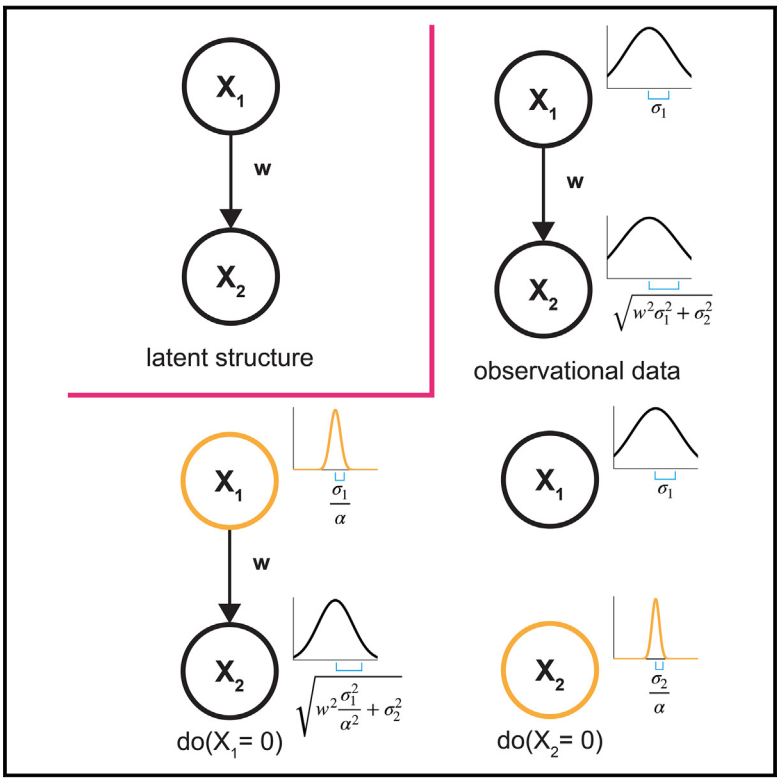
iScience

Article

dotears: Scalable and consistent directed acyclic graph estimation using observational and interventional data

Graphical abstract



Highlights

- Continuous causal structure learning which incorporates interventional data
- Interventions enable robust identifiability of structure
- State-of-the-art performance in simulations and real data
- Estimator is statistically consistent under mild assumptions

Authors

Albert Xue, Jingyou Rao, Sriram Sankararaman, Harold Pimentel

Correspondence

asxue@ucla.edu (A.X.), sriram@cs.ucla.edu (S.S.), hjp@ucla.edu (H.P.)

In brief

Gene network; Bioinformatics; Biocomputational method

Xue et al., 2025, iScience 28, 111673
February 21, 2025 Published by Elsevier Inc.
https://doi.org/10.1016/j.isci.2024.111673

CellPress

iScience

CellPress
OPEN ACCESS

Article

dotears: Scalable and consistent directed acyclic graph estimation using observational and interventional data

Albert Xue,[1,6,*] Jingyou Rao,[2] Sriram Sankararaman,[2,3,4,5,*] and Harold Pimentel[2,3,4,5,*]
[1]Bioinformatics Indepartmental Program, UCLA, Los Angeles, CA 90024, USA
[2]Department of Computer Science, UCLA, Los Angeles, CA 90024, USA
[3]Department of Computational Medicine, David Geffen School of Medicine, UCLA, Los Angeles, CA 90024, USA
[4]Department of Human Genetics, David Geffen School of Medicine, UCLA, Los Angeles, CA 90024, USA
[5]These authors contributed equally
[6]Lead contact
*Correspondence: asxue@ucla.edu (A.X.), sriram@cs.ucla.edu (S.S.), hjp@ucla.edu (H.P.)
https://doi.org/10.1016/j.isci.2024.111673

SUMMARY

New assays such as Perturb-seq link parallel CRISPR interventions to transcriptomic readouts, providing insight into gene regulatory networks. Causal regulatory networks can be represented by directed acyclic graphs (DAGs), but lack of identifiability and a combinatorial solution space complicate learning DAGs from observational data. Score-based methods have improved the practical scalability of inferring DAGs, but are sensitive to error variance structure. Furthermore, correction for error variance is difficult without prior knowledge of structure. We present dotears [doo-tairs], a continuous optimization framework leveraging observational and interventional data to infer causal structure, assuming a linear Structural Equation Model. dotears exploits structural consequences of hard interventions to estimate and correct for error variance structure. dotears is a provably consistent estimator of the true DAG under mild assumptions and outperforms other state-of-the-art methods in varied simulations. In real data, differential expression tests and high-confidence protein-protein interactions validate dotears-inferred edges with higher precision and recall than others.

INTRODUCTION

Understanding gene regulatory networks can identify mechanisms and pathways linking GWAS significant variants to phenotype. Recent efforts to map regulatory networks through *trans*-eQTLs are partly limited by power; for example, the GTEx project finds only 143 *trans*-eQTLs in 838 individuals.[1] On the other hand, Võsa et al. detect almost 60,000 *trans*-eQTLs in ∼31,000 individuals, across more than a third of trait-associated variants.[2] These results imply that *trans*-regulatory relationships are pervasive, but our ability to detect small eQTL effects is often limited by small sample size regimes, observational data, and a high multiple testing burden. Importantly, since rare tissues are unlikely to be sampled at sufficient sample sizes, capturing gene regulation events across a wide array of cell types and tissues requires another experimental method.

High-throughput genomic technologies such as Perturb-seq provide a natural alternative for learning gene regulatory networks. Perturb-seq links high-dimensional transcriptomic readouts to known, highly parallel CRISPR interventions, allows the direct interrogation of causal regulatory relationships, and has scaled genome-wide.[3–5] In particular, the effects of CRISPR gene interventions are large in comparison to QTL effects, facilitating inference of downstream regulatory relationships. Notably, analogous experiments have already mapped gene-gene networks in yeast.[6–8]

The inference of gene regulatory networks can be treated as a causal structure learning problem, which considers learning relationships between variables in the form of a Directed Acyclic Graph (DAG). Here, directedness gives a natural causal interpretation, while acyclicity ensures that the causal interpretation is valid. For example, in the mediator DAG $i \to j \to k$, we understand that gene j has a direct causal regulatory effect on gene k, and similarly that gene i (indirectly) affects gene k only through its direct effect on gene j.

Identifiability and scalability are the primary difficulties in learning DAGs from data. For identifiability, distinct DAGs may contain the same conditional independence relationships in *observational* data, and DAGs are only identifiable up to Markov equivalence.[9–11] For scalability, Zheng et al. introduced DAGs with NO TEARS, a method that allows for continuous optimization through a continuous, differentiable acyclicity constraint.[12] As a result, DAGs with NO TEARS avoids combinatorial characterizations of DAGs, and is a fundamental methodological

building block for many structure learning methods.[13,14] However, NO TEARS and related methods infer DAGs whose topological order follows increasing marginal variance, and re-scaling data can change or reverse their inferences.[15] This poses issues for inferring gene regulatory networks from data, where the scale of exogenous error between genes is likely not uniform.

Fundamentally, this is still an issue of identifiability. Because NO TEARS uses observational data, it must choose a single member of a class of Markov equivalent DAGs. The "tiebreaker" is then a function of the variance. However, we show that interventional data can correct for variance sensitivity in NO TEARS and related methods, and further that this correction is sufficient for the consistent estimation of structure.

Explicitly, exogenous error in the linear SEM drives variance sensitivity. Let X be a p-dimensional random vector (e.g., the distribution of gene expression across p genes), $W_0 \in \mathbb{R}^{p\times p}$ the weighted adjacency matrix of the true DAG, and ϵ a p-dimensional random vector specifying the exogenous error. The linear SEM gives the autoregressive formulation

$$X = XW_0 + \epsilon,$$

where $\Omega_0 := \mathrm{Cov}(\epsilon) = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$. For any node j we then have

$$\mathrm{Var}(X_j) = \sum_{i=1}^{p} w_{ij}^2 \mathrm{Var}(X_i) + \sigma_j^2. \qquad \text{(Equation 1)}$$

$\mathrm{Var}(X_i)$ is itself a linear function of $\sigma_1^2,\ldots,\sigma_p^2$. Critically, parental error variances thus propagate to downstream nodes to provide a signal of structure. However, in observational data the exact relationship between W_0, Ω_0, and $\mathrm{Var}(X)$ is unidentifiable.

Consequently, given Ω_0 then W_0 is recoverable even in observational data.[16] However, the estimation of Ω_0 is difficult without *a priori* knowledge of W_0. Previous methods either ignore exogenous variance structure or use the conditional estimate $\widehat{\Omega}_0|W$[12,13]. We show that both procedures are sensitive to exogenous variance structure even in the simplest two node DAG.

Hard interventions[17,18] remove upstream variance in the linear SEM to allow the marginal estimation of Ω_0. We show that the naive incorporation of interventional data into the NO TEARS framework, without the estimation of Ω_0, is insufficient for structural recovery. Finally, we show correction by this marginal estimate is sufficient for structural recovery.

Accordingly, we present dotears, a novel optimization framework for structure learning. dotears uses **1.)** a novel marginal estimation procedure for Ω_0 using the structural consequences of interventions, and **2.)** joint estimation of the causal DAG from observational and interventional data, given the estimated $\widehat{\Omega}_0$. dotears provides a simple model that we show, by extending results from Loh and Buhlmann,[16] is a provably consistent estimator of the true DAG under mild assumptions. In simulations, dotears corrects for exogenous variance structure and is robust to reasonable violations of its modeling assumptions. We also apply dotears to the Perturb-seq experiment in Replogle et al.[5] dotears infers a sparse set of edges that validate with high precision in differential expression tests and in an orthogonal set of protein-protein interactions. In both simulations and real data, dotears outperforms all other tested methods in all used metrics.

Model

We represent a gene regulatory network as $G = ([p], \mathcal{E})$, a DAG on p nodes with node set $[p] := \{1\ldots p\}$ and edge set $\mathcal{E}$. We represent $\mathcal{E}$ with the weighted adjacency matrix $W \in \mathcal{D} \subset \mathbb{R}^{p\times p}$, where $\mathcal{D} \subset \mathbb{R}^{p\times p}$ is the set of weighted adjacency matrices on p nodes whose support is a DAG. We denote the parent set of node i in the observational setting as $\mathrm{Pa}(i)$. For w_{ij} the i,j entry in W, $|w_{ij}| > 0$ indicates an edge $i \to j$ with weight w_{ij}, equivalently denoted $i \overset{w_{ij}}{\to} j$, or equivalently an inferred gene regulatory event between genes i and j. $k = 0, 1, \ldots, p$ indexes the intervention, where $k = 0$ is reserved for the observational system, and $k \neq 0$ denotes intervention on node k. Similarly, we denote $(\cdot)^{(0)}$ for observational quantities, and $(\cdot)^{(k)}$ for quantities under intervention on node k. For brevity, a variable without a superscript is assumed to be observational; for example, $X \equiv X^{(0)}$. $\boldsymbol{X}$ (bolded) denotes n_0 samples drawn from the p-dimensional random vector X (unbolded), and ϵ (bolded) denotes n_0 samples drawn from the p-dimensional random vector ϵ (unbolded). Similarly, if $X^{(k)}$ is a p-dimensional random vector, then $\boldsymbol{X}^{(k)} \in \mathbb{R}^{n_k \times p}$ represents n_k observations of X. We denote the total sample size $n := \sum_{k=0}^{p} n_k$, and the true weighted adjacency matrix W_0.

The linear SEM is an autoregressive representation of $X^{(k)}$ and weighted adjacency matrix $W_0^{(k)}$,

$$X^{(k)} = X^{(k)} W_0^{(k)} + \epsilon^{(k)}, k = 0, \ldots, p. \qquad \text{(Equation 2)}$$

Here, $W_0^{(k)}$ is permutation-similar to a strictly upper triangular matrix, representing the constraint $W_0^{(k)} \in \mathcal{D}$. For each k, $\epsilon^{(k)}$ is a p-dimensional random vector such that $\mathbb{E}\epsilon^{(k)} = 0_p$, and $\Omega_0^{(k)} := \mathrm{Cov}(\epsilon^{(k)})$. Denote ϵ_i as the ith element of ϵ. Then $\epsilon_i^{(k)}$ is the exogenous error on node i, such that $\mathbb{E}\epsilon_i^{(k)} = 0$ and $\epsilon_i^{(k)} \perp\!\!\!\perp \epsilon_j^{(k)}$ for $i \neq j$. We further define

$$\Omega_0 := \mathrm{Cov}\left(\epsilon^{(0)}\right) = \mathrm{diag}\left(\sigma_1^2, \sigma_2^2, \cdots, \sigma_p^2\right).$$

Motivated by recent work on a genome-wide screen that performs known single interventions on all protein-coding genes,[5] we consider the linear SEM with known single interventions on all p nodes. Accordingly, we obtain a system of $p+1$ structural equations

$$\begin{aligned} X^{(0)} &= X^{(0)} W^{(0)} + \epsilon^{(0)} \\ &\vdots \\ X^{(p)} &= X^{(p)} W^{(p)} + \epsilon^{(p)}. \end{aligned}$$

In this setting we have complicated our problem. Before, with the single data matrix $\boldsymbol{X}$, we inferred a single W; now, with the $p+1$ data matrices $\boldsymbol{X}^{(0)}, \boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(p)}$, it seems that we must infer $p+1$ adjacency matrices $W^{(0)}, W^{(1)}, \ldots, W^{(p)}$. Our model assumes hard interventions, i.e., that an intervention on node k removes causal influences from observational parents of k.[17,18] Hard interventions relax the do operation $do(X_k^{(k)} = 0)$ to allow for

residual noise, modeling the limit of experimental interventional efficacy combined with a noisy readout.

Assumption 1. For an intervention $k \neq 0$, if in the observational setting

$$X_k^{(0)} = \sum_{i \in Pa(k)} W_{ik} X_i^{(0)} + \epsilon_k^{(0)},$$

then upon intervention on k

$$X_k^{(k)} = \epsilon_k^{(k)}.$$

Under Assumption 1, we can relate $W^{(0)}$ to $W^{(k)}$ by setting the kth column of $W^{(0)}$ to $\vec{0}_p$, giving

$$W_{ij}^{(k)} := \begin{cases} W_{ij} & j \neq k \\ 0 & j = k \end{cases}. \quad \text{(Equation 3)}$$

This gives $p+1$ data matrices $\boldsymbol{X}^{(k)}$ to jointly infer a single weighted adjacency matrix W. We now characterize the $p+1$ exogenous variance structures $\Omega_0^{(k)}$. We assume that the exogenous variance of non-targeted nodes $j \neq k$ is invariant, modeling a system where the effects of CRISPR interventions are restricted to the target.

Assumption 2. For an intervention $k \neq 0$, $Var(\epsilon_j^{(k)}) = Var(\epsilon_j^{(0)}) = \sigma_j^2$ for $j \neq k$.

We allow interventions affect the error variance of the target, but require the effect to be uniform across targets, modeling an experimental intervention with uniform effects on noise.

Assumption 3. Let unknown $\alpha \in \mathbb{R}$. Then $\forall\ k$, if $Var(\epsilon_k^{(0)}) = \sigma_k^2$ then $Var(\epsilon_k^{(k)}) = \frac{\sigma_k^2}{\alpha^2}$.

Here, α is shared across interventions. As $\alpha \to \infty$, this interventional model is equivalent to $do(X_k = 0)$.[18] Under these assumptions, the variance of the target is

$$\mathrm{Var}\left(X_k^{(k)}\right) = \frac{\sigma_k^2}{\alpha^2}.$$

Let $\widehat{\mathrm{Var}}$ denote the unbiased sample variance. We then obtain the estimator

$$\widehat{\Omega}_0 = \mathrm{diag}\left(\widehat{\mathrm{Var}}\left(X_1^{(1)}\right), \ldots, \widehat{\mathrm{Var}}\left(X_p^{(p)}\right)\right). \quad \text{(Equation 4)}$$

RESULTS

dotears

DAGs with NO TEARS[12] transforms the combinatorial constraint $W \in \mathcal{D}$ into the continuous constraint $h(W) = 0$, where $\circ$ denotes the Hadamard product and

$$h(W) = \mathrm{tr}[\exp(W \circ W)] - p.$$

Define $\|\cdot\|_1$ as the vector ℓ_1 norm on $\mathrm{vec}(W)$, i.e., $\|\mathrm{vec}(W)\|_1$, and $\|\cdot\|_F$ the Frobenius norm. For some loss function $\mathcal{F}$, the differentiability of h allows for the optimization framework

$$\min_W \mathcal{F}(W, \boldsymbol{X}) + \lambda \|W\|_1 \quad \text{(Equation 5)}$$

$$\text{s.t. } h(W) = 0.$$

We present dotears, a consistent, intervention-aware joint estimation procedure for structure learning. Loh and Buhlmann (2014) showed that the Mahalanobis norm is a consistent estimator of W_0, and is uniquely minimized in expectation by W_0 given Ω_0, but give no estimation procedure for Ω_0[16,19]. Note the Mahalanobis norm's characterization as inverse-variance-weighted by Ω_0,

$$\|(\boldsymbol{X} - \boldsymbol{X}W)\Omega_0^{-\frac{1}{2}}\|_F^2 = \sum_{i=1}^{p} \frac{1}{\sigma_i^2} \|(\boldsymbol{X} - \boldsymbol{X}W)_i\|_F^2. \quad \text{(Equation 6)}$$

dotears solves the following optimization problem:

$$\min_W \frac{1}{p} \sum_{k=0}^{p} \frac{1}{2n_k} \left\| \left(\boldsymbol{X}^{(k)} - \boldsymbol{X}^{(k)} W^{(k)}\right) \widehat{\Omega}_0^{-\frac{1}{2}} \right\|_F^2 + \lambda \|W\|_1 \quad \text{(Equation 7)}$$

$$\text{s.t. } h(W) = 0,$$

where

$$W_{ij}^{(k)} = \begin{cases} W_{ij} & j \neq k \\ 0 & j = k \end{cases}.$$

dotears retains the continuous DAG constraint and ℓ_1 regularization of W from NO TEARS,[12] but incorporates exogenous variance structure through $\widehat{\Omega}_0$ as well as interventional data $(k = 1, \ldots, p)$.

dotears successfully corrects for exogenous variance structure

We show that dotears is robust to exogenous variance structure, and motivate the necessity of the marginal estimation of $\widehat{\Omega}_0$ using the simplest non-trivial DAG $X_1 \xrightarrow{w} X_2$, where gene X_1 regulates gene X_2 with effect size w. Assume $X_1 \xrightarrow{w} X_2$ has true weighted adjacency matrix $W_0 := \begin{pmatrix} 0 & w \\ 0 & 0 \end{pmatrix}$ and SEM

$$\begin{aligned} X_1 &= \epsilon_1, \mathrm{Var}(\epsilon_1) = \sigma_1^2, \\ X_2 &= wX_1 + \epsilon_2, \mathrm{Var}(\epsilon_2) = \sigma_2^2. \end{aligned} \quad \text{(Equation 8)}$$

Let $\gamma := \frac{\sigma_1^2}{\sigma_2^2}$, such that $\Omega_0 = \begin{pmatrix} \gamma & 0 \\ 0 & 1 \end{pmatrix}$ and $W_0 = \begin{pmatrix} 0 & w \\ 0 & 0 \end{pmatrix}$.

The least squares loss used by NO TEARS is minimized in expectation if and only if

$$|w| \geq \sqrt{1 - \frac{1}{\gamma}}, \quad \text{(Equation 9)}$$

which is true if and only if the topological ordering of the DAG follows increasing marginal variance, or equivalently a *varsortable* DAG. For the full proof, see Supplementary Material S1.1, or Reisach et al. and Kaiser et al.[15,19] In Figure 1, we examine the

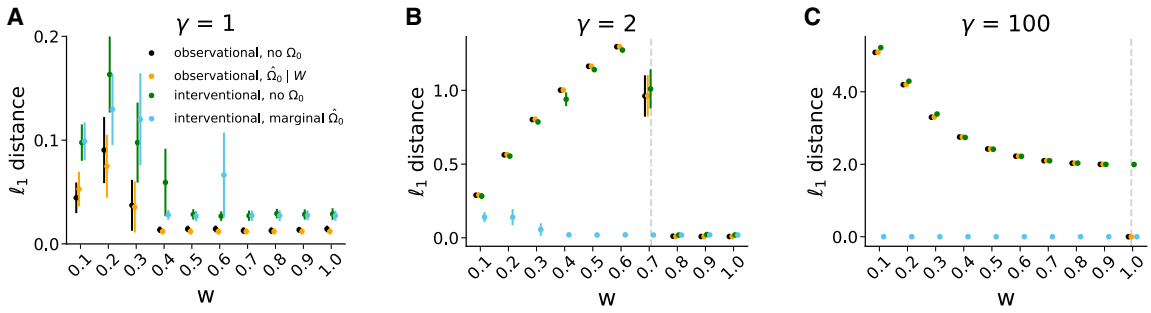


Figure 1. dotears successfully corrects for exogenous variance structure in two-node DAG simulations

Comparison of ℓ_1 distance (lower is better) between true structure and estimates from NO TEARS (black), GOLEM-NV(orange), NO TEARS interventional (green), and dotears (blue). Each method corrects differently for Ω_0. For each $w = 0.1, 0.2, \ldots, 1.0$ and $\gamma = 1,2,100$, we generate Gaussian data from the structure $X_1 \overset{w}{\to} X_2$ such that $\sigma_1^2 = \gamma\sigma_2^2$. For each pair w, γ, we draw 25 simulations at a sample size of $n = 3000$. The dashed gray line represents the varsortability bound $|w| \geq \sqrt{1 - \frac{1}{\gamma}}$. a) Under the equal variance assumption $\gamma = 1$, correction for Ω_0 is unneeded. All methods are sufficient for structure recovery. b) At $\gamma = 2$, NO TEARS (black), GOLEM-NV (orange), and NO TEARS interventional (green) infer correctly on varsortable w. c) As γ grows large, the varsortability bound approaches $|w| \geq 1$. Only dotears estimates correctly for all w. Points represent mean estimates, and bars represent standard errors; some standard errors are too small to see.

performance of four different strategies of correcting for Ω_0 in simulations. NO TEARS (black) uses the least squares loss, which ignores Ω_0, while GOLEM-NV (orange) uses a likelihood loss that estimates $\widehat{\Omega}_0|W$[12,13]. We also include the scenario when $\widehat{\Omega}_0$ is set to I_p in Equation 7. Call this NO TEARS interventional (green), the simplest extension of NO TEARS to interventional data. NO TEARS interventional is aware of hard interventions through the structure $W^{(k)}$, but ignores Ω_0. NO TEARS interventional is thus an ablation study on $\widehat{\Omega}_0$. For each set $(w, \gamma) \in \{0.1, \ldots, 1.5\} \times \{1, 2, 4, 10, 100\}$, we draw 25 simulations of observational and interventional data, with sample size $n = (p+1) * 1000 = 3000$. For observational data, this is 3000 observations from the observational system; for interventional data, this is 1000 observations from each system $k = 0, 1, 2$. For observational data, we draw Gaussian data under the SEM in Equation 8. For interventional methods, we draw Gaussian data under the system of SEMs in Supplementary Material S1.4. To isolate the behavior of the loss, we remove ℓ_1 regularization. Full simulation details and results are given in Supplementary Material S1.4. NO TEARS does not correct for Ω_0 and uses only observational data. As a result, in Figure 1 it estimates correctly only on varsortable pairs of w, γ. Note that the gray dashed line represents the theoretical varsortability cutoff $|w| \leq \sqrt{1 - \frac{1}{\gamma}}$ given in Equation 9. GOLEM-NV uses the maximum likelihood estimate $\widehat{\Omega}|W$ under a Gaussian model. Subsequently, joint estimation is performed over $W, \widehat{\Omega}|W$ using the Gaussian negative log likelihood and profile likelihood for Ω_0, which simplifies to

$$\frac{1}{2}\sum_{i=1}^{p} \log\left(\left\|(X - XW)_i\right\|_F^2\right). \qquad \text{(Equation 10)}$$

This profile likelihood is insufficient to correct for exogenous variance structure, and only infers varsortable structures in Figure 1. This evidence qualitatively holds in simulations on three node topologies (Supplementary Material S1.5), where the behavior of GOLEM-NV remains deterministic in w, γ. Joint estimation of $W, \widehat{\Omega}_0|W$ is thus still sensitive to exogenous variance structure. Through NO TEARS interventional, we also see that interventional data alone, without correction for Ω_0, is insufficient to infer the two node structure. The NO TEARS interventional estimate is also deterministic in w, γ in Figure 1, and behaves almost identically to NO TEARS and GOLEM. Note that since NO TEARS interventional is given interventional data for all nodes, it operates in a fully identifiable setting (where NO TEARS and GOLEM-NV do not). We also show that CoLiDE-NV, which uses observational data to try to correct for Ω_0, is insufficient to infer structure (Figure S1).[20] Thus, neither interventional data nor correction by $\widehat{\Omega}_0|W$ are alone sufficient to infer structure. dotears combines the two to give the marginal estimate of Ω_0 in Equation 4 and a robust estimate of W. We do not imply that other methods using interventional data cannot infer the two node case under interventional data; in fact, many do successfully (see Supplementary Material S1.4). Rather, we use **1.)** an observational procedure that ignores Ω_0, **2.)** an observational procedure that corrects for an estimated $\widehat{\Omega}_0$, **3.)** an interventional procedure that ignores Ω_0, and **4.)** dotears, an interventional procedure that corrects for an estimated $\widehat{\Omega}_0$, to motivate dotears as most parsimonious model robust to exogenous variance structure under this line of thought.

Optimization and consistency

We now wish to show that dotears is a consistent estimator of the true DAG. However, two natural problems arise from our usage of $\widehat{\Omega}_0$. First, we have provided no estimation procedure for α, but $\mathbb{E}\widehat{\Omega}_0 \neq \Omega_0$ for $\alpha \neq 1$. However, $\mathbb{E}\widehat{\Omega}_0 \propto \Omega_0$ for all α, and constant scalings of Ω are rescalings of the loss.[16] $\widehat{\Omega}_0$ is therefore well-specified for inference on observational data $k = 0$. However, if $\alpha \neq 1$ then $\widehat{\Omega}_0$ is still misspecified for interventional data $k \neq 0$. Under Assumption 3

$$\mathrm{Cov}\left(\epsilon^{(k)}\right) = \mathrm{diag}\left(\sigma_1^2, \sigma_2^2, \ldots, \frac{\sigma_k^2}{\alpha^2}, \ldots, \sigma_p^2\right),$$

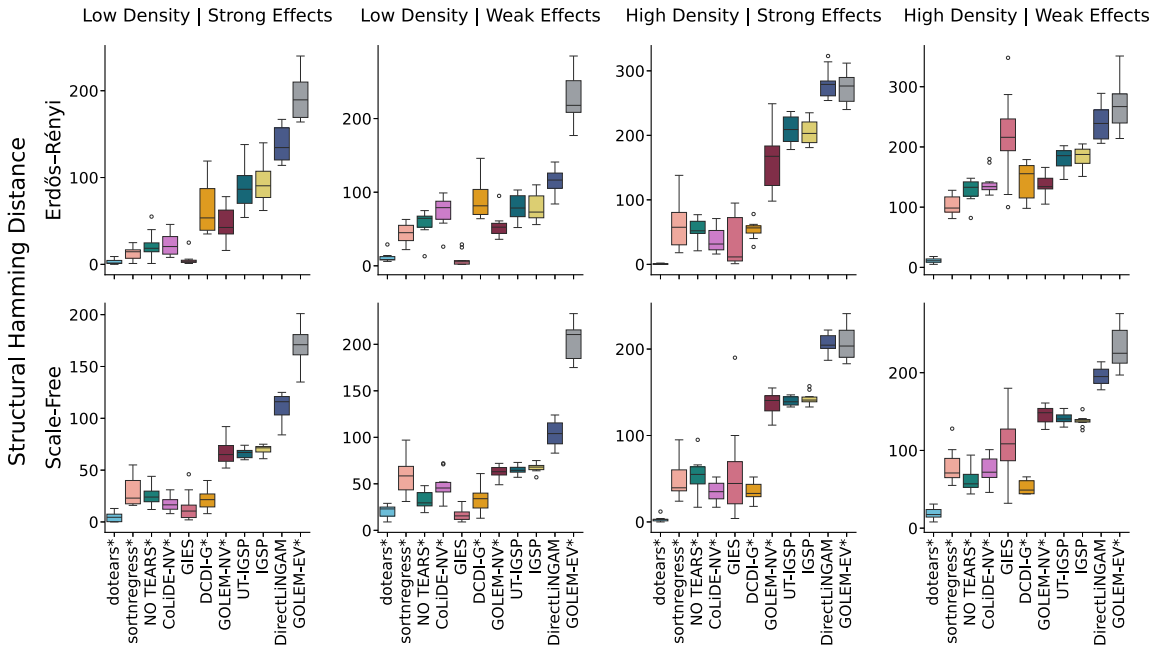


Figure 2. dotears outperforms other methods in large-scale random DAG simulations

Method performance on large random graphs ($p = 40$) using Structural Hamming Distance (lower is better) Rows index Erdős-Rényi or Scale Free topologies. Columns index parameterizations of edge density and weight, ordered in increasing difficulty. For details, see STAR Methods. 10 simulations were drawn for each parameterization with sample size $(p+1) * 100 = 4100$. * indicates cross-validated methods. Methods are sorted by average performance.

and thus $\mathbb{E}\widehat{\Omega}_0 \propto \mathrm{Cov}(\epsilon^{(k)})$. A naive approach might estimate $\widehat{\Omega}_0$ from interventional data only and then estimate $\widehat{W}$ from observational data only. However, this approach ignores a majority of our data and performs substantially worse in simulations (Supplementary Material S1.6). We show that $\widehat{\Omega}_0$ is well-specified even for $k \neq 0$, and the estimation of α is unnecessary (see Corollary 2). Under a sub-Gaussian assumption, Loh and Buhlmann show consistency of the Mahalanobis norm on observational data given Ω_0.[16] We extend these results, using $\widehat{\Omega}_0 \xrightarrow{p} \Omega_0$, to show

$$\arg\min_W \left\| \left(\boldsymbol{X}^{(k)} - \boldsymbol{X}^{(k)} W^{(k)} \right) \widehat{\Omega}_0^{-\frac{1}{2}} \right\|_F^2$$

is a consistent estimator of $W_0^{(k)}$ for each k, where $\xrightarrow{p}$ denotes convergence in probability. A full proof is given in Supplementary Material S1.2.

Simulations

We evaluate structure learning methods across a range of DAG topologies, effects distributions, and generative models. We benchmark methods that leverage interventional data (dotears, GIES, IGSP, UT-IGSP, DCDI) and methods using only observational data (NO TEARS, sortnregress, GOLEM-EV, GOLEM-NV, DirectLiNGAM, CoLiDE-NV).[10,12–15,20–24] dotears outperforms all tested methods in DAG estimation and is robust to reasonable violations of the model. Some methods come with important caveats for evaluation. sortnregress is not intended as a "true" structure learning method, but benchmarks the data's varsortability.[15] UT-IGSP can infer structure with unknown interventional targets, but we constrain to known targets for fairness.[23] Most simulations use Gaussian data, but dotears, NO TEARS, sortnregress, GIES, IGSP, and UT-IGSP do not assume Gaussianity. The non-Gaussianity assumption is violated for DirectLiNGAM, and the equal variance assumption for GOLEM-EV.[13,24] We simulate synthetic data from large Erdős-Rényi and Scale-Free DAGs[25,26] ($p = 40$), with 10 replicates each. We simulate under four parameterizations: {Low Density, High Density} × {Weak Effects, Strong Effects}. Observational and interventional data have matched sample size $n = (p+1) \times 100 = 4100$. Methods using 5-fold cross-validation for hyperparameter tuning are denoted by *. For non-binary methods, edge weights are thresholded. Methods are robust to threshold choice (see STAR Methods).[15] For simulation, cross-validation, thresholding details, and memory and runtime benchmarking, see STAR Methods. We note that the memory usage of DCDI-G was extreme even with no hidden layers. On average, dotears outperforms all tested methods in structural recovery (Structural Hamming Distance (SHD), Figure 2) and edge weight recovery (ℓ_1 distance, Figure S9). Furthermore, dotears outperforms all other methods in most parameterizations. GIES matches dotears in "Low Density" simulations, but performs substantially worse in "High Density" simulations. We hypothesize that the greedy nature of GIES hinders performance in more complex DAGs. The primary modeling assumptions of dotears are **1.)** hard interventions (Assumption 1), **2.)** shared α across interventions (Assumption 3), and **3.)** linearity of the

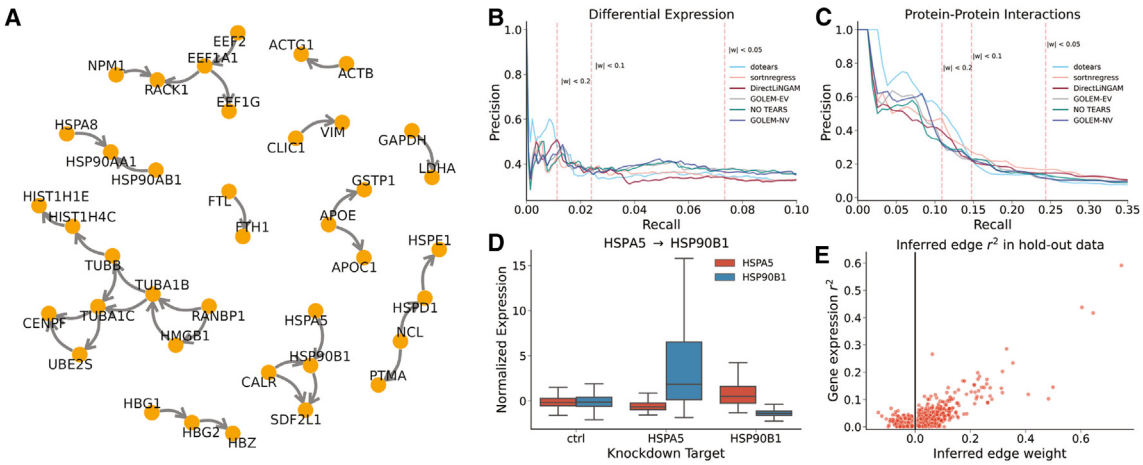


Figure 3. Differential Expression tests and protein-protein interactions validate dotears-inferred regulatory networks in genome-wide Perturb-seq data

(A) dotears-inferred network. Edges with magnitude less than 0.2, and genes without inferred edges, were removed

(B) Precision-recall curves across differential expression calls made by DESeq2. Dashed red lines indicate recall of dotears at thresholds of $|w| < 0.2, 0.1,$ and 0.05 respectively.

(C) Precision-recall curves across high confidence protein-protein interactions nominated by STRING. Dashed red lines indicate recall of dotears at thresholds of $|w| < 0.2, 0.1,$ and 0.05 respectively. d) dotears infers HSPA5 → HSP90B1. HSPA5 knockdown increases expression of HSP90B1, but HSP90B1 knockdown does not change HSPA5 expression.

(D) dotears inferred edges show correlated gene expression in hold-out observational data.

SEM. In Supplementary Material S1.7.1 and S1.7.2 we assess the sensitivity of dotears to each assumption. In addition, in Supplementary Material S1.7.1 we also consider **4.)** simulations under different interventional models. We find that dotears performance can change under violations of the hard intervention assumption, but is robust to violations of its interventional model and linearity. Surprisingly, dotears is the second best performing method under a mean-shift intervention model. Moreover, dotears outperforms the neural network method DCDI-DSF in nonlinear simulations, even with "imperfect" interventions, where the main factor determining performance is denseness of the DAG.

Genome-wide Perturb-seq

We apply all benchmarked methods in simulations to a genome-wide Perturb-seq experiment from Replogle et al.[5] We validate inferred edges through **1.)** differential expression tests in the training data using DESeq2[27,28] and **2.)** an orthogonal set of high-confidence protein-protein interactions from the STRING database.[29] We also examine gene-gene correlations in held-out observational expression data. High-confidence edges inferred by dotears show differential expression and/or protein-protein interactions more frequently than those found by other methods, and dotears outperforms all other methods in precision and recall under reasonable thresholding. Replogle et al. provide both normalized and raw data. We select the top 100 most variable genes in the raw observational data. We then benchmark all methods on the normalized, feature-selected data. Cross-validation is not performed due to low sample sizes in some knockdowns; instead, the ℓ_1 penalty is arbitrarily set to 0.1 for all methods where appropriate. Figure 3A shows the network inferred by dotears, thresholded at $|w| < 0.2$. For full details, see STAR Methods. We use DESeq2 differential expression calls and high-confidence protein-protein interactions from the STRING database to validate inferred edges.[27–29] For differential expression, we call an edge $i \to j$ a true positive if either gene shows differential expression under knockdown of the other. This is because all methods struggle equally with predicting directionality (see Table S7). For protein-protein interactions, we take high-confidence physical interactions as true positives, where here "high-confidence" is defined by STRING as having a confidence level of over 70%.[29] Figure 3B and 3C show precision and recall across thresholds for differential expression calls and the protein-protein interactions, respectively. Vertical lines indicate different thresholding regimes. Observational methods were only given observational data; for results when given both observational and interventional data, see STAR Methods. dotears shows much higher precision than all other methods at equivalent recall. Over 65% of inferred edges validated by either differential expression or high-confidence protein-protein interactions. GIES, IGSP, UT-IGSP, and DCDI are excluded because they infer binary edges. These methods inferred 3038, 3064, 3075, and 2039 out of a possible 4950 edges, respectively; no other method predicted more than 700 even at a weight threshold of 0.05. Accordingly, they had almost random precision - see STAR Methods and Tables S4–S6 for detailed results for all methods at multiple thresholds. These results reinforce concerns about the scalability of GIES to more complex scenarios. For DCDI, we report intermediate results, since convergence was not obtained under 24 h on GPU training. CPU training attempts ran out of memory even after the allocation of 180GB; GPU training attempts also repeatedly ran out of

memory. We re-run dotears on the same data excluding the observational data, and use the held-out observational data to validate inferred edges in the knockdown data. For each pair of genes i,j, Figure 3E shows the inferred edge weight against the r^2 of i,j in the observational data. Note that we take the largest magnitude weight between w_{ij} and w_{ji}. Figure 3E shows a clear relationship between inferred edge weight in knockdown data and the r^2 in observational expression data.

DISCUSSION

We present dotears, a structure learning framework that uses interventional data to estimate exogenous variance structure and subsequently leverages observational and interventional data to learn the causal graph. We showed that dotears is appropriate for Peturb-seq data analysis and can recover high-confidence gene regulation events and show that dotears outperforms all tested state-of-the-art methods in simulations. Finally, we prove that the loss function used by dotears provides consistent DAG estimation under mild assumptions. In simulations, simple methods generally outperform complex methods in structure recovery. In particular, dotears and sortnregress regularly outperform more complex methods, including neural network methods, even under modeling assumption violations and nonlinear data. Their strong performance also shows the effectiveness of using variance patterns to infer structure. In real data, dotears infers gene regulation events supported by knockdown expression in training data, orthogonal high-confidence protein-protein interactions, and gene expression correlations in held out observational data. In general, dotears provides robust inference of relevant gene regulatory events. We show that dotears is robust to error variance structure or model misspecification. Furthermore, dotears-inferred edges validate with higher precision than any other method without sacrificing power. The relatively high precision of dotears-inferred regulatory events provides confidence in identifying targets for potential experimental validation.

Limitations of the study

One limitation of dotears not addressed in simulations is that it assumes that every node has a corresponding intervention in order to estimate the error variance. Without intervention on every node, it is difficult to properly specify Ω_0.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Albert Xue (asxue@ucla.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This article analyzes existing, publicly available data, i.e., the genome-wide Perturb-seq described in Replogle et al.[5] and accessible at the link in the key resources table.
- All original code has been deposited at https://github.com/asxue/dotears/and is publicly available at https://doi.org/10.5281/zenodo.14286216 as of the date of publication.
- Any additional information required to reanalyze the data reported in this article is available from the lead contact upon request.

ACKNOWLEDGMENTS

AX was supported by the NIH Training Grant in Genomic Analysis and Interpretation T32HG002536. This work was partially funded by HHMI Hanna H Gray and Sloan fellows programs to HP. This work was partially funded by NIH grant (R35GM125055) and NSF grants (CAREER-1943497, IIS-2106908) to SS. We thank Nathan LaPierre for helpful conversations.

AUTHOR CONTRIBUTIONS

Conceptualization, A.X., S.S., and H.P.; methodology, A.X., S.S., and H.P.; formal analysis, A.X., J.R., S.S., and H.P.; investigation, A.X., S.S., and H.P.; data curation, A.X.; writing – original draft, A.X.; writing – review and editing, A.X., J.R., S.S., and H.P.; funding acquisition, A.X., S.S., and H.P.; supervision, S.S. and H.P.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- METHOD DETAILS
  - Large random graph simulations
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Statistical significance of SHD distribution difference
  - Genome-wide Perturb-seq
  - Differential expression testing
  - Protein-protein interactions
  - Precision recall tables
  - Inclusion of interventional data for observational methods
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.111673.

Received: November 1, 2024
Revised: November 20, 2024
Accepted: December 19, 2024
Published: December 24, 2024

REFERENCES

1. GTEx Consortium (2020). The gtex consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330.
2. Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M.J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Yazar, S., et al. (2021). Large-scale cis-and trans-eqtl analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat. Genet. *53*, 1300–1310.
3. Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C.P., Jerby-Arnon, L., Marjanovic, N.D., Dionne, D., Burks, T., Raychowdhury, R., et al. (2016). Perturb-seq: dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. cell *167*, 1853–1866.e17.
4. Norman, T.M., Horlbeck, M.A., Replogle, J.M., Ge, A.Y., Xu, A., Jost, M., Gilbert, L.A., and Weissman, J.S. (2019). Exploring genetic interaction

manifolds constructed from rich single-cell phenotypes. Science *365*, 786–793.

5. Replogle, J.M., Saunders, R.A., Pogson, A.N., Hussmann, J.A., Lenail, A., Guna, A., Mascibroda, L., Wagner, E.J., Adelman, K., Lithwick-Yanai, G., et al. (2022). Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. Cell *185*, 2559–2575.e28.
6. Boone, C., Bussey, H., and Andrews, B.J. (2007). Exploring genetic interactions and networks with yeast. Nat. Rev. Genet. *8*, 437–449.
7. Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E.D., Sevier, C.S., Ding, H., Koh, J.L.Y., Toufighi, K., Mostafavi, S., et al. (2010). The genetic landscape of a cell. Science *327*, 425–431.
8. Costanzo, M., VanderSluis, B., Koch, E.N., Baryshnikova, A., Pons, C., Tan, G., Wang, W., Usaj, M., Hanchard, J., Lee, S.D., et al. (2016). A global genetic interaction network maps a wiring diagram of cellular function. Science *353*, aaf1420.
9. Verma, T., and Pearl, J. (1990). On equivalence of causal models. In Proceedings of the Sixth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-90), pp. 220–227.
10. Hauser, A., and Bühlmann, P. (2012). Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. J. Mach. Learn. Res. *13*, 2409–2464.
11. Squires, C., and Uhler, C. (2022). Causal structure learning: a combinatorial perspective. Found. Comut. Math., 1–35.
12. Zheng, X., Aragam, B., Ravikumar, P.K., and Xing, E.P. (2018). Dags with no tears: Continuous optimization for structure learning. Adv. Neural Inf. Process. Syst. *31*.
13. Ng, I., Ghassami, A.E., and Zhang, K. (2020). On the role of sparsity and dag constraints for learning linear dags. Adv. Neural Inf. Process. Syst. *33*, 17943–17954.
14. Brouillard, P., Lachapelle, S., Lacoste, A., Lacoste-Julien, S., and Drouin, A. (2020). Differentiable causal discovery from interventional data. Adv. Neural Inf. Process. Syst. *33*, 21865–21877.
15. Reisach, A., Seiler, C., and Weichwald, S. (2021). Beware of the simulated dag! causal discovery benchmarks may be easy to game. Adv. Neural Inf. Process. Syst. *34*, 27772–27784.
16. Loh, P.-L., and Bühlmann, P. (2014). High-dimensional learning of linear causal networks via inverse covariance estimation. J. Mach. Learn. Res. *15*, 3065–3105.
17. Eberhardt, F., and Scheines, R. (2007). Interventions and causal inference. Philos. Sci. *74*, 981–995.
18. Pearl, J. (2009). Causality (Cambridge university press).
19. Kaiser, M., and Sipos, M. (2022). Unsuitability of notears for causal graph discovery when dealing with dimensional quantities. Neural Process. Lett. *54*, 1587–1595.
20. Saman Saboksayr, S., Mateos, G., and Tepper, M. (2023). Colide: Concomitant linear dag estimation. Preprint at: arXiv. https://doi.org/10.48550/arXiv.2310.02895
21. Yang, K., Katcoff, A., and Uhler, C. (2018). Characterizing and learning equivalence classes of causal dags under interventions. In International Conference on Machine Learning (PMLR), pp. 5541–5550.
22. Wang, Y., Solus, L., Yang, K., and Uhler, C. (2017). Permutation-based causal inference algorithms with interventions. Adv. Neural Inf. Process. Syst. *30*, 5823–5832.
23. Squires, C., Wang, Y., and Uhler, C. (2020). Permutation-based causal structure learning with unknown intervention targets. In Conference on Uncertainty in Artificial Intelligence (PMLR), pp. 1039–1048.
24. Shimizu, S., Inazumi, T., Sogawa, Y., Hyvarinen, A., Kawahara, Y., Washio, T., Hoyer, P.O., Bollen, K., and Hoyer, P. (2011). Directlingam: A direct method for learning a linear non-gaussian structural equation model. J. Mach. Learn. Res. *12*, 1225–1248.
25. Erdős, P., and Rényi, A. (1960). On the evolution of random graphs. Publ. Math. Inst. Hung. Acad. Sci *5*, 17–60.
26. Albert-László, B., and Albert, R. (1999). Emergence of scaling in random networks. Science *286*, 509–512.
27. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. Genome Biol. *15*, 550.
28. Ahlmann-Eltze, C., and Huber, W. (2021). glmgampoi: fitting gamma-poisson generalized linear models on single cell count data. Bioinformatics *36*, 5701–5702.
29. Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., Gable, A.L., Fang, T., Doncheva, N.T., Pyysalo, S., et al. (2023). The string database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res. *51*, D638–D646.
30. Köster, J., and Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. Bioinformatics *28*, 2520–2522.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Genome-scale Perturb-seq experiment data	Replogle et al.[5]	https://plus.figshare.com/articles/dataset/_Mapping_information-rich_genotype-phenotype_landscapes_with_genome-scale_Perturb-seq_Replogle_et_al_2022_processed_Perturb-seq_datasets/20029387
STRING protein-protein interactions database	Szklarczyk et al.[29]	https://string-db.org/
Software and algorithms		
DAGs with NO TEARS	Zheng et al.[12]	https://github.com/xunzheng/notears
Python 3.9	Python Software Foundation	https://www.python.org/
R 4.1.0	R Software	https://www.r-project.org/
DESeq2	Ahlmann-Eltze and Huber.[28]	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
Snakemake	Köster and Rahmann[30]	https://snakemake.readthedocs.io/en/stable/

METHOD DETAILS

Large random graph simulations

Data generation

For evaluation, data is generated for DAGs with $p = 40$ nodes. Structures are drawn from both Erdős-Rényi (ER) and Scale-Free (SF) DAGs.[25,26] We parameterize ER graphs as ER-r and SF graphs as SF-z, where $r \in [0, 1]$ represents the probability of assignment of each individual edge, whereas $z \in \mathbb{Z}$ is the integer number of edges assigned per node. We simulate under two parameterizations $\{\text{Low Density}, \text{High Density}\}$ of the edge densities (r,z). In the "Low Density" parameterization, we give $(r,z) = (0.1,2)$. To evaluate performance on higher density topologies, we also give "High Density" parameterizations, where $(r,z) = (0.2,4)$. Given an edge density scenario, we simulate under two parameterizations of the edge weights. In the "Strong Effects" parameterization, $w \sim \text{Unif}([-2.0, -0.5] \cup [0.5,2.0])$. We also give the "Weak Effects" parameterization, which edge weights are drawn from $w \sim \text{Unif}([-1.0,0.3] \cup [0.3,1.0])$ such that $|w| \leq 1$ is guaranteed. Here, there is no guarantee any two nodes will be varsortable (although they may be in practice, as a function of Ω_0).[15,19] Table S1 summarizes all four possible simulation parameterizations. For each node i, we draw $\sigma_i \sim \text{Unif}([0.5,2.0])$, and draw n_0 observations from $\epsilon_i^{(0)} \sim \mathcal{N}(0,\sigma_i^2)$. For each DAG we generate an instance of observational data, where $n_0 = (p+1) * n_k = 4100$, and an instance of interventional data, where $n_k = 100$ for all $k = 0 \ldots p$ to match sample size. We set the distribution of $\epsilon_i^{(k)}$ according to Assumptions 3 and 2 for $\alpha = 4$. For dotears, NO TEARS, sortnregress, GOLEM-EV, GOLEM-NV, CoLiDE-NV and DCDI-G 5-fold cross-validation was performed to select the regularization parameters. For each drawn DAG, a separate data instance of interventional and observational data was re-drawn from the same distribution specifically for cross-validation. After choosing a λ (or for GOLEM-EV and GOLEM-NV, the set (λ_1,λ_2)) from the data for cross-validation, the methods were evaluated on the original simulated data. For dotears, NO TEARS, sortnregress, and DCDI-G, 5-fold cross-validation was performed across the grid $\lambda \in \{001,.01,.1,1,10,100\}$. For GOLEM-EV and GOLEM-NV, 5-fold cross-validation was performed across the grid $\lambda_1 \times \lambda_2 \in \{001,.01,.1,1,10,100\} \times \{.05,.5,5,50\}$. We threshold at 0.2 for "Weak Effects" simulations, and at 0.3 for "Strong Effects" simulations. Methods are generally robust to thresholding choice. Results for precision and recall on thresholded edges are shown in Figures S7 and S8, respectively.

Benchmarking

In Figure S5, we benchmark wallclock time and memory usage for all methods in $p = 40$ simulations[30] on the UCLA hoffman2 cluster. All continuous optimization methods (dotears, NO TEARS, GOLEM-EV, GOLEM-NV, and CoLiDE-NV) have significantly higher average runtimes than other methods, which is partially explained by cross-validation procedures. dotears has relatively light memory usage, outperformed only by NO TEARS, sortnregress, and CoLiDE-NV. DCDI-G has enormous memory requirements, especially relative to other methods. We note that for DCDI-G, the reported benchmarks do not include memory usage or runtime from cross-validation folds. Instead, we report only the "main" run of DCDI-G, and thus DCDI-G is not denoted with the * indicating cross-validation.

Thresholding on large simulation results

Edge thresholding for weighted adjacency matrices is necessary for accurate evaluation using SHD, but the choice of threshold can feel arbitrary. We find that methods are generally robust to thresholding choice, following similar results from Reisach et al.[15] Figure S6

examines the effect on thresholding of small weights in W in large random DAG simulations, for methods that infer weighted adjacency matrices (dotears, NO TEARS, sortnregress, DirectLiNGAM, GOLEM-NV, GOLEM-EV, and CoLiDE-NV).[12,13,15,24] For simplicity, we summarize simulations in terms of the generative edge distribution. Simulations with "Weak Effects", $w \sim \text{Unif}([-1.0, -0.3] \cup [0.3, 1.0])$, are shown in Figure S6a; simulations with "Strong Effects", $w \sim \text{Unif}([-2.0, -0.5] \cup [0.5, 2.0])$ are shown in Figure S6b. We compare the SHD between the ground truth adjacency matrix and the inferred adjacency matrix for each method at all thresholds between 0 and the absolute lower bound of the true edge weight distribution (0.3 for "Weak Effects", 0.5 for "Strong Effects"). Any edge whose magnitude is below the chosen threshold is set to 0 for SHD evaluation. Figure S6 shows that thresholding is necessary for the evaluation of SHD between weighted adjacency matrices, but also that methods are generally robust to thresholding choice. Without thresholding (equivalently, a threshold of 0), SHD results are inflated, but recover even at low thresholds.

Edge weight estimation for large random simulations

Accurate estimation of edge weights is important for understanding structure. Figure 2 gives results on structural recovery through SHD, but does not inform edge weight recovery. To measure edge weight recovery we use ℓ_1 distance, defined as the vector ℓ_1 norm between the flattened true weighted adjacency matrix and the flattened inferred weighted adjacency matrix. ℓ_1 distance gives information on both structure recovery and edge weight estimation simultaneously. In Figure S9, we benchmark the recovery of edge weights for methods that return a weighted adjacency matrix (dotears, sortnregress, NO TEARS, GOLEM-EV, GOLEM-NV, DirectLiNGAM, CoLiDE-NV).[12,13,15,20,24] For fairness, we exclude methods that only return a binary adjacency matrix. Methods are thresholded in the same manner as in Figure 2 dotears outperforms all other methods in terms of effect size recovery. In addition, the relative ordering of the methods stays consistent with Figure 2.

QUANTIFICATION AND STATISTICAL ANALYSIS

In simulations, we benchmark in terms of Structural Hamming Distance (SHD), the number of edge additions, removals, or reversals needed to change one DAG into another. Figures either depict mean (point) and standard error (bar) (Figure 1) or boxplot quartiles (Figures 2 and 3). Details of statistical tests can be found in the following sections.

Statistical significance of SHD distribution difference

For each method, and across all large random graph simulations, we compare the SHD distributions, marginalized across all simulation parameterizations, against that of dotears. For each method, we perform a one-sided Mann-Whitney U test to test difference between the method's SHD distribution and the SHD distribution of dotears. Here, $n = 80$ is the total number of large random graph simulations. The resulting p-values are reported in Table S2, and are significant for all methods.

Genome-wide Perturb-seq

We benchmark all methods on a single-cell Perturb-seq experiment from Replogle et al.,[5] who provide both raw count data as well as normalized data, where normalized here means a z-scoring relative to the mean and standard deviation of the control dataset, accounting for batch. Here, the control dataset indicates a set of pre-selected control cells, and not simply cells with non-targeting guides. For details, see Replogle et al.[5] We perform feature selection in the raw data. We select the top 100 most variable genes in the raw observational data, excluding **1.)** genes coding for ribosomal proteins and **2.)** genes without knockdown data. Here, the observational data indicates cells incorporating non-targeting control guides. We then take the normalized expression data for these selected 100 genes in **1.)** the observational data, and **2.)** each of the 100 knockdowns. This forms our training set in Figures 3A–3D. In Figure 3E, we perform the same procedure as above, but exclude **1.)** the observational data. In other words, our training set is formed exclusively from expression data in the 100 knockdowns. Cross-validation is not performed due to low sample sizes in some knockdowns; instead, the L1 penalty is arbitrarily set to 0.1 for all methods where appropriate. Otherwise, method settings are the same as simulations. For DCDI, we report intermediate results, since convergence was not obtained under 24 h even on GPU training. CPU training was attempted with 30 cores and 180 GB memory, but was killed by the operating system. GPU training was attempted, but was also repeatedly killed by the operating system for memory constraints. DCDI is run as DCDI-G, since DCDI-DSF would not fit in memory, and is also run with imperfect interventions and known interventions. DCDI reported 2982 total edges, but many of these are both causal and anti-causal; in total, 2039 "interactions" were reported.

Differential expression testing

We use DESeq2 to test for differential expression.[27,28] We calculate size factors across the raw feature-selected data, and use these in all downstream tests. Next, for each knockdown $\text{ko}(i)$, we test for differential expression for all genes j compared to the observational data. For single-cell data, we run DESeq with parameters test = LRT, fitType = glmGamPoi, useT = TRUE, minmu = 1e-6, minReplicatesForReplace = Inf, reduced = 1. We repeat this for each knockdown independently in turn. Here, n is the number of cells in each knockdown, and can be found in Replogle et al.. At the end of this procedure, we have 100^2 unadjusted p-values. We perform Benjamini-Hochberg correction to an FDR level of 0.05. Subsequently, in $\text{ko}(i)$, if gene j has an adjusted p-value less than 0.05, we call that a true edge in downstream calculations of precision and recall.

Protein-protein interactions

As a second validation edge set, we use protein-protein interactions from the STRING database.[29] We pull protein-protein interactions for Homo Sapiens down, and use only physical evidence for interactions. Protein-protein interactions were only used if the STRING confidence score was over 70%.

Precision recall tables

We present precision and recall at three edge threshold levels ($|w| > 0.2$, $|w| < 0.1$, and $|w| < 0.05$) for all methods. We use both differential expression calls and protein-protein interactions as true sets, and separate results for both. Tables S4, S5, and A6 denote results for $|w| > 0.2$, $|w| < 0.1$, and $|w| < 0.05$, respectively. For differential expression calls we ignore causal direction, and instead for two given genes denote any differential expression in either direction as a "true" edge. This is because all methods struggle equally with inferring causal direction under differential expression, and call an equal number of edges in the "correct" causal direction as the "incorrect" anticausal direction (see Table S7).

Inclusion of interventional data for observational methods

In Figure 3, observational methods (sortnregress, DirectLiNGAM, GOLEM-EV, GOLEM-NV, NO TEARS, CoLiDE-NV) are only given observational data. It is reasonable to wonder if their decreased performance relative to dotears is due to a substantial sample size decrease (93691 total to 75328 purely observational). To test this, we also ran the observational methods on observational and interventional data concatenated. Since these are observational methods, there is no way to label these data as interventional; instead, these methods assume they are observational. Results are shown in Figure S21. No substantial relative performance difference was observed in the new trials. All observational methods still perform worse than dotears, especially in high-confidence regimes.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

This study did not include experiments with a specific model or subject.