Supplementary materials

Supplementary material S1. Feasibility of independence test

The algorithm makes the following assumptions:

1. The data follow a multivariate Gaussian distribution.
2. Each observed variable is associated with at most one hidden variable.
3. Hidden variables are not influenced by other variables.

In stage 1, we start from a fully connected network and use the Fisher Z-test as a conditional independence test. The Fisher Z-test is a correlation test, but we assume that the variables follow a multivariate Gaussian distribution, and a property of the multivariate Gaussian distribution is that if the partial correlation coefficient between the variables is 0, they are conditionally independent. Therefore, the Fisher Z-test can be used to perform the conditional independence test. The k-th order partial correlation coefficient of any two variables i, j is $r_{i,j|k}$:

$$r_{i,j|k} = r_{ij\cdot pq} = \frac{r_{ij\cdot p} - r_{iq\cdot p}\, r_{jq\cdot p}}{\sqrt{(1-r_{iq\cdot p}^2)(1-r_{jq\cdot p}^2)}} \quad (S1)$$

Eq. S1 provides partial correlation coefficients of arbitrary order, where p represents any subset within the set K and q is the complement of the subset p within K.

After obtaining $r_{i,j|k}$, we transform it into normal distribution by Fisher Z transformation:

$$Z(i,j|k) = \frac{1}{2}\log\left(\frac{1+r_{i,j|k}}{1-r_{i,j|k}}\right) \quad (S2)$$

We then define the null hypothesis and the alternative hypothesis:

Null hypothesis H_0: $r_{i,j|k} \neq 0$

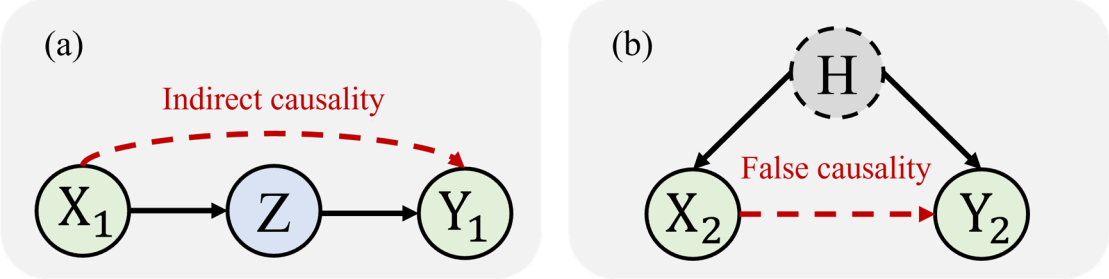
Alternative hypothesis H_1: $r_{i,j|k} = 0$

Given a significance level α ($0 < \alpha < 1$), a two-sided test can be performed using the following formula:

$$|Z(i,j|k)| \leq \Phi^{-1}\left(1-\frac{\alpha}{2}\right) \quad (S3)$$

where Φ represents the cumulative distribution function of the standard normal distribution, and Φ^{-1} is its inverse. If the inequality in Eq. S3 holds, we reject the null hypothesis that variables i and j are significantly correlated given the other variables in K. Then the edge between variables i and j is deleted.

Supplementary material S2. Feasibility of linear regression



Supplementary Figure S1. Indirect causality and false causality caused by hidden variables. (a)

Indirect causality. X_1 and Y_1 are indirectly causality. **(b) False causality.** X_2 is not causally related to Y_2, but X_2 and Y_2 are affected by the hidden variable *H*.

$$Y_1 = f(Z) + \widehat{e_1} \quad (S4)$$
$$Z = g(X_1) + \widehat{e_2} \quad (S5)$$
$$Y_1 = f(g(X_1) + \widehat{e_2}) + \widehat{e_1} \quad (S6)$$

$$X_2 = f(H) + \widehat{e_1} \quad (S7)$$
$$H = f^{-1}(X_2) + \widehat{e_2} \quad (S8)$$
$$Y_2 = g(H) + \widehat{e_3} \quad (S9)$$
$$Y_2 = g(f^{-1}(X_2) + \widehat{e_2}) + \widehat{e_3} \quad (S10)$$

In Supplementary Figure S1, assume that both f and g are functions fitted by regression and f^{-1} is the inverse function of f. When X_1 and Y_1 share an indirect causal relationship, Y_1 can be effectively modeled by X_1 as described by Eq. S6 (Supplementary Figure S1a), allowing X_1 to substitute for Z's influence. If X_1 is removed from the model while Z is present, the impact on predicting Y_1 is minimal because Y_1 can still be directly influenced by Z. Conversely, as depicted in Supplementary Figure S1b, if there's a hidden variable H influencing both X_2 and Y_2, and X_2 and Y_2 are not directly causally related, Y_2 can be modeled by X_2 using Eq. S10 (Supplementary Figure S1b). However, if X_2 is removed under this condition (Supplementary Figure S1b), the hidden nature of H means that no other variable can adequately account for H's influence on Y_2. Therefore, removing X_2 significantly increases the loss in fitting Y_2 (Supplementary Figure S1b).

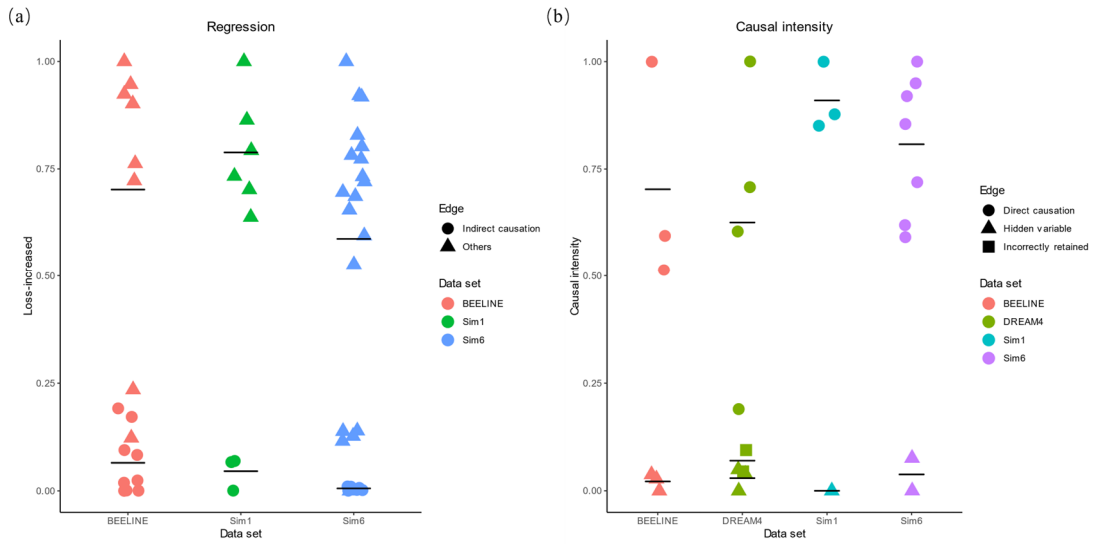
Supplementary material S3. Feasibility of threshold selection

The purpose of this section is to demonstrate that indirect causality can be distinguished from the other two types of causality (direct causality and pseudo-causality due to the effect of hidden variables) based on the change in $\Delta\mathrm{Loss}$ from the linear regression in Stage 2 (Figure S1). Furthermore, direct causality and pseudo-causality can be distinguished in Stage 3 using the causal intensity indicator (Figure S1).

As shown in Supplementary Figure S2a, each point represents the non-zero elements of the A_M1 matrix, which was generated in Stage 1 and reflects the $\Delta\mathrm{Loss}$ between two variables. The x-axis represents different datasets, while the y-axis measures the $\Delta\mathrm{Loss}$ values. Circular points indicate an indirect causal relationship between two variables, whereas triangular points represent other types of relationships. Black horizontal lines in the figure represent the mean value of each type of points across different datasets. Notably, the DREAM4 dataset, having had all the edges of indirect causality removed in the first stage, is absent from Supplementary Figure S2a. The significantly lower $\Delta\mathrm{Loss}$ values associated with indirect causal relationships, compared to other types, confirm the utility of $\Delta\mathrm{Loss}$ in distinguishing between different types of causal connections.

In stage 3, we calculate the causal intensity between neighboring nodes (Supplementary Figure S2b). The x-axis represents different datasets and the y-axis quantifies the causal intensity between each pair of neighboring nodes in the A_M2 matrix. Circular points indicate a true direct causal relationship between two variables, while triangular points indicate that both variables are influenced by a common hidden variable. In the DREAM4 dataset, a few edges are incorrectly retained from Stage 2 are denoted by square dots. Black horizontal lines in the figure represent the

mean value of each type of points. The notably lower causal intensities among variables influenced by hidden variables underscore the effectiveness of the causal intensity indicator in distinguishing between direct causality and pseudo-causality due to hidden variable influence.



Supplementary Figure S2. Increased losses in linear regression and causal intensity. (a) Increased losses in linear regression. The vertical axis represents the $\Delta\mathrm{Loss}$ generated in Stage 2, and points of different colors represent different datasets. Circular dots represent the $\Delta\mathrm{Loss}$ between two variables in indirect causation, while triangular dots represent the $\Delta\mathrm{Loss}$ between two variables in direct causation or spurious causation caused by a hidden variable. **(b) Causal intensity.** The vertical axis represents the causal intensity generated in Stage 3. Circular dots represent the causal intensity between two variables with direct causation. Triangular points represent the causal intensity between two variables in spurious causation caused by a hidden variable. Square points reflect the causal intensity for edges incorrectly retained in Stage 3.

Supplementary material S4. Implementation details of calculating causal intensity

In this paper, we define causal intensity as:

$$\mathrm{CI}(\mathrm{X_i},\mathrm{X_j}) = \frac{\mathrm{I}(\mathrm{X_i},\mathrm{X_j})}{\mathrm{H}(\mathrm{X_i},\mathrm{X_j})} = \frac{\mathrm{H}(\mathrm{X_i})+\mathrm{H}(\mathrm{X_j})-\mathrm{H}(\mathrm{X_i},\mathrm{X_j})}{\mathrm{H}(\mathrm{X_i},\mathrm{X_j})} \quad (S11)$$

Where, $\mathrm{I}(\mathrm{X_i},\mathrm{X_j})$ represents the mutual information and $\mathrm{H}(\mathrm{X_i},\mathrm{X_j})$ represents the joint entropy.

For Gaussian distributions, we calculate the entropy and joint entropy using the formula:

$$\mathrm{H} = \frac{1}{2}\log(2\pi e)^{\mathrm{d}}|\sum| \quad (S12)$$

where, d is the dimension of the random variable, and $\sum$ is the covariance matrix, with $|\sum|$ representing the determinant of this matrix. The dimension d is 1 when computing $\mathrm{H}(\mathrm{X_i})$ or $\mathrm{H}(\mathrm{X_j})$, and d is 2 when computing $\mathrm{H}(\mathrm{X_i},\mathrm{X_j})$.

We can interpret this formula from the definition of information entropy. Information entropy is a measure of the uncertainty of a random variable, the larger it is the greater the uncertainty of the random variable. For a Gaussian distribution, the magnitude of its entropy depends on the dimension of the random variable and the size of the covariance matrix.

Specifically, we can express the probability density function of the multivariate Gaussian

distribution as:

$$P(x) = \frac{1}{(2\pi)^{d/2}|\sum|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^{T}\sum^{-1}(x-\mu)\right) \qquad (S13)$$

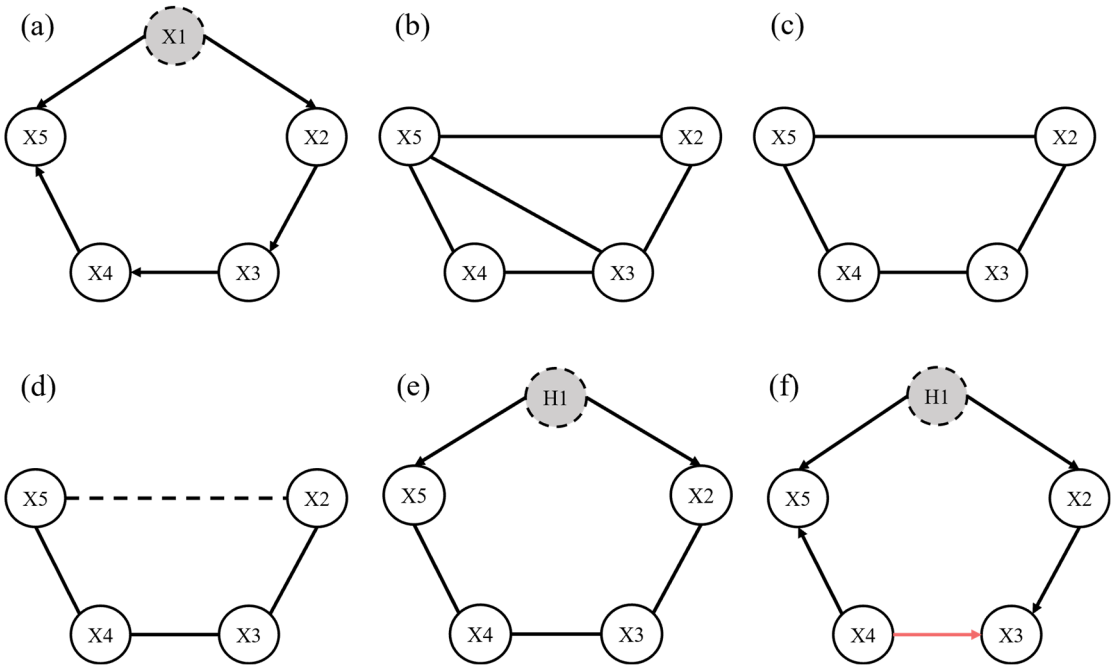
where μ is the mean vector and $\sum^{-1}$ is the inverse of the covariance matrix. We integrate this into the formula for information entropy:

$$H = -\int p(x)\log P(x)dx \qquad (S14)$$

For Gaussian distributions, the integral can be simplified by the permutation method, which ultimately leads to Eq. S12, where $\log(2\pi e)^{d}$ is a constant term that does not affect the uncertainty of the random variable, so it can be omitted. Thus, the entropy of the Gaussian distribution depends only on the size of the covariance matrix.

Supplementary material S5. Specific results for each stage of RLCI for different datasets

S5.1 Sim1



Supplementary Figure S3. Steps in reconstructing the causal network of Sim1. (a) Ground truth. Node X_1 is hidden and treated as a hidden variable, while the remaining nodes are observed variables. **(b)-(c) Stages 1-2.** Black lines indicate predicted edges in each stage. **(d) Stage3.1.** The dashed line indicates the edge with lower causal intensity compared to other edges. **(e) Stage3.2.** Reconstruction of the variable H_1 based on causal intensity. **(f) Stage 4. The final network reconstructed by RLCI.** Black lines indicate correctly predicted edges, while red lines highlight the incorrectly predicted edges.

In the first stage, we used independence test to derive the correlation network shown in Supplementary Figure S3b, which removes edges representing either small correlations or independence between variables from a fully connected network.

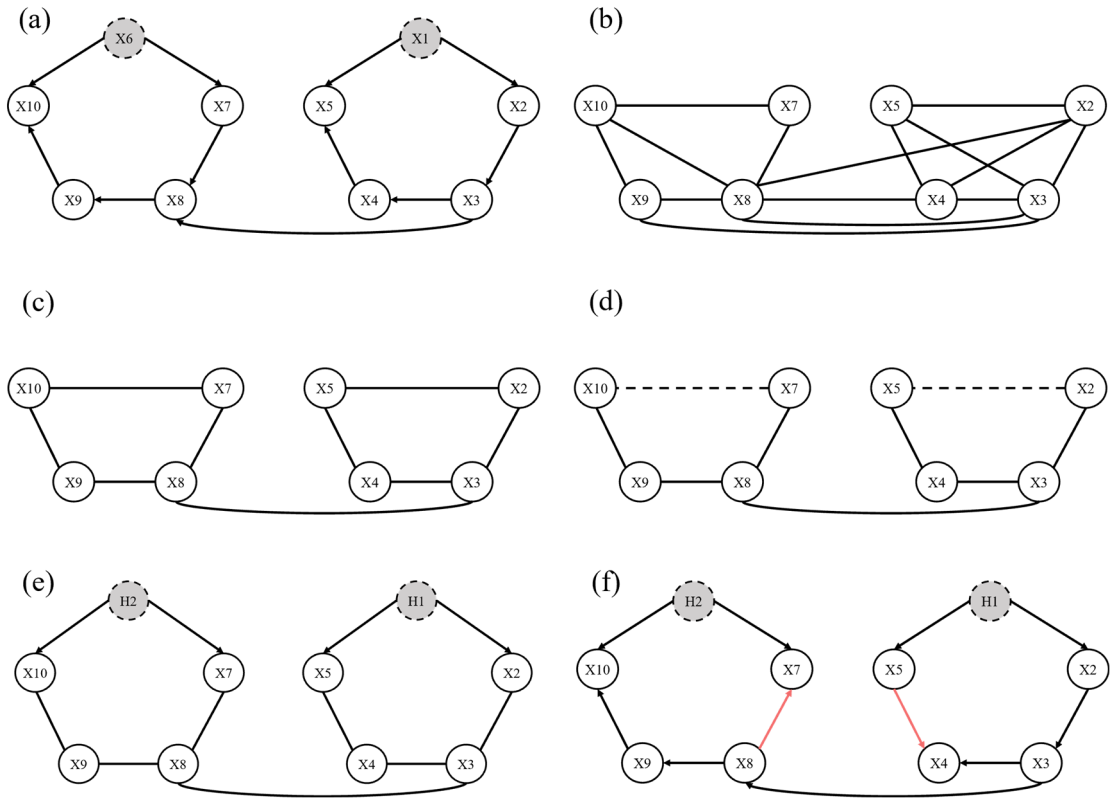
Proceeding to the second stage, we established the pseudo-causal network (Supplementary

Figure S3c). This network excludes all edges representing indirect causality, retaining only those representing direct causality and those caused by hidden variables.

In the third stage, we calculate the causal intensity for each edge retained from the second stage. It was observed that the causal intensity between X_2 and X_5 was significantly smaller compared to the other edges (Supplementary Figure S3d). Based on this finding, it was inferred that the edge 'X_2-X_5' is generated by a hidden variable. Consequently, we introduced a hidden variable, denoted as H_1, connecting both X_2 and X_5 (Supplementary Figure S3e).

Finally, in the fourth stage, the IGCI algorithm was utilized to identify the causal direction and reconstruct the complete causal network (Supplementary Figure S3f). The figure highlights correctly predicted edges in black and incorrectly predicted edges in red.

S5.2 Sim6



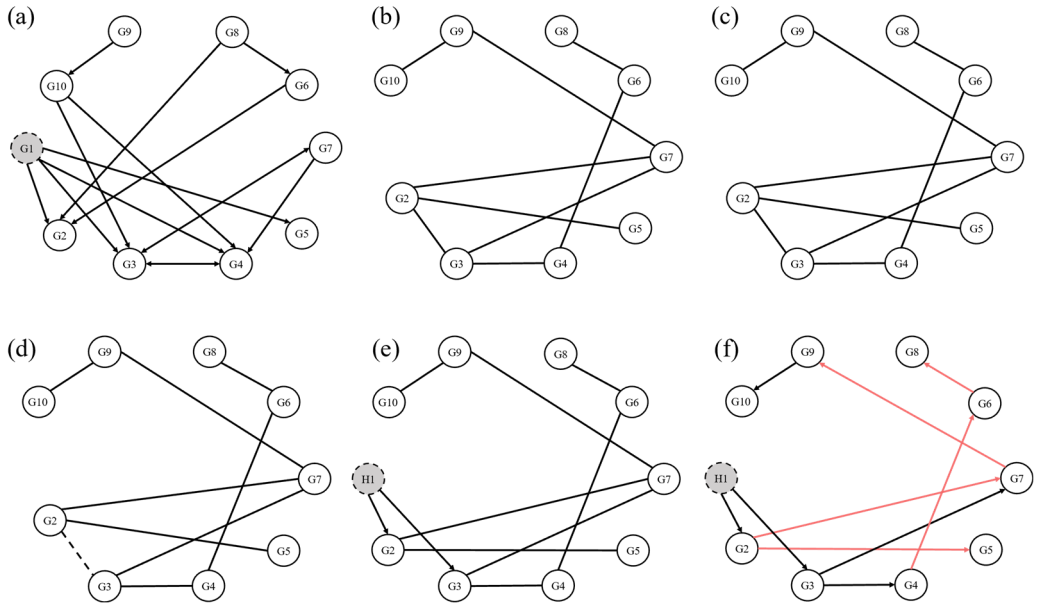
Supplementary Figure S4. Steps in reconstructing the causal network of Sim6. (a) Ground truth. Nodes X_1 and X_6 are hidden and treated as the hidden variables, while the remaining nodes are observed variables. **(b)-(c) Stages 1-2.** Black lines indicate predicted edges in each stage. **(d) Stage 3.1**. The dashed line represents the edge with lower causal intensity compared to other edges. **(e) Stage 3.2**. Reconstruction of the variables H_1 and H_2. **(f) Stage 4. The final network reconstructed by RLCI.** Black lines represent correctly predicted edges, while red lines highlight incorrectly predicted edges.

We begin by deriving the correlation network (Supplementary Figure S4b), which leads to the establishment of the basic skeleton of the pseudo-causal network (Supplementary Figure S4c). During the third stage, as illustrated in Supplementary Figure S4d-e, we identify that the edges '$X_2 - X_5$' and '$X_7 - X_{10}$' are generated by hidden variables based on their causal intensities. In the fourth stage, we employ the IGCI method to identify the causal directions among the observed variables,

and subsequently reconstruct the complete causal network (Supplementary Figure S4f).

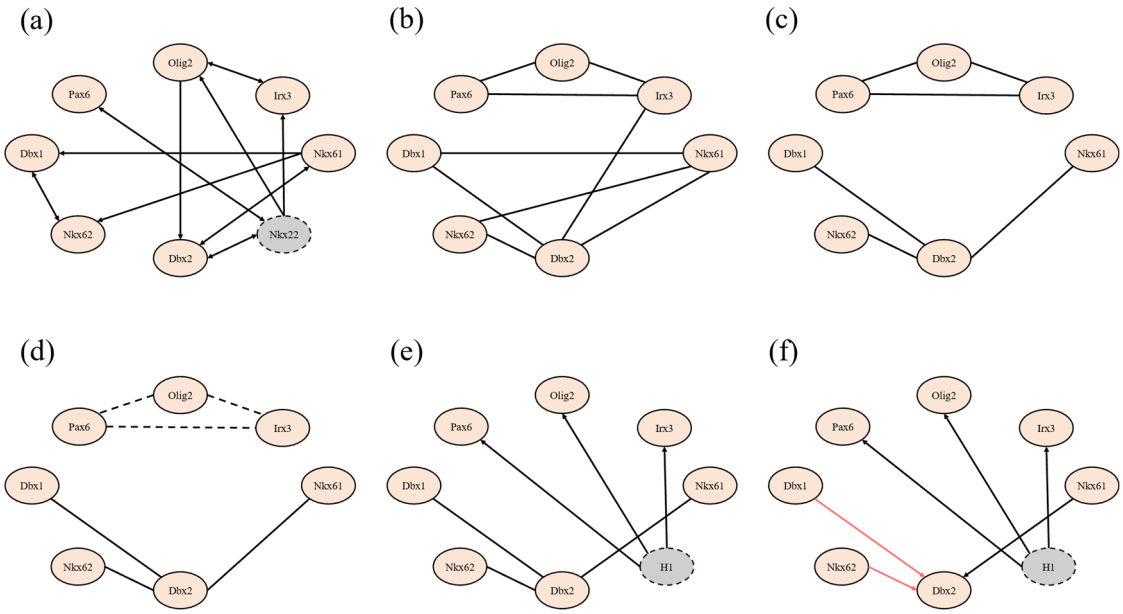
S5.3 DREAM4

The results of the RLCI method for each stage of the DREAM4 dataset are shown in Supplementary Figure S5b-f. In Stage 2, we observed significant $\Delta\mathrm{Loss}$ values, indicating the absence of edges with indirect causality. The outputs from Stage 1 remain unchanged. Furthermore, the '$\mathrm{G}_2 - \mathrm{G}_3$' edge is identified as a pseudo-causality edge (Supplementary Figure S5d). Consequently, it is inferred that a hidden variable simultaneously affects both G_2 and G_3 (Supplementary Figure S5e).



Supplementary Figure S5. Steps in reconstructing the causal network of DREAM4. (a) Ground truth. Node G_1 is the hidden variable, and the remaining nodes are observed variables. **(b)-(c) Stages 1-2.** Black lines represent predicted edges in each stage. **(d) Stage 3.1.** The dashed line denotes the edge with low causal intensity. **(e) Stage 3.2.** Reconstruction of the variable H_1. **(f) Stage 4. The final network reconstructed by RLCI.** Black lines represent correctly predicted edges, while red lines highlight incorrectly predicted edges.

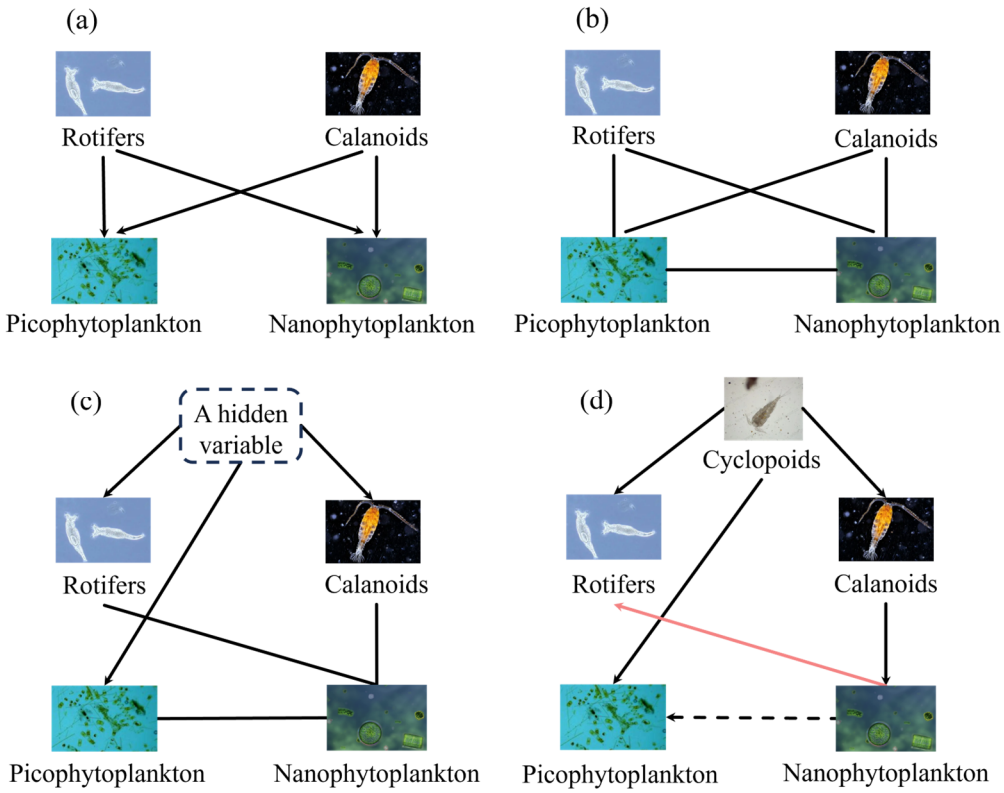
S5.4 BEELINE_VSC



Supplementary Figure S6. Steps in reconstructing the causal network of BEELINE_VSC. (a) Ground truth. Gene Nkx22 is the hidden variable, and the remaining genes are observed variables. **(b)-(c) Stages 1-2.** Black lines represent predicted edges in each stage. **(d) Stage 3.1.** The dashed line represents the edge with low causal intensity. **(e) Stage 3.2.** H_1 is the reconstructed variable. **(f) Stage 4. The final network reconstructed by RLCI.** Black lines represent correctly predicted edges, while red lines highlight incorrectly predicted edges.

Supplementary Figure S6 shows the results of our algorithm at each stage of the BEELINE_VSC dataset. In the third stage, the edges 'Pax6-Olig2', 'Pax6-Irx3', and 'Olig2-Irx3' are identified as pseudo-causality edges generated by the influence of hidden variables (Supplementary Figure S6d). Given our assumption that each observed variable is affected by no more than one hidden variable, we infer the existence of a hidden variable labeled H_1, which simultaneously affects the three genes: Pax6, Olig2, and Irx3 (Supplementary Figure S6e).

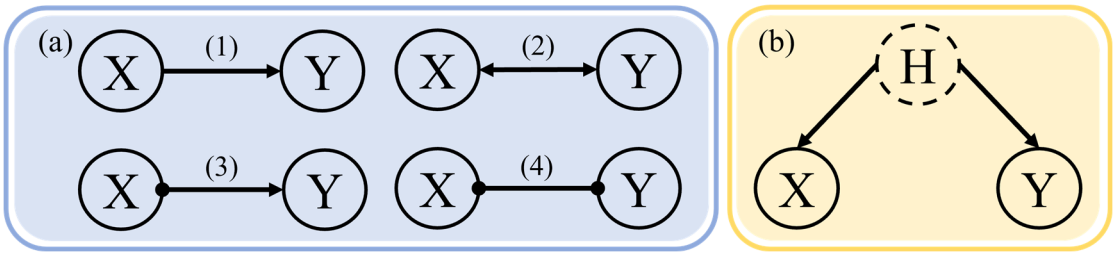
S5.5 Food chain



Supplementary Figure S7. Steps in reconstructing the causal network of the food chain dataset. (a) Ground truth. (b) Stage 2. The black lines represent predicted edges. **(c) Stage 3.** A hidden variable is identified that affects rotifers, calanoids, and picophytoplankton simultaneously. **(d) Stage 4. The final network reconstructed by RLCI.** Cyclopoids are considered as potential influencing factors, serving as the hidden variable. The black lines represent correctly predicted edges and the red lines represent incorrectly predicted edges. The dotted line represents that the accuracy of the predicted outcome is undetermined.

In the food chain dataset, which comprises only four species, accurately determining connections based on correlation alone is challenging. Therefore, the correlation network output from stage 1 is still a fully connected network. In stage 2, the pseudo-causal network is generated with the 'Rotifers-Calanoids' edge removed (Supplementary Figure S7b). In Stage 3, the causal intensity indicator suggests that rotifers, calanoids, and picophytoplankton are likely affected by a hidden variable (Supplementary Figure S7c). Ultimately, we hypothesize that cyclopoids could be this hidden variable and validate this hypothesis by verifying the predation relationships among cyclopoids, rotifers, calanoids, and picophytoplankton (Supplementary Figure S7).

Supplementary material S6. Specific comparison results



Supplementary Figure S8. Different types of edges generated by FCI, GFCI, and RFCI algorithms in reconstructing causal networks. (a) Four kinds of edges: (1) X causes Y. (2) X and Y are not causally related but there is a hidden variable that affects both of them. (3) Y is not an ancestor of X. (4) No set d-separates X and Y. **(b)** Another clear representation of Supplementary Figure S8a (4), where a hidden variable H affects both X and Y.

It is important to note that the FCI, GFCI, and RFCI algorithms identify four distinct types of relationships between any two nodes X and Y: X directly causes Y, both X and Y are influenced by hidden variables, Y is not an ancestor of X, and the relationship between X and Y exists but the causal direction is undetermined (Supplementary Figure S8a). In evaluating these algorithms, we do not consider undirected edges, which are those whose causal direction cannot be determined. The case where X and Y are influenced by a shared hidden variable, suggesting that they are not causally related but are impacted by the same hidden factor, is depicted in Supplementary Figure S8b.

Supplementary Table S1. Evaluation results on Sim1.

Algorithm	Precision	Recall	F1 score
ours	**0.8000**	**0.8000**	**0.8000**
PC	0.3333	0.2000	0.2500
LiNGAM	0.6000	0.6000	0.6000
DirectLiNGAM	0.6000	0.6000	0.6000
FCI	0.5000	0.2000	0.2857
GFCI	0.3333	0.2000	0.2500
RFCI	0.3333	0.2000	0.2500
RCD	0	0	0
CAM-UV	0	0	0

Supplementary Table S2. Evaluation results on Sim1 of 1000 samples.

Algorithm	Precision	Recall	F1 score
ours	**0.8000**	**0.8000**	**0.8000**
PC	0.3333	0.2000	0.2500
LiNGAM	0.4000	0.4000	0.4000
DirectLiNGAM	0.4000	0.4000	0.4000
FCI	0.5000	0.2000	0.2857
GFCI	0.5000	0.2000	0.2857
RFCI	0.5000	0.2000	0.2857
RCD	0	0	0
CAM-UV	0.7500	0.6000	0.6667

Supplementary Table S3. Evaluation results on Sim6.

Algorithm	Precision	Recall	F1 score
ours	**0.8182**	**0.8182**	**0.8182**
PC	0.2500	0.2727	0.2609
LiNGAM	0.3684	0.6364	0.4667
DirectLiNGAM	0.3684	0.6364	0.4667
FCI	0.3846	0.4545	0.4167
GFCI	0.4545	0.4545	0.4545
RFCI	0.2143	0.2727	0.2400
RCD	0	0	0
CAM-UV	0	0	0

Supplementary Table S4. Evaluation results on Sim6 of 1000 samples.

Algorithm	Precision	Recall	F1 score
ours	0.7273	**0.7273**	**0.7273**
PC	0	0	0
LiNGAM	**0.7778**	0.6364	0.7000
DirectLiNGAM	0.2500	0.3636	0.2963
FCI	0.5000	0.2727	0.3529
GFCI	0.8000	0.3636	0.5000
RFCI	0.6667	0.5455	0.6000
RCD	0	0	0
CAM-UV	0.4000	0.3636	0.3810

Supplementary Table S5. Evaluation results on DREAM4.

Algorithm	Precision	Recall	F1 score
ours	**0.5000**	**0.3333**	**0.4000**
PC	0.3333	0.1333	0.1905
LiNGAM	0.1481	0.2667	0.1905
DirectLiNGAM	0.1786	**0.3333**	0.2326
FCI	0.4545	0.2000	0.3846
GFCI	0	0	0
RFCI	0.2500	0.0667	0.1053
RCD	0.2000	0.0667	0.1000
CAM-UV	0.2000	0.1333	0.1600

Supplementary Table S6. Evaluation results on BEELINE_VSC.

Algorithm	Precision	Recall	F1 score
ours	**0.6667**	**0.2667**	**0.3810**
PC	0.3333	0.1333	0.1905
LiNGAM	0.2222	**0.2667**	0.2424
DirectLiNGAM	0.0667	0.0667	0.0667

FCI	0.2000	0.0667	0.1000
GFCI	0.3333	0.1333	0.1905
RFCI	0.3333	0.1333	0.1905
RCD	0	0	0
CAM-UV	0	0	0