Discovering biomedical causality by a generative Bayesian causal network under uncertainty

1st Ting Ye
School of Big Data & Software Engineering
Chongqing University
Chongqing, China
tingye@cqu.edu.cn

2nd Jun Liao
School of Big Data & Software Engineering
Chongqing University
Chongqing, China
liaojun@cqu.edu.cn

3rd Xuewen Yan
School of Big Data & Software Engineering
Chongqing University
Chongqing, China
yanxuewen@cqu.edu.cn

4th Hao Luo
School of Big Data & Software Engineering
Chongqing University
Chongqing, China
20151610@cqu.edu.cn

5th Wenbing Zhang
School of Big Data & Software Engineering
Chongqing University
Chongqing, China
WebbeZhang@cqu.edu.cn

6th Li Liu
School of Big Data & Software Engineering
Chongqing University
Chongqing, China
dcsliuli@cqu.edu.cn
*corresponding author

***Abstract*—With the rapid development of biomedical technology, discovering causality from genes and human physiological and pathological characteristics has become a hot but challenge spot over the past decades. Due to the increment of the amount of biomedical data, discovering causality from observed data becomes more and more difficult to search this large body of knowledge in a meaningful manner. To address the issues in existing causality discovering models, we introduce a generative Bayesian causal network that combines neural network to explicitly characterize these unique causal-effect relationships as a variable number of nodes and links. Particularly, a basic skeleton is generated for node selection to reduce the network size by minimizing the maximum mean discrepancy among variables. In addition, a causal generative neural network model is presented to construct causal network with cause-effect scores between variables. Empirical evaluations on two publicly available biomedical datasets and four synthetic datasets suggest our approach significantly outperforms the state-of-the-art methods in discovering causal relationships among biomedical variables.**

***Index Terms*—casual discovery, Bayesian networks, cause-effect dependency, biomedical application**

I. Introduction

Over the past decade approaches for discovering causal structure directly from large-scale biological datasets such as expression data and gene expression data has become a commonplace research field, given its role in uncovering novel biological insights, whose ultimate goal is to understand all the components of a biological system and how it works. Compared with traditional experimental methods, focusing on the analysis of single-cell functions, such collective observation datasets are much more practicable to analyze causal relationships than experimenting with complete interventions. In an experimental study, one or more variables are often randomly manipulated, and the corresponding effects on other variables are measured. Consequently, experiments on genes, which require complicate interventions, are expensive and difficult to put into force. On the other hand, causal relations between biological variables can be effectively explored and identified from these observation data, which are passively observed from experiments with or without full intervention, that may interfere with the biological system, resulting in guiding decision-making on future experiments or studying a complex organism such as a living organism and revealing the complex interactions between cells.

How to find meaningful relationships from these massive biomedical data, especially causality, is one of the most promising areas of biology and biomedicine. So far causal discovery is receiving widespread attention in this field [5]–[7]. Causality strictly distinguishes between *causal variables* and *effect variables*, playing an important role in revealing the mechanism of occurrence and guiding intervention behaviors [9]. For example, smoking and yellow teeth have a strong correlation with lung cancer, but only smoking is the cause of lung cancer. Quitting smoking can reduce the incidence of lung cancer, while cleaning teeth cannot reduce the risk of lung cancer. Actually in this case yellow teeth is another effect of smoking which has high correlation with lung cancer but is independent of it. Such causal relations among the three variables are theoretically named *v-structure*.

Despite being a very challenging problem, in recent years

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

there has been a rapid growth of interest in causal discovery. Bayesian causal network is the most commonly used causal representation in causal discovery that estimate the causal relationships of all variables in terms of nodes and edges as well as their joint probability distributions under Markov blanket property. The Bayesian network model has many advantages for representing and learning causality from observed data. First, it is flexible that can naturally incorporate prior knowledge. Also, it has considerable ability to derive models from observed data, and it can combine expert knowledge and observed data to improve model performance. Additionally, it can handle incomplete data, which often appear in biomedical applications. Most importantly, its structures and parameters have clear meanings, allowing it explicitly to represent causal relationships. Many advances have been made on Bayesian causal network in the areas of model evaluation and scoring as well as model search [11], since the Bayesian model is expressive and intuitive, which is commonly expected to be a representative of molecular biological processes. For instance, it is described that graphical model is one of the most promising approaches to represent cellular pathways [26]. Bayesian findings of causal networks (with potential variables) have been widely used in systems biology studies, especially in learning from a variety of experiments about gene networks [27], [28].

Existing Bayesian structure-based methods for causality discovery can be roughly divided into three categories: score-based methods, constraint-based methods and function-based methods. The score-based methods, as their names imply, scoring each possible resulting network, are more interpretive but more complex than constraint-based methods. The typical score-based methods, such as GES algorithm [12], GIES algorithm [13], among others, normally search for a graph that represents the correct causality by first increasing the explanatory power through edge adding and then reducing the complexity via edge removing. Subsequently, the explanation with the highest score is found through such search process. However, they have high time complexity, which is NP-hard, due to the involvement of network structure search process. On the other hand, the constraint-based methods are more computationally effective than score-based methods, which mainly includes two stages: causal skeleton learning and causal direction inference. By defining causal Markov assumption, conditional independence is used to test causal skeleton from observation variables, and then v-structure is explored to determine causal directions. The typical constraint-based methods include PC algorithm [10], FCI algorithm [11], and etc. The main limitation of such methods is that they may fail to distinguish the potential causal structures from statistically equivalent structures, which is called *Markov equivalence problem*, and consequently, they return uncertain causal directions. To address such problem in constraint-based methods, causal function models are put forward from the perspective of the data distribution characteristics caused by the causal mechanism. These models are based on the structural equation model (SEM) [14], which is a framework that can be used for multivariate analysis, including random variable sets and equation sets. Random variable sets include both observation variables and implicit error variables. A set of structural equations corresponds to a directed graph of a node as observed variables, which implies the causal structure of the model and the form of the structural equation. Although SEM can be used for multivariate analysis, in many cases, the classical SEM cannot estimate the causal direction of variables. To this end, causal data generation mechanism is incorporated to obtain a causal function model by extending SEM with representative capability. A variety of causal discovery algorithms are improved by leveraging causal function model, such as Linear Non-Gaussian Acyclic Model (LiNGAM) [15], Post-NonLinear (PNL) [16], Additive Noise Model (ANM) [17], and so on. These approaches need to search the entire dataset during the learning process, which are computationally expensive, and would end up being intractable with the growth of variable size.

To address these issues in causal discovery for biomedical applications, we present a generative neural network-based framework for learning casual Bayesian structure from observed biomedical data. In particular, our approach considers a principled way of dealing with the inherent causal variability among nodes associated with biomedical properties, which combines the capability of deep learning with the interpretability of causal models. Briefly speaking, to discover representative nodes in biomedical properties, a basic skeleton is generated by a generative feature selection model from a representative subset which is chosen from original biomedical observation dataset. Specifically, we propose to construct a basic network skeleton by minimizing the *maximum mean discrepancy* between the chosen variables and the observed data. In this way, the network size can be reduce to guarantee computational efficiency while remaining cause-effect dependency. By leveraging a *generative neural network*, causal inference is conducted to orientate undirected edges over the skeleton to obtain a causal network. In addition, a *causal generative neural network model* is presented to further improve the resulting causal network with cause-effect scores between variables. In this way, our causal network-based approach is more capable of discovering the inherit cause-effect dependency in biomedical variables when compared to existing methods, which is also verified during empirical evaluations.

The structure of the paper is as follows. First, in section II, we define the terminology and assumptions associated with our model. Then our model is introduced in section III. Section IV introduces the datasets and evaluation metrics. The experimental results are reported in section V. Conclusions are given in section VI.

II. Definitions and Assumptions

A. Causal Network

A causal network is generally represented by a directed acyclic graph (DAG) with probabilistic dependencies between variables, which can be denoted by a triplet $\mathcal{G} = (X, E, P)$.

$X = \{x_1, x_2, \ldots, x_n\}$ represents the set of all nodes in the network. $E = \{e(x_i, x_j) | x_i, x_j \in X\}$ represents the set of one-way edges between any pair of nodes, where $e(x_i, x_j)$ represents a dependency $x_i \rightarrow x_j$ between x_i and x_j. $P = \left\{P\left(x_i | pa_{x_j}\right) | x_i, pa_{x_i} \in X\right\}$ is a set of conditional probabilities, where $P(x_i | pa_{x_i})$ represents the probabilistic influence of the parent node set pa_{x_i} of x_i.

It can be seen that a causal network is a directed acyclic graph, where nodes represent biomedical variables, while edges between nodes represent direct casual dependencies between variables. In addition, each node is associated with a probability distribution. The root node $r \in X$ is attached to its edge distribution $P(r)$, while a non-root node $x \in X$ is attached to the conditional probability distribution $P(x_i | pa_{x_i})$ of x_i. Note that a casual network is actually a graphical representation of the joint probability distribution $P(x_i, x_2, \ldots, x_n)$.

B. *The d-separation*

The d-separation criterion is an important property for describing the causal relationship between nodes in a causal network. Let $U, V, W \subset X$ be the set of any three disjoint nodes in a directed acyclic graph $\mathcal{G}$. We call the node set W d-separates node sets U and V in graph $\mathcal{G}$, if a path *p* for any node from U to V is blocked by W, that is, a node v_i on path *p* satisfies one of the following conditions:

- v_i has a collision arrow on *p*, namely $\rightarrow v_i \leftarrow$, and neither v_i nor its descendants belong to W.
- v_i does not have a collision arrow on *p*, that is, $\rightarrow v_i \rightarrow$ or $\leftarrow v_i \rightarrow$, and $v_i \in W$.

According to the probability density implication of the d-separation criterion, if the sets U and V are d-separated by the sets W, then U and V are independent given W and conversely, if sets U and V are not d-separated by sets W, then U and V are interdependent given W.

C. *Markov Blankets*

Given any node x_i in $\mathcal{G}$, its parent node set is pa_{x_i} and its child node set is ch_{x_i}. The parent node set of each node in the child node-set are called the Markov blanket of node x_i. A node x_i and the set of nodes in the directed acyclic graph $\mathcal{G}$ that do not belong to the Markov blanket are d-separated by the Markov blanket. That is, the node x_i and nodes in the Bayesian network that does not belong to its Markov blanket are conditional independent of Markov blanket. Under certain conditions, the conditional independent relation in the probability pattern $P(X)$ for the same problem corresponds to the d-separation relation in the Bayesian network $\mathcal{G}$.

D. *Casual Assumption*

Currently, there are three main assumption for causal discovery of biomedical variables, namely, causal sufficiency assumption, causal Markov assumption and causal faithfulness assumption.

A variable set is considered sufficient when all direct causes of any two variables belongs to a variable set. It is called *causal sufficiency assumption*. In other words, there are no common confounders of the observed variables in $\mathcal{G}$.

For a set of variables with causal sufficiency, if all variables are conditionally independent of their non-descendant nodes under the condition of variables' parent nodes, we call this case satisfying *causal Markov assumption*.

Given the variable set $X = (x_1, \ldots, x_n)$, if the variable x_i and x_j are independent or conditionally independent, then in the causal graph $\mathcal{G}$ consisting of the variables and their causal dependencies, all paths between x_i and x_j are d-separation by the appropriate variable in the variable set X. We called that all the joint distribution of random variables P and graph $\mathcal{G}$ satisfies the *causal faithfulness assumption*. The implication of the causal faithfulness assumption is that no additional (conditional) independent relationships between variables occur during causal discovery. Under the causal faithfulness assumption, the model not only contains structural equations defined on variables or variable sets but also in the real situation. The real function form and the real value of coefficients have no additional implicit constraints.

The establishment of a causal network is generally based on its implicit assumptions. This inspires us to present in what follows a model where these networks can be systematically discovered to construct a resulting causal network to discover the cause-effect pairs among biomedical variables.

III. Our Approach

In this section, we present a novel framework that learns multivariate causal network structure under uncertainty for biomedicine. The main procedure of our model is illustrated in Figure 1.

A. *The Framework*

Given a dataset $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N]^{\mathrm{T}} \in \mathbb{R}^{N \times M}$, which is a matrix of N samples with M biomedical properties. It first selects a representative subset based on the original observation dataset by iteratively searching variables (nodes) with generative steps to obtain a basic skeleton with undirected edges. After then, a generative neural network is used for causal network construction to determine directions in the skeleton. Finally, a causal generative neural network model is presented to optimize the resulting causal network with cause-effect scores among variables. The specific process is shown in Figure 1.

B. *Representative Subset Selection*

In active learning, the number of samples is reduced by selecting the most representative sample to represent the whole training set. Uncertainty sampling is a method of active learning, but it is easily influenced by outliers. We present a sampling method based on k-NN to solve the problem caused by outliers by the following definitions:

$$\mathcal{D} = \sum_{\mathbf{x}_i \in \mathbf{X}} I_2(\mathbf{X}, \mathbf{x}_i), \tag{1}$$

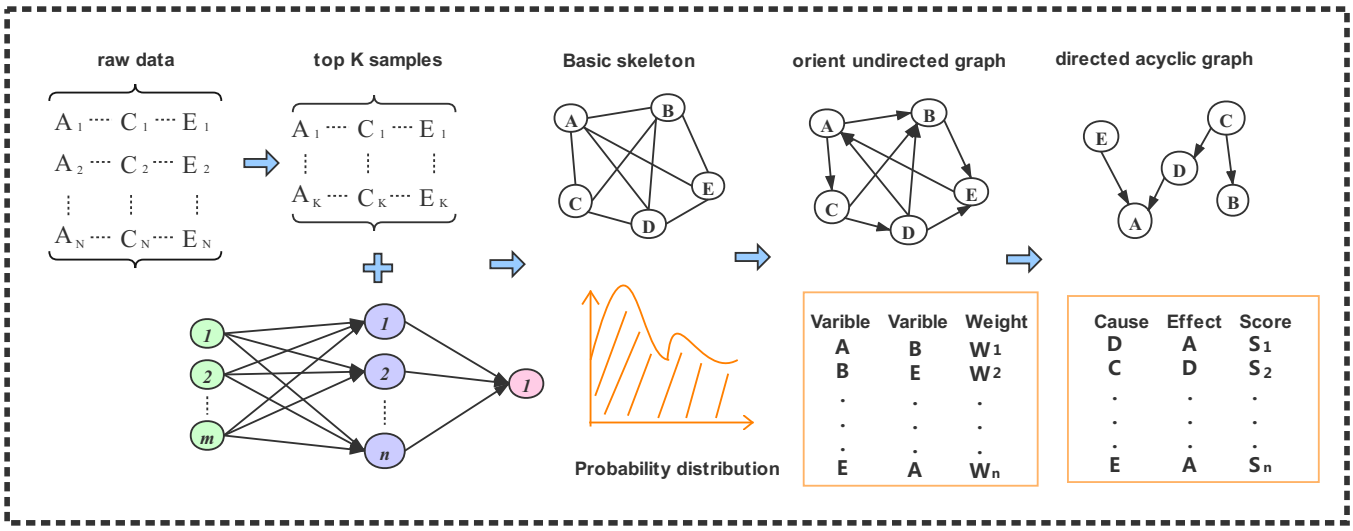

Fig. 1: Our framework of discovering casual network structure with generative neural networks.

Given an unlabeled dataset $\mathbf{X}$, we can use Equation (1) to find the representative subset $\mathcal{D}$ of $\mathbf{X}$, where

$$I_2\left(\mathbf{X}, \mathbf{x}_i\right) = \sum_{\mathbf{x}_i \in \mathrm{TopKdist}(\mathbf{x}_j, \mathbf{X}, k)} I_1\left(\mathbf{X}, \mathbf{x}_i\right), \tag{2}$$

where $\mathrm{TopKdist}\left(\mathbf{x}_j, \mathbf{X}, k\right)$ represents the k most similar samples to $\mathbf{x}_j$ in $\mathbf{X}$, where k is constant. $I_2\left(\mathbf{X}, \mathbf{x}_i\right)$ can measure the probability of finding the k most similar samples using test data in set $\mathbf{X}$ and the strength of the similarity. $I_1(\mathbf{x}_i, \mathbf{x}_j)$ is defined as:

$$I_1(\mathbf{x}_i, \mathbf{x}_j) = \mathrm{dist}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{m=1}^{M} \left(\mathbf{x}_i^m - \mathbf{x}_j^m\right)^2}, \tag{3}$$

where $\mathrm{dist}(\mathbf{x}_i, \mathbf{x}_j)$ represents the similarity between samples $\mathbf{x}_i$ and $\mathbf{x}_j$.

C. Basic Skeleton Identification

Due to the super-exponential complexity of the deep neural networks, we first determine the skeleton of our causal network through a feature selection model from the set of continuous variables $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_M\}$, where a variable $\mathcal{D}_i = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$. Then, the selected variables are taken as candidates for the network, and the undirected edges are determined between variables according to a feature selected score threshold. The process runs iteratively and finally returns feature selection scores for each node, which can be regarded as the weight of the node.

Given a set of continuous observed variables $\mathcal{D} = (\mathcal{D}_1, \ldots, \mathcal{D}_M)$ and $\mathcal{G}_u = (\mathcal{D}, f, \mathcal{E})$ be an undirected graph over $\mathcal{D}$, the relationship among variables satisfies the following formula:

$$\mathcal{D}_i \leftarrow f_i\left(\mathcal{D}_{\mathrm{Pa}(i,\mathcal{G}_u)}, E_i\right), E_i \sim \mathcal{E}, \text{ for } i = 1, \ldots, M \tag{4}$$

These continuous random variables with joint distribution P can be factorized over $\mathcal{G}_u$ as follows:

$$P(\mathcal{D}) = \prod_i P\left(\mathcal{D}_i | \mathcal{D}_{Pa(i;\mathcal{G}_u)}\right), \tag{5}$$

By taking both the ground samples $\mathcal{D}$ and the generated samples $\widehat{\mathcal{D}}$ in any order and returning the score between the two empirical distributions, the score is defined by:

$$\begin{aligned} \widehat{\boldsymbol{\xi}} = & \frac{1}{n^2} \sum_{i,j=1}^{n} k\left(\mathbf{x}_i, \mathbf{x}_j\right) + \frac{1}{n^2} \sum_{i,j=1}^{n} k\left(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j\right) \\ & - \frac{2}{n^2} \sum_{i,j=1}^{n} k\left(\mathbf{x}_i, \hat{\mathbf{x}}_j\right), \end{aligned} \tag{6}$$

where kernel k is computed by $k\left(\mathbf{x}, \mathbf{x}'\right) = \exp\left(-\gamma \left\|\mathbf{x} - \mathbf{x}'\right\|_2^2\right)$, which is usually regarded as the Gaussian kernel. $\widehat{\boldsymbol{\xi}}$ has a quadratic complexity, which has the property that as n increases infinitely it decreases to zero if and only if $P = \hat{P}$.

D. Edge Orientation and Optimization

Based on the above process, we obtain a casual network skeleton $\mathcal{G}_u$, which is an undirected graph. Then we need to determine the direction of the edges between variables. Pairwise variant of our approach models the causal directions $x_i \rightarrow x_j$ and $x_i \leftarrow x_j$ with a 1-hidden layer neural network. The causal direction is considered as the best-fit between the two causal directions.

Algorithm 1: The framework of discovering multivariate causal network structure under uncertainty for biomedical variables.

Input: The set of variables for current batch $\mathbf{X}$, threshold Θ

Output: The resulting network $\mathcal{G}$ for the variables sets $\mathbf{X}$ and the scores $\mathcal{S}$ of causal-effect between two variables in $\mathcal{G}$

Initialization parameters;
$I(\mathbf{x}_i) = \sum_{\mathbf{x}_j \in X} \mathrm{dist}(\mathbf{x}_i, \mathbf{x}_j)$;
$\mathcal{D} = \phi$;
$\mathrm{TopKdist}(\mathbf{x}_i, \mathcal{D}, k) = \phi$;
while $|\mathcal{D}| < n$ **do**
    $\mathbf{x}_{selected} = \mathrm{argmax}(I(\mathbf{x}_i))$;
    for $\mathbf{x}_i \in \mathbf{X}$ **do**
        if $\mathrm{dist}(\mathbf{x}_{\text{selected}}, \mathbf{x}_i) > \mathrm{Kthdist}(\mathbf{x}_i, \mathcal{D}, k)$: **then**
            $\mathrm{TopKdist}(\mathbf{x}_i, \mathcal{D}, k).add(\mathbf{x}_{selected}, \mathbf{x}_i)$;
        if $|\mathrm{TopKdist}(\mathbf{x}_i, \mathbf{X}, k)| > k$: **then**
            //find out the minimal score
            $\mathbf{x}_{removed} = \mathrm{argmin}\,\mathrm{TopKdist}(\mathbf{x}_i, \mathcal{D}, k)$;
Calculate the feature selected score $\mathcal{S}_{ij}$ between two variables x_i and x_j in the subset $\mathcal{D}$;
if *score* $\mathcal{S}_{ij} > \Theta$ **then**
    added x_i and x_j into the network skeleton $\mathcal{G}_u$; edge $x_i - x_j$ is added at the same time with the weight between variables;
Orientate each variable $x_i - x_j$ as $x_i \to x_j$ or $x_j \to x_i$ among $\mathcal{D}$ by selecting associated two-variable approach;
Traverse paths from a random set of nodes with the edges pointing towards a visited node reveal cycles must be reversed until all nodes are reached;
For a number of iterations, reverse the edge that leads to the maximum improvement of the model score $S(\mathcal{G}, \mathcal{D})$;
return $\mathcal{G}$ and $\mathcal{S}$;

There exists $\hat{f} = \left(\hat{f}_1, \ldots, \hat{f}_M\right)$, where $\hat{f}_i$ a 1-hidden layer regression neural network with n_h hidden neurons such that $P(\mathcal{D})$ equals the generative model defined from $(\mathcal{G}, f, \mathcal{E})$:

$$\begin{aligned}\hat{D}_i &= \hat{f}_i\left(\hat{D}_{Pa(i;\mathcal{G})}, E_i\right) \\ &= \sum_{k=1}^{n_h} \bar{w}_k^i \sigma\left(\sum_{j \in Pa(i;\mathcal{G})} \hat{w}_{jk}^i \hat{D}_j + w_k^i E_i + b_k^i\right) + \bar{b}^i,\end{aligned} \tag{7}$$

where n_h is the number of hidden units, $\bar{w}_k^i$, $\hat{w}_{jk}^i$, w_k^i, b_k^i are the parameters of the neural networks, and σ is an activation function.

$$\forall e^{(M)}, \left\| z_M\left(e^{(M)}\right) - \widehat{z_M}\left(e^{(M)}\right)\right\| < \epsilon, \tag{8}$$

Let Z_M be the set of variables with topological order less than M and s_M be its size. For any s_M-dimensional vector of noise values $e^{(M)}$, let $z_M\left(e^{(M)}\right)$ be the vector of values computed in topological order from f. For any $\epsilon > 0$, there exists a set of networks $\hat{f}$ with $\mathcal{G}$.

A n-sample set $\hat{\mathcal{D}}$ sampled after the joint distribution $\hat{P}$ defined by the casual model estimated $(\hat{\mathcal{G}}, \hat{f}, \mathcal{E})$. In fact, the casual model is trained by minimizing $S(\widehat{\mathcal{G}}, \mathcal{D})$, which is a scoring metric on model evaluation defined as:

$$S(\widehat{\mathcal{G}}, \mathcal{D}) = -\widehat{\boldsymbol{\xi}}(\mathcal{D}, \widehat{\mathcal{D}}) - \lambda|\widehat{\mathcal{G}}|,$$

where $|\widehat{\mathcal{G}}|$ means the number of edges in $\widehat{\mathcal{G}}$, and λ is a penalization weight.

We use T to indicate the number of edges in the skeleton, and then define an orient edge optimization problem, the complexity of which is $O(2^T)$. Note that not all orient edges are remained, because the searching process must end up with a directed graph that is DAG. The purpose of this step is to decouple the edge selection task and the edge orientation task, and enable them to be evaluated independently. Any edge $x_i - x_j$ in the skeleton represents a direct dependency between the variables x_i and x_j. We consider the causal Markov condition and the causal faithfulness assumptions, where such direct dependence either reflects a direct causal relationship between two variables ($x_i \to x_j$ or $x_i \leftarrow x_j$), or x_i and x_j acknowledge a potential (unknown) common cause.

The general process is as follows:

- First, consider each $x_i - x_j$ edge individually, and then use our approach to evaluate its direction. Calculate the scores of the two oriented edges $S\left(\mathcal{C}_{x_i \to x_j, \hat{f}}, x_{ij}\right)$ and $S\left(\mathcal{C}_{x_i \leftarrow x_j, \hat{f}}, x_{ij}\right)$ simultaneously, where $x_{ij} = \{[\mathbf{x}_{i,q}, \mathbf{x}_{j,q}] \,|q = (1, \ldots, n)\}$. Keep the minimum score corresponding to the best direction. After this step, the complexity of the initial graph is 2^T.
- Modify the initial diagram to remove all circles. Starting with a set of random variables, traverse all paths until all variables are reached, with the edges pointing to the visited nodes and reversing the edges in a cycle. Finally, we get the DAG as the initial graph of the oriented edge optimization.
- The optimization of DAG structure is accomplished by a hill-climbing algorithm, which aims to optimize the global score $S\left(\mathcal{C}_{\mathcal{G}, \hat{f}}, \mathcal{D}\right)$. Iteratively, i) select an edge $x_i - x_j$ uniformly randomly in the current graph; (ii) consider the graph obtained by inverting this edge (if it is still a DAG and was not previously considered) and retrain the relevant global score; iii) if the graph has a lower global score than the previous graph, it becomes the current graph, and the process is iterated until a (local) optimum is reached. In this paper, we use the method of hill-climbing to achieve a reasonable balance between computational time and accuracy performance.

Table I: Summary of the two publicly available datasets and the four synthetic datasets.

Datasets	No. of observations	No. of variables
Sachs	7465	11
Dream4	100	100
NN	9000	10
PN	9000	10
LIN	9000	10
SIG	9000	10

IV. EXPERIMENTAL SET-UPS

A. *Datasets*

Two real datasets and four synthetic datasets as shown in Table I are used to evaluate the construction of causal network structure in biomedicine.

Sachs dataset [21]: It consists of observational data collected after general perturbation, which relies on simultaneous measurement of single cell expression profiles of 11 pyrophosphate proteins involved in a signaling pathway of human primary T cells. It contains 7465 observation samples.

Dream4 dataset [22]: It is a commonly used dataset which provides five different structured networks that reflects the common topological characteristics of real gene regulatory networks in E.coli or S.cerevisiae, including feedback loops. Each subnetwork of 100-node contains 100 samples.

Synthetic datasets: Given a causal mechanism, we randomly generate data and their corresponding acyclic graph.

- **NN**: 9000 artificial samples are generated with a neural network initialized with its random weights and random distribution for the cause.
- **PN**: 9000 artificial samples are generated by a polynomial causal mechanism. The effect variables are built with post multiplicative noise $(Y = f(X) \times E)$ or pre-multiplicative noise $(Y = f(X \times E))$.
- **LIN**: 9000 artificial samples are generated by a linear causal mechanism. The effect variables are built with post additive noise setting $(Y = f(X) + E)$ or pre-additive noise $(Y = f(X + E))$.
- **SIG**: 9000 artificial samples are generated by a sigmoid-Mix causal mechanism initialized with random weights and random distribution for the cause.

The function used to initiate variables of the graph defaults to a Gaussian Mixture model. The number of variable nodes is set to 10. The proportion of Gaussian noise in the mechanisms is set to 0.4.

B. *Baseline Approaches*

To evaluate the effectiveness of our model, we compared it with other eight competitive causal discovery methods.

- **PC** [10]: It is a typical constraint-based algorithm. We implemented it using pcalg-R [29]. Fisher Z-Score conditional independence test is used as a conditional independence test for determining the skeleton of the graph.
- **GES** [12]: It is a score-based algorithm that searches heuristically the graph which minimizes likelihood scores. It was implemented by pcalg-R [29]. L0-penalized Gaussian maximum likelihood estimator is used for scoring the candidate causal networks.
- **GIES** [13]: It is a variant of GES that it accepts interventional data for its inference.
- **LiNGAM** [15]: It is a SEM method that handles linear structural equation models, where each variable is modeled as $X_j = \sum_k \alpha_k P_a^k(X_j) + E_j, j \in [1, d]$, with $P_a^k(X_j)$ the k-th parent of X_j and α_k a real scalar.
- **GS** [31]: It is a constraint-based algorithm to recover bayesian networks. It consists in two phases, one growing phase in which nodes are added to the markov blanket based on conditional independence and a shrinking phase in which most irrelevant nodes are removed. It is implemented by bnlearn-R [30].
- **IAMB** [32]: It is a constraint-based algorithm to recover Markov blankets in a forward selection and a modified backward selection process.
- **Fast-IAMB** [33]: Similar to IAMB, Fast-IAMB adds speculation to provide more computational performance without affecting the accuracy of markov blanket recovery.
- **Inter-IAMB** [33]: It is another variant of IAMB which has a progressive forward selection minimizing false positives.

C. *Evaluation Metrics*

We used AUPR, SHD, and SID to evaluate the performance of the competing methods on learning causal structures:

Area Under the Precision/Recall Curve(AUPR): AUPR is a single number summary of the information in the precision-recall (PR) curve.

Structural Hamming Distance(SHD) [24] [25]: SHD considers two partially directed acyclic graphs and calculates how many edges do not coincide, i.e., the number of edges that changes to convert one graph to another.

Structural Intervention Distance(SID) [23]: SID estimates the number of equivalent bivariate interventions between the two graphs. It is based solely on graphical standards and quantifies the proximity between two DAGs based on the corresponding causal inference statement.

D. *Experimental Settings*

The neural network structure of our model is designed as a 1-hidden layer network with ReLU activation function. The number of hidden units for each generative neural network is set to 20. The bandwidth γ range of multi-scale Gaussian kernel used in the score function is $(0.005, 0.05, 0.25, 0.5, 1.5, 50)$. Since our method is a probabilistic model, in order to have a stable evaluation of the operation, we run our model 12 times for each test. The distribution $\mathcal{E}$ of the noise variables is set to $\mathcal{N}(0, 1)$. We used Adam tool to train data with 0.01 learning rate until convergence, and evaluated it on the generated samples. All experiments run on an Intel Xeon 2.5GHz CPU, and four NVIDIA GTX 1080Ti GPU.

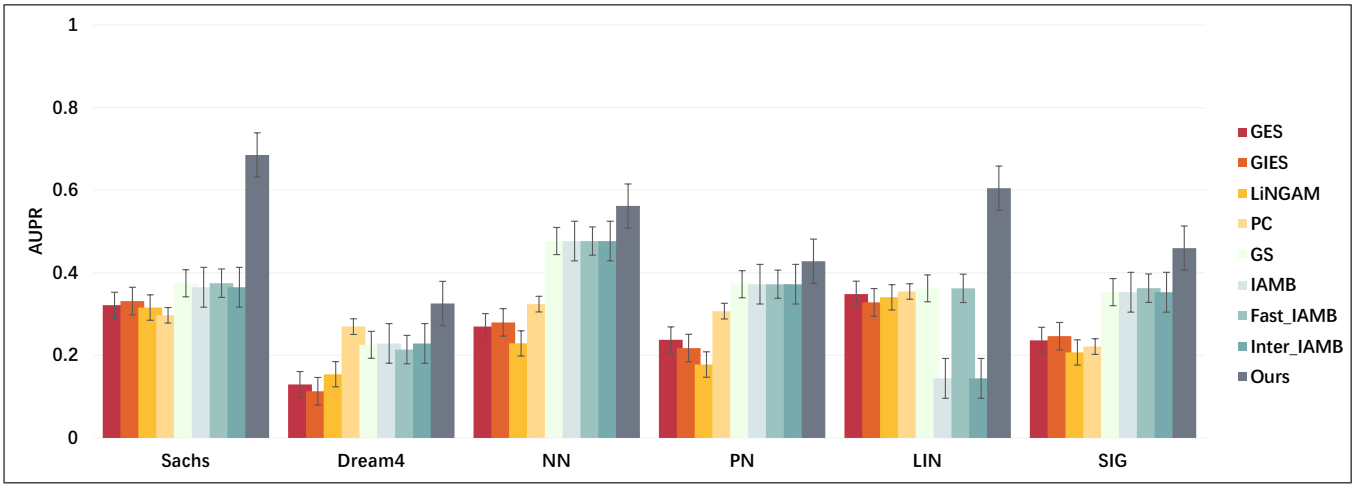

Fig. 2: AUPR comparison.

V. Experimental Results

A. *AUPR Comparison*

Figure 2 shows the comparisons of the averaged AUPR results of the eight competing methods on the six datasets. It can be viewed that on either real biomedical datasets or synthetic datasets, the AUPR score of the causal network generated by our proposed method is better than the others. In general, our proposed model is relatively superior to methods in capturing causal relationships. In addition, it is clear that our method is much more accurate than other methods of causal discovery (about 15%-50% performance improvement). The main reason is that some of these competing approaches (e.g., GES and GIES) use score function based on maximum likelihood estimation, so that the network structure can be fitted to the maximum to obtain the network structure with the highest degree of fit to the dataset. However, the ultimate goal is not to infer causal between variables. On the other hand, other competing methods analyze the probability distribution between nodes based on the latent structure in the network. By analyzing conditional independence tests between nodes, certain rules are used to infer causality. However, such rules are often subjective and do not fully consider the causal relation implied between the variables. Particularly, PC cannot make correct inferences about Markov equivalence classes. It can only discover partial causality, and thus it cannot make accurate inferences about edge directions.

Notably, regardless of the sizes of the datasets, our method outperforms others in terms of AUPR, even in the case of high-dimensional samples. Obviously, when the data dimension becomes higher, the performance of all the competing methods declines gradually, especially for the scores-based methods. To infer the edge direction of the network, these methods search all possible network structures, which increase exponentially with the number of variables, resulting in high time complexity. In addition, the conditional independence test often fails to get an ideal result as the dimension of the conditional set is too high. So these competing methods are only applicable to the network with dozens of dimensions.

B. *SHD and SID Comparison*

As shown in Table II and Table III, compared with other algorithms, our method performs better results than other competing methods. It further emphasizes the fact that when the skeleton is known, using the structure of the graph will yield better results than the methods that only use local information.

Table II: The SHD comparison of causal networks constructed by competing algorithms.

Method	Sachs	Dream4	NN	PN	LIN	SIG
GES	39	522	36	37	16	36
GIES	39	522	36	37	16	36
LiNGAM	18	285	29	31	14	**17**
PC	27	260	19	27	15	25
GS	25	261	18	27	13	24
IAMB	25	266	18	27	14	24
Fast-IAMB	25	260	18	27	13	24
Inter-IAMB	25	263	18	27	14	24
Ours	**16**	**234**	**17**	**26**	**12**	19

Table III: The SID comparison of causal networks constructed by competing algorithms.

Method	Sachs	Dream4	NN	PN	LIN	SIG
GES	76	7628	55	78	**26**	70
GIES	76	7628	55	78	**26**	70
LiNGAM	80	6512	79	72	40	77
PC	82	7309	62	80	45	82
GS	95	5630	83	73	39	76
IAMB	89	5993	83	73	49	76
Fast-IAMB	95	5614	83	73	39	78
Inter-IAMB	89	5932	83	73	49	76
Ours	**72**	**4772**	**43**	**62**	36	**46**

C. *Robust Test*

The two real datasets were used to verify the robustness of our proposed method by perturbing edges in the network

skeleton. The network skeleton was changed by perturbation of 10% and 20% edges. As shown in Table IV, all the methods have lower scores after introducing false edges into the graph skeleton. Our proposed method still performs the best, since our proposed method takes advantage of conditional independence as well as distribution asymmetry. The least robust methods are constraint-based methods, because they rely heavily on the structure of the graph to determine the direction of the edges.

Table IV: Performance under synthetic errors of edges in network skeleton.

Dataset	Method	Original skeleton			Under 10% synthetic error			Under 20% synthetic error		
		AUPR	SHD	SID	AUPR	SHD	SID	AUPR	SHD	SID
Sachs	**PC**	0.31	27	82	0.20	43	119	0.19	73	134
	GES	0.30	39	76	0.24	56	120	0.21	89	178
	GIES	0.30	39	76	0.24	56	120	0.21	89	178
	LiNGAM	0.29	18	80	0.17	49	148	0.12	88	198
	GS	0.38	25	95	0.29	45	167	0.18	87	193
	IAMB	0.37	25	89	0.30	44	179	0.24	68	204
	Fast-IAMB	0.38	25	95	0.29	45	167	0.18	87	199
	Inter-IAMB	0.37	25	89	0.30	44	179	0.24	68	204
	Ours	**0.68**	**16**	**72**	**0.58**	**25**	**102**	**0.51**	**37**	**125**
LIN	**PC**	0.38	15	45	0.33	43	75	0.23	68	94
	GES	0.37	16	**26**	0.34	39	**56**	0.29	67	78
	GIES	0.35	16	**26**	0.34	39	**56**	0.29	67	78
	LiNGAM	0.36	14	40	0.29	37	76	0.22	64	98
	GS	0.37	13	39	0.33	39	58	0.27	65	86
	IAMB	0.17	14	49	0.14	43	84	0.11	68	119
	Fast-IAMB	0.37	13	39	0.33	39	58	0.27	65	86
	Inter-IAMB	0.17	14	49	0.14	43	84	0.11	68	119
	Ours	**0.61**	**12**	36	**0.57**	**34**	**56**	**0.52**	**61**	**73**

VI. Conclusion

In this work we present an effective causal network framework that combines deep neural network to discover cause-effect relations among biomedical variables. Empirical experiments show the effectiveness of our framework on various biomedical datasets. It is verified that our framework is more efficient and robust than existing methods for biomedical causality discovery. As for future work, we will consider improving optimization method in the framework, and we will further investigate the performance of our model on more biomedical tasks.

Acknowledgement

This work was supported by grants from the National Major Science and Technology Projects of China (grant nos. 2018AAA0100703, 2018AAA0100700), the National Natural Science Foundation of China (grant no. 61977012), the Chongqing Provincial Human Resource and Social Security Department (grant no. cx2017092), the Central Universities in China (grant nos. 2019CDJGFDSJ001).

References

[1] Mattmann C A, "Computing:A vision for data science," Nature, vol. 493, pp. 473-475, April 2013.

[2] McAfee A and Brynjolfsson E, "Big data:The management revolution," Harvard Business Review, vol. 90, pp. 60-68, 2012.

[3] Hey T, Tansley S and Tolle K, "The Fourth Paradigm Data Intensive Scientific Discovery," Redmond,USA Microsoft Research, 2009.

[4] McAfee A and Brynjolfsson E, "Big data:The management revolution," Harvard Business Review, vol. 90, pp. 60-68, 2012.

[5] Goto, T., Fernandes, A.F.A., Tsudzuki, M. et al., "Causal phenotypic networks for egg traits in an F2 chicken population," Mol Genet Genomics, pp. 1455C1462, 2019.

[6] Reshef D N, Reshef Y A and Finucane H K, "Detecting novel associations in large data sets," Science, vol. 334, pp. 1518-1524, 2011.

[7] Justin D. Finkle, Jia J. Wu, and Neda Bagheri, "Windowed Granger causal inference strategy improves discovery of gene regulatory networks ," PNAS, vol. 115, 2018.

[8] Brandon L. Pierce and Lin Tong, "Co-occurring expression and methylation QTLsallow detection of common causal variants andshared biological mechanisms," NATURE, 2018.

[9] Pearl J, "Causality Models Reasoning and Inference," Cambridge, 2rd ed., United Kingdom Cambridge University Press, 2009.

[10] P.Spirtes and C.N.Glymour, "An algorithm for fast recovery of sparse causal graphs," Social science computer review, vol. 9, pp. 62-72, 1991.

[11] P.Spirtes, C.N.Glymour and R.Scheines, "Causation,Prediction,and Search," Cambridge, vol. 90, 2rd ed., MIT Press, 2000.

[12] D.M. Chickering, "Optimal structure identification with greedy search," Journal of Machine Learning Research, vol. 3, pp. 507-554, 2002.

[13] A.Hauser and P.Bhlmann, "Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs," Journal of Machine Learning Research, vol. 13, pp. 2409-2464, 2012.

[14] Bollen K A, "Structural Equations with Latent Variables," John Wiley and Sons, 2014.

[15] Shimizu S, Hoyer P O and Hyvgrinen A, "A linear nonłGaussian acyclic model for causal discovery," Journal of Machine Learning Research, vol. 3, pp. 2003-2030, 2006.

[16] Zhang Kun and Hyvgrinen A, "On the identifiability of the postłnonlinear causal model," UAI, 2009.

[17] Hoyer P O, Janzing D and Mooij J M, "Nolinear causal discovery with additive noise models," NIPS, pp. 689-696, 2009.

[18] Rubin DB, "Bayesian inference for causal effects:The role of randomization," The Annals for empirical research, vol. 6, pp. 34-58, 1978.

[19] Pearl J, "Causal diagrams for empirical research," Biometrika, vol. 82, pp. 669-688, 1995.

[20] Lopez-Paz D and Muandet K, "Towards a learning theory of cause-effect inference," ICML, pp. 1452-1461, 2015.

[21] K.Sachs, O.Perez, D.Peer, D.A.Lauffenburger and G.P.Nolan, "Causal proteinsignaling networks derived from multiparameter single-cell data," Science, vol. 308, pp. 523-529, 2005.

[22] A. Greenfield, A. Madar, H. Ostrer and R. Bonneau, "Dream4: Combining genetic and dynamic information to identify biological networks and dynamical models," PloS one, vol. 5, pp. 10, 2010.

[23] J.Peters and P.Bhlmann, "Structural intervention distance (sid) for evaluating causal graphs," arXiv preprint arXiv, 2013.

[24] S.Acid and L.M.de Campos, "Searching for bayesian network structures in the space of restricted acyclic partially directed graphs," Journal of Artificial Intelligence Research, vol. 18, pp. 445-490, 2003.

[25] I.Tsamardinos, L.E.Brown and C.F.Aliferis, "The max-min hill-climbing bayesian network structure learning algorithm," Machine learning, vol. 65, pp. 31-78, 2006.

[26] Friedman N, "Inferring Cellular Networks Using Probabilistic Graphical Models," Science, vol. 303(5659), pp. 799-805, 2004.

[27] Beal M, et al , "A Bayesian approach to reconstructing genetic regulatory networks with hidden factors," Bioinformatics, vol. 21(3), pp. 349-356, 2005.

[28] Yoo C and Cooper G , "An Evaluation of a System that Recommends Microarray Experiments to Perform to Discover Gene-Regulation Pathways," Journal of Artificial Intelligence in Medicine, vol. 31, pp. 169-182, 2004.

[29] M. Kalisch, M. Machler, and D. Colombo, "pcalg: Estimation of cpdag/pag and causal inference using the ida algorithm," URL http://CRAN.R-project.org/package=pcalg.R package version, pp. 1-1, 2010.

[30] M. Scutari, "bnlearn: Bayesian network structure learning, parameter learning and inference," R package version, vol. 3, 2012.

[31] Margaritis D, "Learning Bayesian Network Model Structure from Data," School of Computer Science Carnegie-Mellon University, Pittsburgh, PA, vol. 3, 2003.

[32] Tsamardinos I, Aliferis CF and Statnikov A , "Algorithms for Large Scale Markov Blanket Discovery,",International Florida Artificial Intelligence Research Society Conference, pp. 376-381, 2003.

[33] Yaramakala S, Margaritis D, "Speculative Markov Blanket Discovery for Optimal Feature Selection,", ICDM, pp. 809-812, 2005.