Interpretable Structure Learning for Knowledge Components in Education

YUANG WEI, Shanghai Institute of AI for Education, East China Normal University, China and Department of Computer Science, National University of Singapore, Singapore
YUAN-HAO JIANG, Shanghai Institute of AI for Education, East China Normal University, China and Graduate School, Shanghai Jiao Tong University, China
CHANGYONG QI, Shanghai Institute of AI for Education, East China Normal University, China
WEI ZHANG, School of Computer Science and Technology, East China Normal University, China
BO JIANG*, Shanghai Institute of AI for Education, East China Normal University, China

Structural relationships among Knowledge Components (KCs) are essential for adaptive learning systems, as they support accurate cognitive diagnosis, personalized path planning, and targeted resource recommendation. However, existing approaches frequently capture correlations instead of reliable directional dependency signals and tend to converge prematurely or become inefficient as graph dimensionality grows. These limitations weaken the reliable modeling of KC-level structure, which in turn reduces interpretability and limits downstream benefits for diagnosis, planning, and recommendation. To this end, we propose a novel structure learning framework that integrates psychometric modeling with structural search. First, we design the **I**tem **R**esponse **T**heory (IRT)-based **I**nformation **C**riterion (**IRIC**), an interpretable scoring function that combines information entropy with causal effect estimation grounded in IRT, jointly capturing statistical associations and directionality-sensitive signals under latent ability control. Second, we develop **C**o-**E**volutionary **O**ptimization for **S**tructural **S**earch (**CEO-SS**), a multi-population evolutionary algorithm with a game-inspired co-evolution mechanism that balances exploration and exploitation, avoiding premature convergence and showing robust search behavior as graph dimensionality increases within the evaluated benchmarks. Extensive experiments on three types of datasets—including benchmark causal discovery datasets, the public educational dataset, and real-world classroom data—demonstrate that our framework consistently outperforms strong baselines in accuracy and stability, with especially clear gains in adjacency recovery and more modest improvements in edge-direction recovery. In addition, expert evaluation suggests that the learned structures are more diagnostically useful, more actionable for remediation, and more pedagogically plausible than those produced by alternative scoring methods. Overall, the proposed framework provides an interpretable and practically valuable approach to learning KC structures for adaptive learning.

CCS Concepts: • **Applied computing → Computer-assisted instruction**; • **Computing methodologies → Causal reasoning and diagnostics**; • **Theory of computation → Evolutionary algorithms**.

*Corresponding author.

Authors' Contact Information: Yuang Wei, philrain.cs@gmail.com, Shanghai Institute of AI for Education, East China Normal University, Shanghai, China and Department of Computer Science, National University of Singapore, Singapore, Singapore; Yuan-Hao Jiang, jiangyuanhao@stu.ecnu.edu.cn, Shanghai Institute of AI for Education, East China Normal University, Shanghai, China and Graduate School, Shanghai Jiao Tong University, Shanghai, China; Changyong Qi, changyongqi@stu.ecnu.edu.cn, Shanghai Institute of AI for Education, East China Normal University, Shanghai, China; Wei Zhang, zhangwei.thu2011@gmail.com, School of Computer Science and Technology, East China Normal University, Shanghai, China; Bo Jiang, bjiang@deit.ecnu.edu.cn, Shanghai Institute of AI for Education, East China Normal University, Shanghai, China.

ACM 2157-6912/2026/5-ART
https://doi.org/10.1145/3815188

Additional Key Words and Phrases: AI for education, causal discovery, scoring function, item response theory, structure search algorithm

1 Introduction

Consider a student solving problems on *extreme values*. Although the student can mechanically apply *differentiation rules*, misconceptions about the underlying *function concept* and *domains* lead to systematic errors. As illustrated in Figure 1(c), the prerequisite chain "*Function Concept* ⟶ *Differentiation Rules* ⟶ *Extreme Value*" shows how missing foundations propagate to advanced tasks. An adaptive learning system that overlooks such dependencies may misinterpret the student's knowledge state, provide unsuitable exercises, and ultimately hinder learning progress. This example highlights the importance of uncovering accurate relationships among Knowledge Components (KCs)—the fundamental building blocks of adaptive learning systems. Modeling these relationships enables precise cognitive diagnosis [31, 33, 58, 72], personalized path planning [22, 39, 73, 80], and personalized resource recommendations [39, 77, 84], thereby supporting higher levels of personalization and long-term learning success [66, 70].

KC relationships can take diverse forms, including causal, prerequisite, hierarchical, parallel, integrative, and interdependent [19]. These categories are not mutually exclusive, and a single KC may interact with others in multiple ways. Among them, causal and prerequisite links are especially critical for effective teaching and learning [34], with prerequisite relations typically regarded as a special case of causal dependencies [51]. Current adaptive systems often rely on expert-defined KC graphs [51], but constructing these graphs is costly and difficult to scale [2, 81]. Data-driven methods—such as association rules [11], information-theoretic models [3, 21], and semantic analysis of resources like MOOCs or Wikipedia [35, 43, 68]—reduce manual effort but largely capture correlations rather than reliable directional dependency signals. This limits their ability to explain why students succeed or fail, underscoring the need for structure discovery tailored to educational data [8].

Insights from other fields further highlight this need. In medicine, Bayesian networks have been used to identify precursor factors driving disease progression [29, 53]. In geography, causal analysis supports disaster early-warning by uncovering hidden precursors [32, 61]. Analogously, a student's mastery of foundational concepts such as functions can be seen as a precursor influencing outcomes in advanced topics such as calculus. These examples suggest that KC structure learning in education requires methods capable of uncovering interpretable structural patterns that approximate underlying directional dependencies. The common framework of structure discovery is illustrated in Figure 1(a).

Despite progress in causal discovery research [20, 56, 86], applying it to education faces two acute challenges. **First**, general-purpose scoring functions such as BIC [52] or BDeu [23] overlook psychometric signals embedded in student responses, while black-box models such as NOTEARS [85] and DAG-GNN [79] raise concerns about fairness, transparency, and educational ethics [71]. As a result, they often yield causal graphs that are statistically valid but pedagogically uninformative, limiting their usefulness for teachers and learners. **Second**, structural search remains notoriously difficult: exhaustive enumeration is infeasible for multi-node systems, greedy heuristics quickly fall into local optima, and random exploration wastes computational resources [13, 16]. Existing evolutionary approaches show promise in generic optimization [37], but without mechanisms to preserve diversity and avoid stagnation [88], they struggle to scale to the large, high-dimensional graphs characteristic of KC networks [63]. Together, these limitations leave current methods inadequate for discovering interpretable and practically useful dependency structures in education.

To address these gaps, we propose a domain-sensitive framework that integrates psychometric modeling with causal discovery. We introduce the **Item Response Theory-based Information Criterion (IRIC)**, a new network scoring function that combines information entropy with causal effect estimation grounded in Item Response Theory (IRT). By leveraging both statistical associations and latent traits inferred from student responses,

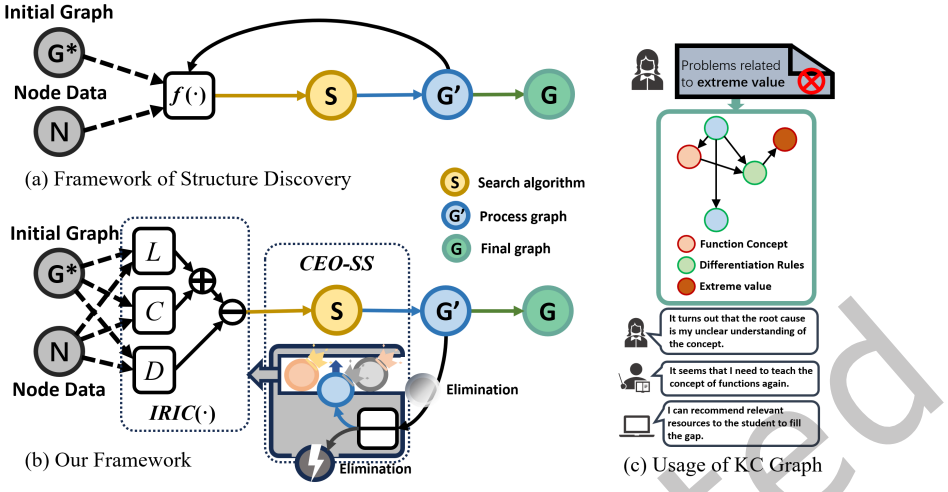


Fig. 1. Framework of structure discovery. (a) General workflow; (b) the scoring function IRIC integrates three components—L for structural likelihood, C for causal effect, and D for structural complexity.

IRIC is, to the best of our knowledge, the first scoring function tailored to uncover interpretable, directionality-sensitive dependency structures among KCs under ability adjustment. Building on IRIC, we develop the **Co-Evolutionary Optimization for Structural Search (CEO-SS)**, a multi-population evolutionary algorithm that employs a game-inspired co-evolution mechanism to balance global exploration and local refinement. Unlike existing heuristic search methods, CEO-SS explicitly prevents premature convergence by maintaining multiple interacting subpopulations, making it particularly effective for high-dimensional KC graphs. Together, IRIC and CEO-SS provide an interpretable and computationally practical framework for KC structure learning with directionality-sensitive signals. An overview of our framework is illustrated in Figure 1(b).

We validate our framework on three types of datasets: (i) general-domain public datasets widely used in causal discovery, (ii) the Junyi benchmark dataset with expert-labeled KC relationships, and (iii) a large-scale real-world dataset collected from schools across multiple regions in China. Results demonstrate consistent and statistically significant improvements in accuracy and stability, together with favorable search behavior under the evaluated settings, compared with strong baselines. Expert evaluation further suggests that the structures discovered by our framework are more diagnostically useful, more actionable for remediation, and more pedagogically plausible than those produced by alternative scoring methods.

Our contributions are summarized as follows:

- We propose **IRIC**, a novel scoring function that integrates information entropy with IRT-grounded causal-effect estimation, enabling interpretable discovery of directionally informative dependencies among KCs.
- We introduce **CEO-SS**, a structural search algorithm based on a game-inspired multi-population co-evolution mechanism, which effectively balances exploration and exploitation and supports robust search in higher-dimensional graph-structure learning settings.
- We conduct extensive experiments on general-domain, benchmark educational, and large-scale real-world datasets, demonstrating consistent and statistically significant improvements over existing causal discovery methods.

- We validate the practical value of the discovered KC structures through expert evaluation, showing that our framework yields structures that are more useful for diagnosis, more actionable for remediation, and more pedagogically plausible than those learned by alternative methods.

2 Related Work

Discovery of relationships among KCs: KCs are the building blocks of adaptive learning systems, and understanding their interrelationships is crucial for tasks such as cognitive diagnosis, resource recommendation, and learning path planning. Early studies mainly focused on identifying prerequisite relationships, which specify sequential dependencies between concepts [34]. Data-driven strategies such as association rule mining and test-based methods have been explored [10, 11], but they often fail to scale to large datasets or complex networks. Information-theoretic and topic-modeling approaches [21] improve interpretability but require substantial manual effort. More recently, deep learning methods have leveraged semantic features from Wikipedia or MOOCs to infer prerequisite graphs [35, 43], yet these approaches suffer from limited interpretability and weak generalization across domains. Although prerequisite relationships provide a partial view of KC dependencies, they can be regarded as one important form of directional dependency in educational knowledge structures. For adaptive learning systems, inferring structurally plausible and pedagogically meaningful dependencies from student learning outcomes is more valuable, as it enables accurate identification of root causes behind students' misconceptions and supports fine-grained personalization [8, 78]. For example, mastering addition is widely regarded as an instructional prerequisite for multiplication; failure to grasp the former leads to difficulties with the latter. Therefore, interpretable structure learning methods tailored to educational data is essential for constructing more reliable knowledge structures than those derived solely from correlations or prerequisite assumptions.

Causal discovery: Causal discovery aims to recover causal structures from observational data and has become an active research area in machine learning and related disciplines [20, 56]. Classical approaches fall into three categories: constraint-based [25], score-based [40], and functional causal model-based methods [86]. Constraint-based methods rely on conditional independence tests but typically require large samples [54]. Score-based methods, such as BIC [18], BDeu [23], and MDL [28], are more suitable for smaller sample sizes common in education and medicine. Recently, continuous optimization techniques (e.g., NOTEARS [85], DAG-GNN [79], GAE [41]) and causal representation learning [17, 74] have further advanced structural discovery, though often at the cost of interpretability. In parallel, evolutionary algorithms have shown promise in large-scale graph structure search, including differential evolution (DE) [27], ant colony optimization (ACO) [60], and multi-population co-evolution [59]. While effective for combinatorial optimization, their direct application to causal discovery remains challenging due to biases in scoring functions and difficulties in balancing global exploration with local exploitation. This motivates the design of domain-specific scoring criteria and tailored search frameworks to achieve both accuracy and interpretability in causal network learning.

Causal inference: Causal inference complements causal discovery by estimating causal effects through interventions or counterfactual reasoning [47]. Classical approaches include propensity score matching [48], difference-in-differences [1], instrumental variables [67], and the do-calculus [44, 46]. Recent extensions such as synthetic DID [4] enhance robustness when randomized experiments are infeasible. In the education domain, these methods provide useful tools for estimating the effect of instructional interventions or curriculum changes. However, they are less suited for discovering latent causal structures among KCs directly from student response data. This gap highlights the need for approaches that integrate psychometric modeling—such as IRT—with causal discovery, enabling interpretable and domain-specific structure learning tailored for adaptive learning systems.

3 Methodology

3.1 Network Structure Scoring Function via IRT

Causal dependencies among KCs are often more complex and fine-grained than simple statistical correlations. In mathematics learning, for example, mastery of addition typically serves as a prerequisite for understanding multiplication, reflecting a dependency with clear directionality. However, conventional information-theoretic scoring criteria primarily characterize undirected associations between variables and therefore struggle to distinguish such cause-effect relationships. To address this limitation, we propose IRIC, which unifies structural likelihood, causal effect estimation, and structural parsimony within a single scoring function:

$$\text{score}_{IRIC}(\mathcal{G} : \mathcal{D}) = \ell(\hat{\theta}_{\mathcal{G}} : \mathcal{D}) + \text{ATE}(\text{edge}_{\mathcal{G}}) - \frac{\log N}{2} Dim[\mathcal{G}], \tag{1}$$

where $\mathcal{G}$ denotes a candidate graph, $\mathcal{D}$ the dataset, N the sample size, and $Dim[\mathcal{G}]$ the number of free parameters implied by the structure $\mathcal{G}$, used to characterize model complexity.

Structural likelihood term. The first component aggregates local conditional likelihoods:

$$\ell(\hat{\theta}_{\mathcal{G}} : \mathcal{D}) = \sum_{i=1}^{n} \left[\sum_{u_i \in Val(Pa_{X_i}^{\mathcal{G}})} \sum_{x_i} M[x_i, u_i] \log \hat{\theta}_{x_i|u_i} \right]. \tag{2}$$

In Eq.2, n represents the total number of nodes in the graph, Pa denotes the set of parent nodes, u_i represents a specific parent node, x_i represents the current node, $M[\cdot]$ stands for the count of occurrences where x_i and u_i appear together, and θ is the maximum likelihood estimate of $P(x_i|u_i)$ when the current node is x_i and the parent node is u_i.

Now, considering only one term within the summation brackets, let $U_i = Pa_{X_i}$:

$$\sum_{u_i \in Val(Pa_{X_i}^{\mathcal{G}})} \sum_{x_i} M[x_i, u_i] \log \hat{\theta}_{x_i|u_i} \rightarrow \frac{1}{N} \sum_{u_i} \sum_{x_i} M[x_i, u_i] \log \hat{\theta}_{x_i|u_i} \quad . \tag{3}$$

Let $\hat{P}$ be the empirical distribution induced by the dataset $\mathcal{D}$, with $\hat{P}(x, y)$ denoting the empirical joint probability of x and y. It follows that $M[x, y] = N\hat{P}(x, y)$, $M[y] = N\hat{P}(y)$, and the maximum likelihood estimator satisfies $\hat{\theta}_{x_i|u_i} = \hat{P}(x_i \mid u_i)$, yielding:

$$\frac{1}{N} \sum_{u_i} \sum_{x_i} M[x_i, u_i] \log \hat{\theta}_{x_i|u_i} = I_{\hat{P}}(X_i; Pa_{X_i}) - H_{\hat{P}}(X_i), \tag{4}$$

where $I_{\hat{P}}(X_i; Pa_{X_i})$ denotes the mutual information between X_i and its parent set Pa_{X_i} under the distribution $\hat{P}$, and $H_{\hat{P}}(X_i)$ denotes the Shannon entropy of X_i under $\hat{P}$.

Causal-effect term. To capture directionality, IRIC adds the average treatment effect (ATE) between each node and its parents:

$$\text{ATE}(\text{edge}_{\mathcal{G}}) \;=\; N \sum_{i=1}^{n} \text{ATE}(X_i; Pa_{X_i}). \tag{5}$$

To identify $\text{ATE}(X; Pa_X)$ from observed learning logs, we adopt a back-door adjustment formulation. Its validity relies on the following standard causal assumptions [24, 45]: (1) Consistency / Stable Unit Treatment Value Assumption (SUTVA), namely that the observed outcome equals the corresponding potential outcome under the realized values of the parent KCs; (2) Positivity, meaning that, conditional on the adjustment variables, different values of the parent KCs have sufficient support in the observed logs; and (3) Conditional exchangeability, such

that, given the adjustment set Z, the parent KC values Pa_X are approximately conditionally independent of the potential outcomes $Y_X(pa)$.

Under these assumptions, let Z be an adjustment set satisfying the back-door criterion: Z contains no descendants of Pa_X and blocks all back-door paths from Pa_X to X. Then,

$$\mathbb{E}[Y_X \mid do(Pa_X{=}pa)] = \sum_z \mathbb{E}[Y_X \mid Pa_X{=}pa, Z{=}z]\ P(Z{=}z), \tag{6}$$

and $\mathrm{ATE}(X; Pa_X)$ is obtained from contrasts across pa values.

It is worth noting that during structural search, the candidate graph G evolves continuously. Strictly following the definition, reconstructing and validating an adjustment set $Z(G)$ that satisfies the back-door criterion for every candidate graph would incur substantial computational overhead. Moreover, under a purely observational learning-log setting, it is difficult to empirically verify whether all back-door paths are fully blocked. Leveraging prior knowledge from educational measurement, we attribute the primary source of cross—KC confounding to stable individual differences among learners, and characterize this variability using latent ability parameters estimated via Item Response Theory (IRT). Accordingly, throughout the structural search process we adopt a fixed adjustment variable $Z \triangleq \hat{\vartheta}$—namely, the IRT-estimated ability $\hat{\vartheta}$—as the primary back-door variable. This choice controls for confounding effects arising from learners' abilities simultaneously influencing performance across multiple KCs.

Under this setting, Z remains invariant across candidate graphs G, thereby avoiding the complexity of repeatedly constructing graph-specific adjustment sets $Z(G)$. Consequently, the ATE term should be interpreted as a directionality-sensitive effect measure under ability adjustment, serving as a directional regularization signal for structure search rather than as a definitive proof of true pedagogical causal prerequisite relations.

IRT-driven expectation. Applying the *do*-calculus in Eq. (6) requires estimating the conditional expectation $\mathbb{E}[Y_X \mid Pa_X, Z]$ needed for back-door adjustment, where Y_X denotes the learning performance variable associated with KC X. It is important to note that Y_X is not directly observable; instead, it is indirectly manifested through students' responses to individual items. Consequently, a probabilistic modeling approach is required to map item-level response outcomes to KC-level performance.

To this end, we adopt IRT as a measurement model to characterize the relationship between students' latent abilities and their probabilities of correctly answering items. Specifically, we employ the two-parameter logistic (2PL) IRT model, which jointly captures individual differences in latent ability as well as item discrimination and difficulty. For a student with latent ability ϑ, the probability of correctly answering an item associated with KC X is given by:

$$\hat{Y} = \mathrm{estimate}(Y = 1 \mid \vartheta, a, b) = \frac{e^{Da(\vartheta - b)}}{1 + e^{Da(\vartheta - b)}}, \tag{7}$$

where a and b denote the item discrimination and difficulty parameters, respectively, and $D = 1.702$ is an empirically validated scaling constant that allows the logistic function to better approximate the normal ogive model [7]. Both the latent ability ϑ and the item parameters (a, b) can be estimated from student response data via the EM algorithm [69].

To align the item-level predicted probabilities produced by IRT with the KC-level variable Y_X in Eq. (6), we interpret Y_X as the expected correctness rate over a set of items associated with KC X. Accordingly, given parent KC values Pa_X and the back-door variable Z, the conditional expectation $\mathbb{E}[Y_X \mid Pa_X, Z]$ can be approximated by aggregating the item-level predicted probabilities $\hat{Y}$ obtained from IRT. In this sense, IRT is not treated as a causal generative model, but rather as a probabilistic measurement and denoising tool. By controlling for students' latent ability differences, it provides stable estimates of the conditional expectation $\mathbb{E}[Y_X \mid Pa_X, Z]$ required for back-door adjustment, thereby supporting the subsequent computation of ATEs between KCs.

Expanded IRIC. By integrating structural likelihood, causal effects, and a model complexity penalty, the final expanded scoring function of IRIC is defined as:

$$\text{score}_{IRIC}(\mathcal{G} : \mathcal{D}) = N \sum_{i=1}^{n} \left(I_{\hat{P}}(X_i; Pa_{X_i}) + \text{ATE}(X_i; Pa_{X_i}) \right) - N \sum_{i=1}^{n} \left(H_{\hat{P}}(X_i) + H_{\tilde{Y}}(\tilde{X}_i) \right) - \frac{\log N}{2} Dim[\mathcal{G}]. \quad (8)$$

Although these terms originate from different mathematical formulations, they can all be mapped onto a unified information-theoretic scale. We align them through deterministic normalization and monotonic transformations and then combine them additively, thereby avoiding the introduction of tunable trade-off weights. This design choice is intentional: introducing adjustable weights would weaken the interpretability of IRIC and reduce its diagnostic value in subsequent analyses. The complete derivation and parameter sensitivity analysis are provided in Appendix A.1.

3.2 Structural Learning Framework

Existing search strategies for causal discovery, such as greedy heuristics or single-population evolutionary algorithms, often suffer from premature convergence and limited robustness as graph dimensionality increases. To overcome these limitations, we propose a novel meta-heuristic algorithm, termed CEO-SS. The core innovation of CEO-SS lies in its multi-population co-evolution mechanism with rank-adaptive modification, which is specifically tailored to improve search quality, robustness, and reliability in causal structure learning for educational data.

Population design. In complex search spaces, maintaining a single global population often leads to the premature loss of population diversity and an imbalance between global exploration and local exploitation. Recent studies have shown that dividing the population into multiple functionally complementary subpopulations and establishing directed information exchange mechanisms can effectively alleviate variable coupling in large-scale optimization, prevent evolutionary stagnation [76], and improve search efficiency as well as convergence accuracy in high-dimensional optimization problems [9, 83]. Inspired by these findings, CEO-SS divides the overall population into three complementary subpopulations to facilitate efficient information transmission and collaborative search. Specifically, CEO-SS consists of a superior subpopulation (*SupPop*), a developing subpopulation (*DevPop*), and an eliminated subpopulation (*EliPop*). The *SupPop* preserves the best-performing individuals to strengthen the exploitation of promising regions; the *DevPop* adaptively modifies individuals according to their fitness ranking to maintain population diversity and encourage exploration; and the *EliPop* removes poorly performing individuals to improve overall search efficiency. Within *DevPop*, each individual is assigned a ranking ratio $R \in [0, 1]$. A higher R implies fewer modifications to its corresponding graph structure (*GS*), enabling refined search in near-optimal regions, whereas a lower R allows larger structural modifications to promote broader exploration of the search space.

To operationalize this ranking-based modification strategy, we introduce a rank-based adaptive mutation operator for individuals in *DevPop*. For an individual with decision vector dimension D, the mutation intensity is determined by its fitness rank, and the number of mutated dimensions is computed as $m = \lfloor (1 - R)D \rfloor$. The algorithm then randomly selects m indices from the decision vector and reassigns the corresponding entries to one of three discrete states: forward connection, reverse connection, or no connection. This operation modifies the topological states of the selected edges in the candidate graph.

Optimization objective. CEO-SS is a general score-based structural search algorithm whose core objective is to identify a directed acyclic graph (DAG) that maximizes a structural scoring function $S(A)$ within the candidate graph space. Specifically, each candidate structure is represented by its adjacency matrix A (or equivalently by the flattened decision vector DV derived from its upper-triangular part ΔA). The fitness of an individual during the search phase is defined as:

$$F_{\text{score}}(DV) = S(A). \quad (9)$$

Accordingly, the structural search problem can be formulated as:

$$\min \ loss(DV) = -F_{\text{score}}(DV) \quad \Leftrightarrow \quad \max_{A \in \mathcal{G}_{\text{DAG}}} S(A), \tag{10}$$

where the structure with the highest score $S(A)$ is returned as the final output after the search process.

In benchmark settings where expert-annotated structures are available, we additionally compute the F1-score between candidate structures and the expert-annotated graph as a proxy evaluation metric during the search phase. This proxy fitness is used to analyze the consistency between the evolving structures and expert priors throughout the search process, and is defined as:

$$\min \ loss_{\text{proxy}}(DV) = 1 - F_{\text{proxy}}(DV). \tag{11}$$

It is important to emphasize that this F1 metric is used solely for benchmark evaluation and behavioral analysis, and is not employed for final structure selection. In all experimental settings, both the retention of elite candidates and the final structural output are determined exclusively by the scoring function $S(A)$.

Encoding and evaluation. Each individual is encoded as a vector of length $E_{\text{dag}} = \frac{n(n-1)}{2}$, where n denotes the number of nodes. The value set is $\{0, 1, -1\}$, representing no edge, a directed edge following the indexed direction, and a directed edge in the opposite direction, respectively. After mutation and crossover, a repair mechanism is applied to ensure the acyclicity and validity of the graph.

Specifically, the repair mechanism traverses all nodes in the graph using a depth-first search strategy. If a successor node is found to point back to the starting node of the current traversal path, a cycle is detected. In such cases, the algorithm deterministically removes the back-edge responsible for forming the cycle. This procedure strictly guarantees the directed acyclic property of the mutated network structure while preserving the original topological characteristics as much as possible.

In practical scenarios without expert priors, the fitness of each individual equals the structural score computed by $S(A)$, which is then used for ranking, selection, and evolution. In benchmark experiments where expert-annotated structures are available, we additionally compute the F1-score between the candidate structure and the expert-annotated graph as a consistency evaluation and comparability metric (i.e., proxy fitness):

$$F_{\text{proxy}}(DV) = \text{F1}(A, A^{GT}) = 2 \times \frac{\text{Precision}(A) \cdot \text{Recall}(A)}{\text{Precision}(A) + \text{Recall}(A)}, \tag{12}$$

where Precision and Recall are computed by comparing the predicted edges with the expert-annotated ground-truth edges. It is important to emphasize that, regardless of whether expert-annotated structures are available, both elite preservation and final structure selection in the algorithm are consistently determined by the scoring function $S(A)$.

Control parameters. The co-evolutionary process is regulated by several hyper-parameters. The ambient pressure $AP \in [0, 1]$ controls the proportion of $SupPop$ in the overall population. The maximum number of function evaluations $maxFE$ defines the stopping criterion, with a counter FE incremented at each evaluation until $FE \geq maxFE$. The population size $PopSize$ specifies the number of individuals in each generation. These parameters jointly balance convergence speed, computational cost, and population diversity. The full pseudocode of CEO-SS is provided in Appendix Algorithm 1.

In summary, CEO-SS adopts the structural scoring function $S(A)$ as the core criterion for search and selection. In scenarios without expert priors, the algorithm performs structure search autonomously and returns the structure with the highest $S(A)$. In benchmark experiments where expert-annotated structures are available, we additionally report the F1-score as a proxy metric for consistency evaluation, enabling comparable comparisons across different search algorithms without affecting the final structural decision.

4 Results

4.1 Experimental Settings

This study comprises two experimental segments. The first segment is dedicated to assessing the performance of the structural search algorithm, highlighting its accuracy, robustness, and convergence behavior in addressing structural search challenges. The second part focuses on validating the performance of the IRIC method across both generic datasets and real-world datasets. This aims to demonstrate the effectiveness of IRIC in learning directionally informative KC structures in educational settings.

4.1.1 Datasets. We evaluated our framework on three types of datasets. First, we used the Alarm datasets [82], a benchmark suite for the Causal Discovery competition at PCIC 2021 [65]. It comprises three graphs with 18, 24, and 25 nodes (referred to as Alarm18, Alarm24, and Alarm25), each containing several hundred causal relationships. These datasets provide controlled settings for validating causal discovery algorithms, and we followed the official configurations without parameter modifications.

Second, we conducted experiments on the publicly available Junyi dataset[1], which records over 16 million exercise attempts from more than 72,000 students on the Junyi Academy platform. Distinct from other educational datasets such as EdNet[2] and ASSISTments[3], Junyi includes expert-annotated KC relationships, capturing both hierarchical and prerequisite structures that are essential for evaluating causal models.

Finally, to assess real-world applicability, we collected a dataset from mathematics courses in 77 classes across 19 primary and 7 middle schools in China. This dataset spans grades 4 to 7 and includes KC annotations with expert-defined interrelationships, providing a robust testbed for evaluating our approach in authentic educational contexts.

4.1.2 Baselines. We compared CEO-SS with a wide range of baselines from both causal discovery and meta-heuristic optimization.

Causal discovery methods. We included nine representative algorithms: **PC** [56], **DirectLiNGAM** [55], **ICALiNGAM** [55], **GES** [12], **NOTEARS** [85], **ADM4** [87], **MLE-SGL** [75], **PCMCI** [49], and **THP** [6].

Meta-heuristic algorithms. To verify the effectiveness of CEO-SS against other evolutionary approaches, we further compared with four recent meta-heuristics: **CMMO** [38], **EESHHO** [30], **IMODE** [50], and **ADSAPSO** [36]. In addition, we included two classical baselines—**PSO** [15] and **DE** [57]—due to their strong convergence and generalization performance. IMODE and ADSAPSO represent improved variants of these classical algorithms.

Scoring functions. In the causal structure learning experiments on educational datasets, we compared the proposed IRIC scoring method with three widely used scoring functions: BIC [52], BDeu [23], and K2 [14]. All scoring methods were evaluated under the same structural search frameworks to ensure fair comparison. Specifically, the structural search frameworks include the traditional GES [12] and our proposed CEO-SS method.

4.1.3 Implementation Details. All experiments were implemented using the PlatEMO framework [62, 64] and executed on a server equipped with dual Intel Xeon Gold 6330 CPUs, 512GB RAM, running Ubuntu 22.04. No GPU acceleration was used. To mitigate randomness, each experiment was repeated nine times with different random seeds, and the average results are reported. Random seeds were fixed to ensure reproducibility. For CEO-SS, we set the population size to 100, the maximum number of function evaluations ($maxFE$) to 10,000, and following the recommended settings for the ambient pressure parameter in prior studies [26], AP was set to 0.2 in our experiments. A higher AP retains more outstanding individuals in each iteration but discards more inferior ones. Unless otherwise specified, all other evolutionary operators (mutation, crossover, and selection strategies)

[1] https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=1198
[2] https://github.com/riiid/ednet
[3] https://sites.google.com/site/assistmentsdata/datasets

Table 1. Structural pattern metrics used for causal discovery methods [42]

Metric	Description
Missing edges (ME)	Edges present in the expert-referenced structure but missing in the learned graph
Extra edges (EE)	Edges present in the learned graph but absent from the expert-referenced structure
Correct adjacencies (AD)	Undirected adjacencies shared by the learned graph and the expert-referenced structure
Correct directed edges (CDE)	Directed edges in the learned graph with directions consistent with expert annotations
Incorrect directed edges (IDE)	Directed edges in the learned graph with directions inconsistent with expert annotations
Total Count (TC)	Overall graph evaluation score
Adjacency precision	Adj P = AD / (number of predicted adjacencies)
Adjacency recall	Adj R = AD / (number of true adjacencies)
Arrowhead precision	Arrhd P = CDE / (number of predicted arrowheads, relative to expert-annotated instructional dependencies)
Arrowhead recall	Arrhd R = CDE / (number of true arrowheads, relative to expert-annotated instructional dependencies)

followed the default settings in PlatEMO. The algorithm terminated strictly when *maxFE* was reached, and no early stopping was applied.

For general benchmark datasets (e.g., Alarm), we consistently adopt the standard BDeu scoring function to evaluate the general performance of CEO-SS as a structure search algorithm without introducing domain-specific priors, thereby avoiding confounding effects from differences in scoring functions when comparing search capabilities. For educational datasets, we further report results from cross-combinations of different structure search algorithms and scoring functions. This experimental design allows the effects of search strategies and scoring functions to be disentangled and analyzed under the same data and evaluation criteria, enabling simultaneous assessment of the effectiveness of structure search methods and the additional gains brought by the domain-sensitive scoring function (IRIC). All comparative methods were executed under identical hardware environments and data partitioning settings. Hyperparameters were set according to recommendations in the original papers or default configurations in publicly available implementations, ensuring fairness and reproducibility of the experimental results.

4.1.4 Evaluation Metrics. In our comparative study on the effectiveness of structural search within general domain datasets, we utilized the commonly adopted F1-score metric as our primary evaluation criterion. To evaluate the performance of IRIC in causal structure learning within the educational domain, we adopt a widely accepted Bayesian network structural evaluation protocol, with minor semantic adaptations to accommodate the educational context. The specific definitions of the evaluation metrics are summarized in Table 1.

To ensure that the learned knowledge structures possess genuine educational utility, we additionally conducted an expert-based evaluation to assess the pedagogical usefulness of the inferred structures. Three domain experts independently reviewed ten local subgraphs (each containing 3—6 KCs) extracted from structures learned using four different scoring functions. For each evaluation instance, experts were provided with a brief student scenario describing the learner's mastery levels on the relevant knowledge concepts, together with four anonymized candidate graphs presented in randomized order. Experts rated each graph on a five-point Likert scale along three dimensions: (i) Diagnostic Usefulness (DU), (ii) Remediation Actionability (RA), and (iii) Pedagogical Plausibility / Explainability (PE). In addition, experts were asked to select the single most useful graph (Top-1) among the four candidates.

The specific evaluation questions were as follows:

- **DU**: Does this structure help trace a student's learning difficulty to potential prerequisite weaknesses?
- **RA**: Based on this structure, can you design a reasonable remediation sequence or practice plan?
- **PE**: Does this structure align with instructional order and knowledge dependency logic, and is it easy to explain to students?

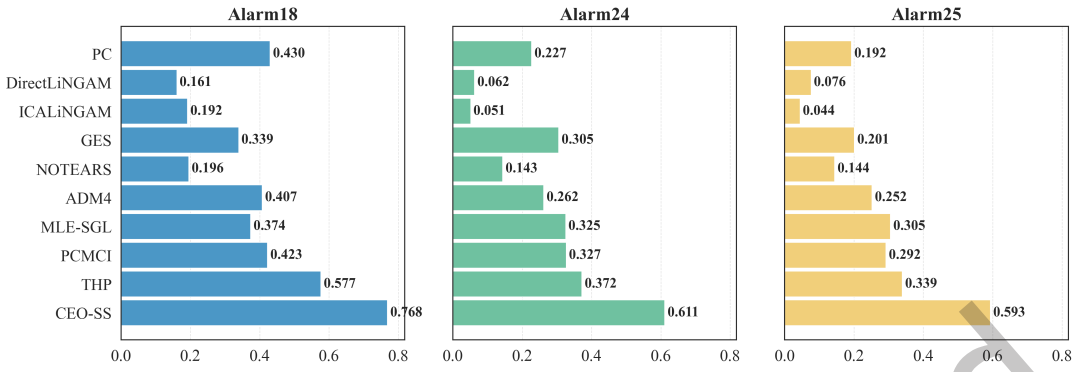


Fig. 2. F1-scores for comparison algorithms with causal structure learning methods.

Table 2. Experimental results of meta-heuristic algorithms.

Dataset	*D*	*loss*	CEO-SS	CMMO		ADSAPSO		EESHHO		IMODE		PSO		DE	
Alarm18	153	Mean	**2.33E-01**	5.61E-01	[+]	6.13E-01	[+]	4.09E-01	[+]	***4.08E-01***	[+]	4.86E-01	[+]	4.32E-01	[+]
		Std.	***1.63E-02***	2.22E-02		2.27E-02		8.06E-02		6.07E-02		**1.27E-02**		3.14E-02	
Alarm24	276	Mean	**3.68E-01**	6.26E-01	[+]	6.54E-01	[+]	5.62E-01	[+]	5.89E-01	[+]	***5.60E-01***	[+]	5.62E-01	[+]
		Std.	***1.32E-02***	2.66E-02		3.67E-02		3.99E-02		**1.16E-02**		3.54E-02		1.67E-02	
Alarm25	300	Mean	**4.02E-01**	6.52E-01	[+]	6.90E-01	[+]	5.55E-01	[+]	6.14E-01	[+]	5.73E-01	[+]	***5.49E-01***	[+]
		Std.	***7.71E-03***	1.06E-02		1.01E-02		8.82E-03		**8.15E-04**		1.56E-02		3.30E-02	
+/-/=		-	-	3 / 0 / 0		3 / 0 / 0		3 / 0 / 0		3 / 0 / 0		3 / 0 / 0		3 / 0 / 0	

- **Top-1**: Which of the four graphs is the most suitable for diagnosis and remediation during instruction?

4.2 Analysis of Structural Search Algorithm

4.2.1 Comparative Experiments with Causal Structure Learning Methods. Compared to other algorithms [6, 12, 49, 55, 56, 75, 85, 87], as illustrated in Figure 2, the F1-score of CEO-SS consistently outperforms other benchmark algorithms. Therefore, the results suggest that CEO-SS is a highly competitive approach and achieves statistically significant improvements over multiple baselines (significance tests are reported in Appendix B.1).

Benefiting from the multi-subpopulation collaborative mechanism, CEO-SS is able to capture complex dependencies among knowledge structures, enabling the accurate identification of graph structures. In contrast, methods such as ADM4 and GES show less favorable performance as problem size increases within the evaluated benchmarks, while DirectLiNGAM also struggles with high-dimensional data. Although THP is among the strongest recent baselines, its heavier computational demands make it less practical under the evaluated settings. Beyond scalability, CEO-SS benefits from its meta-heuristic strategy, which reframes knowledge structure learning as a dual optimization problem. This approach eliminates the need for strong assumptions or prior knowledge structures, offering broader applicability than methods such as PC, which rely heavily on independence assumptions, or PCMCI, which requires extensive parameter tuning. Finally, CEO-SS exhibits notable robustness. By effectively balancing population diversity and individual superiority, it resists noise more effectively than NOTEARS and handles sparse graph structures more reliably than MLE-SGL.

Table 3. Ablation study of CEO-SS.

Model	Alarm18 ($D = 153$)	Alarm24 ($D = 276$)	Alarm25 ($D = 300$)
CEO-SS w/o *DevPop*	2.88E-01 (1.57E-02)	4.18E-01 (8.95E-03)	4.33E-01 (1.47E-02)
CEO-SS w/o *EliPop*	6.24E-01 (1.46E-02)	6.84E-01 (1.08E-02)	7.03E-01 (1.97E-02)
CEO-SS w/o *SupPop*	***5.66E-01 (1.75E-02)***	***6.67E-01 (1.28E-02)***	***6.71E-01 (3.34E-02)***
CEO-SS (Full)	**2.33E-01 (1.63E-02)**	**3.68E-01 (1.28E-02)**	**4.02E-01 (3.34E-02)**

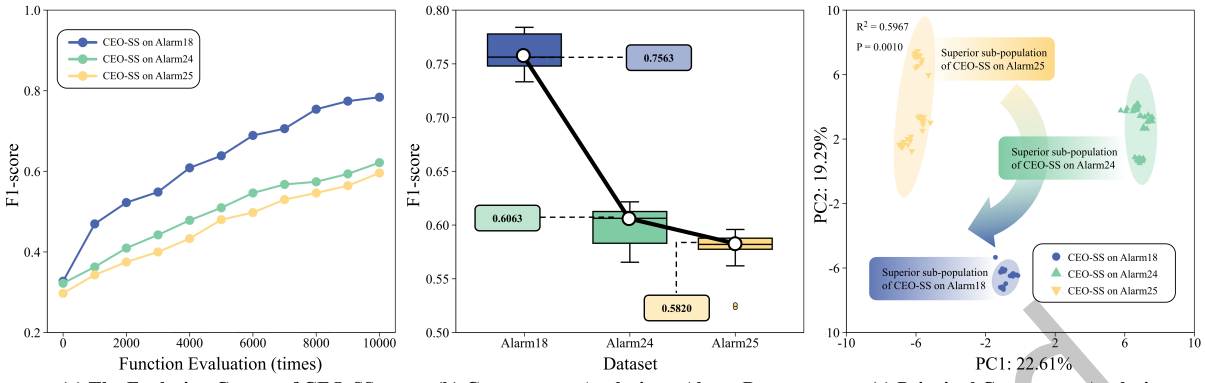
Note. "w/o" indicates removing the corresponding subpopulation. D denotes the dimensionality of the structural search problem (i.e., the number of decision variables). Results are reported as mean (standard deviation). The best and second-best results for each dataset are highlighted in **bold** and ***bold-italic***, respectively.

4.2.2 Comparative Experiments with Meta-heuristic Algorithms. In the previous section, we experimentally demonstrated that the proposed algorithm achieves notable advantages over existing causal structure learning methods. To further evaluate its effectiveness, we now conduct a comprehensive comparison with advanced meta-heuristic algorithms, focusing on the ability to recover benchmark reference graph structures and the overall performance achieved. The results of these comparative experiments are reported in Table 2. Specifically, we present the mean and standard deviation of the *loss* obtained by each algorithm. Following common practice, the symbols [+], [-], and [=] indicate whether the proposed algorithm performs better, worse, or equal to a given baseline, respectively. From the results, the proposed algorithm consistently achieves the lowest *loss* values across all datasets, indicating that the graph structures identified by CEO-SS exhibit the smallest deviation from the expert-annotated structures. Furthermore, CEO-SS attains the second-lowest standard deviation in all cases, suggesting that the distribution of results is relatively concentrated. This concentration contributes to more convergent outcomes and mitigates the adverse impact of randomness. Although CEO-SS does not always deliver the strongest convergence among all baselines, its consistently lowest *loss* values demonstrate that the algorithm effectively balances convergence and diversity. We attribute its superior overall performance to this balance.

4.2.3 Ablation experiment. To further validate the effectiveness and necessity of each module in the multi-subpopulation collaborative mechanism, we conducted an ablation study using the same experimental design as in the previous subsection. Specifically, three variant versions were constructed by removing each of the three subpopulations from the full CEO-SS algorithm in turn for comparative analysis. As shown in Table 3, the full version consistently achieved the lowest mean loss values across all test datasets. This finding strongly confirms that the three subpopulations each play an indispensable role in the co-evolutionary process, and together they contribute to the superior performance of the algorithm in optimizing complex network structures.

The results in the table clearly reveal the differentiated impact of each mechanism on the overall performance. When the *EliPop* subpopulation is removed, the algorithm exhibits the most severe performance degradation, with the highest loss values among all variants across all datasets. This strongly suggests that, in high-dimensional decision spaces, timely elimination of low-quality individuals is crucial for maintaining overall search efficiency. Similarly, removing *SupPop* also leads to a substantial decline in performance, directly demonstrating the central role of the elite-preservation mechanism in guiding the population toward stable convergence to optimal structures. In addition, although the variant without *DevPop* achieves the second-best performance, its loss values remain consistently higher than those of the full algorithm. This further confirms that the rank-based adaptive mutation mechanism plays an irreplaceable role in maintaining population diversity and preventing the algorithm from becoming trapped in local optima.

4.2.4 Convergence analysis. As an evolutionary computation method, CEO-SS has previously shown strong performance in structural search tasks. However, like other stochastic search algorithms, its outcomes may



(a) The Evolution Curves of CEO-SS **(b) Convergence Analysis on Alarm Datasets** **(c) Principal Component Analysis**

Fig. 3. Convergence analysis of the CEO-SS. In this figure, sub-figure (a) shows the obtained evolutionary curves of the optimal individuals of CEO-SS on three datasets of different sizes. Further, all individuals in the superior sub-populations obtained by repeating the run 9 times on the three datasets are extracted for further analysis. On this basis, sub-figure (b) shows the boxplot of the F1-scores of all individuals in the superior sub-populations, and sub-figure (c) shows the results of the principal component analysis performed on these individuals.

be influenced by randomness. To assess its stability and convergence, we conducted convergence experiments and plotted all superior individuals obtained from nine repeated runs on the three datasets (see Figure 3). The results reveal several noteworthy patterns. First, as problem scale decreases, the structural search tasks become easier and CEO-SS achieves better performance, a trend consistent with other algorithms in Figure 2. As shown in Figure 3(a), CEO-SS exhibits effective search performance across all datasets. Owing to the game-inspired co-evolution among sub-populations, no obvious late-stage stagnation is observed in the search process. This can be attributed to the dynamic interaction between the superior sub-population (*SupPop*) and the developing sub-population (*DevPop*), which preserves individual superiority while maintaining diversity. In this study, because directly measuring structural diversity would incur substantial computational overhead, we use fitness diversity [5] as a practical surrogate for characterizing population diversity.

Second, with increasing task complexity, the fitness diversity within *SupPop* grows (Figure 3(b)). This indicates that CEO-SS adapts by intensifying exploration through *DevPop*, transferring superior individuals to *SupPop*, and promoting fitness diversity across sub-populations. Such adaptive dynamics help the algorithm balance exploration and exploitation, preventing premature stagnation.

Finally, PCA analysis of all superior individuals from the nine runs (Figure 3(c)) further supports these findings. The results show that individuals from different datasets are well separated, suggesting that CEO-SS performs targeted learning based on dataset-specific histories. Moreover, smaller graphs yield more concentrated distributions, consistent with lower task difficulty. Together, these results demonstrate that CEO-SS achieves superior convergence and robustness: despite inherent randomness, repeated runs yield consistently well-clustered F1-scores, underscoring the algorithm's resilience to stochastic variations.

4.3 Analysis of Network Structure Learning

4.3.1 Comparison Evaluation. In our causal graph structure learning experiments, we applied multiple scoring functions while consistently using the same structural search algorithm to ensure fair comparison. We first evaluated three simple reference structures. As shown in Figure 4, each circle denotes a KC, and directed edges represent prerequisite relations. The accompanying table reports the structure scores assigned by different scoring

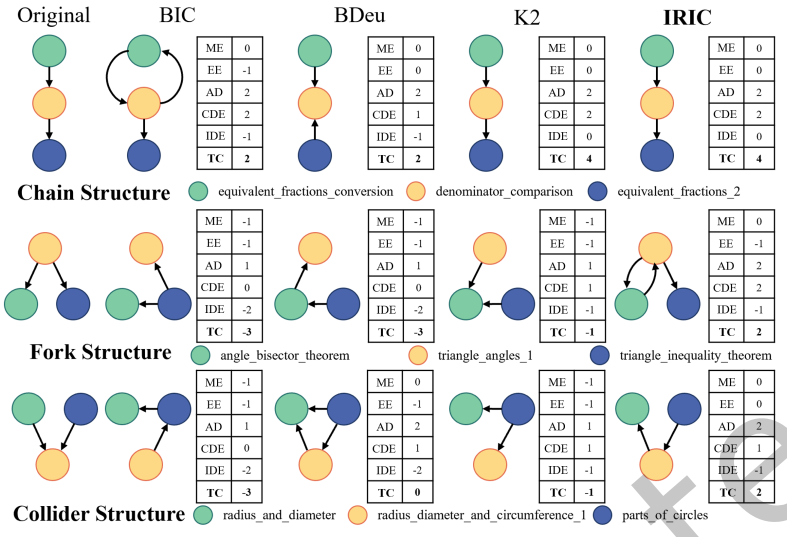


Fig. 4. Learning outcomes for randomly selected simple structures from the Junyi dataset.

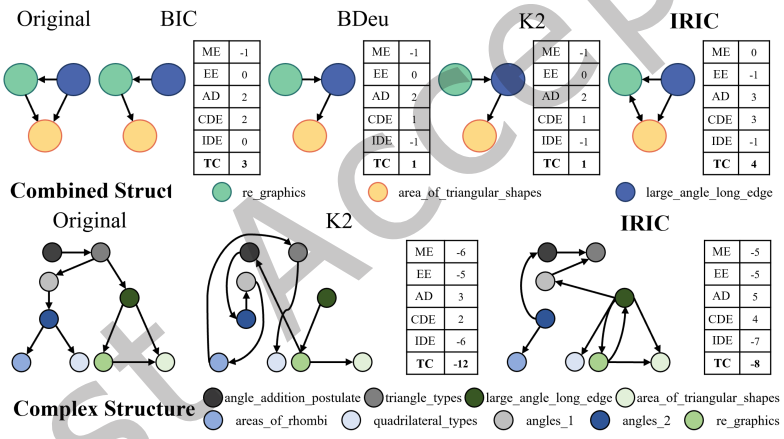


Fig. 5. Learning outcomes for randomly selected complex structures from the Junyi dataset.

functions. IRIC-based structural search demonstrated a stronger ability to recover structures consistent with observed data patterns and expert priors, under ability-adjusted directional signals.

We then extended the evaluation to more complex structures (Figure 5). Here IRIC proved particularly effective at uncovering plausible directional dependencies not captured by other scoring functions. This advantage can be attributed to two factors: its use of do-calculus-inspired, ability-adjusted effect estimation, which provides a more informative directional signal, and its integration of correlation- and causality-based components, which enhances the identification of directionally plausible dependencies.

Finally, Table 4 summarizes the quantitative comparison results, which comprise three levels of controlled analyses: (1) Search algorithm comparison: under a fixed scoring function, different search strategies are evaluated to assess their structural search capabilities under the same scoring criterion; (2) Scoring function comparison:

Table 4. Comparison of IRIC with other Methods on Junyi and Real—world datasets.

Search methods	Scoring functions	Junyi				Real—world			
		Adj P	Adj R	Arrhd P	Arrhd R	Adj P	Adj R	Arrhd P	Arrhd R
GES	BIC	0.514	0.523	0.162	0.183	0.584	0.461	0.262	0.224
	BDeu	0.564	0.550	0.265	0.246	0.643	0.523	0.252	0.212
	K2	0.556	0.573	0.232	0.224	0.613	0.497	0.188	0.159
	IRIC	0.644	0.756	0.320	0.564	0.742	0.801	0.331	0.462
CEO-SS	BIC	0.539	0.574	0.211	0.244	0.623	0.509	0.319	0.283
	BDeu	0.601	0.606	0.321	0.325	0.696	0.578	0.310	0.258
	K2	0.598	0.618	0.295	0.313	0.674	0.534	0.232	0.201
	IRIC w/o Structural Likelihood	0.648	0.782	0.327	0.592	0.748	0.836	0.333	0.525
	IRIC w/o ATE	0.672	0.742	0.338	0.523	0.776	0.804	0.344	0.498
	IRIC	**0.689**	**0.811**	**0.363**	**0.651**	**0.802**	**0.868**	**0.372**	**0.553**

under a fixed search algorithm, different scoring functions are compared to isolate the effect of the scoring criterion itself on structure learning performance; and (3) IRIC ablation analysis: under the same search algorithm (CEO-SS), the full IRIC is further compared with its variants that remove the structural-likelihood term (IRIC w/o Structural Likelihood) and the causal-effect term (IRIC w/o ATE), in order to validate the independent contributions of IRIC's constituent components in practical structure learning.

Overall, IRIC achieves the best performance on both the Junyi benchmark and the real-world datasets, with particularly strong results in learning adjacency relationships. At the same time, IRIC exhibits certain limitations in modeling prerequisite directions. Nevertheless, its ability to capture deeper directional dependencies enables it to recover adjacency relations missed by other methods, albeit sometimes at the cost of producing more complex graph structures. This trade-off highlights IRIC's advantage in recovering dependency structures that are more consistent with expert-referenced relations.

4.3.2 Detailed Analysis. To showcase the capability of IRIC in learning relationships between KCs and to highlight its advantages and differences compared to other scoring methods, the quantified relationship values between the basic and composite structures from Fig.4 and Fig.5 are listed in Table 5. In the table, the first row indicates the adjacency relationships and precedence directions between three nodes, presenting the existence of edges in the original structure (O-edge) and the quantified relationship values between KCs calculated using different scoring methods. The scores obtained by each method are compared internally to identify the most significant node relationships.

The experimental results across four sets of examples lead to three main observations. **(i) Higher accuracy.** IRIC consistently preserves a larger proportion of valid edges than other methods. Since the scores within each method are relative, bold entries indicate preserved edges, and IRIC yields more valid relations overall. **(ii) Sensitivity to weak correlations.** By combining mutual information (Mu) with causal effects (Ca), IRIC can surface expert-annotated links even when the correlation signal is weak. For Set 2, the expert-annotated edge $C \rightarrow B$ is preferred despite modest association (Mu = 0.1817, Ca = 0.3499), score 0.3285) over its reverse $B \rightarrow C$ (score 0.3126; Δ=0.0159), which is missed by two of three baselines (BIC and K2). Likewise in Set 4, IRIC recovers $A \rightarrow B$ with very small correlation(Mu=0.0898) thanks to a non-trivial causal-effect component (Ca = 0.3261; score = 0.3037). **(iii) Suppression of spurious correlations.** IRIC down-weights directions that exhibit strong association but relatively weak ability-adjusted directional effect signals. A representative case is Set 2 $B \rightarrow A$: although its association is relatively large (Mu = 0.3254), the causal-effect term is small (Ca = 0.1756), yielding

Table 5. The calculated results of quantifying relationships between KCs.

Relation	**A → B**		**B → A**		**A → C**		**C → A**		**B → C**		**C → B**	
equivalent_fractions_conversion(A), denominator_comparison(B), equivalent_fractions_2(C)												
Method	O-edge	score	O-edge	score	O-edge	score	O-edge	score	O-edge	score	O-edge	score
BIC	**TRUE**	**0.8449**	FALSE	**0.8461**	FALSE	0.3369	FALSE	0.2679	**TRUE**	**0.7459**	FALSE	0.7123
BDeu	**TRUE**	**0.8160**	FALSE	0.4252	FALSE	0.5530	FALSE	0.3877	**TRUE**	0.3428	FALSE	**0.8458**
K2	**TRUE**	**0.7319**	FALSE	0.5577	FALSE	0.4191	FALSE	0.2768	**TRUE**	**0.8337**	FALSE	0.1751
IRIC	Mu:0.5223 Ca:0.5531	**0.5515**	Mu:0.5667 Ca:0.3616	0.3549	Mu:0.1060 Ca:0.2859	0.2769	Mu:0.0832 Ca:0.3062	0.2950	Mu:0.4133 Ca:0.3986	**0.3993**	Mu:0.3542 Ca:0.1745	0.1811
angle_bisector_theorem(A), triangle_inequality_theorem(B), triangle_angles_1(C)												
Method	O-edge	score	O-edge	score	O-edge	score	O-edge	score	O-edge	score	O-edge	score
BIC	FALSE	0.5848	FALSE	**0.6026**	FALSE	0.5299	**TRUE**	0.5532	FALSE	**0.8598**	**TRUE**	0.5458
BDeu	FALSE	0.4498	FALSE	**0.7474**	FALSE	**0.8992**	**TRUE**	0.6903	FALSE	0.1124	**TRUE**	0.4630
K2	FALSE	0.5918	FALSE	**0.8578**	FALSE	0.7254	**TRUE**	**0.8137**	FALSE	0.4167	**TRUE**	0.1244
IRIC	Mu:0.1625 Ca:0.1788	0.1761	Mu:0.3254 Ca:0.1756	0.1809	Mu:0.1276 Ca:0.4021	**0.3691**	Mu:0.1668 Ca:0.3807	**0.3534**	Mu:0.0758 Ca:0.3373	0.3126	Mu:0.1817 Ca:0.3499	**0.3285**
radius_and_diameter(A), parts_of_circles(B), radius_diameter_and_circumference_1(C)												
Method	O-edge	score	O-edge	score	O-edge	score	O-edge	score	O-edge	score	O-edge	score
BIC	FALSE	0.5003	FALSE	**0.8550**	**TRUE**	0.4026	FALSE	0.3767	**TRUE**	0.5962	FALSE	**0.6898**
BDeu	FALSE	0.2326	FALSE	**0.7901**	**TRUE**	0.6592	FALSE	**0.8754**	**TRUE**	**0.8071**	FALSE	0.2731
K2	FALSE	0.1068	FALSE	0.7769	**TRUE**	0.5001	FALSE	0.5151	**TRUE**	**0.8647**	FALSE	0.4968
IRIC	Mu:0.1492 Ca:0.1131	0.1144	Mu:0.1375 Ca:0.0961	0.0978	Mu:0.4231 Ca:0.1228	0.1361	Mu:0.3949 Ca:0.1972	**0.2038**	Mu:0.2792 Ca:0.3294	**0.3157**	Mu:0.2874 Ca:0.1743	0.1779
large_angle_long_edge(A), re_graphics(B), area_of_triangular_shapes(C)												
Method	O-edge	score	O-edge	score	O-edge	score	O-edge	score	O-edge	score	O-edge	score
BIC	**TRUE**	**0.7196**	FALSE	0.6505	**TRUE**	0.0315	FALSE	0.33518	**TRUE**	**0.8498**	FALSE	0.6106
BDeu	**TRUE**	0.3994	FALSE	**0.7694**	**TRUE**	**0.9271**	FALSE	0.6479	**TRUE**	0.3352	FALSE	0.5502
K2	**TRUE**	0.2293	FALSE	**0.8339**	**TRUE**	**0.8840**	FALSE	0.7537	**TRUE**	0.6336	FALSE	0.1903
IRIC	Mu:0.0898 Ca:0.3261	**0.3037**	Mu:0.2015 Ca:0.2487	0.2414	Mu:0.1387 Ca:0.2742	**0.2610**	Mu:0.0481 Ca:0.2288	0.2161	Mu:0.3115 Ca:0.3540	**0.3380**	Mu:0.0946 Ca:0.3873	**0.3553**

Note: Mu: mutual information; Ca: causal effect; O-edge: whether the edge exists in the original graph.

a moderate total score (0.1809) compared with the expert-referenced directional relations in the same set (e.g., $C \rightarrow A$ 0.3534 and $C \rightarrow B$ 0.3285). In contrast, baselines that do not explicitly separate correlation from causation rank $B \rightarrow A$ highly (e.g., BIC 0.6026, BDeu 0.7474, K2 0.8578), contradicting the expert-labeled structure.

These case studies suggest that IRIC provides useful directional signals beyond pure association-based scoring. However, consistent with the results in Table4, these gains are stronger for adjacency recovery than for prerequisite-direction recovery, and therefore should not be interpreted as definitive evidence of pedagogical prerequisite direction.

4.3.3 Human Evaluation Results. During the expert evaluation, we first conducted a reliability analysis to assess the consistency of ratings provided by the three educational experts across the three evaluation dimensions. The results indicate good internal consistency among experts for all three questions, with Cronbach's α coefficients ranging from 0.76 to 0.78. This suggests that the designed evaluation items reliably capture the latent construct of the usability of knowledge structures for diagnosis and remediation. Based on this finding, the subsequent analyses use the average score across the three questions as a composite subjective evaluation metric.

Table 6. Summary of human evaluation results.

Scoring Function	DU	RA	PE	Overall	Top-1 Votes
BIC	3.33±1.32	3.30±1.29	3.33±1.32	3.32±1.31	4 (13.3%)
BDeu	2.63±0.85	2.60±0.86	2.53±0.86	2.59±0.84	2 (6.7%)
K2	2.80±1.06	2.80±1.06	2.73±1.01	2.78±1.04	8 (26.7%)
IRIC	**3.53±1.17**	**3.53±1.17**	**3.47±1.14**	**3.51±1.15**	**16** (53.3%)
Scale reliability (Cronbach's α)	0.760	0.770	0.783	—	—

Note: Values are reported as mean ± standard deviation across all expert ratings. Overall denotes the average score across the three evaluation dimensions (Diagnostic Usefulness, Remediation Actionability, and Pedagogical Plausibility / Explainability).

Table 6 summarizes the mean and standard deviation of expert ratings for the four scoring functions across the three evaluation dimensions and their overall average. Overall, the knowledge structures learned by IRIC achieve higher mean ratings than the baseline methods across all three dimensions, and attain the highest overall score. This indicates that, compared with scoring functions based purely on statistical assumptions, the educational semantic constraints introduced by IRIC are more effective in generating knowledge structures that align with pedagogical reasoning and are easier to apply in learning diagnosis and remediation. It is worth noting that although BIC achieves mean ratings comparable to IRIC on certain dimensions, its standard deviations are relatively large, suggesting greater disagreement among experts regarding the usability of its learned structures. In contrast, IRIC exhibits a more concentrated rating distribution, reflecting better overall stability.

In addition to continuous ratings, experts were also asked to select, for each group of candidate structures, the single knowledge structure they considered most suitable for student diagnosis and remediation. The results show that, out of a total of 30 expert selections, IRIC-based structures were chosen as the "best structure" most frequently, demonstrating a clear advantage over the other methods. Taken together, both the subjective ratings and the top-structure selection results indicate that knowledge structures generated by IRIC not only receive higher evaluations on average, but are also more consistently regarded by experts as the most diagnostically and pedagogically valuable. These findings provide direct expert-based evidence supporting the advantages of IRIC in personalized learning diagnostic scenarios.

5 Conclusion and Future Work

In this study, we present IRIC, an IRT-grounded and interpretable scoring method for uncovering directional dependency structures among KCs. By integrating information entropy with causal-effect estimation under latent ability control and coupling this score with our Co-Evolutionary Optimization for Structural Search, the framework achieves strong performance across three diverse datasets. The learned KC graphs may support more reliable cognitive diagnosis, personalized learning path planning, and targeted resource recommendation, thereby contributing to more transparent adaptive learning pipelines. At the same time, several limitations should be noted. Although the framework shows clear advantages in adjacency recovery, edge-direction inference remains more challenging, with residual errors concentrated in near-tie pairs. In addition, because the current method is learned from observational response logs, inferred directions should be interpreted cautiously, especially in educational settings where expert-annotated KC graphs are better viewed as pedagogical reference structures rather than uniquely identifiable causal ground truth. Evidence for computational efficiency and scalability also remains limited to benchmark settings with relatively small Alarm graphs and does not yet include direct runtime or memory comparisons. Future work will therefore focus on strengthening directionality through global ordering and acyclicity priors as well as interventional proxies, improving robustness under noisy or incomplete logs,

validating the educational utility of learned KC graphs in authentic adaptive learning scenarios, and developing more actionable, human-centered explanations for teachers and students.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China under Grant 62477012, Natural Science Foundation of Shanghai under Grant 23ZR1418500, and AI for Science Program, Shanghai Municipal Commission of Economy and Informatization under Grant 2025-GZL-RGZN-BTBX-01014.

References

[1] Alberto Abadie. 2005. Semiparametric difference-in-differences estimators. *The review of economic studies* 72, 1 (2005), 1–19.
[2] Rawaa Alatrash, Mohamed Amine Chatti, Nasha Wibowo, and Qurat Ul Ain. 2026. Inferring Prerequisite Knowledge Concepts in Educational Knowledge Graphs: A Multi-criteria Approach. In *Knowledge Graphs*. Springer Nature Singapore, Singapore, 288–303.
[3] Olivier Allègre, Amel Yessad, and Vanda Luengo. 2023. Discovering Prerequisite Relationships between Knowledge Components from an Interpretable Learner Model. In *Proceedings of the 16th International Conference on Educational Data Mining*. International Educational Data Mining Society, Bangalore, India, 490–496.
[4] Dmitry Arkhangelsky, Susan Athey, David A Hirshberg, Guido W Imbens, and Stefan Wager. 2021. Synthetic difference-in-differences. *American Economic Review* 111, 12 (2021), 4088–4118.
[5] Edmund K Burke, Steven Gustafson, and Graham Kendall. 2004. Diversity in genetic programming: An analysis of measures and correlation with fitness. *IEEE Transactions on Evolutionary Computation* 8, 1 (2004), 47–62.
[6] Ruichu Cai, Siyu Wu, Jie Qiao, Zhifeng Hao, Keli Zhang, and Xi Zhang. 2022. THPs: Topological Hawkes processes for learning causal structure on event sequences. *IEEE Transactions on Neural Networks and Learning Systems* 35, 1 (2022), 479–493.
[7] Gregory Camilli. 1994. Teacher's corner: origin of the scaling constant d= 1.7 in item response theory. *Journal of Educational Statistics* 19, 3 (1994), 293–295.
[8] Erika R Carlson. 2022. Fundamental Components of Personalized Coaching Models for Faculty: Addressing Inequities in Learning Outcomes Data. *Journal of Assessment in Higher Education* 3, 1 (2022), 1–20.
[9] Lei Chen, Yiu-Ming Cheung, Hai-Lin Liu, and Yutao Lai. 2025. MOTEA-II: A Collaborative Multiobjective Transformation-Based Evolutionary Algorithm for Bilevel Optimization. *IEEE Transactions on Evolutionary Computation* 29, 2 (2025), 474–489. doi:10.1109/TEVC.2025.3538611
[10] Yetian Chen, José P. González-Brenes, and Jin Tian. 2016. Joint Discovery of Skill Prerequisite Graphs and Student Models. In *Proceedings of the 9th International Conference on Educational Data Mining*. International Educational Data Mining Society, Raleigh, NC, USA, 8 pages.
[11] Yang Chen, Pierre-Henri Wuillemin, and Jean-Marc Labat. 2015. Discovering Prerequisite Structure of Skills through Probabilistic Association Rules Mining. In *Proceedings of the 8th International Conference on Educational Data Mining*. International Educational Data Mining Society, Madrid, Spain, 117–124.
[12] David Maxwell Chickering. 2002. Optimal structure identification with greedy search. *Journal of machine learning research* 3 (2002), 507–554. Issue Nov.
[13] Sathiyaraj Chinnasamy, M Ramachandran, M Amudha, and Kurinjimalar Ramu. 2022. A review on hill climbing optimization methodology. *Recent Trends in Management and Commerce* 3, 1 (2022), 1–7.
[14] Gregory F Cooper and Edward Herskovits. 1992. A Bayesian method for the induction of probabilistic networks from data. *Machine learning* 9 (1992), 309–347.
[15] Russell Eberhart and James Kennedy. 1995. A New Optimizer Using Particle Swarm Theory. In *Proceedings of the Sixth International Symposium on Micro Machine and Human Science (MHS'95)*. IEEE, Piscataway, NJ, USA, 39–43.
[16] Zhuangyan Fang, Shengyu Zhu, Jiji Zhang, Yue Liu, Zhitang Chen, and Yangbo He. 2023. On low-rank directed acyclic graphs and causal structure learning. *IEEE Transactions on Neural Networks and Learning Systems* 35, 4 (2023), 4924–4937.
[17] Tomer Galanti, Ofir Nabati, and Lior Wolf. 2020. A critical view of the structural causal model.
[18] Dan Geiger and David Heckerman. 1994. Learning Gaussian Networks. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence (UAI-94)*. Morgan Kaufmann, Seattle, WA, USA, 235–243.
[19] Robert Glaser. 1984. Education and thinking: The role of knowledge. *American psychologist* 39, 2 (1984), 93.
[20] Clark Glymour, Kun Zhang, and Peter Spirtes. 2019. Review of causal discovery methods based on graphical models. *Frontiers in genetics* 10 (2019), 524.
[21] Jonathan Gordon, Linhong Zhu, Aram Galstyan, Prem Natarajan, and Gully Burns. 2016. Modeling Concept Dependencies in a Scientific Corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, 866–875.

[22] Subhajit Haldar, Souvik Sengupta, and Asit Kumar Das. 2025. Personalized Learning Path Recommendation using Graph Reinforcement Learning. *Procedia Computer Science* 258 (2025), 3480–3489.
[23] David Heckerman, Dan Geiger, and David M Chickering. 1995. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine learning* 20 (1995), 197–243.
[24] Miguel A Hernán and James M Robins. 2010. *Causal inference*. CRC Boca Raton, FL, California, USA.
[25] Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. 2020. Causal discovery from heterogeneous/nonstationary data. *The Journal of Machine Learning Research* 21, 1 (2020), 3482–3534.
[26] Yuan-Hao Jiang, Zi-Wei Chen, Cong Zhao, Kezong Tang, Jicong Duan, and Yizhou Zhou. 2025. Explainable Learning Outcomes Prediction: Information Fusion Based on Grades Time-Series and Student Behaviors. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, 1–11. doi:10.1145/3706599.3721212
[27] Yuan-Hao Jiang, Shang Gao, Yu-Hang Yin, Zi-Fan Xu, and Shao-Yong Wang. 2023. A control system of rail-guided vehicle assisted by transdifferentiation strategy of lower organisms. *Engineering applications of artificial Intelligence* 123 (2023), 106353.
[28] Diviyan Kalainathan, Olivier Goudet, Isabelle Guyon, David Lopez-Paz, and Michèle Sebag. 2022. Structural agnostic modeling: Adversarial learning of causal graphs. *The Journal of Machine Learning Research* 23, 1 (2022), 9831–9892.
[29] Piotr Ladyzynski, Maria Molik, and Piotr Foltynski. 2022. Dynamic Bayesian networks for prediction of health status and treatment effect in patients with chronic lymphocytic leukemia. *Scientific Reports* 12, 1 (2022), 1811.
[30] ChenYang Li, Jun Li, HuiLing Chen, and Ali Asghar Heidari. 2021. Memetic Harris Hawks Optimization: Developments and perspectives on project scheduling and QoS-aware web service composition. *Expert Systems with Applications* 171 (2021), 114529.
[31] Sheng Li, Quanlong Guan, Liangda Fang, Fang Xiao, Zhenyu He, Yizhou He, and Weiqi Luo. 2022. Cognitive Diagnosis Focusing on Knowledge Concepts. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*. ACM, Atlanta, GA, USA, 3272–3281.
[32] Xuechun Li, Paula M. Bürgi, Wei Ma, Hae Young Noh, David Jay Wald, and Susu Xu. 2023. DisasterNet: Causal Bayesian Networks with Normalizing Flows for Cascading Hazards Estimation from Satellite Imagery. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, Long Beach, CA, USA, 4391–4403.
[33] Xiaoyu Li, Shaoyang Guo, Jin Wu, and Chanjin Zheng. 2025. An interpretable polytomous cognitive diagnosis framework for predicting examinee performance. *Information Processing & Management* 62, 1 (Jan. 2025), 103913. doi:10.1016/j.ipm.2024.103913
[34] Chen Liang, Jianbo Ye, Shuting Wang, Bart Pursel, and C. Lee Giles. 2018. Investigating Active Learning for Concept Prerequisite Learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, Vol. 32. AAAI Press, New Orleans, LA, USA, 7913–7919.
[35] Chen Liang, Jianbo Ye, Zhaohui Wu, Bart Pursel, and C. Lee Giles. 2017. Recovering Concept Prerequisite Relations from University Course Dependencies. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, Vol. 31. AAAI Press, San Francisco, CA, USA, 4786–4791.
[36] Jianqing Lin, Cheng He, and Ran Cheng. 2022. Adaptive dropout for high-dimensional expensive multiobjective optimization. *Complex & Intelligent Systems* 8, 1 (2022), 271–285.
[37] Jing Liu, Ruhul Sarker, Saber Elsayed, Daryl Essam, and Nurhadi Siswanto. 2024. Large-scale evolutionary optimization: A review and comparative study. *Swarm and Evolutionary Computation* 85 (2024), 101466.
[38] Fei Ming, Wenyin Gong, Ling Wang, and Liang Gao. 2022. Balancing convergence and diversity in objective and decision spaces for multimodal multi-objective optimization. *IEEE Transactions on Emerging Topics in Computational Intelligence* 7, 2 (2022), 474–486.
[39] Amir Hossein Nabizadeh, José Paulo Leal, Hamed N Rafsanjani, and Rajiv Ratn Shah. 2020. Learning path personalization and recommendation methods: A survey of the state-of-the-art. *Expert Systems with Applications* 159 (2020), 113596.
[40] Ignavier Ng, Yujia Zheng, Jiji Zhang, and Kun Zhang. 2021. Reliable causal discovery with improved exact search and weaker assumptions. *Advances in Neural Information Processing Systems* 34 (2021), 20308–20320.
[41] Ignavier Ng, Shengyu Zhu, Zhitang Chen, and Zhuangyan Fang. 2019. A graph autoencoder approach to causal structure learning.
[42] Ana Rita Nogueira, Andrea Pugnana, Salvatore Ruggieri, Dino Pedreschi, and João Gama. 2022. Methods and tools for causal discovery and causal inference. *WIREs Data Mining and Knowledge Discovery* 12, 2 (2022), e1449. doi:10.1002/widm.1449
[43] Liangming Pan, Chengjiang Li, Juanzi Li, and Jie Tang. 2017. Prerequisite Relation Learning for Concepts in MOOCs. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, 1447–1456.
[44] Judea Pearl. 1995. Causal diagrams for empirical research. *Biometrika* 82, 4 (1995), 669–688.
[45] Judea Pearl. 2009. *Causality*. Cambridge university press, Cambridge, United Kingdom.
[46] Judea Pearl and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic Books, New York, NY, USA.
[47] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, USA.

[48] Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983), 41–55.
[49] Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. 2019-11. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances* 5, 11 (2019-11), eaau4996. doi:10.1126/sciadv.aau4996
[50] Karam M. Sallam, Saber M. Elsayed, Ripon K. Chakrabortty, and Michael J. Ryan. 2020. Improved Multi-Operator Differential Evolution Algorithm for Solving Unconstrained Problems. In *Proceedings of the 2020 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, Piscataway, NJ, USA, 1–8.
[51] Richard Scheines, Elizabeth Silver, and Ilya M. Goldin. 2014. Discovering Prerequisite Relationships Among Knowledge Components. In *Proceedings of the 7th International Conference on Educational Data Mining*. International Educational Data Mining Society, London, United Kingdom, 355–356.
[52] Gideon Schwarz. 1978. Estimating the Dimension of a Model. *The Annals of Statistics* 6, 2 (1978), 461–464.
[53] Flávio Luiz Seixas, Bianca Zadrozny, Jerson Laks, Aura Conci, and Débora Christina Muchaluat Saade. 2014. A Bayesian network decision model for supporting the diagnosis of dementia, Alzheimer's disease and mild cognitive impairment. *Computers in biology and medicine* 51 (2014), 140–158.
[54] Rajen D. Shah and Jonas Peters. 2020. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics* 48, 3 (2020), 1514 – 1538. doi:10.1214/19-AOS1857
[55] Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvarinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, Kenneth Bollen, and Patrik Hoyer. 2011. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research-JMLR* 12 (2011), 1225–1248. Issue Apr.
[56] Peter Spirtes, Clark N Glymour, and Richard Scheines. 2000. *Causation, prediction, and search*. MIT Press, Cambridge, MA, USA.
[57] Rainer Storn and Kenneth Price. 1997. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization* 11 (1997), 341–359.
[58] Xinjie Sun, Qi Liu, Kai Zhang, Shuanghong Shen, Fei Wang, Yan Zhuang, Zheng Zhang, Weiyin Gong, Shijin Wang, Lina Yang, et al. 2025. HCD: A Hierarchy Constraint-Aware Neural Cognitive Diagnosis Framework. *Neural Networks* 190 (2025), 107668.
[59] Ke Tang, Shengcai Liu, Peng Yang, and Xin Yao. 2021. Few-Shots Parallel Algorithm Portfolio Construction via Co-Evolution. *IEEE Transactions on Evolutionary Computation* 25, 3 (2021), 595–607. doi:10.1109/tevc.2021.3059661
[60] Kezong Tang, Xiong-Fei Wei, Yuan-Hao Jiang, Zi-Wei Chen, and Lihua Yang. 2023. An adaptive ant colony optimization for solving large-scale traveling salesman problem. *Mathematics* 11 (2023), 4439.
[61] Xinlu Tang, Chencheng Zhang, Rui Guo, Xinling Yang, and Xiaohua Qian. 2023. A causality-aware graph convolutional network framework for rigidity assessment in parkinsonians. *IEEE Transactions on Medical Imaging* 43, 1 (2023), 229–240.
[62] Ye Tian, Ran Cheng, Xingyi Zhang, and Yaochu Jin. 2017. PlatEMO: A MATLAB platform for evolutionary multi-objective optimization. *IEEE Computational Intelligence Magazine* 12, 4 (2017), 73–87.
[63] Ye Tian, Langchun Si, Xingyi Zhang, Ran Cheng, Cheng He, Kay Chen Tan, and Yaochu Jin. 2021. Evolutionary large-scale multi-objective optimization: A survey. *ACM Computing Surveys (CSUR)* 54, 8 (2021), 1–34.
[64] Ye Tian, Weijian Zhu, Xingyi Zhang, and Yaochu Jin. 2023. A practical tutorial on solving optimization problems via PlatEMO. *Neurocomputing* 518 (2023), 190–205.
[65] Peking University and Huawei Noah's Ark Lab. 2021. PCIC 2021: Causal Discovery. https://competition.huaweicloud.com/information/1000041487/dataset
[66] Jon Wade, Steven Dow, Hortense Gerardo, Richard Gessner, and Jace Hargis. 2022. Closed-Loop Dynamically Adaptive Educational Systems. In *INCOSE International Symposium*, Vol. 32. Wiley Online Library, Hoboken, NJ, USA, 109–118.
[67] Miao Wang, Chunchen Liu, and Geng Zhi. 2018. Statistical methods for causal inference. *Science China Mathematics* 48, 12 (2018), 1753–1778.
[68] Shuting Wang, Alexander Ororbia, Zhaohui Wu, Kyle Williams, Chen Liang, Bart Pursel, and C Lee Giles. 2016. Using prerequisites to extract concept maps fromtextbooks. In *Proceedings of the 25th acm international on conference on information and knowledge management*. ACM, Indianapolis, IN, USA, 317–326.
[69] Tianyou Wang and Lingjia Zeng. 1998. Item parameter estimation for a continuous response model using an EM algorithm. *Applied Psychological Measurement* 22, 4 (1998), 333–344.
[70] Wenyi Wang, Juanjuan Zheng, Lihong Song, Yukun Tu, and Peng Gao. 2021. Test Assembly for Cognitive Diagnosis Using Mixed-Integer Linear Programming. *Frontiers in Psychology* 12 (Feb. 2021), 14 pages. doi:10.3389/fpsyg.2021.623077
[71] Xiong-Fei Wei, Kezong Tang, Xiao-Bin Chen, Yuan-Hao Jiang, and Xinyun Wang. 2024. EduEvolve: An Application Framework of Evolutionary Computation in Assessing and Improving Learning Outcomes. In *Enhancing Educational Practices: Strategies for Assessing and Improving Learning Outcomes*, Yuang Wei, Changyong Qi, Yuan-Hao Jiang, and Ling Dai (Eds.). Nova Science Publishers, New York, NY, USA, 67–85. https://doi.org/10.52305/RUIG5131
[72] Yuang Wei and Bo Jiang. 2024. Interpretable Cognitive State Prediction via Temporal Fuzzy Cognitive Map. *IEEE Transactions on Learning Technologies* 17 (2024), 514–526. doi:10.1109/TLT.2023.3307565

[73] Yuang Wei, Yizhou Zhou, Yuan-Hao Jiang, and Bo Jiang. 2024. Enhancing Explainability of Knowledge Learning Paths: Causal Knowledge Networks. In *Joint Proceedings of the Human-Centric eXplainable AI in Education and the Leveraging Large Language Models for Next Generation Educational Technologies Workshops (HEXED-L3MNGET 2024) co-located with 17th International Conference on Educational Data Mining (EDM 2024)*. International Educational Data Mining Society, Atlanta, Georgia, USA, 9–17. doi:10.48550/arXiv.2406.17518

[74] Pengzhou Wu and Kenji Fukumizu. 2020. Causal mosaic: Cause-effect inference via nonlinear ica and ensemble method. In *International Conference on Artificial Intelligence and Statistics*. PMLR, Palermo, Sicily, Italy, 1157–1167.

[75] Hongteng Xu, Mehrdad Farajtabar, and Hongyuan Zha. 2016. Learning granger causality for hawkes processes. In *International conference on machine learning*. PMLR, New York City, New York, USA, 1717–1726.

[76] Ying Xu, Chong Xu, Huan Zhang, Lei Huang, Yiping Liu, Yusuke Nojima, and Xiangxiang Zeng. 2023. A Multi-Population Multi-Objective Evolutionary Algorithm Based on the Contribution of Decision Variables to Objectives for Large-Scale Multi/Many-Objective Optimization. *IEEE Transactions on Cybernetics* 53, 11 (2023), 6998–7007. doi:10.1109/TCYB.2022.3180214

[77] Ziqian Xu and Sheng Jiang. 2022. Study on personalized recommendation algorithm of online educational resources based on knowledge association. *Computational intelligence and neuroscience* 2022, 1 (2022), 2192459.

[78] Hong Qing Yu and Stephan Reiff-Marganiec. 2022. Learning disease causality knowledge from the web of health data. *International Journal on Semantic Web and Information Systems (IJSWIS)* 18, 1 (2022), 1–19.

[79] Yue Yu, Jie Chen, Tian Gao, and Mo Yu. 2019. DAG-GNN: DAG structure learning with graph neural networks. In *International Conference on Machine Learning*. PMLR, Long Beach, California, USA, 7154–7163.

[80] Feng Zhang, Xuguang Feng, and Yibing Wang. 2024. Personalized process–type learning path recommendation based on process mining and deep knowledge tracing. *Knowledge-Based Systems* 303 (2024), 112431.

[81] Juntao Zhang, Hai Lan, Xiandi Yang, Shuaichao Zhang, Wei Song, and Zhiyong Peng. 2022. Weakly supervised setting for learning concept prerequisite relations using multi-head attention variational graph auto-encoders. *Knowledge-Based Systems* 247 (2022), 108689.

[82] Keli Zhang. 2024. Gcastle-Hub Dataset. https://github.com/gcastle-hub/dataset

[83] Lei Zhang, Huabin Zhang, Zihao Chen, Sibo Liu, Haipeng Yang, and Hongke Zhao. 2024. A Multi-Population Based Evolutionary Algorithm for Many-Objective Recommendations. *IEEE Transactions on Emerging Topics in Computational Intelligence* 8, 2 (2024), 1969–1982. doi:10.1109/TETCI.2024.3359093

[84] Yixuan Zhang and Yanyi Wang. 2025. A personalized recommendation algorithm for English exercises incorporating fuzzy cognitive models and multiple attention mechanisms. *Scientific Reports* 15, 1 (2025), 11531.

[85] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. 2018. DAGs with NO TEARS: Continuous Optimization for Structure Learning. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc., Montréal, Canada.

[86] Yujia Zheng, Ignavier Ng, and Kun Zhang. 2022. On the identifiability of nonlinear ica: Sparsity and beyond. *Advances in Neural Information Processing Systems* 35 (2022), 16411–16422.

[87] Ke Zhou, Hongyuan Zha, and Le Song. 2013. Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In *Artificial Intelligence and Statistics*. PMLR, Scottsdale, Arizona, USA, 641–649.

[88] MengChu Zhou, Meiji Cui, Dian Xu, Shuwei Zhu, Ziyan Zhao, and Abdullah Abusorrah. 2024. Evolutionary optimization methods for high-dimensional expensive problems: A survey. *IEEE/CAA Journal of Automatica Sinica* 11, 5 (2024), 1092–1105.

A Method Details

A.1 Derivation of the IRIC Criterion

Notation. Let $\mathcal{G}$ be a candidate graph (DAG) over n KCs $\{X_1, \dots, X_n\}$ and $\mathcal{D}$ be a dataset of size N. For X_i with parents $Pa^{\mathcal{G}}_{X_i}$, denote realizations by x_i and u_i for X_i and $U_i \equiv Pa^{\mathcal{G}}_{X_i}$, respectively. Let $M[\cdot]$ be empirical counts and $\hat{P}$ the empirical distribution.

Conventions. All logarithms are natural (units in nats). For any set of discrete variables V, $Val(V)$ denotes the Cartesian product of their value sets. We write $Pa^{\mathcal{G}}_{X_i}$ for the parent set of X_i under $\mathcal{G}$ and abbreviate it as Pa_{X_i} when unambiguous. The logistic function is $\sigma(z) = 1/(1+e^{-z})$. The clipping operator is $\text{clip}(z, a, b) = \min\{\max\{z, a\}, b\}$. Unless otherwise stated, empirical probabilities $\hat{P}$ are estimated with Dirichlet (Laplace) smoothing to avoid zeros in MI/entropy estimators.

IRIC definition.

$$\text{score}_{IRIC}(\mathcal{G} : \mathcal{D}) = \ell(\hat{\theta}_{\mathcal{G}} : \mathcal{D}) + \text{ATE}(\text{edge}_{\mathcal{G}}) - \frac{\log N}{2} Dim[\mathcal{G}] \tag{13}$$

Local conditional likelihood.

$$\ell(\hat{\theta}_{\mathcal{G}} : \mathcal{D}) = \sum_{i=1}^{n} \left[\sum_{u_i \in Val(Pa^{\mathcal{G}}_{X_i})} \sum_{x_i} M[x_i, u_i] \log \hat{\theta}_{x_i|u_i} \right] \tag{14}$$

where $\hat{\theta}_{x_i|u_i} = \hat{P}(x_i \mid u_i)$ and $M[x_i, u_i] = N\hat{P}(x_i, u_i)$, with $\hat{P}$ computed under a Dirichlet prior (add-one smoothing) to ensure nonzero support.

$$\sum_{u_i \in Val(Pa^{\mathcal{G}}_{X_i})} \sum_{x_i} M[x_i, u_i] \log \hat{\theta}_{x_i|u_i} \to \frac{1}{N} \sum_{u_i} \sum_{x_i} M[x_i, u_i] \log \hat{\theta}_{x_i|u_i} \tag{15}$$

Let $\hat{P}$ be the empirical distribution where $M[x, y] = N \cdot \hat{P}(x, y)$ and $\hat{\theta}_{x_i|u_i} = \hat{P}(x_i \mid u_i)$. Then

$$\frac{1}{N} \sum_{u_i} \sum_{x_i} M[x_i, u_i] \log \hat{\theta}_{x_i|u_i} \tag{16}$$

$$= \sum_{u_i} \sum_{x_i} \hat{P}(x_i, u_i) \log \hat{P}(x_i|u_i) \tag{17}$$

$$= \sum_{u_i} \sum_{x_i} \hat{P}(x_i, u_i) \log \left(\frac{\hat{P}(x_i, u_i)\hat{P}(x_i)}{\hat{P}(u_i)\hat{P}(x_i)} \right) \tag{18}$$

$$= \sum_{u_i} \sum_{x_i} \hat{P}(x_i, u_i) \log \frac{\hat{P}(x_i, u_i)}{\hat{P}(u_i)\hat{P}(x_i)} + \sum_{x_i} \left(\sum_{u_i} \hat{P}(x_i, u_i) \right) \log \hat{P}(x_i) \tag{19}$$

$$= I_{\hat{P}}(X_i; U_i) - \sum_{x_i} \hat{P}(x_i) \log \frac{1}{\hat{P}(x_i)} \tag{20}$$

$$= I_{\hat{P}}(X_i; U_i) - H_{\hat{P}}(X_i) \tag{21}$$

If $Pa_{X_i} = \emptyset$, then $I_{\hat{P}}(X_i; Pa_{X_i}) = 0$. Using (21), the local conditional likelihood $\ell(\hat{\theta}_{\mathcal{G}} : \mathcal{D})$ can be rewritten as $N \sum_i \big(I_{\hat{P}}(X_i; Pa_{X_i}) - H_{\hat{P}}(X_i) \big)$. We further include the IRT-driven latent entropy $H_{\hat{Y}}(\widetilde{X}_i)$ to capture predictive uncertainty in responses, yielding the expanded IRIC below.

Expanded IRIC.

$$\begin{aligned}\text{score}_{IRIC}(\mathcal{G} : \mathcal{D}) = N \sum_{i=1}^{n} \Big(I_{\hat{P}}(X_i; Pa_{X_i}) + \text{ATE}(X_i; Pa_{X_i}) \Big) \\ - N \sum_{i=1}^{n} \Big(H_{\hat{P}}(X_i) + H_{\hat{Y}}(\tilde{X}_i) \Big) - \frac{\log N}{2} Dim[\mathcal{G}]\end{aligned} \tag{22}$$

where $I_{\hat{P}}(X_i; Pa_{X_i})$ is mutual information under the empirical pmf $\hat{P}$, $H_{\hat{P}}(X_i)$ is the Shannon entropy of X_i under $\hat{P}$, and $H_{\hat{Y}}(\widetilde{X}_i)$ is the Bernoulli entropy computed from IRT-predicted correctness $\hat{Y}$ (defined below). Pa_{X_i} denotes the parent set of X_i, and $Dim[\mathcal{G}]$ is the number of free parameters implied by $\mathcal{G}$. The dataset size is N.

All components are mapped to a common information-theoretic scale:

Association and observation entropies. Let $\hat{P}$ be the empirical distribution estimated from $\mathcal{D}$. We compute $I_{\hat{P}}(X_i; Pa_{X_i}) = \sum_{x_i, p_i} \hat{P}(x_i, p_i) \log \frac{\hat{P}(x_i, p_i)}{\hat{P}(x_i)\hat{P}(p_i)}$, $\quad H_{\hat{P}}(X_i) = -\sum_{x_i} \hat{P}(x_i) \log \hat{P}(x_i)$, and aggregate them with a factor N so that the terms are $\mathcal{O}(N)$. Multiplying by N aligns the order of magnitude with sample-aggregated information, so that all additive terms scale consistently with dataset size.

IRT-driven latent entropy. For each response (student s, item i), the 2PL IRT gives $\hat{Y}_{si} = \sigma\big(D\, a_i(\vartheta_s - b_i)\big)$ with D=1.702. We define the Bernoulli entropy and average across observed responses:

$$H_{\hat{Y}}(\widetilde{X}_i) = -\sum_{x_i \in X_i} \hat{P}(x_i) \, \frac{1}{N_{x_i}} \sum_{s \in \mathcal{S}_{x_i}} h(\hat{Y}_{s\, x_i}), \tag{23}$$

$$h(x) = -x \log x - (1 - x) \log(1 - x), \tag{24}$$

where $\mathcal{S}_i$ is the set of students who attempted item i and N_{x_i}=$|\mathcal{S}_i|$.

Causal-effect term and normalization. We estimate an ATE-like effect for node X_i given its current parent set Pa_{X_i} via back-door adjustment (Sec. 3.1), obtaining a *risk difference* $\text{ATE}_i \in [-1, 1]$. To place it on an information scale and avoid introducing extra hyper-parameters, we use a symmetric logit mapping with clipping:

$$\tilde{a}_i = \text{clip}\Big(\frac{\text{ATE}_i + 1}{2}, \, \delta, \, 1 - \delta \Big), \quad \text{CE}_i = \log \frac{\tilde{a}_i}{1 - \tilde{a}_i}, \tag{25}$$

Here, CE_i denotes the causal effect score of node X_i given its parent set Pa_{X_i}. This score is obtained by applying probabilistic normalization and a logit mapping to the ATE, and is used to characterize directional causal influence on an information-theoretic scale. The parameter δ is set to a small value (e.g., 10^{-6}) to prevent divergence near the boundaries. Let $m = \min_j \text{CE}_j$ and $M = \max_j \text{CE}_j$. If $M > m$, we normalize CE_i as $\text{CE}_i \leftarrow (\text{CE}_i - m)/(M - m)$; otherwise, all CE_i are set to 0. Finally, we aggregate $N \sum_i \text{CE}_i$ so that its magnitude is comparable to that of the association and entropy terms.

Table 7. Sensitivity analysis of the parameter δ in IRIC

δ	10^{-12}	10^{-9}	10^{-6}	10^{-4}	10^{-3}
Number of edges	12	12	12	12	12
ΔE (relative to $\delta = 10^{-6}$)	0	0	0	0	0

Sensitivity analysis on δ shows that varying δ across several orders of magnitude does not change the causal structure selected by IRIC under the same structural search procedure, compared with the structure obtained

when $\delta = 10^{-6}$. As shown in Table 7, ΔE denotes the number of edge differences between the structure obtained with the current δ and the structure obtained when $\delta = 10^{-6}$.
Complexity penalty. We adopt the standard $\frac{\log N}{2} Dim[\mathcal{G}]$ form as in BIC to penalize model size.

A.2 CEO-SS Algorithm

The complete procedure of CEO-SS is summarized in Algorithm 1.

Algorithm 1 Co-Evolutionary Optimization for Structural Search

Require: The ambient pressure AP, the $PopSize$ and the $maxFE$ of the population.
Ensure: The obtained best Graph Structure GS_{best}.

Initialization: $FE \leftarrow 0$;
Randomly initialize the population Pop;
Pop.Evaluate() `// Compute the fitness of each individual in the initial population.`
$FE \leftarrow FE + PopSize$;
$EdagNum \leftarrow (PointNum \cdot (PointNum - 1))/2$;
while $FE \leq maxFE$ **do**
  Pop.Sort() `// Sort all individuals in the population from best to worst.`
  $SupPop \leftarrow Pop[1 : \lfloor AP \cdot PopSize \rfloor]$ `// Get sub-populations.`
  $DevPop \leftarrow Pop[1 : PopSize - \lfloor AP \cdot PopSize \rfloor]$;
  $EliPop \leftarrow Pop \setminus DevPop$;
  for $i = 1$ **to** len($DevPop$) **do**
    Calculate ranking ratio R of $DevPop(i)$ in $DevPop$;
    Randomly select $\lfloor R \cdot EdagNum \rfloor$ pairs of nodes in $DevPop(i)$ to mutate;
    $DevPop(i)$.Repair();
  end for
  $DevPop$.Evaluate(); `// Compute the fitness of individuals in the growing sub-population.`
  $FE \leftarrow FE +$ len($DevPop$);
  $EliPop$.Delete();
  $Pop \leftarrow SupPop \cup DevPop$ `// Offspring for next iteration.`
end while
Pop.Sort();
$I_{best} \leftarrow Pop$.Best() `// Return the best individual.`
$GS_{best} \leftarrow I_{best}$.Decode();

To further verify the significant advantage of the CEO-SS algorithm in avoiding evolutionary stagnation, we conducted an intuitive horizontal comparison between the evolutionary curves of our method and those of the three best-performing baseline algorithms—DE, PSO, and IMODE—which each achieved the second-best result on one of the three adopted sub-datasets. From the comparative evolution curves, it can be clearly observed that the baseline algorithms exhibit typical premature convergence and local stagnation phenomena. Specifically, the PSO algorithm stops achieving substantive fitness improvements at an early stage of evolution and falls into a prolonged plateau across all datasets. In addition, although the DE and IMODE algorithms do not stagnate as early or as completely as PSO, their evolutionary trajectories still present clear step-like plateau periods. In contrast, throughout the entire cycle of 10,000 function evaluations, the F1 score of CEO-SS maintains a stable and continuous upward trend across all three datasets of different scales. This intuitive comparison strongly demonstrates that, through its multi-subpopulation collaborative mechanism, CEO-SS successfully and efficiently

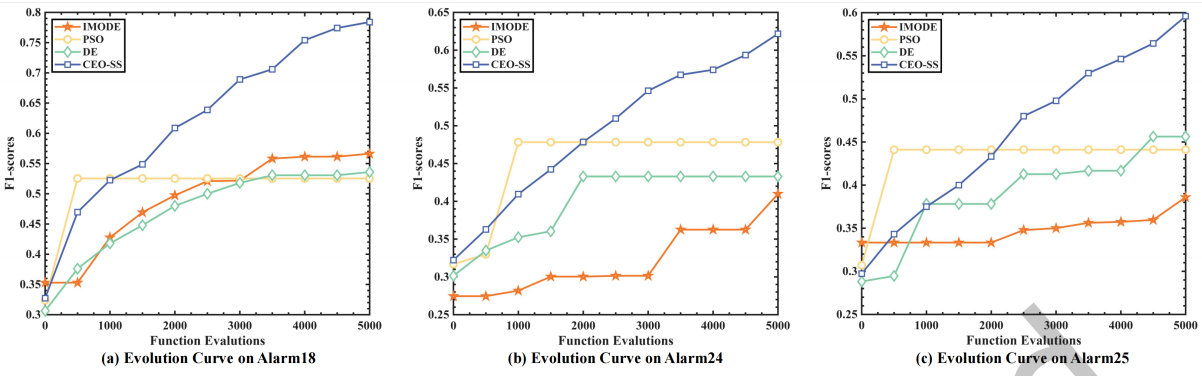


Fig. 6. Convergence analysis of baseline algorithms. (a), (b), and (c) present the evolutionary curves (EC) of the proposed CEO-SS and baseline methods on the Alarm18, Alarm24, and Alarm25 datasets, respectively.

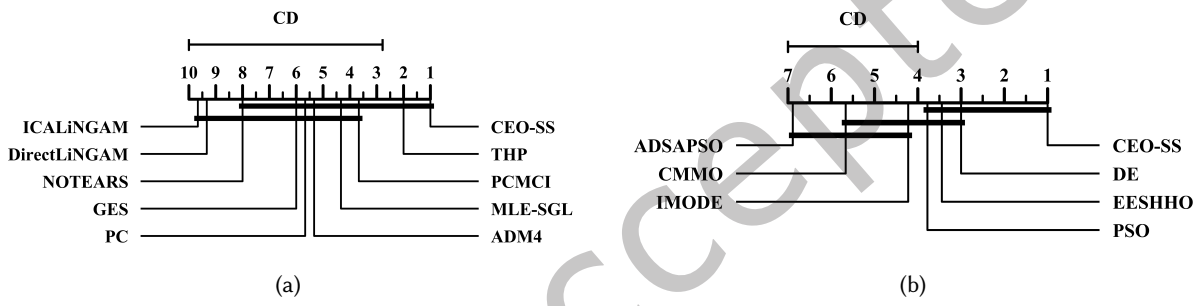


Fig. 7. Friedman test results. (a) Comparison of causal structure learning methods based on F1-scores across all datasets. (b) Comparison of meta-heuristic algorithms based on the fitness of optimal individuals across datasets. In both subfigures, CD denotes the critical difference, and a smaller rank indicates better performance.

maintains exploration momentum within the search space, thereby enabling continuous optimization of the global network structure.

B Supplementary Experiments

B.1 Friedman Tests for Statistical Significance

To complement the empirical comparisons in the main text, we conducted Friedman tests to assess whether the observed performance differences across datasets were statistically significant.

Comparison with causal structure learning methods. Based on the F1-scores of all causal structure learning methods (Figure 7a), CEO-SS achieved the best average rank, followed by THP. Apart from these two methods, no statistically significant differences were found among the remaining baselines. These results further support the consistent advantage of CEO-SS over existing causal structure learning approaches.

Comparison with meta-heuristic algorithms. We further applied the Friedman test to the meta-heuristic baselines (Figure 7b). CEO-SS again ranked first, confirming its overall superiority. Among the remaining methods, DE showed the strongest performance and was not significantly different from EESHHO, PSO, IMODE, or CMMO at the 0.05 level. In contrast, CEO-SS significantly outperformed ADSAPSO, CMMO, and IMODE. Although

IMODE showed smaller variance and faster convergence on some datasets, CEO-SS achieved a better balance between convergence and fitness diversity.

Received 19 October 2025; revised 22 March 2026; accepted 20 April 2026