IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 21 August 2024, accepted 3 September 2024, date of publication 9 September 2024, date of current version 30 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3456233

RESEARCH ARTICLE

Causal Discovery Evaluation Framework in the Absence of Ground-Truth Causal Graph

TINGPENG LI[1], LEI WANG[2], DANHUA PENG[1], JUN LIAO[2], LI LIU[2], (Member, IEEE), AND ZHENDONG LIU[3]

[1]State Key Laboratory of Complex Electromagnetic Environmental Effects on Electronics and Information System, Luoyang 471003, China
[2]School of Big Data and Software Engineering, Chongqing University, Chongqing 401331, China
[3]iFLYTEK Big Data College, Chongqing City Vocational College, Chongqing 402160, China

Corresponding authors: Zhendong Liu (lzd033353@cqcvc.edu.cn) and Li Liu (dcsliuli@cqu.edu.cn)

This work was supported in part by the State Key Laboratory of Complex Electromagnetic Environment Effects on Electronics and Information System under Grant CEMEE2023G0202, in part by the National Major Science and Technology Projects of China under Grant 2022YFB3303302, in part by the National Natural Science Foundation of China under Grant 62207007, and in part by the Natural Science Foundation of Chongqing under Grant CSTB2022NSCQ-MSX1256.

ABSTRACT In causal learning, discovering the causal graph of the underlying generative mechanism from observed data is crucial. However, real-world data for causal discovery is scarce and expensive, leading researchers to rely on synthetic datasets, which may not accurately reflect real-world performance. To address this, we propose a novel method for evaluating causal discovery algorithms without needing real causal graphs. Specifically, our method employs deep learning evaluation strategies and ensemble learning techniques to robustly assess the performance of causal discovery methods. To elaborate, our approach emulates deep learning validation strategies by dividing the data into training and testing sets. We perform causal discovery on the training set and subsequently use the testing set to conduct Markov blanket tests on the node set and causal direction determination on the edge set. Moreover, we employ multiple ensemble strategies to ensure a comprehensive evaluation of the algorithms. Furthermore, experiments on both synthetic and real datasets demonstrate our method's effectiveness in accurately and comprehensively validating causal discovery algorithms. Our results show that our proposed method can reflect the performance of causal discovery methods in practice with reasonable error.

INDEX TERMS Causal discovery, Markov blanket, causal graphical models, cause effect identification, condition independence testing.

I. INTRODUCTION

Deep learning has achieved remarkable success across various domains, revolutionizing fields such as computer vision [1], natural language processing [2], and healthcare [3]. The capacity of deep learning models to discern intricate patterns from extensive datasets has driven artificial intelligence to unparalleled levels of progress. The emergence of large models has further elevated artificial intelligence to new heights [4], [5], [6]. However, behind this success lies a crucial question: do these models truly comprehend the underlying causal relationships in the data, or are they merely capturing correlations [7]? While deep learning models excel at learning correlations, their reliance solely on correlations poses limitations when it comes to understanding causal relationships [7], [8], [9]. Correlation does not imply causation, and blindly relying on correlations can lead to erroneous conclusions and suboptimal decisions [10]. For instance, consider a model trained to predict patient outcomes based on medical records. It may learn that certain symptoms or treatments are correlated with improved outcomes, but without understanding the underlying causal mechanisms, it could misguide medical interventions.

To address these limitations and move towards more robust and interpretable learning systems, there is a growing interest in causal learning. Causal learning is a branch of

The associate editor coordinating the review of this manuscript and approving it for publication was Jolanta Mizera-Pietraszko.

statistics focused on understanding and identifying cause-and-effect relationships from data. Unlike traditional machine learning, which primarily identifies correlations and patterns, causal learning aims to discern how changes in one variable can directly influence another [11]. As data-driven decision-making becomes increasingly prevalent, the ability to accurately determine causation is critical for reliable and actionable insights [12], [13]. Causal learning can be roughly divided into causal discovery methods, which reveal the underlying causal structure of data, and causal inference methods employ intervention and counterfactual inference on the variables of interest. This paper primarily discusses the issues of causal discovery and its evaluation criteria.

The goal of causal discovery is to explore the causal relationships between observed variables and represent these relationships using directed acyclic graphs (DAGs) [14]. This approach is advantageous because it helps us understand causal structures in complex systems, enabling interventions and counterfactual reasoning on causal graphs, which in turn leads to more accurate and reliable predictions and decisions. However, despite the theoretical potential of causal discovery methods, they face several practical challenges and limitations. One major challenge is the absence of robust evaluation metrics. Although randomized controlled trials (RCTs) [15] are considered the gold standard for evaluating causal discovery due to their ability to minimize bias and establish causation, they are often impractical due to selection bias and high costs. As a result, existing evaluation criteria mainly rely on simulated causal mechanisms and real-world dataset evaluations. Most causal discovery research methods adopt simulated evaluations [16], [17], [18], [19] because collecting real-world datasets, especially in fields with sparse data such as the military and meteorology, is challenging. However, relying solely on synthetic or semi-synthetic datasets (simulated causal mechanisms) for assessing causal discovery methods can limit their applicability to the complexities and diversity of real-world scenarios [20]. Therefore, there is a pressing need for more comprehensive and effective evaluation measures that can more accurately assess the performance of causal discovery methods and the reliability of their inference results.

In this paper, we propose a method to evaluate causal discovery in the absence of real causal graphs. Inspired by the evaluation methods in deep learning, we adopt the idea of dividing data into training and testing sets. We first run causal discovery algorithms on the training set to obtain corresponding causal graphs, and then validate the learned causal graphs on the testing set. The intuition is that theoretically, the underlying causal structure from the same batch of data remains unchanged. Additionally, we conduct Markov blanket tests on the testing set to detect the rationality of the learned causal graphs. Subsequently, we extract each edge from the causal graph to form a set of causal pairs, and utilize four causal pair identification algorithms to determine the causal directions using the testing set. Finally, we employ three ensemble strategies to integrate the results of the four identification algorithms and obtain the final validation accuracy. To the best of our knowledge, this is the first paper to propose a method for evaluating causal discovery in the absence of real causal graphs. Our proposed evaluation framework is shown in Figure 1, and the contributions are summarized as follows:

- We propose a method to evaluate the performance of causal discovery algorithms in the absence of real causal graphs. This method divides data into training and testing sets and comprehensively evaluates the performance of causal learning algorithms.
- The comprehensive evaluation of the learned causal graph includes using multiple causal identification algorithms and result integration strategies. This method not only evaluates the causal graph at the overall level but also at the level of each edge, thereby obtaining a more comprehensive evaluation result.
- Through effectiveness testing on synthetic and real datasets, we verify that our method can reflect the true performance of causal learning algorithms to a certain extent. Additionally, we perform multiple ensembles, employing a balanced strategy between stability and true accuracy and recall, for evaluation. Experimental results demonstrate that our proposed method can effectively evaluate the accuracy and generalization of causal graphs.

II. RELATED WORKS

A. CAUSAL DISCOVERY

In past research, causal discovery methods can be roughly categorized into five types based on their assumptions, optimizations, and other characteristics. (1) Constraint-based causal discovery relies on conditional independence assumptions and the V-structures [12] defined by structural causal models to identify causal relationships among variables. Representative methods include PC [21], FCI [21], and CD-NOD [16]. (2) Score-based causal discovery combines specific scoring functions (such as BIC [22], BEDU [23], AIC [24], etc.) with conditional independence to iteratively manipulate the edges of an initialized causal skeleton. These methods calculate the score of causal graphs using scoring functions to select the optimal causal graph. Representative methods include GES [25], SGES [26], and RL-BLC [27]. (3) Function-based causal discovery builds causal relationships among variables by constructing function causal models and then utilizes function properties or distributional properties of variables to uncover causal relationships. Representative methods include LiNGAM [17], ANM [18], and CAM [28]. (4) Continuous optimization-based causal discovery is an improvement over score-based methods. While score-based methods search for optimization in discrete Markov equivalence class spaces, continuous optimization-based methods leverage neural networks to construct a continuous search space by combining acyclic constraints and optimization problems. This integration with neural networks harnesses their powerful representation learning capabilities.

Representative methods include NOTEARS [19], DAG-GNN [29], BayesDAG [30], BCDNet [31], SAM [32] and ENCO [33]. (5) Prior knowledge-based and hybrid methods allow the incorporation of domain expertise during the causal discovery process and involve combining multiple types of methods to form multi-stage causal discovery approaches. Representative methods include MMHC [34] and JCI [35].

B. RESEARCH ON CAUSAL DISCOVERY BENCHMARKS

One of the most significant challenges in the field of causal discovery in practice is the lack of benchmark datasets, especially in the absence of real causal graphs. Researchers have made the following contributions to address this issue: (1) Simulation of causal mechanisms: Due to the scarcity of real causal datasets, synthetic datasets generated by simulating causal mechanisms are often widely used as benchmark datasets for evaluating causal discovery methods. For example, datasets simulated through Erdős-Rényi(ER) models [36] and Scale-Free(SF) models [37], synthetic datasets collected in Bayesian network repositories,[1] synthetic datasets used for studying gene regulatory networks such as DREAM3 [38] and CausalTime [39], and time series datasets generated by the causalLens framework [40]. (2) Semi-real datasets: These are benchmark datasets generated by simulating causal mechanisms in specific domains, where the foundation of these datasets lies in real data. For instance, simulated BOLD fMRI datasets [41], double pendulum datasets [42], synthetic twin birth datasets [43], alarm message systems used for patient monitoring [33], [44], neural activity data [42], and gene expression data [45]. (3) Real datasets: While naturally collected datasets for causal discovery are rare in reality, they are valuable as they often reflect the complexity and diversity of the real world, which synthetic datasets struggle to fully simulate. Common real datasets include CauseEffectPairs (CEP) dataset [46] collected by experts for studying paired causality, CauseMe dataset [47] containing real-world Earth science data, protein and phospholipid expression level dataset in human immune system cells [19], [48] by a Markov process designed for modeling lung cancer and its associated causal relationships, and real datasets constructed by causalAssembly [20] and CausalBench [49] in manufacturing and biomedical fields. Although the above benchmarks for causal discovery exist, they often come with significant limitations. For example, synthetic/semi-real datasets and certain carefully designed real datasets for specific domains may exhibit deliberate artificial traces, where causal mechanisms are relatively simple or dynamics can be easily captured. This doesn't reflect the true performance of causal discovery algorithms in the face of complex and diverse real-world scenarios. Therefore, this paper explores a method to measure the real performance of causal discovery methods without relying on specific datasets.

[1] https://www.bnlearn.com/bnrepository/

III. FOUNDATIONS OF GRAPHICAL MODELS AND CAUSAL EFFECT IDENTIFICATION

A. TRAIN/TEST DATA SPLIT

In deep learning, dividing data into training and testing sets is crucial. Such division not only helps evaluate the model's generalization capability but also effectively prevents overfitting and facilitates parameter tuning. In our proposed evaluation approach, we focus on leveraging the invariance of causal structures on the training and testing sets to assess the performance of causal discovery methods. Specifically, we partition the given dataset X into training set X_{Train} and testing set X_{Test}. During the training phase, we conduct causal discovery using X_{Train} to obtain a causal graph. Subsequently, in the testing phase, we employ the causal graph obtained during training, along with the testing set X_{Test}, to evaluate the accuracy and applicability of the causal graph. This approach not only helps validate the effectiveness of the model but also provides a reliable foundation for further causal inference.

B. GRAPHICAL MODELS, CAUSAL GRAPH AND MARKOV BLANKETS

1) GRAPHICAL MODELS

Graphical models [50] are visual mathematical analysis tools represented in the form of edges and nodes, widely employed in fields such as statistics, artificial intelligence, and biomedical research. It is categorized into directed graph models and undirected graph models based on the type of edges. Directed graph models, also known as Bayesian networks, consist of nodes and directed edges, where each node represents a system variable, and the directed edges denote probabilistic dependencies between variables. Nodes can be classified into observed variables, directly observable in the data, and latent variables, inferred from the data. Bayesian networks are commonly used in medical diagnosis, probabilistic decision-making, and causal inference. Undirected graph models, also referred to as Markov random fields, consist of nodes connected by undirected edges. In these models, edges represent correlations between nodes without explicit causal relationships. Typically, potential functions are employed in undirected graph models to describe the relationships between nodes, quantifying the interactions between variables. Markov random fields find widespread applications in fields like image segmentation, social network analysis, and natural language processing. Through graphical models, we can comprehend and analyze the dependencies between random variables in data, facilitating a better understanding of the data generation process and enabling effective prediction and decision-making.

2) CAUSAL GRAPH

Causal graphs [51] are directed acyclic graph models, differing from Bayesian networks in that they represent causal relationships between node variables rather than probabilistic dependencies. Node variables in causal graph models can be classified into exogenous variables and

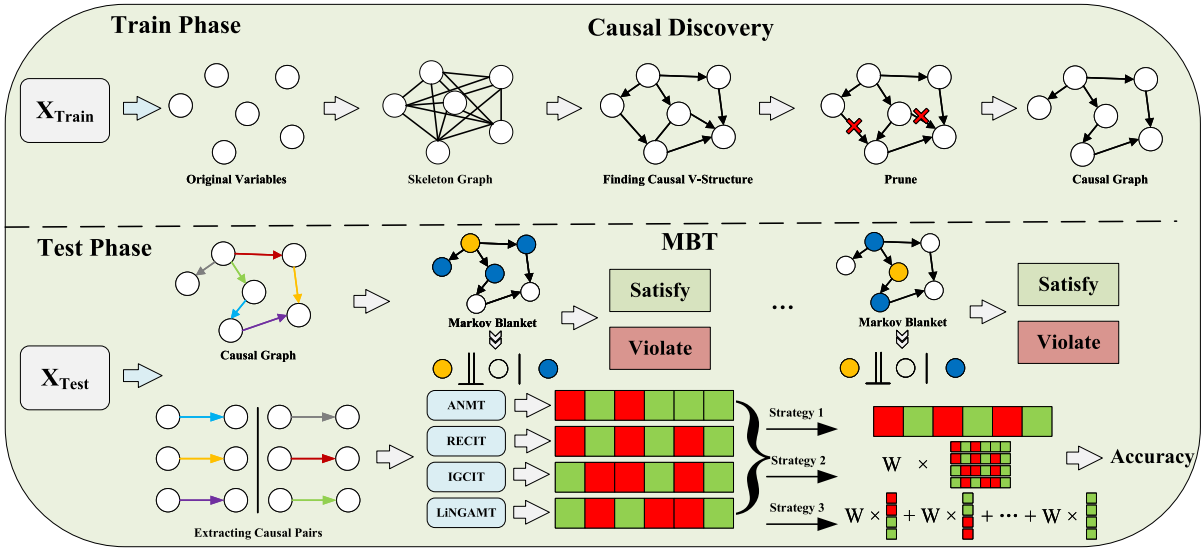

FIGURE 1. Our evaluation method is primarily divided into two stages: training stage and testing stage. In the training stage, causal discovery is performed using the training dataset (approximately following the PC algorithm depicted in the diagram) to obtain a causal graph. In the testing stage, the causal graph is initially subjected to Markov Blanket Testing (MBT), followed by extraction of learned causal pairs from the causal graph. Subsequently, causal pair identification is carried out using ANM, RECI, IGCI and LiNGAM algorithms.

endogenous variables, determined by the constructed causal model. Exogenous variables are assumed to be independent of other variables within the model. In other words, the values of exogenous variables are not influenced by other variables within the model and are often considered as "driving factors" or "determining factors" in the model. Exogenous variables are typically considered to be determined by factors outside the model, such as policy changes, natural disasters, or individual preferences. Endogenous variables, on the other hand, are assumed to depend on other variables within the model. In other words, the values of endogenous variables are influenced by other variables within the model and are often viewed as "outcomes" or "responses" in the model. Endogenous variables are typically inferred or predicted through the rules or equations of the model, and their changes are determined by the changes of exogenous variables and other endogenous variables. Causal discovery is a method of exploring (learning) causal graphs from data, hence causal graphs learned from causal discovery usually involve endogenous variables, while exogenous variables are defined by humans in causal models. Endogenous variables, similar to Bayesian networks, can also be further classified into observable variables and hidden variables. During the process of learning causal graphs from observational data, the presence of hidden intermediate variables (unobserved variables between cause variables and effect variables) and confounding variables (unobserved variables that simultaneously influence cause variables and effect variables) in the data often poses a challenge, making it one of the core challenges in causal discovery.

3) MARKOV BLANKETS

The Markov blanket [52] can be used to describe, in a directed graph model, a node that contains the smallest set of variables from which no additional information about the node can be obtained. Specifically, the Markov blanket of a node variable X refers to its parent nodes, child nodes, and other parents of its child nodes (spouse nodes). These nodes collectively form the Markov blanket of variable X. As shown in Figure 2, an example of a Bayesian network, the node Y is depicted in yellow, with its corresponding Markov blanket consisting of Y's parent node A, child nodes G, C, B, and the parent node (spouse node) P of node B (depicted in blue). Due to the fact that all information carried by a node can be obtained from its Markov blanket, a node is conditionally independent of other system variables given its Markov blanket.

Definition: In a Bayesian network, for any node X_i, we define its Markov blanket $MB(X_i)$ as the set of nodes consisting of X_i's parents, children, and any other nodes that share a child with X_i.

Proposition: With this definition, if X_i is conditionally independent of all other nodes $\forall X_j \in X \setminus (MB(X_i) \cup X_i)$ in the network, given its Markov blanket $MB(X_i)$, then we have:

$$X_i \perp\!\!\!\perp X_j \mid MB(X_i) \tag{1}$$

In the evaluation method proposed in this paper, we utilize the causal graph learned from the training set to obtain the Markov blanket of each node. Subsequently, we conduct Markov Blanket Testing (MBT) on the test set data according to Equation (1). For each causal discovery method, we report the overall test result. Specifically, if there is a node in the learned causal graph that does not pass MBT, we return "Violate;" otherwise, if all nodes pass, we return "Satisfy."

C. IDENTIFICATION OF CAUSAL EFFECTS METHODS

1) ADDITIVE NOISE MODEL (ANM)

ANM [18] is a causal discovery method for identifying causal directions between pairs of variables. This method models

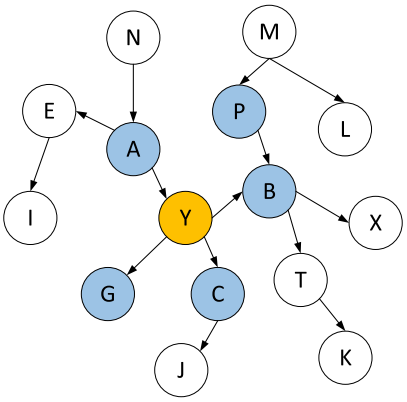

FIGURE 2. Example of Bayesian network and Markov blanket. The class variable *Y* is in orange and its *MB* are in blue.

causal relationships between system variables as functional relationships. Specifically, the method models the outcome variable x_i as an additive function of the causal variable x_{pa_i} and a noise (disturbance) variable e_i, where Pa represents the parents of node i, and as shown below:

$$x_i = f_i(x_{pa_i}) + e_i \tag{2}$$

where f_i can be arbitrary function, and for each x_i, f_i may be different. Similarly, the noise e_i can follow any arbitrary distribution, and noises are mutually independent. The main idea of ANM is to determine causal directions between cause and effect variables based on the independence of exogenous variables (noise) of the cause and effect variables. The noise e_i and x_{pa_i} are theoretically independent in the correct causal direction, but not in the non-causal direction. Therefore, the discriminative process of ANM can be summarized into the following steps (using two observed variables X and Y as an example): Firstly, a statistical independence test is used to determine whether X and Y are independent. If they are independent, it indicates that X and Y have no causal relationship. If they are not independent, nonlinear regression is used to fit Y given X. The result of the model fitting is denoted as Y'. Then, the residual E_Y between the true Y and the fitted Y' is computed, and the independence between the residual E_Y and X is tested. If they are independent, it indicates that accepting X pointing to Y as the causal model is possible. Similarly, given Y, regression is used to fit X, and the result of the model fitting is denoted as X'. Then, the residual E_X between the true X and the fitted X' is computed, and the independence between the residual E_X and Y is tested. If they are independent, it indicates that accepting Y pointing to X as the causal model is possible. Therefore, based on the above process, the conclusions drawn by ANM can have the following four scenarios: (1) X and Y are independent; (2) X is the cause of Y; (3) Y is the cause of X; (4) Unable to determine. If the fourth scenario occurs, it indicates that the ANM model cannot determine the causal direction, and more complex causal models need to be constructed for determination.

2) INFORMATION GEOMETRY CAUSAL INFERENCE (IGCI)

The IGCI [53] is a causal discovery method that utilizes the distribution structure of variables and mapping functions to identify the direction of causality. When the assumption function is differentiable and invertible, the following intuition holds:

Postulate (Indep. of Input and Function [53]). If $X \to Y$, then the distribution of X and the function f that maps X to Y are independent, as they correspond to independent natural mechanisms.

In the IGCI model, the asymmetry is reflected in the fact that if X causes Y, $P(X)$ and $P(Y|X)$ represent independent natural mechanisms, and therefore do not contain information about each other. The causal direction identification of IGCI can be determined according to the following criteria:

$$C_{X\to Y} := D\left(p_X \| \varepsilon_X\right) - D\left(p_Y \| \varepsilon_Y\right) \tag{3}$$

where ε_X and ε_Y are associated with exponential families, serving as "smooth" reference distributions for X and Y, while p_X and p_Y denote the probability density of X and Y respectively, and D represents the KL divergence. If $C_{X\to Y} < 0$, it indicates that X causes Y; if $C_{X\to Y} > 0$, it suggests that Y causes X; and if $C_{X\to Y} = 0$, it implies that the causal direction cannot be determined.

3) REGRESSION ERROR BASED CAUSAL INFERENCE (RECI)

RECI [54] is a causal discovery method that utilizes the asymmetry between causal and non-causal models to infer the directionality of causation between paired variables. Specifically, it involves fitting the outcome variable given the causal variable, and then calculating the Mean Squared Error (MSE) between the fitted outcome and the true outcome as a measure of fit. After computing in both directions (from cause to outcome and from outcome to cause), the direction with the smaller MSE is considered the true causal direction. This process is similar to the ANM algorithm [18], but with several differences. That is, given a causal function model in the following form:

$$x_i = \phi(x_{pa_i}) + \alpha e_i \tag{4}$$

where $\alpha \in \mathbb{R}^+$ is a parameter controlling the level of noise. This causal function model does not require statistical independence between the causal variable x_{Pa} and the noise e, but imposes certain assumptions on the function ϕ and the causal variables x as shown below.

- The invertible function assumption posits that the function ϕ should be a continuously increasing and differentiable function.
- The compact support assumption asserts that the distribution of the variables x is confined within a finite range.
- The unbiased noise assumption mandates that the expected value of the conditional expectation given x is zero.
- The unit variance assumption stipulates that the expected value of the conditional variance of the noise, given the cause, is unity.

- The independence assumption ensures that the function ϕ does not encode any information about the underlying distribution P_x, thereby guaranteeing that the conditional probability $P_{x|x_{Pa}}$ remains uninfluenced by P_x.

4) LINEAR NON-GAUSSIAN ACYCLIC MODEL (LiNGAM)

The Linear Non-Gaussian Acyclic Model (LiNGAM) [17] is the first causal inference method that assumes the system variables are modeled as a function. Specifically, this method assumes that the causal relationships between system variables are composed of linear functions of the cause variables and a noise term:

$$x_i = \sum_{x_k \in pa(x_i)} b_{ik} pa(x_k) + e_i \tag{5}$$

where b_{ij} are the coefficients of the cause variables (can be understood as the causal weight matrix), and ϵ_i is a non-zero variance non-Gaussian noise term. The intuition behind LiNGAM for identifying causal direction is that when the function of the causal model is linear and the noise terms are non-Gaussian distributed, the noise terms of the cause variables and the result variables are independent. This is because the incorrect causal direction cannot satisfy the independence condition between noise and causal variables. The goal of LiNGAM is to solve for the causal weight matrix b in equation (5), which can be achieved through the ideas of Independent Component Analysis (ICA) [55] algorithm.

IV. THE PROPOSED EVALUATION METHOD

Based on the above basis, our proposed evaluation method is implemented through four steps, and the overall flowchart is shown in Figure 1. Given the data $X \in \mathcal{R}^{n \times d}$, where n is the number of samples and d is the number of features (number of nodes in the network), we divide it into training set X_{Train} and test set X_{Test}, with a ratio of 7:3. The specific steps are as follows:

A. STEP 1 OBTAIN THE CAUSAL GRAPH G USING THE TRAINING SET

In the absence of true labels, we divide the data into training and test sets. In the first step, we need to use the training set X_{Train} for causal learning to obtain the learned causal graph G, and then proceed with the following steps.

B. STEP 2 MARKOV BLANKET TESTING (MBT)

The Markov blanket test plays a crucial role in causal relationship verification by examining whether the Markov blanket of a specific node meets certain conditions to determine the causal relationships between the node and other nodes. Specifically, the Markov blanket is a subset consisting of a node's parent nodes, child nodes, and nodes that share the same parent nodes with its child nodes. If the Markov blanket of a node includes all nodes that have a direct influence on it, and the states of these nodes are known, then the node can be considered conditionally independent from other nodes. This conditional independence is a key indicator of a causal relationship. The Markov blanket test ensures that in the predicted causal graph, the causal relationships of each node are reasonable and conform to the independence assumption by verifying the Markov blanket of each node. The specific process involves first checking whether the Markov blanket of a node fully covers all the factors that directly affect the node, and then verifying whether these factors sufficiently explain the changes in the node's state. If these conditions are met, it indicates that the causal graph is locally reasonable. The advantage of the Markov blanket test is that it not only verifies the causal relationship of a single node but also checks the structural reasonableness of the local network through the relationships of its neighboring nodes. This test can serve as a basic criterion for evaluating the causal graph obtained by a causal discovery algorithm, as it ensures that the causal relationships of each node in the causal graph have been rigorously verified, providing reliability and reasonableness. Therefore, passing the Markov blanket test is the most fundamental requirement for the causal graph generated by a causal discovery algorithm. For a causal graph $G = (V, E)$, where V is the set of nodes and E is the set of edges, we first find its Markov blanket MB. Then, we test whether each node satisfies conditional independence with the other nodes in the graph under its Markov blanket, which can be expressed by the following formula:

$$v \perp\!\!\!\perp u | MB(v), \forall v \in V, u \neq v, \forall u \in V \setminus MB(v) \tag{6}$$

For each node, we perform Markov blanket testing, The confidence level is set to 0.05. For a causal graph G, if all nodes satisfy MBT, we output "Satisfy;" conversely, if any node fails to satisfy MBT, we return "Violate."

C. STEP 3 CAUSE EFFECT IDENTIFICATION

After performing MBT, we use three causal pair identification algorithms to test each edge (causal pair) learned from the causal graph on the test set, denoted as ANMT, RECIT, IGCIT and LiNGAMT. The intuition behind this approach is that although we divide a dataset into two parts, the causal structure does not change. Therefore, we can test each edge learned from the causal graph, not only to validate the correctness of causal learning algorithms but also to test the generalization ability of causal learning algorithms (similar to deep learning test sets). For a causal graph G, we extract each edge $v \in V$ to form a set of causal pairs, denoted as CEP (Cause-Effect Pair). For each causal pair set, we use the test set X_{Test} to perform ANM, RECI, IGCI and LiNGAM methods respectively. For the ANM algorithm, we set its regression algorithm to Gaussian Process Regression (GPR) [56], and for RECI, we set it to linear regression. The results are denoted as ANM_{disc}, $RECI_{disc}$, $IGCI_{disc}$ and $LiNGAM_{disc}$ respectively. The result of each algorithm is a binary vector, and the dimension of the result is the same as the number of edges in the causal graph G.

D. STEP 4 RESULT ENSEMBLE

For the discriminative results learned in step three, we adopt three ensemble strategies to obtain the final accuracy. Specifically, **Strategy 1:** We vote for each dimension (corresponding to each edge in the causal graph) of ANM_{disc}, $RECI_{disc}$, $IGCI_{disc}$, and $LiNGAM_{disc} \in \mathcal{R}^{1\times V}$ to obtain the final result, where V is the number of edges learned by each causal discovery algorithm. If there are three or more vectors that vote in favor (True or 1) for each dimension, the final result is 1; otherwise, it is 0. For the final discriminative vector (denoted as $Ensemble_{disc}$), we take the average as the final accuracy. **Strategy 2:** We adopt a weighted ensemble strategy, setting the weight matrix as $W \in \mathcal{R}^{1\times 4} = \{42.20, 60.20, 68.30, 56.57\}$, where the values are determined by the accuracy of the four causal pair algorithms on the CauseEffectPair dataset [46] (results sourced from [57]). For the discriminative results of each algorithm, we first calculate the average accuracy, and then obtain the final accuracy by separately weighting the normalized weights and the four algorithms. This strategy focuses on evaluating each algorithm's contribution to the causal graph. **Strategy 3:** We adopt a weighted ensemble strategy at the sample level. Using the aforementioned weight matrix, for the discriminative results of each algorithm, we weight each dimension to obtain the weighted result. Then, we sum the weighted discriminative results of the four algorithms, and then set a threshold θ for each dimension result. If it is greater than θ, the result is 1; otherwise, it is 0. For the final result, we take the average as the final accuracy. This strategy focuses more on the discrimination of each edge in the causal graph.

The above steps first use the training set for causal discovery to obtain the causal graph. Then, by using the test set for Markov blanket testing, we ensure that the learned causal graph is correct. This step also tests whether the causal discovery method has an impact on the correctness of causal graph learning under the same causal structure with generalized data. Next, by employing four causal pair identification methods, we assess the correctness of each causal pair extracted from the causal graph on the test set. This step evaluates the identification accuracy of each causal pair in the graph by causal learning algorithms. Finally, we integrate the results from the four causal pair identification methods, utilizing a combination of voting and two weighted ensemble strategies to obtain the final outcome and calculate accuracy. Through these steps, we not only ensure the correctness of causal graph learning without labels but also test its accuracy and generalization performance on real-world data. The algorithm steps are presented in Algorithm 1 and Algorithm 2.

Algorithm 1 Get Markov Blanket

Require: Causal graph G, node v

```
Initialize MB as an empty set
for p ∈ Pa(v) do
    Add p to MB
end for
for c ∈ Child(v) do
    Add c to MB
end for
for n = Child(v) ∈ Pa(n) do
    if Pa(n) is not v then
        Add n to MB
    end if
end for
return MB
```

V. EXPERIMENTS

A. DATASETS

In our experiment, we used six synthetic dataset and the Sachs dataset to evaluate the reliability of the proposed evaluation method. The details of the datasets are as follows:

- **Synthetic Dataset:** For synthetic datasets, we utilized the six synthetic datasets described in [32]. Specifically, they are as follows:
  - *Linear Synthetic Mechanism:* $X_i = \sum_{j\in pa(i)} a_{i,j}X_j + E_i$, where $a_{i,j} \sim \mathcal{N}(0, 1)$
  - *Polynomial Synthetic Mechanism:* $f(x) = \sum_{j=0}^{d} coeff_j \cdot x^j$, where $coeff_j$ are the coefficients for the polynomial, d is the degree of the polynomial.
  - *Sigmoid Addition Mechanism:* $X_i = \sum_{j\in pa(i)} f_{i,j}\left(X_j\right) + E_i$, where $f_{i,j}\left(x_j\right) = a \cdot \frac{b\cdot(x_j+c)}{1+\left|b\cdot(x_j+c)\right|}$ with $a \sim \text{Exp}(4) + 1$, $b \sim \mathcal{U}([-2, -0.5] \cup [0.5, 2])$ and $c \sim \mathcal{U}([-2, 2])$.
  - *Sigmoid Mix Mechanism:* $X_i = f_i\left(\sum_{j\in pa(i)} X_j + E_i\right)$, where f_i is as in the previous bullet-point.
  - *Gaussian Process Addition Mechanism:* $X_i = \sum_{j\in pa(i)} f_{i,j}\left(X_j\right) + E_i$, where where $f_{i,j}$ is an univariate Gaussian process with a Gaussian kernel of unit bandwidth.
  - *Gaussian Process Mix Mechanism:* $X_i = f_i\left(\left[X_{pa(i)}, E_i\right]\right)$, where f_i is a multivariate Gaussian process with a Gaussian kernel of unit bandwidth.

  To generate the synthetic datasets, we first define the underlying structure of the data using directed acyclic graphs with 10 nodes. For each mechanism described above, we simulate the relationships between variables according to the specific functional form and noise distribution. We then generate a dataset of 5000 samples based on these mechanisms. The generated data is split into a training set and a test set in a 7:3 ratio.
- **Sachs dataset:** [58] The Sachs dataset measures the expression levels of phospholipids and proteins in human cells. It simultaneously measures 11 phospholipids and phosphorylated proteins from thousands of immune system cells. The Sachs dataset is continuous, representing concentrations of the molecules under study. The standard approach in literature assumes concentrations follow Gaussian distributions and uses GBN to construct protein signaling networks. It consists

Algorithm 2 Our Proposed Evaluation Method

Require: $X = \{X_{\text{Train}}, X_{\text{Test}}\}$, *CausalDiscoveryMethods*, W

```
# Step 1: Causal Discovery
G ← CausalDiscovery(X_Train)
# Step 2: Markov Blanket Testing
for each node v ∈ V do
    MB ← Get Markov Blanket(G, v)
    pass ← True
    for each node u ∈ V do
        if u ≠ v and u ∉ MB then
            indep ← CIT(X_Test, MB) ⇔ v ⫫ v|MB
            if indep is False then
                pass ← False
                break
            end if
        end if
    end for
    if pass is True then
        Output "Satisfy" for node v
    else
        Output "Violate" for node v
    end if
end for
if all nodes v is "Satisfy" then
    Output "Satisfy"
else
    Output "Violate"
end if
# Step 3: Cause Effect Identification
for causal pairs CEP formed by each edge e ∈ E. do
    ANM_disc ← ANMT(X_Test, CEP, G)
    RECI_disc ← RECIT(X_Test, CEP, G)
    IGCI_disc ← IGCIT(X_Test, CEP, G)
    LiNGAM_disc ← LiNGAM(X_Test, CEP, G)
end for
# Step 4: Result Ensemble
# Strategy 1:
Ensemble_disc = vote(ANM_disc, RECI_disc, IGCI_disc, LiNGAM_disc)
Accuracy = Mean(Ensemble_disc)
# Strategy 2:
Ensemble_disc ← Stack(ANM_disc, RECI_disc, IGCI_disc, LiNGAM_disc)
Accuracy_disc = Mean(Ensemble_disc)
Accuracy = Σ_{i=1}^{n} W_i × Acc^i_disc
# Strategy 3:
ANM_w = Σ_{i=1}^{len(ANM_disc)} W_i × ANM^i_disc
RECI_w = Σ_{i=1}^{len(RECI_disc)} W_i × RECI^i_disc
IGCI_w = Σ_{i=1}^{len(IGCI_disc)} W_i × IGCI^i_disc
LiNGAM_w = Σ_{i=1}^{len(LiNGAM_disc)} W_i × LiNGAM^i_disc
Ensemble_disc = (ANM_w + RECI_w + IGCI_w + LiNGAM_w) > θ
Accuracy = Mean(Ensemble_disc)
return Accuracy
```

of 11 nodes, 17 edges, and 7466 instances, with an average Markov blanket size of 3.09. The ground-truth network of Sachs is shown in the figure 3.

TABLE 1. Using strategy 1 of the proposed method to validate causal discovery methods on synthetic datasets. The table presents the difference of accuracy, true precision (*Diff*$_P$), and true recall (*Diff*$_R$), along with the mean values across each dataset (*MDiff* denotes mean difference).

Datasets	Methods	$Diff_P$	$Diff_R$	$MDiff_P$	$MDiff_R$
Linear	PC	0.3125	0.2125	0.1666	0.1217
	GES	0.1696	0.3196		
	LiNGAM	0.1904	0.1571		
	NOTEARS	0.2750	0.0500		
	DAG-GNN	0.1182	0.0318		
	CORL	0.0476	0.0571		
	GOLEM	0.0943	0.1591		
	MCSL	0.0416	0.0583		
	GAE	0.2500	0.0500		
Polynomial	PC	0.0989	0.0989	0.1202	0.1180
	GES	0.1190	0.2948		
	LiNGAM	0.1333	0.0641		
	NOTEARS	0.0000	0.0898		
	DAG-GNN	0.2000	0.2000		
	CORL	0.0727	0.0308		
	GOLEM	0.0333	0.1025		
	MCSL	0.1349	0.0470		
	GAE	0.2895	0.1336		
Sigmoid Add	PC	0.1277	0.3480	0.1867	0.2401
	GES	0.0065	0.2615		
	LiNGAM	0.2662	0.2937		
	NOTEARS	0.4333	0.5231		
	DAG-GNN	0.1556	0.1898		
	CORL	0.1250	0.1971		
	GOLEM	0.0079	0.0299		
	MCSL	0.1582	0.0105		
	GAE	0.4000	0.3077		
Sigmoid Mix	PC	0.0471	0.0615	0.1992	0.2252
	GES	0.1891	0.1186		
	LiNGAM	0.2464	0.4650		
	NOTEARS	0.1806	0.0096		
	DAG-GNN	0.2091	0.2231		
	CORL	0.1594	0.2876		
	GOLEM	0.0741	0.1713		
	MCSL	0.3302	0.2673		
	GAE	0.3571	0.4231		
GP Add	PC	0.2007	0.3221	0.1344	0.2108
	GES	0.0237	0.3441		
	LiNGAM	0.0980	0.1493		
	NOTEARS	0.2250	0.0865		
	DAG-GNN	0.0654	0.1197		
	CORL	0.0184	0.0905		
	GOLEM	0.3334	0.4052		
	MCSL	0.1174	0.0636		
	GAE	0.1278	0.3163		
GP Mix	PC	0.1528	0.2596	0.3012	0.2747
	GES	0.0485	0.3100		
	LiNGAM	0.5455	0.4615		
	NOTEARS	0.4583	0.2179		
	DAG-GNN	0.4000	0.2846		
	CORL	0.5364	0.4385		
	GOLEM	0.3398	0.3077		
	MCSL	0.0072	0.1410		
	GAE	0.2223	0.0513		

B. BASELINE ALGORITHMS

We tested our proposed evaluation method using nine causal discovery approaches, including baseline methods such as constraint-based, score-based, function-based, and continuous optimization-based causal discovery methods.[2] Here are the detailed descriptions:

[2]https://github.com/huawei-noah/trustworthyAI

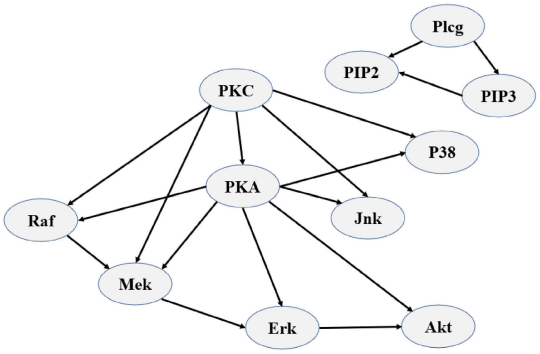

FIGURE 3. Ground-truth network of the SACHS dataset.

- **PC algorithm** [21]: A classic constraint-based causal discovery method that utilizes conditional independence to search for Completed partially directed acyclic graphs (CPDAGs) representing causal V-structures.
- **GES algorithm** [25]: A score-based causal discovery method that optimizes the best causal graph using a scoring function, iteratively adding (forward edge addition) or removing (backward edge deletion) directed edges from the skeleton graph while evaluating the causal graph to select the optimal causal graph as the final output.
- **LiNGAM algorithm** [17]: A causal discovery method that constructs variables as causal functions, utilizing a discriminant mechanism based on the non-Gaussian independence of variables and the noise independence in the correct causal direction under linear functions to identify causal directions.
- **NOTEARS algorithm** [19]: Extends the discrete causal graph search optimization space of score-based causal discovery methods to continuous optimization space based on acyclic constraints and combines deep learning for causal discovery.
- **DAG-GNN algorithm** [29]: Utilizes variational inference and a pair of parameterized encoder-decoder along with specially designed graph neural networks for causal discovery.
- **CORL algorithm** [59]: A ranking-based causal discovery algorithm that formulates the ranking search problem as a multi-step Markov decision process using reinforcement learning to learn the causal graph.
- **GOLEM algorithm** [60]: A method that learns linear DAG models using gradient DAG penalty learning and is a likelihood-based causal learning approach.
- **MCSL algorithm** [61]: A causal discovery algorithm based on masked gradient that shows if the original SEM is identifiable, then under mild conditions, the adjacency matrix can be identified to the supergraph of the true causal graph.
- **GAE algorithm** [62]: A gradient-based causal structure learning method that employs a graph autoencoder framework to handle nonlinear structural equation models.

TABLE 2. Using Strategy 2 of the proposed method to validate causal discovery methods on synthetic datasets. The table presents the difference of accuracy, true precision ($Diff_P$), and true recall ($Diff_R$), along with the mean values across each dataset (*MDiff* denotes mean difference).

Datasets	Methods	$Diff_P$	$Diff_R$	$MDiff_P$	$MDiff_R$
Linear	PC	0.0534	0.1534	0.2224	0.2743
	GES	0.1970	0.0470		
	LiNGAM	0.1701	0.2034		
	NOTEARS	0.1404	0.3654		
	DAG-GNN	0.2602	0.4102		
	CORL	0.2971	0.2876		
	GOLEM	0.2914	0.2266		
	MCSL	0.3746	0.3579		
	GAE	0.2172	0.4172		
Polynomial	PC	0.1340	0.1340	0.2111	0.2500
	GES	0.1738	0.0020		
	LiNGAM	0.0728	0.1420		
	NOTEARS	0.3267	0.4165		
	DAG-GNN	0.4913	0.4913		
	CORL	0.1162	0.1581		
	GOLEM	0.2339	0.3031		
	MCSL	0.3097	0.2218		
	GAE	0.0417	0.3814		
Sigmoid Add	PC	0.1119	0.1084	0.2315	0.2000
	GES	0.2419	0.0261		
	LiNGAM	0.0913	0.0638		
	NOTEARS	0.5717	0.6615		
	DAG-GNN	0.4117	0.4459		
	CORL	0.1633	0.0912		
	GOLEM	0.2158	0.1938		
	MCSL	0.2701	0.1224		
	GAE	0.0057	0.0866		
Sigmoid Mix	PC	0.2323	0.1237	0.2587	0.2420
	GES	0.4035	0.0958		
	LiNGAM	0.0046	0.2232		
	NOTEARS	0.0743	0.2453		
	DAG-GNN	0.3790	0.3930		
	CORL	0.0862	0.0420		
	GOLEM	0.2604	0.1632		
	MCSL	0.4248	0.3619		
	GAE	0.4637	0.5297		
GP Add	PC	0.2128	0.0914	0.2073	0.1431
	GES	0.3566	0.0112		
	LiNGAM	0.1886	0.1373		
	NOTEARS	0.0032	0.1353		
	DAG-GNN	0.3501	0.2958		
	CORL	0.2111	0.1390		
	GOLEM	0.0044	0.0674		
	MCSL	0.4073	0.3535		
	GAE	0.1313	0.0572		
GP Mix	PC	0.1714	0.0646	0.1483	0.0855
	GES	0.2810	0.0195		
	LiNGAM	0.1036	0.0196		
	NOTEARS	0.1421	0.0983		
	DAG-GNN	0.0387	0.0767		
	CORL	0.1759	0.0780		
	GOLEM	0.0731	0.1052		
	MCSL	0.3096	0.1758		
	GAE	0.0391	0.1319		

C. EXPERIMENT RESULTS

We conducted evaluation experiments on synthetic datasets and real-world datasets, respectively. Additionally, we once again employed an ensemble strategy to further integrate the results of the three strategies, forming a more stable and robust outcome. Moreover, we analyzed the relationship

between the accuracy of the proposed method and the true precision, recall.

TABLE 3. Evaluation and execution time of the causal discovery methods on Sachs dataset.

Methods	MBT	Accuracy	True Recall	True Prcision	Time
PC	Satisfy	0.1290	0.2778	0.1923	32.8024
GES	Satisfy	0.3158	0.6111	0.2821	38.0469
LiNGAM	Satisfy	0.1786	0.2222	0.2000	28.8320
NOTEARS	Satisfy	0.1786	0.2778	0.2500	29.5161
DAG-GNN	Satisfy	0.1500	0.2222	0.1481	21.3239
CORL	Satisfy	0.1333	0.2778	0.2083	31.3998
GOLEM	Satisfy	0.3750	0.2222	0.2500	25.8666
MCSL	Satisfy	0.2593	0.1667	0.1034	28.3861
GAE	Satisfy	0.2500	0.0556	0.5000	4.3961

1) EXPERIMENTAL EVALUATION SETUP

To comprehensively evaluate the effectiveness of the causal discovery algorithms, we employed two evaluation strategies: quantitative evaluation and visualization-based evaluation.

For the quantitative evaluation, we used two metrics: $Diff_P$ and $Diff_R$. $Diff_P$ represents the difference between the proposed evaluation method and the true precision, while $Diff_R$ represents the difference between the proposed evaluation method and the true recall. These metrics were selected to quantitatively assess the accuracy of the proposed method by measuring how closely it approximates the true precision and recall. The smaller the values of $Diff_P$ and $Diff_R$, the closer the proposed method is to the true performance metrics, indicating higher evaluation accuracy.

In the visualization-based evaluation, we compared the evaluation metric curves of multiple causal discovery methods on a single dataset. The comparison involved two key curves: one representing the proposed evaluation method and the other representing the true evaluation metric. The assessment was focused on the closeness and trend of these curves. By visually analyzing the proximity and similarity of the curves, we aimed to gain deeper insights into the relative performance of different causal discovery methods, capturing nuances that quantitative metrics might overlook.

By integrating both quantitative and visualization-based evaluations, our approach provides a thorough assessment of the causal discovery algorithms, ensuring both accuracy and interpretability in evaluating their effectiveness.

2) SYNTHETIC DATASET EVALUATION

We conducted experiments on six common synthetic datasets, and presented the differences between the accuracy of the proposed method and the true precision and recall in Tables 1 and 2, as well as the average differences on true precision and recall for each dataset.

a: RESULTS ANALYSIS

When using Strategy 3 with a threshold set to 0.6, we found consistent conclusions with Strategy 1. This is because Strategy 3 functions as a continuous decision process akin to a voting mechanism, where with a threshold of 0.6, the weights of any three methods exceed 0.6, thus yielding the same conclusion as Strategy 1. Overall, from Tables 1 and 2, we observe that Strategy 1 and Strategy 3 perform well on all datasets except the GP Mix synthetic dataset, with differences from true values averaging between 0.1 and 0.2. Strategy 2 performs better than the other two strategies on the GP Mix synthetic dataset. Specifically, using Strategy 1 and Strategy 3, the differences from true precision mostly range from 0.1 to 0.2, with significant variations observed in individual methods, particularly in those based on continuous optimization. The same conclusion applies to the comparison with true recall. This is because our proposed method mimics the evaluation approach of deep learning, assessing not only the accuracy of causal graph learning methods but also their robustness to data under the same causal structure. Typically, causal learning methods are sensitive to data, with this sensitivity being more pronounced in methods based on continuous optimization, as they usually rely on deep learning frameworks and may experience significant performance variations without sufficient or high-quality training data. The performance of Strategy 2 is inferior to that of Strategy 1 and Strategy 3 because the weighted results of Strategy 2 are smoother and more stable, reducing the variability between different methods. Therefore, for synthetic datasets, we recommend using Strategy 1 or Strategy 3 with an appropriate threshold (0.6 or 0.7).

3) REAL DATASET EVALUATION

We also conducted experiments on the real dataset Sachs, presenting the accuracy, execution time, and MBT test results of the proposed method in Table 3. The accuracy and true precision results of Strategy 1 to 3 and Strategy 3 under different threshold settings are presented in Figure 4.

a: RESULT ANALYSIS

From Figure 4, we can clearly see that the trend of the proposed method and true precision deficiencies are quite similar overall, with the curves closest in the cases of Strategy 1 and Strategy 3 (thresholds of 0.6 or 0.7). For the proposed validation method, we can assess its effectiveness in two ways: first, if it exhibits the same trend as the true indicators and the numerical difference falls within a reasonable range; second, if it closely matches the numerical values of the true indicators. Our proposed method nearly meets the first criterion in several strategies and satisfies both criteria in Strategy 1 and Strategy 3 (thresholds of 0.6 or 0.7), which aligns with the conclusions drawn from synthetic datasets. Furthermore, our method still exhibits significant fluctuations in methods based on continuous optimization, indicating the high sensitivity of continuous optimization-based causal discovery methods to data compared to other types of methods.

4) MULTI-LEVEL ENSEMBLE EXPERIMENT AND EVALUATION ANALYSIS

In addition to the two conventional experiments mentioned above, we further integrated the three strategies by performing an equal-weight integration of their results. For

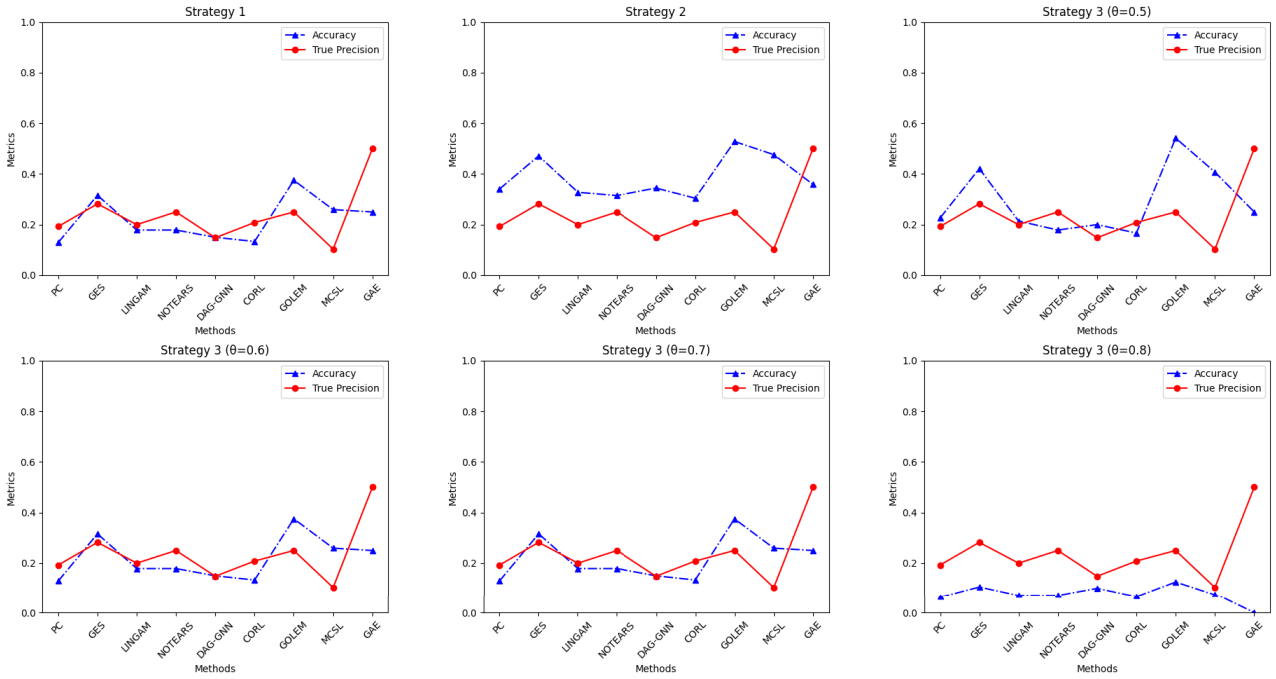

FIGURE 4. Evaluation results for the Sachs dataset.

TABLE 4. Experiment results of integrating three strategies while balancing true precision and true recall. Here, Diff(α:β) represents the average disparity of each dataset under different values of α and β.

Datasets	Sachs	Linear	Polynomial	Sigmoid Add	Sigmoid Mix	GP Add	GP Mix	Average
Diff(0:1)	0.1232	0.0932	0.1275	0.2155	0.1858	0.1871	**0.1945**	0.1610
Diff(0.1:0.9)	0.1119	0.0862	0.1213	0.2105	0.1779	0.1790	0.1972	0.1549
Diff(0.2:0.8)	0.1006	0.0792	0.1151	0.2055	0.1701	0.1708	0.1998	0.1487
Diff(0.3:0.7)	0.0894	**0.0753**	0.1089	0.2005	0.1635	0.1627	0.2025	0.1432
Diff(0.4:0.6)	0.0799	0.0762	**0.1027**	0.1955	0.1625	0.1545	0.2051	0.1395
Diff(0.5:0.5)	**0.0709**	0.0786	0.1059	0.1905	**0.1618**	0.1463	0.2078	0.1374
Diff(0.6:0.4)	0.0716	0.0811	0.1091	0.1855	0.1632	0.1382	0.2104	**0.1370**
Diff(0.7:0.3)	0.0723	0.0835	0.1122	0.1805	0.1647	0.1300	0.2155	**0.1370**
Diff(0.8:0.2)	0.0745	0.0860	0.1154	**0.1762**	0.1661	0.1218	0.2211	0.1373
Diff(0.9:0.1)	0.0826	0.0903	0.1186	0.1772	0.1675	**0.1212**	0.2285	0.1408
Diff(1:0)	0.0906	0.0961	0.1218	0.1781	0.1689	0.1224	0.2399	0.1454

Strategy 3, we used four threshold settings 0.5, 0.6, 0.7, 0.8, resulting in a total of six strategies being integrated. Furthermore, we compared the results with a trade-off metric between true precision and true recall. Specifically, we assigned two weights, $\alpha \in \{0, 0.1, \cdots, 1\}$ and $\beta \in \{1, 0.9, \cdots, 0\}$, resulting in eleven cases. When α is 0 and β is 1, it indicates a focus solely on true recall, while the reverse is true when α is 1 and β is 0, focusing solely on true precision. Our results are presented in Table 4.

a: RESULT ANALYSIS

From Table 4, we observe that in each dataset, the smallest disparities often occur when true recall and true precision are balanced, and they are notably smaller than when only one evaluation metric is considered. Therefore, we can conclude that our proposed validation method not only focuses on verifying the accuracy of causal discovery methods but also considers their recall. Recall, to a certain extent, reflects the generalization performance of a model. Thus, our method also validates the generalization performance of causal discovery methods.

VI. CONCLUSION

In this paper, we propose an algorithm for evaluating causal discovery in the absence of real causal graphs, by leveraging the consistency of causal graphs on training and testing sets to assess causal discovery methods. Since the testing set can not only validate the accuracy of a method but also assess the algorithm's generalization performance, our approach enables a comprehensive evaluation of causal discovery algorithms. We also validate the effectiveness of our proposed method on synthetic and real datasets in experiments. Additionally, in further experiments involving strategy integration, we demonstrate that hierarchical result

integration can better assess the performance of a causal discovery method. Future work will focus on more comprehensive and accurate research on causal discovery methods.

ACKNOWLEDGMENT

(Tingpeng Li and Lei Wang contributed equally to this work.)

REFERENCES

[1] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Comput. Intell. Neurosci.*, vol. 2018, no. 1, pp. 1–13, 2018.

[2] L. Deng and Y. Liu, *Deep Learning in Natural Language Processing*. Cham, Switzerland: Springer, 2018.

[3] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: Review, opportunities and challenges," *Briefings Bioinf.*, vol. 19, no. 6, pp. 1236–1246, Nov. 2018.

[4] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, and Z. Dong, "A survey of large language models," 2023, *arXiv:2303.18223*.

[5] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nature Med.*, vol. 29, no. 8, pp. 1930–1940, 2023.

[6] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, and Y. Wang, "A survey on evaluation of large language models," *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 3, pp. 1–45, 2024.

[7] R. K. Vasudevan, M. Ziatdinov, L. Vlcek, and S. V. Kalinin, "Off-the-shelf deep learning is not enough, and requires parsimony, bayesianity, and causality," *NPJ Comput. Mater.*, vol. 7, no. 1, p. 16, Jan. 2021.

[8] Y. Luo, J. Peng, and J. Ma, "When causal inference meets deep learning," *Nature Mach. Intell.*, vol. 2, no. 8, pp. 426–427, Aug. 2020.

[9] Z. Deng, X. Zheng, H. Tian, and D. Dajun Zeng, "Deep causal learning: Representation, discovery and inference," 2022, *arXiv:2211.03374*.

[10] J. D. Angrist and J.-S. Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ, USA: Princeton Univ. Press, 2009

[11] J. Pearl and D. Mackenzie, *The Book of Why: The New Science of Cause and Effect*. New York, NY, USA: Basic Books, 2018.

[12] J. Pearl, "Causal inference in statistics: An overview," *Statist. Surv.*, vol. 3, pp. 96–146, Jan. 2009.

[13] M. E. Sobel, "Causal inference in the social sciences," *J. Amer. Stat. Assoc.*, vol. 95, no. 450, pp. 647–651, 2000.

[14] C. Glymour, K. Zhang, and P. Spirtes, "Review of causal discovery methods based on graphical models," *Frontiers Genet.*, vol. 10, p. 524, Jun. 2019.

[15] B. Sibbald and M. Roland, "Understanding controlled trials: Why are randomised controlled trials important?" *Brit. Med. J.*, vol. 316, no. 7126, p. 201, Jan. 1998.

[16] B. Huang, K. Zhang, J. Zhang, J. Ramsey, R. Sanchez-Romero, C. Glymour, and B. Schölkopf, "Causal discovery from heterogeneous/nonstationary data," *J. Mach. Learn. Res.*, vol. 21, no. 89, pp. 1–53, 2020.

[17] S. Shimizu, P. O. Hoyer, A. Hyvärinen, A. Kerminen, and M. Jordan, "A linear non-Gaussian acyclic model for causal discovery," *J. Mach. Learn. Res.*, vol. 7, no. 10, pp. 2003–2030, 2006.

[18] P. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf, "Nonlinear causal discovery with additive noise models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 21, 2008, pp. 689–696.

[19] X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing, "DAGs with NO TEARS: Continuous optimization for structure learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.

[20] K. Göbler, T. Windisch, M. Drton, T. Pychynski, M. Roth, and S. Sonntag, "causalAssembly: Generating realistic production data for benchmarking causal discovery," in *Proc. 3rd Conf. Causal Learn. Reasoning*, vol. 236, F. Locatello and V. Didelez, Eds., Apr. 2024, pp. 609–642. [Online]. Available: https://proceedings.mlr.press/v236/gobler24a.html

[21] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*. Cambridge, MA, USA: MIT Press, 2001.

[22] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, Mar. 1978.

[23] W. Buntine, "Theory refinement on Bayesian networks," in *Uncertainty Proceedings 1991*. Amsterdam, The Netherlands: Elsevier, 1991, pp. 52–60.

[24] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Springer Series in Statistics*. Cham, Switzerland: Springer, 1992, pp. 610–624.

[25] D. M. Chickering, "Optimal structure identification with greedy search," *J. Mach. Learn. Res.*, vol. 3, pp. 507–554, Nov. 2002.

[26] D. M. Chickering and C. Meek, "Selective greedy equivalence search: Finding optimal Bayesian networks using a polynomial number of score evaluations," 2015, *arXiv:1506.02113*.

[27] S. Zhu, I. Ng, and Z. Chen, "Causal discovery with reinforcement learning," 2019, *arXiv:1906.04477*.

[28] P. Bühlmann, J. Peters, and J. Ernest, "CAM: Causal additive models, high-dimensional order search and penalized regression," *Ann. Statist.*, vol. 42, no. 6, pp. 2526–2556, Dec. 2014.

[29] Y. Yu, J. Chen, T. Gao, and M. Yu, "DAG-GNN: Dag structure learning with graph neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7154–7163.

[30] Y. Annadani, N. Pawlowski, J. Jennings, S. Bauer, C. Zhang, and W. Gong, "BayesDAG: Gradient-based posterior inference for causal discovery," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 1738–1763.

[31] C. Cundy, A. Grover, and S. Ermon, "BCD nets: Scalable variational approaches for Bayesian causal discovery," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 7095–7110.

[32] D. Kalainathan, O. Goudet, I. Guyon, D. Lopez-Paz, and M. Sebag, "Structural agnostic modeling: Adversarial learning of causal graphs," *J. Mach. Learn. Res.*, vol. 23, no. 219, pp. 1–62, 2022.

[33] P. Lippe, T. Cohen, and E. Gavves, "Efficient neural causal discovery without acyclicity constraints," 2021, *arXiv:2107.10483*.

[34] I. Tsamardinos, L. E. Brown, and C. F. Aliferis, "The max-min hill-climbing Bayesian network structure learning algorithm," *Mach. Learn.*, vol. 65, no. 1, pp. 31–78, Oct. 2006.

[35] J. M. Mooij, S. Magliacane, and T. Claassen, "Joint causal inference from multiple contexts," *J. Mach. Learn. Res.*, vol. 21, no. 1, pp. 1–108, Jan. 2020.

[36] P. Erdős and A. Rényi, "On the evolution of random graphs," *Publ. Math. Inst. Hung. Acad. Sci.*, vol. 5, no. 1, pp. 17–60, 1960.

[37] B. Bollobás, C. Borgs, J. T. Chayes, and O. Riordan, "Directed scale-free graphs," in *Proc. SODA*, vol. 3. Baltimore, MD, USA, 2003, pp. 132–139.

[38] R. J. Prill, D. Marbach, J. Saez-Rodriguez, P. K. Sorger, L. G. Alexopoulos, X. Xue, N. D. Clarke, G. Altan-Bonnet, and G. Stolovitzky, "Towards a rigorous assessment of systems biology models: The DREAM3 challenges," *PLoS ONE*, vol. 5, no. 2, Feb. 2010, Art. no. e9202.

[39] Y. Cheng, Z. Wang, T. Xiao, Q. Zhong, J. Suo, and K. He, "CausalTime: Realistically generated time-series for benchmarking of causal discovery," in *Proc. 12th Int. Conf. Learn. Represent.*, 2024. [Online]. Available: https://openreview.net/forum?id=iad1yyyGme

[40] A. R. Lawrence, M. Kaiser, R. Sampaio, and M. Sipos, "Data generating process to evaluate causal discovery techniques for time series data," 2021, *arXiv:2104.08043*.

[41] M. Nauta, D. Bucur, and C. Seifert, "Causal discovery with attention-based convolutional neural networks," *Mach. Learn. Knowl. Extraction*, vol. 1, no. 1, pp. 312–340, Jan. 2019.

[42] E. De Brouwer, A. Arany, J. Simm, and Y. Moreau, "Latent convergent cross mapping," in *Int. Conf. Learn. Represent.*, 2020.

[43] T. Geffner, J. Antoran, A. Foster, W. Gong, C. Ma, E. Kiciman, A. Sharma, A. Lamb, M. Kukla, N. Pawlowski, M. Allamanis, and C. Zhang, "Deep end-to-end causal inference," in *Proc. NeurIPS Workshop Causality Real-World Impact*, 2022. [Online]. Available: https://openreview.net/forum?id=6DPVXzjnbDK

[44] M. Scutari, "Learning Bayesian networks with thebnlearnRPackage," *J. Stat. Softw.*, vol. 35, no. 3, pp. 1–22, 2010. [Online]. Available: https://www.jstatsoft.org/index.php/jss/article/view/v035i03

[45] T. Van den Bulcke, K. Van Leemput, B. Naudts, P. van Remortel, H. Ma, A. Verschoren, B. De Moor, and K. Marchal, "SynTReN: A generator of synthetic gene expression data for design and analysis of structure learning algorithms," *BMC Bioinf.*, vol. 7, no. 1, pp. 1–12, Dec. 2006.

[46] J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf, "Distinguishing cause from effect using observational data: Methods and benchmarks," *J. Mach. Learn. Res.*, vol. 17, no. 32, pp. 1–102, 2016. [Online]. Available: http://jmlr.org/papers/v17/14-518.html

[47] J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Glymour, M. Kretschmer, M. D. Mahecha, J. Muñoz-Marí, E. H. Van Nes, J. Peters, R. Quax, M. Reichstein, M. Scheffer, B. Schölkopf, P. Spirtes, G. Sugihara, J. Sun, K. Zhang, and J. Zscheischler, "Inferring causation from time series in Earth system sciences," *Nature Commun.*, vol. 10, no. 1, p. 2553, Jun. 2019. [Online]. Available: https://www.nature.com/articles/s41467-019-10105-3
[48] I. Guyon, C. Aliferis, G. Cooper, A. Elisseeff, J.-P. Pellet, P. Spirtes, and A. Statnikov, "Design and analysis of the causation and prediction challenge," in *Proc. Workshop Causation Predict. Challenge (WCCI)*, vol. 3, I. Guyon, C. Aliferis, G. Cooper, A. Elisseeff, J.-P. Pellet, P. Spirtes, and A. Statnikov, Eds., Hong Kong, Jun. 2008, pp. 1–33. [Online]. Available: http://proceedings.mlr.press/v3/guyon08a.html
[49] M. Chevalley, Y. Roohani, A. Mehrjou, J. Leskovec, and P. Schwab, "CausalBench: A large-scale benchmark for network inference from single-cell perturbation data," 2023, *arXiv:2210.17283*.
[50] S. L. Lauritzen, *Graphical Models*, vol. 17. Oxford, U.K.: Clarendon Press, 1996.
[51] J. Pearl, *Models, Reasoning and Inference*, vol. 19, no. 2. Cambridge, U.K.: Cambridge Univ. Press, 2000, p. 3.
[52] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA, USA: Morgan Kaufmann, 1988.
[53] P. Daniušis, D. Janzing, J. Mooij, J. Zscheischler, B. Steudel, K. Zhang, and B. Schölkopf, "Inferring deterministic causal relations," in *Proc. 26th Conf. Uncertainty Artif. Intell.*, 2010, pp. 143–150.
[54] P. Blöbaum, D. Janzing, T. Washio, S. Shimizu, and B. Schölkopf, "Analysis of cause-effect inference by comparing regression errors," *PeerJ Comput. Sci.*, vol. 5, p. 169, Jan. 2019. [Online]. Available: https://europepmc.org/articles/PMC7924496
[55] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Netw.*, vol. 13, nos. 4–5, pp. 411–430, Jun. 2000.
[56] C. K. Williams and C. E. Rasmussen, *Gaussian Processes for Machine Learning*, vol. 2, no. 3. Cambridge, MA, USA: MIT Press, 2006.
[57] L. Wang, S. Huang, S. Wang, J. Liao, T. Li, and L. Liu, "A survey of causal discovery based on functional causal model," *Eng. Appl. Artif. Intell.*, vol. 133, Jul. 2024, Art. no. 108258.
[58] K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan, "Causal protein-signaling networks derived from multiparameter single-cell data," *Science*, vol. 308, no. 5721, pp. 523–529, Apr. 2005.
[59] X. Wang, Y. Du, S. Zhu, L. Ke, Z. Chen, J. Hao, and J. Wang, "Ordering-based causal discovery with reinforcement learning," 2021, *arXiv:2105.06631*.
[60] I. Ng, A. Ghassami, and K. Zhang, "On the role of sparsity and DAG constraints for learning linear DAGs," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 17943–17954.
[61] I. Ng, S. Zhu, Z. Fang, H. Li, Z. Chen, and J. Wang, "Masked gradient-based causal structure learning," in *Proc. SIAM Int. Conf. Data Mining (SDM)*, 2022, pp. 424–432.
[62] I. Ng, S. Zhu, Z. Chen, and Z. Fang, "A graph autoencoder approach to causal structure learning," 2019, *arXiv:1911.07420*.

TINGPENG LI was born in 1987. He received the Ph.D. degree in mechanical engineering from the National University of Defense Technology (NUDT), in 2016, and Ph.D. degree in information engineering, in 2023. He has published a total of more than 30 papers in domestic and international authoritative journals. His research interests include intelligent testing and evaluation and intelligent mining of electromagnetic environment effect mechanisms.

LEI WANG received the master's degree from Shanxi University, China, in 2019. He is currently pursuing the Ph.D. degree with the School of Big Data and Software Engineering, Chongqing University, China. His research interests include causal learning and time series analysis.

DANHUA PENG was born in 1988. She received the Doctor of Engineering degree in computer science from the University of Rostock, Germany, in 2017. She has published a total of more than 20 academic papers in domestic and international authoritative journals. Her research interests include system simulation, simulation model standardization, and model reuse.

JUN LIAO received the master's degree from Guizhou University, China, in 2017. She is currently working as a Researcher and an Experimenter with the School of Big Data and Software Engineering, Chongqing University, China. Her research interests include the application of machine learning, data analysis, and causal intelligent analysis.

LI LIU (Member, IEEE) received the Ph.D. degree in computer science from Universite Paris-Sud, in 2008. He is currently working as a Professor with Chongqing University. He is also working as a Senior Research Fellow with the School of Computing, National University of Singapore. He aims to contribute in interdisciplinary research of computer science and human related disciplines. He has been the Principal Investigator of several funded projects from government and industry. He has published widely in conferences and journals with more than 100 peer-reviewed publications. His research interests include pattern recognition, data analysis, and their applications on human behaviors.

ZHENDONG LIU is currently working as a Professor with Chongqing City Vocational College. He works as an Expert in electronics and information with the Vocational Education Industry Guiding Committee and an Expert in think tank and providing policy consultation. He aims to contribute to promoting the positive impact of artificial intelligence and big data technology on vocational education. His research interests include artificial intelligence application and big data analysis.