf{1}_{:,j}). \tag{6}$$

For linear case, the surrogate model can be achieved by the consistent parametric regressor, where $\mathbf{W}_{ij}$ can be represented by the regression coefficients from $\mathbf{x}_i$ to $\mathbf{x}_j$. In nonlinear settings, the consistent non-parametric regressor (e,g, random forest) is feasible, and $\mathbf{W}_{ij}$ can be represented by permutation importance (Hastie, Tibshirani, and Friedman 2009).

Bounds of alignment. We introduce a strict assumption about the dependency relations in constraints matrix, that determines whether flawed prior can be aligned.

Assumption 1. *(Consistent constraints.) For the ground-truth DAG $\mathbf{B}^*$, constraints matrix $\mathbf{B}^c$ is consistent if it satisfies $\forall X_i \in \boldsymbol{V}_j^{1,0}, X_i \perp\!\!\!\perp X_j | \boldsymbol{V}_j^{1,1}$ and $\forall X_i \in \boldsymbol{V}_j^{0,1}, X_k \in \boldsymbol{V}_j^{1,1}, X_i \perp\!\!\!\perp X_k$, where $\boldsymbol{V}_j^{\alpha,\beta} = \{X_k \mid \mathbf{B}^c_{kj} = \alpha, \mathbf{B}^*_{kj} = \beta\}$, $\alpha, \beta \in \{0, 1\}$.*

Theorem 1. *If $\mathbf{B}^c$ is consistent, there always exists $\tau > 0$ such that the probability limit of $\mathbf{W}^c$ from Eq.* (5) *satisfies:*

1. $\forall X_i \in \boldsymbol{V}_j^{1,0}$, *then* $plim_{n_d \to \infty} \mathbf{W}_{ij}^c < \tau$.
2. $\forall X_i \in \boldsymbol{V}_j^{1,1}$, *then* $plim_{n_d \to \infty} \mathbf{W}_{ij}^c > \tau$.

The detailed proof is provided in Appendix B. Theorem 1 demonstrates that if the constraint matrix $\mathbf{B}^c$ is consistent, the surrogate model $\boldsymbol{\mathcal{M}}$ successfully recovers true edges while simultaneously rejecting the false positive edges that were incorrectly specified in the $\mathbf{B}^c$. The resulting weight matrix $\mathbf{W}^c$ from prior alignment will accurately reflect the *partial* ground-truth structure $\mathbf{B}^*$. And the following discuss in this section is all based on consistent $\mathbf{B}^c$.

Knowledge-driven optimization objective. After prior alignment, $\mathbf{W}^c = \mathcal{M}_{\mathbf{X}}([\mathbf{B}^c]_{\neq 0})$ can serve for the modeling of knowledge-driven optimization objective. This objective aims to promote the non-parametric weighted adjacency matrix $W(\boldsymbol{\theta})$ towards to the refined DAG encoded in $\mathbf{W}^c$. Intuitively, this purpose can be achieved by minimizing the $\ell 1$ norm of the difference between their binarized structures, an objective that exclusively evaluates discrepancies at the locations specified by the original constraint mask $\mathbf{B}^c$

$$\min_{\boldsymbol{\theta}} \left\| [\mathbb{I}(W(\boldsymbol{\theta}) - s > 0) - \mathbb{I}(\mathbf{W}^c - \tau > 0)] \circ \mathbf{B}^c \right\|_1. \quad (7)$$

$\mathbb{I}(\cdot)$ denotes the Heaviside step function, which maps its input to $\{0, 1\}$ based on the specified thresholds s and τ. The $\circ \mathbf{B}^c$ localizes the penalty to the constrained entries. However, the discontinuous nature of $\mathbb{I}(\cdot)$ renders this objective non-differentiable and thus unamenable to standard gradient-based optimization methods.

To facilitate tractable optimization, we introduce a subdifferentiable form of Eq. (7) by substituting the $\mathbb{I}(\cdot)$ with a continuous sigmoid function $\sigma(\cdot)$, which acts as a smooth approximation. This yields the following objective

$$\min_{\boldsymbol{\theta}} \left\| [\sigma(W(\boldsymbol{\theta}) - s) - \sigma(\mathbf{W}^c - \tau)] \circ \mathbf{B}^c \right\|_1, \quad (8)$$

and we denote it as $\mathcal{C}(W(\boldsymbol{\theta}), \mathbf{W}^c)$. For the linear case, Eq. (8) reduces to a more concise form where $s = \tau$ and the sigmoid function $\sigma(\cdot)$ is omitted in favor of a parametric regressor. We can theoretically show that this formulation achieves a lower error bound than fixed thresholding methods, more detailed is provided in Appendix C.

Asymptotic consistency. The following theorem establishes that under a single optimization architecture, which integrates the data-driven optimization objective $\mathcal{F}_{\mathbf{X}}(\boldsymbol{\theta}) + \lambda_2 \mathcal{H}(W(\boldsymbol{\theta}))$, and our knowledge-regularization term $\mathcal{C}$, is asymptotically consistent.

Theorem 2. *Consider the continuous optimization problem defined as:*

$$\min_{\boldsymbol{\theta}} \mathcal{F}_{\mathbf{X}}(\boldsymbol{\theta}) + \lambda_2 \mathcal{H}(W(\boldsymbol{\theta})) + \lambda_3 \mathcal{C}(W(\boldsymbol{\theta}), \mathbf{W}^c). \quad (9)$$

Let $\hat{\boldsymbol{\theta}}$ be the optimal solution to the above problem. As the number of samples $n_d \to \infty$, the graph structure induced by $W(\hat{\boldsymbol{\theta}})$ converges in probability to the ground-truth DAG $\mathbf{B}^*$

$$\mathbb{I}(W(\hat{\boldsymbol{\theta}}) > s) \xrightarrow{p} \mathbf{B}^*. \quad (10)$$

The detailed proof is provided in Appendix B.

Dilemma under non-consistent constraints. According to Theorem 2, if the imperfect constraints $\mathbf{B}^c$ are consistent, the knowledge-driven objective aligns with the data-driven objective in large-sample settings. However, a significant gap exists between this asymptotic ideal and practical application. First, verifying the consistency of given constraints is often intractable, as it would require a relatively accurate understanding of the ground-truth structure $\mathbf{B}^*$. Second, the introduction of prior knowledge is to improve the learning accuracy under small sample size, where theoretical guarantees are weakest.

Consequently, the data-driven term $\mathcal{F}_{\mathbf{X}}(\boldsymbol{\theta}) + \lambda_2 \mathcal{H}(W(\boldsymbol{\theta}))$ and the knowledge-regularization term $\mathcal{C}(W(\boldsymbol{\theta}), \mathbf{W}^c)$ often remain in conflict, further contributing to a highly nonconvex optimization landscape (Reisach, Seiler, and Weichwald 2021; Ng, Huang, and Zhang 2024). This inherent tension necessitates a more sophisticated mechanism to mediate between data and imperfect constraints.

Conflict resolution

We propose a MTL framework designed to balance these two conflicting objectives. Formally, the two optimization tasks are defined as

$$\begin{cases} \min_{\boldsymbol{\theta}} \mathcal{F}_{\mathbf{X}}(\boldsymbol{\theta}) + \lambda_2 \mathcal{H}(W(\boldsymbol{\theta})) \\ \min_{\boldsymbol{\theta}} \mathcal{C}(W(\boldsymbol{\theta}), \mathbf{W}^c). \end{cases} \quad (11)$$

Here we assign the equal preference to both tasks, thus, the parameter λ_3 for the second task is omitted.

Solve the MTL problem. We employ the MGDA to solve MOO problem in Eq. (11) (Sener and Koltun 2018), as it efficiently identifies a single Pareto-stationary point, instead of the entire Pareto front, which is not friendly to decision-makers. Another advantage is that it can adaptively adjust the weights of the two optimization goals, which is crucial for navigating the conflict between data-driven evidence and imperfect constraints. $\boldsymbol{\theta}$ is updated according to

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \eta \boldsymbol{d}_t, \quad (12)$$

where η is the learning rate, and $\boldsymbol{d}_t$ is defined from

$$\begin{aligned} &(\boldsymbol{d}_t, \kappa_t) = \operatorname*{argmin}_{\boldsymbol{d}, \kappa} \kappa + \tfrac{1}{2} \|\boldsymbol{d}\|_2^2 \\ &s.t.\ \Phi_\alpha(\boldsymbol{\theta}_t, \mathbf{X}) = \nabla [\mathcal{F}_{\mathbf{X}}(\boldsymbol{\theta}_t) + \lambda_2 \mathcal{H}(W(\boldsymbol{\theta}_t))]^\top \boldsymbol{d}^{(1)} \leq \kappa \\ &\quad \Phi_\beta(\boldsymbol{\theta}_t, \mathbf{W}^c) = \nabla \mathcal{C}(W(\boldsymbol{\theta}_t), \mathbf{W}^c)^\top \boldsymbol{d}^{(2)} \leq \kappa, \end{aligned} \quad (13)$$

where $\boldsymbol{d}^{(k)}$ denotes the gradient direction of k-th task, and $\kappa \in \mathbb{R}$ is a scalar that indicates the convergence status across all tasks. Furthermore, the following proposition holds (Fliege and Svaiter 2000)

Corollary 1. *If $\boldsymbol{\theta}_t$ is Pareto optimal, then it is a stationary point where $\boldsymbol{d}_t = \mathbf{0}$ and $\kappa_t = 0$. If $\boldsymbol{\theta}_t$ is not Pareto optimal, then $\boldsymbol{d}_t$ is a valid descent direction, and κ_t is strictly negative, satisfying*

$$\begin{aligned} &\kappa_t \leq -\tfrac{1}{2} \|\boldsymbol{d}_t\|_2^2 \leq 0 \\ &\Phi_\alpha(\boldsymbol{\theta}_t, \mathbf{X}) \leq \kappa_t,\ \Phi_\beta(\boldsymbol{\theta}_t, \mathbf{W}^c) \leq \kappa_t. \end{aligned} \quad (14)$$

Corollary 1 clarifies that when $\boldsymbol{d}_t = \mathbf{0}$, the data-driven and knowledge-driven objectives cannot be improved simultaneously. Conversely, if $\boldsymbol{\theta}_t$ is not optimal, non-zero $\boldsymbol{d}_t$ guarantees that a direction exists to concurrently improve both objectives. According to KKT condition, it satisfies

$$\begin{aligned} &\boldsymbol{d}_t = -\lambda_\alpha \Phi_\alpha(\boldsymbol{\theta}_t, \mathbf{X}) - \lambda_\beta \Phi_\beta(\boldsymbol{\theta}_t, \mathbf{W}^c) \\ &s.t. \quad \lambda_\alpha + \lambda_\beta = 1. \end{aligned} \tag{15}$$

The dual problem of Eq. (15) is

$$\min_{\lambda_\alpha} -\tfrac{1}{2}\|\lambda_\alpha \Phi_\alpha(\boldsymbol{\theta}_t, \mathbf{X}) + (1-\lambda_\alpha)\Phi_\beta(\boldsymbol{\theta}_t, \mathbf{W}^c)\|_2^2. \tag{16}$$

The quadratic program (QP) presented in Eq. (16) is equivalent to find the minimum-norm vector in the convex hull of the task gradients. And its solution satisfies (for notational simplicity, we omit the variables in $\Phi(\cdot)$)(Lin et al. 2019):

$$\lambda_\alpha = \begin{cases} 1 & \Phi_\alpha^\top \Phi_\beta \geq \Phi_\alpha^\top \Phi_\alpha \\ 0 & \Phi_\alpha^\top \Phi_\beta \geq \Phi_\beta^\top \Phi_\beta \\ \frac{(\Phi_\beta - \Phi_\alpha)^\top \Phi_\beta}{\|\Phi_\alpha - \Phi_\beta\|_2^2} & \text{otherwise.} \end{cases} \tag{17}$$

Normalization method. Data-driven and knowledge-driven objectives have disparate scales, and the latter requires only sparse parameter modifications and is thus easier to optimize. This imbalance biases QP solution towards neglecting the data-driven task $\lambda_\alpha \approx 0$. To ensure both objectives contribute meaningfully, we normalize the gradients in the following ways

$$\begin{aligned} \Phi_\alpha &= \Phi_\alpha \cdot [(\mathcal{F}_{\mathbf{X}}(\boldsymbol{\theta}_t) + \lambda_2 \mathcal{H}(W(\boldsymbol{\theta}_t))) \cdot \|\Phi_\alpha\|_2]^{-1} \\ \Phi_\beta &= \Phi_\beta \cdot [\mathcal{C}(W(\boldsymbol{\theta}_t), \mathbf{W}^c) \cdot \|\Phi_\beta\|_2]^{-1}. \end{aligned} \tag{18}$$

We discuss other normalization methods in Appendix D.

Overall algorithm. Alg. 1 details the RoaDs. It performs a warm-up stage (lines 2-4), using only the data-driven objective for t_s iterations to find an initial solution. Consistent with the mainstream continuous optimization for causal discovery (Yu et al. 2019; Fang et al. 2024), the main loop uses the Adam optimizer and adjusts the parameters of the acyclicity constraint to accelerate convergence (lines 9-11). We analyze time complexity of Alg. 1 in Appendix E.

Algorithm 1: RoaDs

Input: Dataset $\mathbf{X}$, Imperfect priors $\mathbf{B}^c$.

Output: Optimal weighted matrix $\hat{\mathbf{W}}$.

1: Align the priors as $\mathbf{W}^c = \mathcal{M}_{\mathbf{X}}([\mathbf{B}^c]_{\neq 0})$, set $\boldsymbol{\theta}_0 = \mathbf{0}$
2: **while** $t \leq t_s$ **do**
3: $\quad \boldsymbol{\theta_{t+1}} = \boldsymbol{\theta}_t + \eta(\Phi_\alpha(\boldsymbol{\theta}_t, \mathbf{X}))$
4: **end while**
5: **while** $t > t_s$ and $h(W(\boldsymbol{\theta}_t)) \neq 0$ **do**
6: $\quad$ Normalize $\Phi_\alpha(\boldsymbol{\theta}_t, \mathbf{X}), \Phi_\beta(\boldsymbol{\theta}_t, \mathbf{W}^c)$
7: $\quad$ Compute λ_α and $\boldsymbol{d}_t$ according to Eq. (15) and (17)
8: $\quad \boldsymbol{\theta_{t+1}} = \boldsymbol{\theta}_t + \eta \boldsymbol{d}_t$
9: $\quad$ **if** $h(W(\boldsymbol{\theta}_t)) > c \cdot h(W(\boldsymbol{\theta}_{t-1}))$ **then**
10: $\quad\quad$ Update the parameters in $\mathcal{H}(W(\boldsymbol{\theta}_t))$
11: $\quad$ **end if**
12: **end while**
13: **return** the weighted matrix $\hat{\mathbf{W}} = W(\boldsymbol{\theta})$.

Experiment

Experimental settings

Graphs and datasets. We generate synthetic graphs using Erdős–Rényi (ER) and Scale-Free (SF). Each graph consists of n_v nodes and kn_v edges, denoted as ER-k or SF-k. n_d data is then generated based on SEM defined on these graphs. For linear conditions, the weighted adjacency matrix is sampled randomly from $(-2.0, -0.5] \cup [0.5, 2.0)$. Exogenous noise variables are drawn from Gaussian, Exponential, Gumbel, and Uniform, with settings for both equal variance (EV) and non-equal variance (NV) (Ng, Huang, and Zhang 2024). For nonlinear settings, we generate data using either MLP or Gaussian Processes (GP).

Imperfect constraints usage. We sample $p_a \cdot kn_v$ true edges from the ground-truth graph as positive constraints and $p_c \cdot p_a \cdot kn_v$ non-existent edges as negative constraints. Then, we randomly select a fraction p_b of sampled edges and flip their values to simulate imperfect constraints (i.e., a positive constraint is changed to negative, and vice versa).

Baselines and metrics. We compare RoaDs against baselines from both continuous and combinatorial methods. The former is founded on GOLEM (linear) and NOTEARS-MLP (nonlinear) (Zheng et al. 2020; Ng, Ghassami, and Zhang 2020). We compare with their extensions under priors, including NTS-B (a type of algorithms incorporating the priors as hard constraints, (Sun et al. 2023; Wang et al. 2024)) and ECA (Chen et al. 2025a). The latter includes PC-stable and LiNGAM (Kalisch and Bühlman 2007; Shimizu et al. 2011). Performance is evaluated using the F1-score and the Structural Hamming Distance (SHD) against the ground-truth DAGs (Zhang et al. 2021).

Implementation details. We set $s = 0.3$ and $\tau = 0.01$ in Eq. (8), and the other parameters are default in GOLEM and NOTEARS-MLP. Each experiment was repeated ten times. More details about experimental implementation and code link can be referred in Appendix F.

Results and analysis

Linear SEM (EV). As demonstrated in Table 1, imperfect constraints severely mislead the causal discovery, and LiNGAM introduces too many spurious edges to satisfy them. The performance of PC is hampered by the small sample size, which causes less reliable conditional independence tests. The strong performance of continuous optimization methods (ECA, NTS-B, and RoaDs) is attributed to the less non-convex optimization landscape of the linear EV setting (Reisach, Seiler, and Weichwald 2021). However, NTS-B and ECA rigidly adhere to potentially flawed priors, but RoaDs can harness the benefits of correct priors while resisting misleading ones via prior alignment, resulting in an average F1-score improvement of approximately 4.4% and 17.0% decrease in SHD compared to GOLEM-EV.

Method	Gauss (ER)		Exp (ER)		Gauss (SF)		Exp (SF)	
	F1(↑)	SHD(↓)	F1(↑)	SHD(↓)	F1(↑)	SHD(↓)	F1(↑)	SHD(↓)
PC-stable	0.397	29.5	0.381	30.2	0.374	30.8	0.403	29.3
LiNGAM	0.220	47.1	0.267	46.3	0.204	50.9	0.272	47.7
NTS-B	0.787	13.2	0.745	16.6	0.734	15.9	0.681	19.7
ECA	0.661	24.0	0.638	25.4	0.608	26.6	0.569	29.1
RoaDs (Ours)	**0.821**	**11.4**	**0.777**	**14.6**	**0.750**	**15.2**	**0.734**	**14.1**
GOLEM-EV	0.807	12.1	0.728	17.4	0.701	18.2	0.672	20.0

Table 1: Comparison under EV noise (gauss and exp) for linear SEM on the ER-2 and SF-2 ($n_v = 20$, $n_d = 2n_v$, $p_a, p_b, p_c = 0.3, 0.3, 1$) (↑: higher is better, **bold** indicates the best performance) .

Method	Gauss (ER)		Exp (ER)		Gauss (SF)		Exp (SF)	
	F1(↑)	SHD(↓)	F1(↑)	SHD(↓)	F1(↑)	SHD(↓)	F1(↑)	SHD(↓)
PC-stable	**0.397**	**29.5**	0.381	30.2	0.374	30.8	**0.403**	**29.3**
LiNGAM	0.142	51.0	0.185	48.4	0.124	52.9	0.161	49.1
NTS-B	0.300	36.6	0.360	32.9	0.318	33.6	0.300	35.4
ECA	0.362	38.4	0.391	36.4	0.330	39.2	0.365	36.8
RoaDs (Ours)	0.384	32.7	**0.434**	**30.0**	**0.402**	**30.2**	0.370	33.2
GOLEM-NV	0.301	35.4	0.336	33.9	0.281	35.0	0.371	36.3

Table 2: Comparison under NV noise for linear SEM on the ER-2 and SF-2 ($n_v = 20$, $n_d = 2n_v$, $p_a, p_b, p_c = 0.3, 0.3, 1$).

Linear SEM (NV). As shown in Table 2, the linear NV setting introduces a highly non-convex optimization landscape (Ng, Huang, and Zhang 2024), causing a sharp performance decline for most continuous optimization methods. In contrast, PC remains robust as it is less sensitive to noise variances. Notably, our RoaDs maintains performance competitive with PC, demonstrating its superior resilience in navigating this challenging scenario.

Nonlinear SEM. Under nonlinear conditions (Table 3), PC remains robust due to its non-parametric nature, whereas LiNGAM fails as linearity assumption is violated. NTS-B and ECA, exhibit a significant decline in SHD. They are forced to incorporate an excessive number of edges (over 100) to minimize the least-squares loss while simultaneously adhering to flawed constraints. In this challenging environment, RoaDs achieves remarkable performance, with its F1-score surpassing ECA by an average of 14.5% and NTS-B by 15.4%. Furthermore, RoaDs demonstrates its resilience in settings with GP noise, while NOTEARS-MLP achieves a F1-score below 0.1, which indicates a near-complete failure to identify the correct causal edges.

Further comparisons are provided in Appendix G, covering different noise types (Gumbel and Normal), numbers of variables (n_v), numbers of edges (k), and sample sizes (n_d).

Influence of constraints. Figure 3 and 4 investigate the influence of both the quantity and quality of prior knowledge on continuous optimization methods. When p_a increases, as more imperfect constraints is introduced, ECA exhibits overfitting to the flawed priors. NTS-B performs comparably to GOLEM-EV. In stark contrast, RoaDs demonstrates the ability to effectively filter this information, as its SHD decreases substantially with a higher p_a. When increasing the error rate p_b within the constraints, ECA proves highly sensitive, with its SHD increasing dramatically. NTS-B shows a more gradual performance decline. Our proposed RoaDs distinguishes itself by maintaining a stable and low SHD even at high error rates. This superior robustness stems from its prior alignment mechanism, which mitigates the impact of priors that are inconsistent with the observation data.

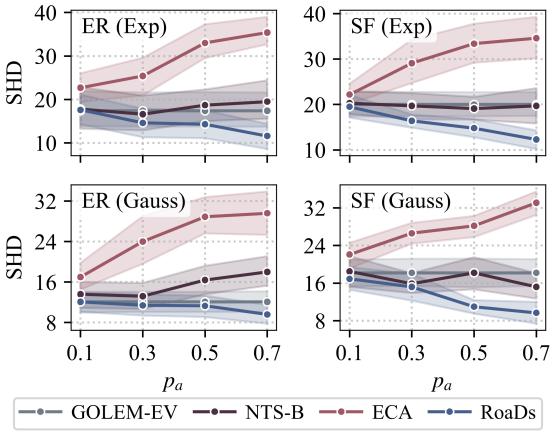

Figure 3: Influence of positive constraints rate p_a for continuous methods ($n_v = 20, n_d = 2n_v, p_b, p_c = 0.3, 1$).

More detailed comparison is provided in Appendix H, including results under other settings and sensitivity for p_c.

Method	MLP (ER)		GP (ER)		MLP (SF)		GP (SF)	
	F1(↑)	SHD(↓)	F1(↑)	SHD(↓)	F1(↑)	SHD(↓)	F1(↑)	SHD(↓)
PC-stable	0.343	31.4	0.323	33.7	0.370	32.7	0.303	35.9
LiNGAM	0.172	39.6	0.065	37.2	0.171	38.6	0.079	37.0
NTS-B	0.321	113.2	0.277	118.7	0.324	110.1	0.264	119.9
ECA	0.344	107.4	0.272	119.0	0.335	106.7	0.271	118.0
RoaDs (Ours)	**0.578**	**25.9**	**0.358**	**32.4**	**0.520**	**28.1**	**0.347**	**32.9**
NOTEARS-MLP	0.489	31.9	0.057	35.9	0.445	30.3	0.054	35.7

Table 3: Comparison under nonlinear SEM on the ER-2 and SF-2 ($n_v = 20$, $n_d = 2n_v$, $p_a, p_b, p_c = 0.3, 0.3, 1$).

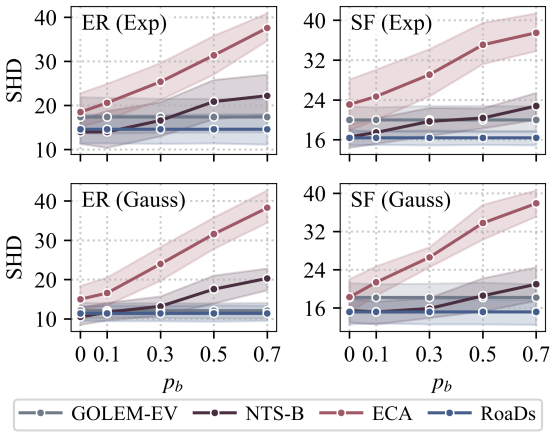

Figure 4: Influence of imperfect constraints rate p_b for continuous methods ($n_v = 20, n_d = 2n_v, p_a, p_c = 0.3, 1$).

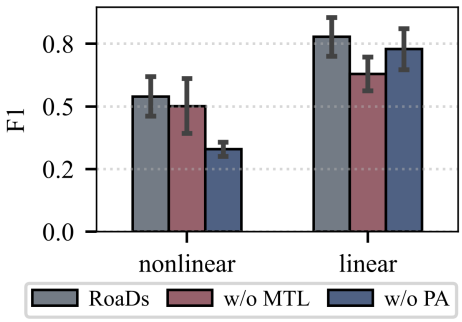

Figure 5: Ablation study on ER-2 ($n_v = 20$, $n_d = 2n_v$,$p_a, p_b, p_c = 0.3, 0.3, 1$, PA indicates prior alignment).

Method	F1	SHD	Precison	Recall
PC_stable	0.333	14.0	0.384	0.291
LiNGAM	-	-	-	-
NTS-B	0.384	14.0	0.500	0.235
ECA	0.414	17.0	0.500	**0.353**
RoaDs	**0.480**	**12.0**	0.750	**0.353**
No Priors	0.364	13.0	**0.800**	0.235

Table 4: Comparison under Saches dataset (thres $= 0.1$).

Ablation study. Figure 5 presents our ablation study on the contributions of Prior Alignment (PA) and Multi-Task Learning (MTL). In the linear case, MTL is more critical: its removal reduces the F1-score by 14.8%, whereas removing PA causes only a 4.9% drop. This suggests that in relatively convex landscapes, effective optimization strategy is more important than the objective's formulation. Conversely, in the highly non-convex nonlinear case, PA becomes dominant. Its removal leads to a 21.1% F1-score decrease, compared to just 3.8% for MTL. This indicates that in such complex landscapes, establishing a well-formed optimization objective is more fundamental than the subsequent optimization strategy.

Further evaluation on other components, including different normalization methods, various surrogate models, and running time comparison, is provided in Appendix I.

Case study. We evaluated our method on the Sachs dataset (Sachs et al. 2005), a widely-used benchmark for causal discovery from human protein-signaling networks. For our experiments, we used its 853 sample observational data (11 variables) and simulated imperfect domain knowledge with parameters $p_a, p_b, p_c = 0.3, 0.3, 1$. As summarized in table 4 (for a threshold of 0.1), RoaDs significantly outperforms all competing approaches by achieving the highest F1-score and lowest SHD. Detailed DAG visualizations and results for other thresholds are provided in Appendix J.

Conclusion

We proposed RoaDs, a novel framework that utilizes the dataset to align priors and employs MTL to resolve the conflict between data-driven and knowledge-driven optimization goals under imperfect structural constraints. Empirical evaluation demonstrates the robustness of RoaDs across both linear (EV and NV) and nonlinear SEMs, as well as its effectiveness under various noise types and constraint rates.

However, this work use MGDA to randomly identify the solution on Pareto front, which may not align with decision-maker's specific preferences. Therefore, future work could focus on developing a Pareto set learning model to generate DAGs adaptable to arbitrary preferences (Navon et al. 2021), or extending RoaDs to incorporate interventional data.

Acknowledgments

This work was supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (GRF Project No. CityU 11215723), by National Natural Science Foundation of China (Project No: 62276223), and by Young Scientists Fund of the National Natural Science Foundation of China (Project No: 52402453).

References

Ban, T.; Chen, L.; Lyu, D.; Wang, X.; Zhu, Q.; and Chen, H. 2025a. LLM-Driven Causal Discovery via Harmonized Prior. *IEEE Trans. Knowl. Data Eng.*, 37(4): 1943–1960.

Ban, T.; Chen, L.; Lyu, D.; Wang, X.; Zhu, Q.; Tu, Q.; and Chen, H. 2025b. Integrating large language model for improved causal discovery. *IEEE Trans. Artif. Intell.*

Ban, T.; Rong, C.; Wang, X.; Chen, L.; Wang, X.; Lyu, D.; Zhu, Q.; and Chen, H. 2025c. Differentiable Structure Learning with Ancestral Constraints. In *Proceeding of the 42nd International Conference on Machine Learning, ICML 2025*.

Bello, K.; Aragam, B.; and Ravikumar, P. 2022. Dagma: Learning dags via m-matrices and a log-determinant acyclicity characterization. In *Proceeding of the 35th Advances in Neural Information Processing Systems, NeurIPS 2022*.

Brehmer, J.; De Haan, P.; Lippe, P.; and Cohen, T. S. 2022. Weakly supervised causal representation learning. In *Proceeding of the 35th Advances in Neural Information Processing Systems, NeurIPS 2022*.

Brouillard, P.; Squires, C.; Wahl, J.; Kording, K. P.; Sachs, K.; Drouin, A.; and Sridhar, D. 2024. The Landscape of Causal Discovery Data: Grounding Causal Discovery in Real-World Applications. *CoRR*, abs/2412.01953.

Caruana, R. 1993. Multitask learning: A knowledge-based source of inductive bias1. In *Proceedings of the 10th International Conference on Machine Learning, ICML 1993*.

Chen, E. Y.-J.; Shen, Y.; Choi, A.; and Darwiche, A. 2016. Learning Bayesian networks with ancestral constraints. In *Proceeding of the 29th Advances in Neural Information Processing Systems, NeurIPS 2016*.

Chen, L.; Ban, T.; Lyu, D.; Sun, Y.; Hu, K.; Wang, X.; and Chen, H. 2025a. Continuous Structure Constraint Integration for Robust Causal Discovery. In *Proceeding of the 28th International Conference on Artificial Intelligence and Statistics, AISTATS 2025*.

Chen, L.; Ban, T.; Wang, X.; Lyu, D.; and Chen, H. 2025b. Mitigating Prior Errors in Causal Structure Learning: A Resilient Approach Via Bayesian Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Chen, Z.; Badrinarayanan, V.; Lee, C.-Y.; and Rabinovich, A. 2018. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proceeding of the 35th International conference on machine learning, ICML 2018*.

Colombo, D.; and Maathuis, M. H. 2014. Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.*, 15(1): 3741–3782.

Constantinou, A. C.; Guo, Z.; and Kitson, N. K. 2023. The impact of prior knowledge on causal structure learning. *Knowl. Inf. Syst.*, 65(8): 3385–3434.

de Campos, L. M.; and Castellano, F. J. G. 2007. Bayesian network learning algorithms using structural restrictions. *Int. J. Approx. Reason.*, 45(2): 233–254.

Fang, Z.; Zhu, S.; Zhang, J.; Liu, Y.; Chen, Z.; and He, Y. 2024. On Low-Rank Directed Acyclic Graphs and Causal Structure Learning. *IEEE Trans. Neural Networks Learn. Syst.*, 35(4): 4924–4937.

Fliege, J.; and Svaiter, B. F. 2000. Steepest descent methods for multicriteria optimization. *Math. Methods Oper. Res.*, 51(3): 479–494.

Glymour, C.; Zhang, K.; and Spirtes, P. 2019. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10: 524.

Hasan, U.; and Gani, M. O. 2022. Kcrl: A prior knowledge based causal discovery framework with reinforcement learning. In *Prceeding of Machine Learning for Healthcare Conference*.

Hastie, T.; Tibshirani, R.; and Friedman, J. H. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition*. Springer Series in Statistics. Springer. ISBN 9780387848570.

Hernán, M. A.; and Robins, J. M. 2010. Causal inference.

Hoyer, P.; Janzing, D.; Mooij, J. M.; Peters, J.; and Schölkopf, B. 2008. Nonlinear causal discovery with additive noise models. In *Proceeding of the 21st Advances in neural information processing systems, NeurIPS 2008*.

Kalisch, M.; and Bühlman, P. 2007. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, 8(3).

Khatibi, E.; Abbasian, M.; Yang, Z.; Azimi, I.; and Rahmani, A. M. 2024. ALCM: Autonomous LLM-Augmented Causal Discovery Framework. *CoRR*, abs/2405.01744.

Kiciman, E.; Ness, R. O.; Sharma, A.; and Tan, C. 2024. Causal Reasoning and Large Language Models: Opening a New Frontier for Causality. *Trans. Mach. Learn. Res.*, 2024.

Kitson, N. K.; Constantinou, A. C.; Guo, Z.; Liu, Y.; and Chobtham, K. 2023. A survey of Bayesian Network structure learning. *Artif. Intell. Rev.*, 56(8): 8721–8814.

Koller, D.; and Friedman, N. 2009. *Probabilistic graphical models: principles and techniques*. MIT press.

Lachapelle, S.; Brouillard, P.; Deleu, T.; and Lacoste-Julien, S. 2020. Gradient-Based Neural DAG Learning. In *Proceeding of the 8thInternational Conference on Learning Representations, ICLR 2020*.

Li, W.; Zhang, W.; Zhang, Q.; Zhang, X.; and Wang, X. 2024. Weakly Supervised Causal Discovery Based on Fuzzy Knowledge and Complex Data Complementarity. *IEEE Trans. Fuzzy Syst.*, 32(12): 7002–7014.

Lin, B.; and Zhang, Y. 2023. LibMTL: A Python Library for Deep Multi-Task Learning. *J. Mach. Learn. Res.*, 24: 209:1–209:7.

Lin, X.; Zhang, X.; Yang, Z.; Liu, F.; Wang, Z.; and Zhang, Q. 2024. Smooth Tchebycheff Scalarization for Multi-Objective Optimization. In *Proceeding of the 41st International Conference on Machine Learning, ICML 2024*.

Lin, X.; Zhen, H.-L.; Li, Z.; Zhang, Q.-F.; and Kwong, S. 2019. Pareto multi-task learning. In *Proceeding of the 32nd Advances in neural information processing systems, NeurIPS 2019*.

Miettinen, K. 1998. *Nonlinear multiobjective optimization*, volume 12 of *International series in operations research and management science*. Kluwer. ISBN 978-0-7923-8278-2.

Navon, A.; Shamsian, A.; Fetaya, E.; and Chechik, G. 2021. Learning the Pareto Front with Hypernetworks. In *Proceeding of the 9th International Conference on Learning Representations, ICLR 2021*.

Ng, I.; Ghassami, A.; and Zhang, K. 2020. On the role of sparsity and dag constraints for learning linear dags. In *Proceeding of the 34th Advances in Neural Information Processing Systems, NeurIPS 2020*.

Ng, I.; Huang, B.; and Zhang, K. 2024. Structure learning with continuous optimization: A sober look and beyond. In *Proceeding of the 3rd Causal Learning and Reasoning, CLeaR 2024*.

Pearl, J. 2009. *Causality*. Cambridge university press.

Peters, J.; Janzing, D.; and Schölkopf, B. 2017. *Elements of causal inference: foundations and learning algorithms*. The MIT press.

Reisach, A.; Seiler, C.; and Weichwald, S. 2021. Beware of the simulated dag! causal discovery benchmarks may be easy to game. In *Proceeding of the 34th Advances in Neural Information Processing Systems, NeurIPS 2021*.

Rolland, P.; Cevher, V.; Kleindessner, M.; Russell, C.; Janzing, D.; Schölkopf, B.; and Locatello, F. 2022. Score matching enables causal discovery of nonlinear additive noise models. In *Proceeding of the 39th International Conference on Machine Learning, ICML 2022*.

Sachs, K.; Perez, O.; Pe'er, D.; Lauffenburger, D. A.; and Nolan, G. P. 2005. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721): 523–529.

Schölkopf, B.; Locatello, F.; Bauer, S.; Ke, N. R.; Kalchbrenner, N.; Goyal, A.; and Bengio, Y. 2021. Towards Causal Representation Learning. *CoRR*, abs/2102.11107.

Sener, O.; and Koltun, V. 2018. Multi-task learning as multi-objective optimization. In *Proceeding of the 31st Advances in neural information processing systems, NeurIPS 2018*.

Shimizu, S.; Inazumi, T.; Sogawa, Y.; Hyvärinen, A.; Kawahara, Y.; Washio, T.; Hoyer, P. O.; and Bollen, K. 2011. DirectLiNGAM: A Direct Method for Learning a Linear Non-Gaussian Structural Equation Model. *J. Mach. Learn. Res.*, 12: 1225–1248.

Spirtes, P.; Glymour, C.; and Scheines, R. 2000. *Causation, Prediction, and Search, Second Edition*. Adaptive computation and machine learning. MIT Press. ISBN 978-0-262-19440-2.

Sun, X.; Schulte, O.; Liu, G.; and Poupart, P. 2023. NTS-NOTEARS: Learning Nonparametric DBNs With Prior Knowledge. In *Proceeding of the 26th International Conference on Artificial Intelligence and Statistics, AISTATS 2023*.

Vowels, M. J.; Camgöz, N. C.; and Bowden, R. 2023. D'ya Like DAGs? A Survey on Structure Learning and Causal Discovery. *ACM Comput. Surv.*, 55(4): 82:1–82:36.

Wang, X.; Ban, T.; Chen, L.; Lyu, D.; Zhu, Q.; and Chen, H. 2025. Large-Scale Hierarchical Causal Discovery via Weak Prior Knowledge. *IEEE Trans. Knowl. Data Eng.*, 37(5): 2695–2711.

Wang, Z.; Gao, X.; Liu, X.; Ru, X.; and Zhang, Q. 2024. Incorporating structural constraints into continuous optimization for causal discovery. *Neurocomputing*, 595: 127902.

Wang, Z.; Gao, X.; Yang, Y.; Tan, X.; and Chen, D. 2021. Learning Bayesian networks based on order graph with ancestral constraints. *Knowl. Based Syst.*, 211: 106515.

Wang, Z.; Gao, X.; and Zhang, Q. 2025. Uncertain Priors for Graphical Causal Models: a Multi-objective Optimization Perspective. *IEEE Transactions on Knowledge and Data Engineering*.

Yu, Y.; Chen, J.; Gao, T.; and Yu, M. 2019. DAG-GNN: DAG structure learning with graph neural networks. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*.

Zhang, K.; Zhu, S.; Kalander, M.; Ng, I.; Ye, J.; Chen, Z.; and Pan, L. 2021. gCastle: A Python Toolbox for Causal Discovery. *CoRR*, abs/2111.15155.

Zhang, Q.; and Li, H. 2007. MOEA/D: A Multiobjective Evolutionary Algorithm Based on Decomposition. *IEEE Trans. Evol. Comput.*, 11(6): 712–731.

Zhang, X.; Zhao, L.; Yu, Y.; Lin, X.; Chen, Y.; Zhao, H.; and Zhang, Q. 2024. LibMOON: A gradient-based multiobjective optimization library in PyTorch. In *Proceeding of the 37th Advances in Neural Information Processing Systems, NeurIPS 2024*.

Zhang, Y.; and Yang, Q. 2022. A Survey on Multi-Task Learning. *IEEE Trans. Knowl. Data Eng.*, 34(12): 5586–5609.

Zhao, H.; and Gordon, G. J. 2022. Inherent Tradeoffs in Learning Fair Representations. *J. Mach. Learn. Res.*, 23: 57:1–57:26.

Zheng, X.; Aragam, B.; Ravikumar, P. K.; and Xing, E. P. 2018. Dags with no tears: Continuous optimization for structure learning. In *Proceeding of the 31st Advances in neural information processing systems, NeurIPS 2018*.

Zheng, X.; Dan, C.; Aragam, B.; Ravikumar, P.; and Xing, E. 2020. Learning sparse nonparametric dags. In *Proceeding of the 23rd International conference on artificial intelligence and statistics, AISTATS 2020*.

Zhou, Y.; Wu, X.; Huang, B.; Wu, J.; Feng, L.; and Tan, K. C. 2024. CausalBench: A Comprehensive Benchmark for Causal Learning Capability of Large Language Models. *CoRR*, abs/2404.06349.