BMC Bioinformatics

RESEARCH

Open Access

Causal models and prediction in cell line perturbation experiments

James P. Long[1*], Yumeng Yang[2], Shohei Shimizu[3], Thong Pham[3] and Kim-Anh Do[1*]

*Correspondence:
jplong@mdanderson.org;
kimdo@mdanderson.org

[1] Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA
[2] Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA
[3] Faculty of Data Science, Shiga University, Hikone, Shiga, Japan

Abstract

In cell line perturbation experiments, a collection of cells is perturbed with external agents and responses such as protein expression measured. Due to cost constraints, only a small fraction of all possible perturbations can be tested *in vitro*. This has led to the development of computational models that can predict cellular responses to perturbations *in silico*. A central challenge for these models is to predict the effect of new, previously untested perturbations that were not used in the training data. Here we propose causal structural equations for modeling how perturbations effect cells. From this model, we derive two estimators for predicting responses: a Linear Regression (LR) estimator and a causal structure learning estimator that we term Causal Structure Regression (CSR). The CSR estimator requires more assumptions than LR, but can predict the effects of drugs that were not applied in the training data. Next we present Cellbox, a recently proposed system of ordinary differential equations (ODEs) based model that obtained the best prediction performance on a Melanoma cell line perturbation data set (Yuan et al. in Cell Syst 12:128–140, 2021). We derive analytic results that show a close connection between CSR and Cellbox, providing a new causal interpretation for the Cellbox model. We compare LR and CSR/Cellbox in simulations, highlighting the strengths and weaknesses of the two approaches. Finally we compare the performance of LR and CSR/Cellbox on the benchmark Melanoma data set. We find that the LR model has comparable or slightly better performance than Cellbox.

Keywords: Causal inference, Prediction, Perturbation biology, Systems biology

Introduction

In cell line perturbation experiments, a collection of cells is perturbed with gene knockdowns, overexpression, or pharmaceutical drugs and responses such as cell survival and gene and protein expression are measured. The results of these experiments play an important role in our understanding of cellular biology and in development of treatments for complex diseases such as cancer [2–5].

There are a huge number of possible perturbations that can be applied to a cell line. For example, in human cell lines there are $\sim 20,000$ genes that could be perturbed (e.g. knocked out). Thus there are ~ 200 million perturbations of gene pairs (double knockouts). Further each perturbation may be applied across hundreds of cell lines (e.g. cells of different types of cancer). Thus in practice even large-scale experiments can only test a small fraction of all possible perturbations.

This limitation has led to the development of *in silico* perturbation response prediction models [1, 4, 6–10]. Models are typically trained on a set of perturbations that are experimentally tested in a laboratory and where cellular responses to the perturbation are known (up to technical replicate variability). These *in silico* models can then be used to make response predictions for untested perturbations. Predicted responses of biological interest, e.g. a perturbation which is predicted to suppress growth in a tumor cell line, can then be experimentally validated *in vitro*.

[1] proposed Cellbox, a perturbation prediction model based on a system of Ordinary Differential Equations (ODEs). Cellbox was benchmarked against several competitors on a Melanoma cell line in which cells were perturbed with 12 drugs, given at varying concentrations in each experiment. Cellbox achieved the best performance both on predicting cellular responses to drugs used in the training set and on predicting responses to drugs not used in the training set. This latter form of prediction, termed Leave One Drug Out (LODO) validation, is both challenging and scientifically impactful because it implies that the model can extrapolate to predict the effect of new drugs. In this work, we make the following contributions:

1. We propose a causal Structural Equation Model (SEM) for modeling the effect of drug perturbation on cell lines. From this model, we derive two estimators, Linear Regression (LR) and Causal Structural Regression (CSR), for predicting cellular responses to perturbations. These results show that CSR, which explicitly estimates the coefficients in a causal graph, can extrapolate to predict cellular responses to untested (in training set) drugs.
2. We derive analytic results which show that the linear version of Cellbox is equivalent to CSR. This provides a formal causal interpretation for the Cellbox model which was not discussed in [1].
3. We compare LR and CSR/Cellbox in simulations. These simulations demonstrated the strengths/weaknesses of the two approaches, including sensitivity of CSR/Cellbox to misspecification of the direct effect of interventions.
4. We show that LR has comparable or superior performance to Cellbox on the Melanoma benchmark data set. Our results reinforce the finding in other works that simple benchmark models may obtain equal or better performance than complex models in cell line perturbation response prediction [11].

All code and data for reproducing the results in this work is publicly available.[1]

Overview of data, prediction problem, and connections to existing work

In this work we consider perturbation experiments on a RAFi-resistant melanoma cell line SkMel-133 originally collected in [4]. The data structure is depicted in Fig. 1a. The cell line was treated with 89 drug perturbations (rows). Perturbations are defined by the concentrations of 12 drugs (12 columns of blue matrix). Drugs were applied as a single

[1] https://github.com/longjp/causal-pred-drug-code

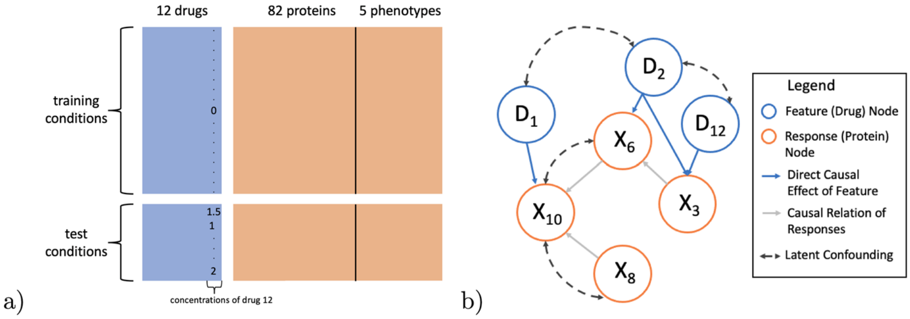


Fig. 1 **a** Overview of perturbation data with a leave-one-drug-out (LODO) training / test set split. Drug 12 is never used in training. **b** Causal graphical model for subset of drugs and responses (proteins and phenotypes). Drugs are exogenous variables with known targets, e.g. it is assumed known that drug 12 directly effects protein X_3. Black dashed lines represent hidden confounding. For example unmeasured cell cycle may effect responses X_{10} and X_8

agent and in combinations of two drugs. Since at most two drugs were used in any experiment, each row of the blue matrix contains either 1 (if perturbation uses single drug) or 2 (if perturbation uses two drugs) non-zero values. In each perturbation experiment, the expression of 82 proteins was measured 24 h after perturbation using Reverse Phase Protein Arrays [12]. In addition, five cell phenotypes were measured, quantifying cell-cycle progression and cell viability (orange columns). For this work, we use the data as supplied by [1].[2]

In perturbation prediction, drug concentrations (blue columns) are used to predict protein and phenotype responses (orange columns). Data is divided into training and test sets. Using only the training data, a model is constructed which can predict protein/phenotypes from the drug concentrations. The performance of the model is then evaluated using the test set drug concentrations and protein/phenotypes. When the test set is a simple random sample of all perturbations, this setup matches the standard approach to fitting and validating predictive models. Models such as linear regression of response variables on the drug concentrations can be used.

In practice one would like to construct a model which can accurately predict the effect of untested drugs and in doing so identify perturbations with interesting responses for further follow up. Random Fold (RF) cross validation, in which the test conditions are a simple random sample of all conditions, does not represent this use case well because all drugs are used in training. A more challenging form of validation, leave-one-drug-out (LODO), more closely aligns with the intended scientific uses of the perturbation prediction model. Figure 1a depicts a LODO training-test set split. Here drug 12 is left out of the training set i.e. the concentration of drug 12 in the training data is always 0 because drug 12 was never used in training perturbations. The test perturbations all use drug 12 so column 12 of the drug matrix is never 0 in test. LODO prediction is challenging for regression models because there is no way for the model to determine the effect of drug 12 on the response variables. For example, coefficients in a linear regression of response on drugs will not be defined because the gram matrix is not invertible.

[2] Available at https://github.com/sanderlab/CellBox.

The direct targets of drugs are often known a priori. For example, a mitogen-activated protein kinase (MEK) inhibitor drug should directly reduce the expression of the MEK protein. Other changes in the system, could then be assumed to be a downstream effect of MEK inhibition. Using this information about drug targets, a causal model, which infers causal relations among the protein and phenotype response variables, can be used to predict responses in LODO validation. The approach is graphically summarized in Fig. 1b. For clarity only a small number of the drug and response variables are shown. Drugs (blue nodes) are known to target (blue arrows) particular proteins (orange nodes). For example drug D_{12} targets protein X_3. The causal relations among proteins is unknown a priori (grey arrows). Training set perturbations can be used to identify and estimate causal effects among the proteins. Then the effect of an untested perturbation, e.g. drug 12, can be determined by first assuming that the direct effect of drug 12 will be on protein 3, and then propagating this effect through the inferred protein network.

Related work and innovation

[1] developed an ordinary differential equation (ODE) model termed Cellbox and tested it on the Melanoma cell line data, both using RF and LODO validation. Cellbox outperformed all competing algorithms in both forms of model validation. In the following sections, we derive analytic results relating Cellbox to causal structure learning models, providing a new causal interpretation of Cellbox.

Several works have developed models for predicting responses to previously tested perturbations in new settings. For example, [7, 8], and [13] developed deep-learning autoencoder models to predict responses to previously untested perturbation-cell type combinations. [6] considered a similar problem and developed prediction model, SI-A, derived from the synthetic control literature in causal inference. [10] developed GEARS to predict the effect of double knockouts/knockdowns using *in vitro* responses to single knockouts/knockdowns.

In these works, test perturbations were applied to training conditions in different cell types or cell lines. These prediction problems are less challenging than the LODO Melanoma prediction because in LODO the held-out treatment has not been applied in the training set. These methods cannot be directly applied to the LODO setting. There is a large literature on inferring causal relations among genes, termed gene regulatory networks [14–18]. However these methods generally do not use interventional data to estimate the network. More importantly, these works view the regulatory network (graph) as the target for inference while in this work we are primarily interested in using the inferred network to make predictions about how interventions (drugs) will affect the system.

Environment-based causal estimators such as Invariant Causal Prediction (ICP) and the Causal Dantzig (CD) have been used to predict the effect of untested gene-knockdowns in yeast cell lines [19–21]. These methods assume the existence of different data collection environments, such as an observational environment where samples are collected without any perturbation and an interventional environment where samples are perturbed. Causal effects among the response variables are estimated based on the principle of invariance. These methods do not leverage information on the direct targets of interventions, which is used by CSR and Cellbox considered in this work.

Causal perturbation model and estimators

Let $X \in \mathbb{R}^p$ be a vector of protein and phenotype responses (row of orange matrix in Fig. 1a) and $D \in \mathbb{R}^q$ be a vector of drug concentrations (row of blue matrix in Fig. 1a). There are n training observations $\{(D_i, X_i)\}_{i=1}^n$. The objective is to predict a test response $X^{te} \in \mathbb{R}^p$ when drugs $D^{te} \in \mathbb{R}^q$ are applied for an observation not in the training data. Subscripts j and k will be used to denote specific elements of X and D.

We propose a causal Structural Equation Model (SEM) for how drugs effect response variables and then derive two estimators using the model. Figure 1b contains a graphical representation of the model. Formally:

$$\begin{aligned} &D \leftarrow f_D \\ &\epsilon \leftarrow f_\epsilon \\ &X = AX + g(D) + \epsilon. \end{aligned} \tag{1}$$

The distribution f_D models selection of drug concentrations to apply to cells. As indicated by black dashed lines in the figure, concentrations may be dependent. The term $\epsilon \in \mathbb{R}^p$ is a vector of errors with $\mathbb{E}[\epsilon] = 0$. Elements of ϵ may be correlated, representing hidden confounding among response variables X. Confounding may be caused by factors such as temperature, cell cycle, and variations in laboratory conditions. Figure 1b represents this hidden (latent) confounding with dashed lines connecting orange nodes. We assume $\epsilon \perp\!\!\!\perp D$ which is generally well justified because the choice of drug concentrations D is independent of the unmeasured factors modeled by ϵ. $A \in \mathbb{R}^{p\times p}$ is a matrix of coefficients where A_{jk} represents the causal effect of a one-unit change of X_k on X_j. The term $g(D) \in \mathbb{R}^p$ represents the direct effects of drug concentrations D on the response variables (protein concentrations).

In this work, we assume that $(I - A) \succ 0$ (positive definite). By Sylvester's criteria, $I - A \succ 0$ whenever A represents an directed acyclic graph (DAG). Since $I - A \succ 0$ implies $I - A$ is invertible, Equation (1) may be rewritten as

$$X = (I - A)^{-1} g(D) + \underbrace{(I - A)^{-1}\epsilon}_{\equiv \delta}. \tag{2}$$

The prediction target is the mean response when drug concentrations $D \in \mathbb{R}^q$ are applied to the system:

$$f(D) \equiv \mathbb{E}[X|do(D)] = \mathbb{E}[X|D] = (I - A)^{-1} g(D). \tag{3}$$

The equality between $\mathbb{E}[X|do(D)]$ and $\mathbb{E}[X|D]$ is justified by the fact that there are no backdoor paths from D to X [22, 23]. Note that this is only true for the vector $D \in \mathbb{R}^q$. For any individual drug concentration, e.g. D_{12}, latent confounding between drugs may induce backdoor paths. We now discuss two approaches to estimating f and their relative strengths and weaknesses.

Regression

One can assume f belongs to a class of functions $\mathcal{F}$ and then select an f which minimizes loss. For example consider

$$\widehat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \sum_{i=1}^{n} ||X_i - f(D_i)||_2^2 + \gamma(f, \lambda) \tag{4}$$

where γ is a penalty function and λ controls the degree of penalization with $\gamma(f, 0) = 0$. The predicted response for the test observation is $\widehat{f}(D^{te})$. In the case where the direct effect of drugs on phenotypes is linear, i.e. $g(D) = BD$ for some $B \in \mathbb{R}^{p \times q}$, then the class of functions $\mathcal{F}$ is linear because $f(D) = (I - A)^{-1}BD$ and can be parameterized by $R = (I - A)^{-1}B$. In this case, the objective function may be written as

$$\widehat{R} = \underset{R}{\operatorname{argmin}} \sum_{i=1}^{n} ||X_i - RD_i||_2^2 + \gamma(R, \lambda) \tag{5}$$

with predicted response

$$\widehat{X}_{LR} = \widehat{R}D^{te}.$$

We term $\widehat{X}_{LR}$ the Linear Regression (LR) prediction.

For LODO validation with drug j held out, recall that $D_{ij} = 0$ for all $i \in \{1, \ldots, n\}$ (the training data), but $D_j^{te} \neq 0$. In this case, the regression estimators will not be consistent. For LR (Equation (5)), without regularization ($\lambda = 0$), the minimizer of the objective function is unique if and only if

$$\begin{pmatrix} D_1^T \\ \vdots \\ D_n^T \end{pmatrix}$$

is full rank. This will not be the case with LODO validation because every element of the j^{th} column is 0. If sparsity inducing regularization is used, e.g. $\gamma(R, \lambda) = \lambda \sum |R_{jk}|$, then the j^{th} column of $\widehat{R}$ will be 0. In this case, the effect of drug j on all response variables is estimated to be 0 and hence inconsistent. Qualitatively, regression is unsuitable because LODO requires an extreme form of extrapolation where feature j (drug concentration j) is always 0 in training but non-zero in the test set.

Causal structure regression (CSR)

Suppose g (the direct targets and strength of effects of the drugs) is known. In this case, it is possible to estimate A (or equivalently $(I - A)^{-1}$). One possible estimator of A is obtained by regressing X on $g(D)$. Specifically

$$\widehat{A} = \underset{\{A: I - A \succ 0\}}{\operatorname{argmin}} \sum_{i=1}^{n} ||X_i - (I - A)^{-1}g(D_i)||_2^2 + \gamma(A, \lambda). \tag{6}$$

Consistency of this estimator does not require non-gaussianity, equal variance, or no hidden confounding assumptions common in the causal discovery literature [24–26]. Instead, the interventions (drug concentrations D) act as exogeneous variables which identify the causal structure matrix A. The predicted response to drug concentrations D^{te} is

$$\widehat{X}_{CSR} = (I - \widehat{A})^{-1} g(D^{te}).$$

We term $\widehat{X}_{CSR}$ the Causal Structure Regression (CSR) prediction because the prediction is based on a regression estimate of the causal structure A. The j, k element of matrix $(I - A)^{-1}$ is the total effect of response X_k on X_j. In practice, it may be simpler to directly estimate $T = (I - A)^{-1}$, check whether the resulting $\widehat{T}$ is positive definite, and then use the estimate $\widehat{T}$ to predict $\mathbb{E}[X|D]$. When implementing this approach without regularization we have

$$\widehat{T} = \underset{T \in \mathbb{R}^{p \times p}}{\operatorname{argmin}} \sum_{i=1}^{n} ||X_i - Tg(D_i)||_2^2 \tag{7}$$

and

$$\widetilde{X}_{CSR} = \widehat{T} g(D^{te}). \tag{8}$$

Now consider CSR for estimating $f(D^{te})$ with LODO validation with drug j held out. The estimator $\widehat{T}$ (Equation (7)) will be uniquely defined if and only if

$$\begin{pmatrix} g(D_1)^T \\ \vdots \\ g(D_n)^T \end{pmatrix}$$

is full column rank. This rank condition may be satisfied even in the LODO setting. In particular, the rank condition for CSR requires that the training drugs directly target every response variable, not that every drug is used in the training set.

The relative merits of the Regression and CSR predictions are summarized as follows:

- **LODO Validation**: As discussed, regression is inconsistent for LODO validation.
- **p versus q**: CSR estimates A, which consists of p^2 parameters corresponding to direct effects of all response variables on each other. LR estimates R, which consists of qp parameters corresponding to the total effect of each of the q drugs on the p response variables. Thus when $q < p$, regression requires estimating fewer parameters.
- g **Assumption**: CSR requires knowledge of g, the direct effects of interventions on response variables. If g is unknown or contains a large amount of uncertainty, the regression estimator may be preferred.
- **Interpretability:** CSR is more interpretable because it estimates matrix A which encodes how response variables causally effect each other, providing biological insight on how cells function.

Causal structure regression and cellbox

We now discuss Cellbox, an ODE model introduced in [1], which obtained state-of-the-art prediction performance on the Melanoma cell line perturbation experiments. First we summarize the Cellbox modeling and fitting procedure, modifying notation in certain instances for clarity.[3] We then derive results which demonstrate a close connection between Cellbox and CSR.

Cellbox uses a system of ODEs to model how proteins and phenotypes influence each other across time. Let $x_i(t, \theta) \in \mathbb{R}^p$ be the log-normalized change at time t (relative to time 0) of a set of p proteins and phenotypes under perturbation condition i. The unknown parameters θ control how proteins influence each other. For observation i, drug concentrations $D_i \in \mathbb{R}^q$ are applied. Define $u_i = g(D_i)$, the direct effect of applying drug concentrations D_i to the system. Since D_i and g are assumed known, u_i is known as well.

Response j (protein or phenotype) under condition i is modeled by

$$\frac{\partial x_{ij}(t,\theta)}{\partial t} = \tau_j \phi\left(\sum_{k \neq j} w_{jk} x_{ik}(t,\theta) + u_{ij}\right) + w_{jj} x_{ij}(t,\theta). \tag{9}$$

The unknown model parameter is $\theta = (W, \tau)$ with w_{jk} for $j \neq k$ representing the interaction between x_j and x_k, w_{jj} characterizes the effect of decay (the tendency of protein j to return to the original level before perturbation), and τ_j controls the saturation effect of the protein. Cellbox can be fit with several envelope functions ϕ including identity, clipped linear, and sigmoid.

In [1] Cellbox was fit with response variables measured at a single time point 24 h after perturbation initiation. It was assumed that by this time, the system has reached steady state. The steady state (equilibrium) changes implied by the model is

$$x_{ij}(\theta) \equiv \lim_{t \to \infty} x_{ij}(t, \theta). \tag{10}$$

To estimate parameters θ, discrepancy between model predicted responses ($x_{ij}(\theta)$) and experimental responses (X_{ij}) was computed with a L_1 (lasso) penalty term to induce sparsity on the off-diagonal elements of W. Specifically

$$L(\theta) = \sum_{j=1}^{p} \sum_{i=1}^{n} |X_{ij} - x_{ij}(\theta)|^2 + \lambda ||W - diag(W)||_1$$

where $diag(W) \in \mathbb{R}^{p \times p}$ is the diagonal component of W. Given a candidate θ, an ODE solver can be used to approximate $x_{ij}(\theta)$. Subsequently, θ are updated using gradient descent with automatic differentiation to determine

$$\widehat{W}, \widehat{\tau} = \underset{\theta=(W,\tau)}{\operatorname{argmin}} L(\theta). \tag{11}$$

To predict the response X^{te} for some combination of drugs D^{te}, first the direct effects of the drug are determined ($u_i = g(D^{te})$), followed by an ODE solver to approximate steady state expression levels (Equation (10)) using parameters $\widehat{W}$ and $\widehat{\tau}$.

We now show that using a linear envelope function ϕ and setting $\tau = 1$, Cellbox is equivalent to the linear CSR model.

[3] See Model Configuration section of METHOD DETAILS in Star Methods of [1] for original exposition of model.

Theorem 1 *Suppose ϕ is a linear envelope function, $\tau = 1$, and $W \prec 0$. We have the following results*

1. *The equilibrium state of the Cellbox model (Equation (10)) has closed form*

$$x_i(\theta) = (x_{i1}(\theta), \ldots, x_{ip}(\theta))^T = -W^{-1}g(D_i) \tag{12}$$

and the Cellbox predicted response for test drugs D^{te} is

$$\widehat{X}_C = -\widehat{W}^{-1}g(D^{te}). \tag{13}$$

2. *The Cellbox parameter optimization (Equation (11)) may be expressed as*

$$\widehat{W} = \underset{W:W \prec 0}{\operatorname{argmin}} \sum_{i=1}^{n} ||X_i - (-W^{-1})g(D_i)||_2^2 + \lambda ||W - diag(W)||_1. \tag{14}$$

3. *If the penalty function $\gamma(A, \lambda) = \lambda ||A - diag(A)||_1$ is used in CSR in Equation (6), then*

$$\widehat{A} = \widehat{W} + I \tag{15}$$

and $\widehat{X}_{CSR} = \widehat{X}_C$.

See Section A.1 for a proof. Equations (12) and (14) show that for the linear version of Cellbox (ϕ identity and $\tau = 1$), ODE solvers are not necessary for estimating parameters $\widehat{W}$ and making predictions. Equation (15) shows that linear Cellbox is a reparameterized version of CSR and test predictions $\widehat{X}_C$ and $\widehat{X}_{CSR}$ are identical. This provides a causal interpretation for Cellbox in terms of structural equations. The general principal of this result, that causal structural equation models are steady state limits of dynamical systems, has been derived in several previous works [27, 28]. The assumption that $W \prec 0$ is weak because 0 or positive eigenvalues in W imply the system is not converging to any steady state.

We note that the implementation of Cellbox in [1] set elements of W to 0 which represent phenotype to protein causal effects. This is accomplished by restricting the domain of the parameter optimization in Equation (11). This enforces the domain knowledge that proteins may influence phenotypes but not vice versa. For clarity of exposition, we do not impose the conditions here or in the simulations since they are not directly relevant for understanding the relationship between regression and causal predictive models. However in the application to the Melanoma cell line, we follow [1] and impose the restriction.

Simulation

We conduct a simulation to compare the performance of the regression estimator ($\widehat{X}_{LR}$) and CSR/Cellbox ($\widetilde{X}_{CSR}$). We simulate from Causal SEM (1)

$$X = AX + g(D) + \epsilon$$

using $p = 5$ response variables ($X \in \mathbb{R}^5$) and $q = 15$ drugs ($D \in \mathbb{R}^{15}$). Drugs are assumed to have a linear effect on response variables so $g(D) = BD$. The structure of B and A are specified in Equation (16). Five drugs target a single response variable and 10 drugs target two of the response variables. Drugs with a single target have a strength of 1 while drugs with 2 targets have a strength of 0.5 for each target. All possible combinations of 2 drugs are applied to the system, thus there are a total of $n = \binom{15}{2} = 105$ observations with each D_i having exactly two non-zero entries equal to 1. The variable X_1 has a causal effect of 1.6 on X_2 and 1.2 on X_3. The variable X_3 has a causal effect of 2 on X_4. All other causal effects among the response variables are 0. The exogenous error ϵ is distributed $N(0, 0.1^2)$.

$$B = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0.5 & 0.5 & 0.5 & 0.5 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0 & 0.5 & \cdots & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0.5 & \cdots & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0.5 & 0 & 0 & \cdots & 0.5 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0.5 & 0 & \cdots & 0.5 \end{pmatrix} \in \mathbb{R}^{5 \times 15} \qquad A = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 1.6 & 0 & 0 & 0 & 0 \\ 1.2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{16}$$

Since $q > p$ (number of drugs is greater than number of response variables), regularization is not necessary (λ is set to 0 in all the simulations). We fit the LR and CSR estimators under three settings:

- **Random Fold (RF):** The data is divided randomly into 2/3 training and 1/3 test. Since the training-test set split is random, every drug is used in training.
- **RF with B Misspecified:** The training-test set split is identical to RF. However the B matrix (direct effect of drugs) is misspecified. Instead of using the correct B, the 10 drugs with 2 targets are assumed to influence their targets with a strength of 1 (rather than the correct value of 0.5).
- **Leave-one-drug-out (LODO):** For each condition in the test set, one (of the two) drugs used in the condition is selected at random. The 2/3 of the training data is subset to only use conditions where the selected drug is not used. For LR, the coefficient on the held out drug is set to 0.

Results are summarized in Fig. 2. The true response values are plotted on the x-axis and the predicted response values are plotted on the y-axis. High correlations imply that the estimator is performing well in the respective setting. For Random Fold (RF) cross validation, both LR and CSR perform well. In the RF with B Misspecified setting, LR performs well and in fact makes identical predictions to RF validation because the LR estimator does not depend on B. In contrast, CSR performs poorly because it uses an incorrectly specified B. Finally in LODO, LR performs poorly because it incorrectly infers that the effect of the left out drug on all response variables 0. In contrast, CSR performs well because it models the causal relations among the response variables which enables it to generalize predictions to untested drugs.

The true A and the estimated $\widehat{A}$ for each setting (RF, RF with B Misspecified, LODO) are displayed graphically in Fig. 3. Note that CSR in LODO estimates A for each test observation. We plot only one of them here. Edge widths are proportional to size of the coefficient estimate. For visual clarity, small effects (coefficients less than 0.2 in absolute size) are not displayed. The Random Fold $\widehat{A}$ in Fig. 3b and the LODO $\widehat{A}$ in Fig. 3d closely resemble the true A in Fig. 3a. In contrast, Fig. 3c shows that when B is misspecified the resulting $\widehat{A}$ is a poor estimate.

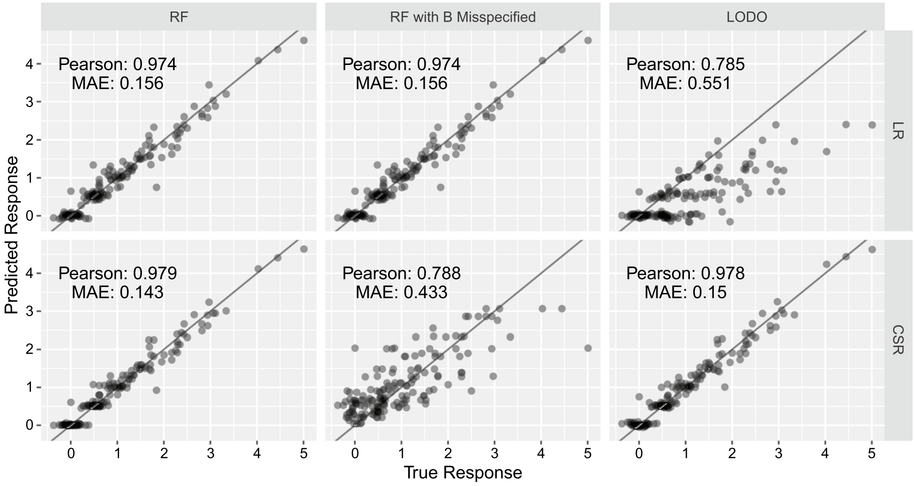


Fig. 2 Comparison of performance of LR and CSR/Cellbox on simulated data. The x-axis is the true response and the y-axis is the predicted response. LR and CSR perform similarly for RF validation. For RF with *B* misspecified, CSR, which uses *B*, performs poorly. LR is unaffected by misspecified *B* because *B* is not used in the LR estimator. For LODO validation, LR performs poorly because it cannot model the effect of the left out drug on the responses

Melanoma cell line perturbation prediction

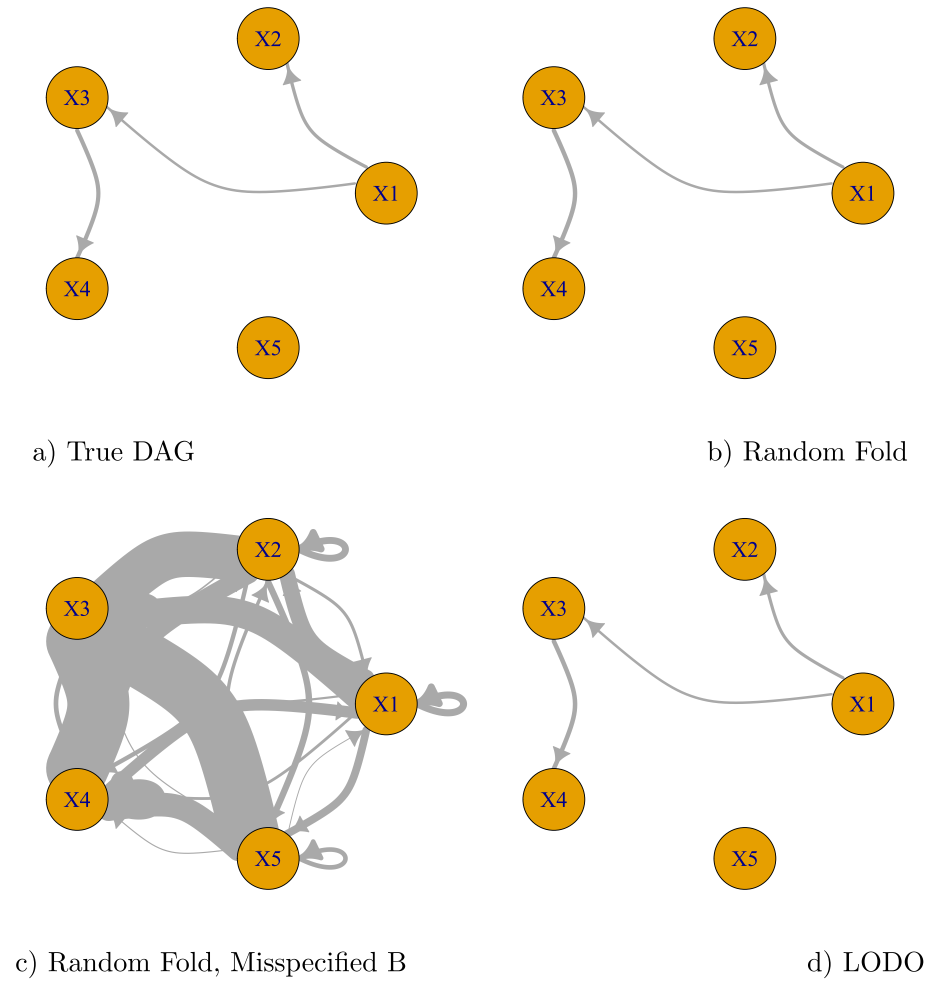
We compare Cellbox and LR for prediction of protein and phenotype responses in the Melanoma data set introduced in Sect. . The two validation procedures we describe below follow the procedures in [1]. Cellbox is implemented with a sigmoid activation function. See [1] for details on tuning parameter selection, choice of *g*, and the (causal) graph/network estimated by the model (Figure 5).

Random fold cross validation

The 89 experimental conditions are split into 70% training (62 conditions) and 30% testing (27 conditions). Models are fit on the training set and used to predict the responses on the test set. This process is repeated 1000 times and the predictions averaged across these runs. Predicted responses versus experiment responses are plotted for Cellbox in Fig. 4a and LR in Fig. 4b. LR predictions show a stronger correlation with the response than Cellbox (Pearson's correlation of 0.947 versus 0.926) and lower mean absolute error (0.093 versus 0.105). RF cross validation favors regression models (relative to LODO) because the regression model estimates fewer parameters and does not require regularization.

Leave one drug out

We now consider the more challenging Leave One Drug Out (LODO) validation where a drug is held out of training. For example, if the drug aMEK is held out, the training data is all conditions with aMEK concentration equal to 0 and the test set is all conditions where aMEK has been applied, either as monotherapy or in combination with other drugs. Since there are 12 drugs, there are 12 training-test set pairs.



a) True DAG

b) Random Fold

c) Random Fold, Misspecified B

d) LODO

Fig. 3 True network and estimated networks for different simulation settings. For Random Fold **b** and LODO **d**, the estimated *A* is quite close to the true DAG *A* **a**. For Random Fold with Misspecified *B* **c**, the estimated *A* contains many erroneous edges

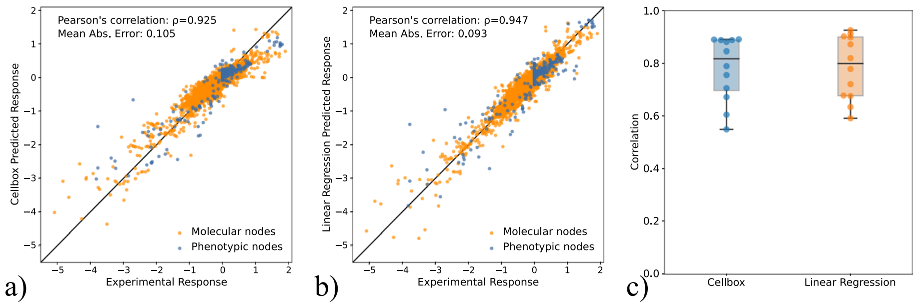


Fig. 4 LR outperforms Cellbox in Random Fold (RF) CV while the two methods have similar performance in LODO. **a** Cellbox performance on Random Fold CV **b** LR performance on Random Fold CV **c** Comparison of performance of Cellbox and LR on LODO validation

For LR, we set the coefficient for the left out drug to 0 and fit the unregularized estimator to the remaining columns of **D**. This approach does not require tuning parameter selection. Qualitatively this assumes that the drug held out has no effect on any of the response variables. This represents a crude benchmark model rather than an empirically motivated model assumption.

For each test set, we compute the correlation between the true responses and the predicted responses. This results in 12 correlations for Cellbox and Linear Regression. Figure 4c displays these correlations. LR (mean correlation coefficient 0.784) and Cellbox (mean correlation coefficient 0.780) have similar performance.

Discussion

The field of causal inference has historically focused on parameter estimation and hypothesis testing. Recently, several works have explored using causal models for prediction [1, 21, 29, 30]. Prediction performance is an important metric for measuring the quality of causal models in many scientific applications, including cell line perturbation experiments. For causal models to provide meaningful scientific insight, it is critical to understand their relationship with regression estimators and appropriately benchmark models when assessing performance.

Here we proposed a causal SEM for modeling responses in cell line perturbation experiments. We derived two estimators based on this model, LR and CSR. We derived analytic results demonstrating a close relationship between CSR and a recently proposed prediction model, Cellbox. The analytic results and simulations facilitated an improved understanding of the strengths and weaknesses of the two approaches to prediction. In brief, regression models, such as LR, are simpler to fit but lack an ability to extrapolate to new data settings, such as prediction of response to a drug not used in the training set. Causal modeling (CSR/Cellbox) is sensitive to the assumption of that the direct targets of the perturbations are known, making it most suitable for knockdown/knockout interventions with fewer off target effects than drug interventions.

Cellbox obtained state of the art performance on a Melanoma cell line perturbation data set, outperforming a Belief Propagation algorithm, a deep learning Neural Network (NN), and a co-expression model. Here we demonstrated that Cellbox, and hence all the competitor methods, failed to outperform LR in either RF or LODO validation. The latter finding is particularly surprising because this is a setting which favors causal modeling approaches. These results highlight that simple modeling strategies can be the most effective and are critically important when benchmarking performance of new models.

The Melanoma perturbation data set used here is relatively small, lacking any information on the temporal dynamics of responses to perturbations. Larger perturbation experiments test hundred or thousand of perturbations across dozens of cell lines with responses measured at multiple time points [2, 3, 31]. These data sets are likely to be more favorable to a model such as Cellbox, as they may contain sufficient information to identify and constrain model parameters. A recent generalization of Cellbox to simulated perturbations with responses measured across time showed promising performance [32].

Appendix A Proofs

A.1 Proof of Theorem 1

1. With the identity envelope $\phi(\cdot) = \cdot$ and $\tau_i = 1$, Equation (9) simplifies to

$$\frac{\partial x_i(t,\theta)}{\partial t} = Wx_i(t,\theta) + g(D_i). \tag{A1}$$

Now rewrite the ODE to explicitly include drug effects in the system as time constant terms. Define

$$y_i(t,\theta) = \begin{pmatrix} g(D_i) \\ x_i(t,\theta) \end{pmatrix} \in \mathbb{R}^{2p}$$

where

$$\frac{\partial y_i(t,\theta)}{\partial t} = \underbrace{\begin{pmatrix} 0 & 0 \\ I & W \end{pmatrix}}_{\equiv A \in \mathbb{R}^{2p\times 2p}} y_i(t,\theta) = \begin{pmatrix} 0 \\ Wx_i(t,\theta) + g(D_i) \end{pmatrix}. \tag{A2}$$

There is a closed form solution for the system at time t, specifically

$$y_i(t,\theta) = e^{At} y_i(0,\theta).$$

See [33] (Section 9.5 Theorem 2) for a derivation of this result. Further by Lemma 2, if W is invertible then,

$$y_i(\theta) \equiv \lim_{t\to\infty} y_i(t,\theta) = \begin{pmatrix} g(D_i) \\ -W^{-1}g(D_i) \end{pmatrix}.$$

Thus $x_i(\theta) = -W^{-1}g(D_i)$.

2. We have

$$\begin{aligned}
\widehat{W} &= \underset{W}{\text{argmin}} \sum_i \sum_j |X_{ij} - x_{ij}(\theta)|^2 + \lambda ||W - diag(W)||_1 \\
&= \underset{W}{\text{argmin}} \sum_i ||X_{ij} - x_i(\theta)||_2^2 + \lambda ||W - diag(W)||_1 \\
&= \underset{W}{\text{argmin}} \sum_i ||X_{ij} - (-W^{-1})g(D_i)||_2^2 + \lambda ||W - diag(W)||_1.
\end{aligned}$$

3. Recall that the Cellbox and CSR optimization problems are

$$\widehat{W} = \underbrace{\underset{\{W:W\prec 0\}}{\text{argmin}}}_{\equiv S_1} \sum_{i=1}^{n} \underbrace{||X_i - (-W^{-1})g(D_i)||_2^2 + \lambda ||W - diag(W)||_1}_{\equiv h_1(W)}$$

$$\widehat{A} = \underbrace{\underset{\{A:I-A\succ 0\}}{\text{argmin}}}_{\equiv S_2} \underbrace{\sum_{i=1}^{n} ||X_i - (I-A)^{-1}g(D_i)||_2^2 + \gamma(A,\lambda)}_{\equiv h_2(A)}$$

Note that $h_1(W) = h_2(I+W)$ and $W \in S_1 \Leftrightarrow I + W \in S_2$. Therefore $\widehat{A} = \widehat{W} + I$. Finally

$$\widehat{X}_C = -\widehat{W}^{-1}g(D^{te}) = (I - \widehat{A})^{-1}g(D^{te}) = \widehat{X}_{CSR}.$$

A.2 Proof of Lemmas

Lemma 2

$$\lim_{t\to\infty} e^{At} = \begin{pmatrix} I & 0 \\ -W^{-1} & 0 \end{pmatrix}.$$

Proof Note that

$$\begin{aligned} e^{At} &= I + \sum_{i=1}^{\infty} \frac{A^i t^i}{i!} \\ &= I + \sum_{i=1}^{\infty} \frac{\begin{pmatrix} 0 & 0 \\ W^{i-1} & W^i \end{pmatrix} t^i}{i!} \\ &= \begin{pmatrix} I & 0 \\ \sum_{i=1}^{\infty} \frac{W^{i-1}t^i}{i!} & \sum_{i=0}^{\infty} \frac{W^i t^i}{i!} \end{pmatrix} \\ &= \begin{pmatrix} I & 0 \\ W^{-1}(e^{Wt} - I) & e^{Wt} \end{pmatrix}. \end{aligned}$$

We now show that $\lim_{t\to\infty} e^{Wt} = 0$ which implies the desired result. Let c_j for $j = 1, \dots, p$ be an eigenbasis for W such that $Wc_j = \lambda_j c_j$. Note that since $W \prec 0$, $\lambda_j < 0$ for all j. Consider any $r \in \mathbb{R}^p$ with basis decomposition $r = \sum_j \gamma_j c_j$. We have

$$\begin{aligned} e^{Wt} r &= \sum_{j=1}^{p} \left(\sum_{i=0}^{\infty} \frac{W^i t^i}{i!} \right) \gamma_j c_j \\ &= \sum_{j=1}^{p} \left(\sum_{i=0}^{\infty} \frac{\lambda_j^i t^i}{i!} \right) \gamma_j c_j \\ &= \sum_{j=1}^{p} e^{\lambda_j t} \gamma_j c_j \\ &\to 0. \end{aligned}$$

Since this is true for any r, $\lim_{t\to\infty} e^{Wt} \to 0$. □

Acknowledgements

The authors gratefully acknowledge Jessica Swann for her scientific editing of the manuscript.

Author Contributions

J.P.L. and K.A.D. conceived the project. J.P.L., S.S., T.P., and K.A.D. developed methodology. J.P.L. and Y.Y. carried out simulations and application of methods to Melanoma perturbation data set. J.P.L. and Y.Y. wrote the manuscript. All authors reviewed the manuscript.

Funding

J.P.L was partially supported by NIH grants P30CA016672, CCTS UL1TR003167, P50CA140388, and P50CA127001. K.A.D. was partially supported by NIH grants P30CA016672, CCTS UL1TR003167, and P50CA140388.

Availability of data and materials

All data and code to reproduce results in this work are available at https://github.com/longjp/causal-pred-drug-code

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interest

None

Received: 20 February 2024 Accepted: 27 December 2024
Published online: 07 January 2025

References

1. Yuan B, Shen C, Luna A, Korkut A, Marks DS, Ingraham J, Sander C. Cellbox: interpretable machine learning for perturbation biology with application to the design of cancer combination therapy. Cell Syst. 2021;12(2):128–40.
2. Zhao W, Li J, Chen M-JM, Luo Y, Ju Z, Nesser NK, Johnson-Camacho K, Boniface CT, Lawrence Y, Pande NT, et al. Large-scale characterization of drug responses of clinically relevant proteins in cancer cell lines. Cancer Cell. 2020;38(6):829–43.
3. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, Gould J, Davis JF, Tubelli AA, Asiedu JK, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. Cell. 2017;171(6):1437–52.
4. Korkut A, Wang W, Demir E, Aksoy BA, Jing X, Molinelli EJ, Babur Ö, Bemis DL, Sumer SO, Solit DB, et al. Perturbation biology nominates upstream-downstream drug combinations in raf inhibitor resistant melanoma cells. Elife. 2015;4:04640.
5. Rampášek L, Hidru D, Smirnov P, Haibe-Kains B, Goldenberg A. Dr. Vae: improving drug response prediction via modeling of drug perturbation effects. Bioinformatics. 2019;35(19):3743–51.
6. Squires C, Shen D, Agarwal A, Shah D, Uhler C. Causal imputation via synthetic interventions. In: Conference on Causal Learning and Reasoning, p. 688–711 (2022). PMLR
7. Lotfollahi M, Wolf FA, Theis FJ. Scgen predicts single-cell perturbation responses. Nat Methods. 2019;16(8):715–21.
8. Lotfollahi M, Naghipourfar M, Theis FJ, Wolf FA. Conditional out-of-distribution generation for unpaired data using transfer vae. Bioinformatics. 2020;36(Supplement–2):610–7.
9. Ji Y, Lotfollahi M, Wolf FA, Theis FJ. Machine learning for perturbational single-cell omics. Cell Syst. 2021;12(6):522–37.
10. Roohani Y, Huang K, Leskovec J. Predicting transcriptional outcomes of novel multigene perturbations with gears. Nat Biotechnol, p. 1–9 (2023)
11. Ahlmann-Eltze C, Huber W, Anders S. Deep learning-based predictions of gene perturbation effects do not yet outperform simple linear methods. BioRxiv, 2024–09 (2024)
12. Tibes R, Qiu Y, Lu Y, Hennessy B, Andreeff M, Mills GB, Kornblau SM. Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. Mol Cancer Ther. 2006;5(10):2512–21.
13. Lotfollahi M, Klimovskaia Susmelj A, De Donno C, Hetzel L, Ji Y, Ibarra IL, Srivatsan SR, Naghipourfar M, Daza RM, Martin B, et al. Predicting cellular responses to complex perturbations in high-throughput screens. Mol Syst Biol. 2023;19(6):11517.
14. Huynh-Thu V.A, Sanguinetti G. Gene regulatory network inference: an introductory survey. Gene regulatory networks: Methods Protoc. 2019;1–23
15. Lachmann A, Giorgi FM, Lopez G, Califano A. Aracne-ap: gene network reverse engineering through adaptive partitioning inference of mutual information. Bioinformatics. 2016;32(14):2233–5.
16. Äijö T, Lähdesmäki H. Learning gene regulatory networks from gene expression measurements using non-parametric molecular kinetics. Bioinformatics. 2009;25(22):2937–44.
17. Hill SM, Lu Y, Molina J, Heiser LM, Spellman PT, Speed TP, Gray JW, Mills GB, Mukherjee S. Bayesian inference of signaling network topology in a cancer cell line. Bioinformatics. 2012;28(21):2804–10.
18. Michailidis G, d'Alché-Buc F. Autoregressive models for gene regulatory network inference: sparsity, stability and causality issues. Math Biosci. 2013;246(2):326–34.
19. Peters J, Bühlmann P, Meinshausen N. Causal inference by using invariant prediction: identification and confidence intervals. J Royal Stat Soc: Ser B (Stat Methodol). 2016;78(5):947–1012.
20. Rothenhäusler D, Bühlmann P, Meinshausen N, et al. Causal dantzig: fast inference in linear structural equation models with hidden variables under additive interventions. Annals Stat. 2019;47(3):1688–722.
21. Meinshausen N, Hauser A, Mooij JM, Peters J, Versteeg P, Bühlmann P. Methods for causal inference from gene perturbation experiments and validation. Proc Natl Acad Sci. 2016;113(27):7361–8.
22. Rothman K. Modern epidemiology. Lippincott Williams & Wilkins (2008)
23. Pearl J, et al. Causal inference in statistics: an overview. Stat Surv. 2009;3:96–146.
24. Shimizu S, Hoyer P.O, Hyvärinen A, Kerminen A, Jordan M. A linear non-gaussian acyclic model for causal discovery. J Mach Learn Res 2006;7(10)
25. Chen W, Drton M, Wang YS. On causal discovery with an equal-variance assumption. Biometrika. 2019;106(4):973–80.
26. Wang YS, Drton M. Causal discovery with unobserved confounding and non-gaussian data. J Mach Learn Res. 2023;24(271):1–61.
27. Dash D. Restructuring dynamic causal systems in equilibrium. In: International Workshop on Artificial Intelligence and Statistics, p. 81–88 (2005). PMLR
28. Lacerda G, Spirtes PL, Ramsey J, Hoyer PO. Discovering cyclic causal models by independent components analysis. arXiv preprint. 2012. arXiv:1206.3273
29. Versteeg P, Mooij JM. Boosting local causal discovery in high-dimensional expression data. In: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), p. 2599–2604 (2019). IEEE

30. Long JP, Ha MJ. Sample selection bias in evaluation of prediction performance of causal models. Stat Anal Data Min: ASA Data Sci J. 2022;15(1):5–14.
31. Peidli S, Green TD, Shen C, Gross T, Min J, Garda S, Yuan B, Schumacher LJ, Taylor-King JP, Marks DS, et al. scperturb: harmonized single-cell perturbation data. Nat Methods. 2024;21(3):531–40.
32. Ji W, Yuan B, Shen C, Regev A, Sander C, Deng S. Inference of cell dynamics on perturbation data using adjoint sensitivity. arXiv preprint. 2021. arXiv:2104.06467
33. Adkins WA, Davidson MG. Ordinary Differential Equations. New York: Springer; 2015

Publisher's Note