Causal Discovery as Semi-Supervised Learning

Chris J. Oates, Newcastle University,
Steven M. Hill, University of Cambridge,
Duncan A. Blythe, DZNE, Bonn,
& Sach Mukherjee, DZNE, Bonn.

Abstract

We frame causal discovery as a semi-supervised machine learning task. The idea is to allow direct learning of a causal graph by treating indicators of causal influence between variables as “labels”. Available data on the variables of interest are used to provide features for the labelling task. Background knowledge or any available interventional data provide labels on some edges in the graph and the remaining edges are treated as unlabelled. To illustrate the key ideas, we develop a distance-based approach (based on simple bivariate histograms) within a semi-supervised manifold regularization framework. We present empirical results on three different biological datasets (including data where causal effects can be verified by experimental intervention), which demonstrate the efficacy and highly general nature of the approach as well as its simplicity from a user’s point of view.

1 Introduction

Causal discovery is concerned with identifying causal relationships between variables. Consider a set of \( p \) variables indexed \( V = \{1 \ldots p\} \). The version of the causal discovery task we focus on here is to determine, for each pair \((i, j) \in V \times V\), whether or not \( i \) exerts a causal influence on \( j \). We focus on the binary “detection” problem (of determining whether or not \( i \) exerts a causal influence on \( j \)) rather than estimation of the magnitude of any causal effect.

Models with explicit causal semantics, such as (causal) directed acyclic graphs (DAGs) are a natural and popular choice for causal discovery [Pearl 2009; Maathuis et al. 2009; Spencer et al. 2015]. Such formulations specify a multivariate data-generating model with structure and parameters that encode causal relationships. In a different vein, there have been some interesting recent efforts in the direction of labelling pairs of variables as causal or otherwise, such as in Lopez-Paz et al. (2015) and Mooij et al. (2016). Our work is in this latter vein. Specifically, we frame causal discovery as a task in semi-supervised learning. We address essentially the same task as a traditional causal discovery – that of estimating a causal graph over a defined set of vertices – but via a “direct” machine learning approach that allows the inclusion of any available information concerning known cause-effect relationships.

In general terms the idea is as follows: Let \( \mathbf{X} \) denote the available data and \( \Phi \) denote any available knowledge on causal relationships among the variables indexed in \( V \) (e.g. based on background knowledge or experimental intervention). Then, the causal discovery task amounts to constructing an estimator of the form \( \hat{G}(\mathbf{X}, \Phi) \), where \( \hat{G} \) is a directed graph with vertex set \( V \) and edge set \( E(\hat{G}) \), with \((i, j) \in E(\hat{G})\) corresponding to the claim that variable \( i \) has a causal influence
on variable \(j\) (we make these notions more precise below). Let \(A(G)\) be the binary adjacency matrix corresponding to the unknown true causal graph \(G\). The idea is to treat the entries in \(A\) as binary “labels” in a machine learning sense (as in Lopez-Paz et al., 2015). From this point of view, the task of constructing the estimator \(\hat{G}(X, \Phi)\) is essentially one of learning the labels \(A_{ij}\) given data and some known labels (derived from \(\Phi\)). We treat this task within a semi-supervised learning framework.

Causal discovery is usually addressed either using explicitly causal approaches (e.g. following Pearl, 2009) or by constraining models using details of the underlying science (e.g. suitable non-linearities, as in Oates et al., 2014). In contrast, it may seem crude to attempt to directly learn the causal graph. However, we note that the claim that there exists an approach – based on any modeling framework – which allows causal discovery using available data and knowledge is equivalent to the claim that there exists an estimator of the form \(\hat{G}(X, \Phi)\) which performs well at the labelling task. If such an estimator exists, direct learning may be effective.

The remainder of the paper is organised as follows. We first introduce some notation and discuss in more detail how causal discovery can be viewed as a semi-supervised learning task. Our formulation is general and shows how essentially any existing semi-supervised approach can be adapted for causal discovery. To fix ideas, we then propose a specific approach, based on manifold regularization using a simple bivariate featurization. Using this specific approach – which we call Semi-Supervised Causal Discovery (SSCD) – we present empirical results using three biological datasets. The results cover a range of scenarios and include examples with explicitly interventional data that allow us to objectively assess the performance of SSCD.

2 Methods

2.1 Notation

Let \(V = \{1 \ldots p\}\) index a set of variables whose mutual causal relationships are of interest. The object of estimation is a directed graph \(G\) with vertex set \(V\) and edge set \(E\), with \((i, j) \in E\) indicating that variable \(i\) has a causal influence on \(j\). That is, \(G\) is the unknown true causal graph. In general, the graph \(G\) need not be acyclic. We restrict discussion to the case where a variable \(i\) either does or does not have such an influence on \(j\). The precise nature of causal influences to be considered will depend on the application but could include, for example, direct or indirect causes. Where useful, we use \(V(G), E(G)\) to denote the vertex and edge sets respectively of a graph \(G\).

Let \(A(G)\) be the binary adjacency matrix corresponding to graph \(G\). To make the mapping between causal discovery and machine learning more transparent, we introduce linear indexing by \(k\) of the pairs \((i, j)\). That is, \(k\) refers to a pair \((i, j) \in V \times V\). Where needed, we make the correspondence explicit, denoting by \((i(k), j(k))\) the variable pair corresponding to linear index \(k\) and by \(k(i, j)\) the linear index for pair \((i, j)\). Let \(Y_k = A_{i(k), j(k)}\) be a binary variable corresponding to an entry in the adjacency matrix \(A\). These \(Y_k\)'s are the labels or outputs to be learned. The available data are denoted \(X\). Depending on the specific problem \(X\) may comprise observational and/or interventional data. Available knowledge about causal relationships between the variables \(V\) is denoted \(\Phi\).
2.2 Causal Discovery as Semi-Supervised Learning

Thus, the task is to estimate the \( Y_k \)'s using \( X \) and \( \Phi \). This is done using a semi-supervised estimator \( \hat{Y}(X, \Phi) \) (we make the connection to semi-supervised learning explicit shortly). For now assume availability of such an estimator (we discuss one specific approach below). Then from \( \hat{Y} \) we have an estimate of the graph of interest as \( \hat{G}(X, \Phi) = (V, E(\hat{G}(X, \Phi))) \) (recall that the vertex set \( V \) is known) with the edge set specified via the semi-supervised learner as

\[
(i, j) \in E(\hat{G}(X, \Phi)) \iff \hat{Y}(X, \Phi)_{k(i,j)} = 1. \tag{1}
\]

Background knowledge \( \Phi \) could be based on relevant science or on available interventional data. For example, in a given scientific setting, certain cause-effect information may be known from previous work or theory. Alternatively, if some interventional data are available in the study at hand, this gives information on some cause-effect pairs. Whatever the source of the information, assume that it is known that certain pairs \( (i, j) \) are either causal pairs (positive information) or not causal pairs (negative information). Using the notation above, this amounts to knowing, for some pairs \( k \), the value of \( Y_k \). In the language of semi-supervised learning, the pairs whose causal status is known correspond to the labelled objects and the remaining pairs are the unlabelled objects.

For each pair \( k \), some of the data, or some transformation thereof will be used as predictors or inputs, denote these generically as \( g_k(X) \). That is, \( g_k \) is a featurization of the data, with the featurization specific to variables \( i(k), j(k) \). Let \( K \) be the set of linear indices (i.e. \( k \in K \) is a variable pair), \( L \subset K \) be the variable pairs with labels available (via \( \Phi \)) and \( U = K \setminus L \) be the set of unlabelled pairs. Let \( Y^L \) be a binary vector comprising the \( m_L = |L| \) available labels and \( Y^U \) be an unknown binary vector of length \( m_U = |U| \). The available labels are determined by the background information \( \Phi \) and we can write \( Y^L(\Phi) \) to make this explicit. A semi-supervised learner gives estimates for the unlabelled objects, given the data and available labels, i.e. an estimate of the form \( \hat{Y}^U(g(X), Y^L(\Phi)) \). With these in hand we have estimates for all labels and therefore for the edge set of the causal graph via (1).

Formulated in this way, it is clear that essentially any combination of featurization \( g \) and semi-supervised learner can be used to provide a causal discovery approach. Below, as a practical example, we explore graph-based manifold learning (following Belkin et al., 2006) combined with a simple bivariate featurization.

2.3 A Bivariate Featurization

For distance-based learning, we require a distance measure between objects (here, variable pairs) \( k, k' \in K \). The simplest candidate distance between variable pairs \( k, k' \) is based only on the bivariate distribution for the variables comprising the pairs (we make this notion precise below). Proofs of propositions appearing in this Section are provided in Appendix A.

2.3.1 Distance metrics between variable pairs

Let \( X \) denote the \( p \)-dimensional random variable whose \( n \) realizations comprise the rows of the data matrix \( X \). Assume \( X \in \mathcal{X}_p = [a, b]^p \) and that \( \mathcal{X}_p \) is endowed with the Borel \( \sigma \)-algebra \( \mathcal{B}_p = \mathcal{B}(\mathcal{X}_p) \). Let \( \mathcal{P} \) be the set of all probability measures on \( (\mathcal{X}_2, \mathcal{B}_2) \) that admit a twice continuously differentiable density function \( \pi = d\Pi/d\Lambda_2 \) with respect to Lebesgue measure \( \Lambda_2 \). Let \( \Pi_k \) be the bivariate (marginal) distribution for components \( i(k), j(k) \in V \) of \( X \).
Assumption 1. Each $\Pi_k$ admits a density function $\pi_k \in \mathcal{P}$.

If available, the distributions $\Pi_k, \Pi_{k'}$ could be used to define a distance between the pairs $k, k'$. Let $d_\mathcal{P} : \mathcal{P} \times \mathcal{P} \rightarrow [0, \infty)$ denote a pseudo-metric defined on the space $\mathcal{P}$. Since we do not have access to the underlying distributions, we construct an analogue using the available data $X$. Let $S_n := [a, b]^{2n}$ denote the space of possible bivariate samples (the sample size is $n$) and $S_k \in S_n$ denote the subset of the data for the variable pair $k$, i.e. $S_k = \{X_{l,i(k)}, X_{l,j(k)}\}_{l=1...n}$. Let $\kappa : S_n \rightarrow \mathcal{P}$ be a density estimator (DE). For notational simplicity we use $\kappa(S)$ to refer both to the probability distribution and (if it exists) the associated density; the intended meaning will be clear from context.

We consider sample quantities of the form $d_S = d_\mathcal{P} \circ (\kappa \times \kappa)$. That is, given data $S_k, S_{k'} \in S_n$ on two pairs $k, k'$, the DE is applied separately to produce density estimates $\kappa(S_k)$ and $\kappa(S_{k'})$, that are compared using $d_\mathcal{P}$ to give $d_S(S_k, S_{k'}) = d_\mathcal{P}(\kappa(S_k), \kappa(S_{k'}))$. This construction ensures that $d_S$ is a pseudo-metric without assumptions on the DE $\kappa$:

**Proposition 1.** Assume that $d_\mathcal{P}$ is a pseudo-metric on $\mathcal{P}$. Then $d_S$ is a pseudo-metric on $S_n$. If, in addition, $\kappa$ is injective and $d_\mathcal{P}$ is a metric on $\mathcal{P}$, then $d_S$ is a metric on $S_n$.

### 2.3.2 Choice of distance metric

For semi-supervised learning we need a metric under which causal pairs are relatively “close” to each other. The metric we consider is:

$$d_\mathcal{P}(\Pi_k, \Pi_{k'}) = \|\pi_k - \pi_{k'}\|_2 = \left(\int_{X_2} |\pi_k(x) - \pi_{k'}(x)|^2 \, d\Lambda_2(x)\right)^{1/2}.$$ 

The right hand side exists since the integrand is continuous on a compact set and thus bounded. This can be contrasted with the kernel embedding that was proposed for supervised causal discovery in [Lopez-Paz et al. (2015)].

**Proposition 2.** $d_\mathcal{P}$ is a metric on $\mathcal{P}$.

The main requirement that we have of the DE is that it provides consistent estimation in the $\| \cdot \|_2$ norm when $\Pi \in \mathcal{P}$. Specifically, consider a sequence $S^{(n)}$ in $S_n$ indexed by the number $n$ of data points. That is, $S^{(n)}$ is built from $n$ independent draws from $\Pi$ (the shorthand notation $S^{(n)} \sim \Pi$ will be used). Let $\pi$ be the density function for $\Pi$. Then $\kappa$ is said to be “consistent” if $\|\pi - \kappa(S^{(n)})\|_2 = o_P(1)$ holds for $S^{(n)} \sim \Pi$ and all fixed $\Pi \in \mathcal{P}$.

**Proposition 3.** Suppose $\kappa$ is consistent. Then, for $S^{(n)} \sim \Pi$, $\tilde{S}^{(n)} \sim \tilde{\Pi}$, we have that $d_S(S^{(n)}, \tilde{S}^{(n)}) = d_\mathcal{P}(\Pi, \tilde{\Pi}) + o_P(1)$.

Thus $d_S$ approximates the idealised metric $d_\mathcal{P}$ in the limit of draws from $\Pi$ and $\tilde{\Pi}$.

For the experiments in this paper, motivated by computational ease, we used a simple bivariate histogram as the DE $\kappa$. To this end, partition $X_2$ into an $m \times m$ regular grid whose $(i, j)$th element is

$$B_{i,j} = \left[ a + (b - a) \frac{i - 1}{m}, a + (b - a) \frac{i}{m} \right] \times \left[ a + (b - a) \frac{j - 1}{m}, a + (b - a) \frac{j}{m} \right].$$

Recall that a pseudo-metric $d$ satisfies all of the properties of a metric with the exception that $d(x, y) = 0 \implies x = y$. 

---

4
The standard bandwidth notation $h = m^{-1}$ will also be used. For a scatter plot $S \in \mathcal{S}_n$, let $b_{i,j}$ denote the number of elements that belong to the set $B_{i,j}$. Then the histogram estimator is

$$\kappa(S)(x) = \sum_{i,j=1}^{m} \frac{b_{i,j}}{n} \frac{1}{h^2} [x \in B_{i,j}].$$

This DE is consistent in the sense of Proposition 3. Indeed:

**Proposition 4.** Let the bandwidth parameter $h$ of the histogram estimator $\kappa$ be chosen such that $nh^2 \to \infty$. Then $\kappa$ is consistent. Moreover, an optimal choice of $h \asymp n^{-1/4}$ leads to $\|\pi - \kappa(S(n))\|_2 = O_P(n^{-1/4})$.

We note that this histogram DE is not rate optimal for the class $\mathcal{P}$ (for comparison, kernel DEs attain a rate of $O_P(n^{-2/3})$ over the same class $\mathcal{P}$ of twice continuously differentiable bivariate densities considered here, see [Wand and Jones, 1994]). However, an important advantage of the histogram DE is that the subsequent evaluation of $\kappa(S)$ is $O(1)$, compared with $O(n)$ for the kernel DE.

### 2.3.3 Implementation

The above arguments support the use of a bivariate histogram to provide a simple featurization for variable pairs. In practice, for all examples below, the data were standardised, then truncated to $[-3,3]^2$, following which a bivariate histogram with bins of fixed width 0.2 was used. The dimension of the resulting feature matrix was then reduced (to 100) using PCA.

### 2.4 Semi-Supervised Learning Using Manifold Regularization

Recall that the goal is to estimate binary labels $Y^U$ for a subset $U \subset K$ of variable pairs given available data $X$ and known labels $Y^L(\Phi)$ for a subset $L = K \setminus U$ (these are taken to be obtained from available interventional experiments and/or background knowledge). For any two pairs $k,k' \in K$, we also have available a distance $d_S(S_k, S_{k'})$. This is a task in semi-supervised learning (see e.g. [Belkin et al., 2006; Fergus et al., 2009]) and a number of formulations and methods could be used for estimation in this setting. Here we describe a specific approach in detail, using manifold regularization methods discussed in [Belkin et al., 2006].

Recall that $m_L = |L|$ is the number of available labels and $m_U = |U|$ the number of unlabelled pairs. Let $m = m_U + m_L (= |K|)$ be the total number of pairs. Using the distance function $d_S$ we first define a $m \times m$ similarity matrix $W$ with entries

$$W_{k,k'} = \exp \left(-\frac{1}{2\sigma^2} d_S(S_k, S_{k'})^2 \right).$$

Here $\sigma > 0$ must be specified and the exponential form is a heuristic only. We will use a partition of the matrix corresponding to the sets $U, L$ as follows

$$W = \begin{bmatrix} W_{LL} & W_{LU} \\ W_{UL} & W_{UU} \end{bmatrix}$$
where we have assumed, without loss of generality, that the variable pairs are ordered so that the labelled pairs appear in the first \(m_L\) places, followed by the \(m_U = m - m_L\) unlabelled pairs. Correspondingly let
\[
y = \begin{bmatrix} y^L \\ y^U \end{bmatrix} \in \{0, 1\}^{m \times 2}
\]
denote a label matrix, where ones appear in the second column for those pairs \(k\) for which \(Y_k = 1\) (i.e. the columns correspond to “non-causal” and “causal” and the rows to variable pairs). The matrix \(y^U\) is unknown and is the object of estimation.

Let \(D\) be the \(m \times m\) diagonal matrix with \(k\)th diagonal entry \(\sum_{k' \in K} W_{k,k'} W_{k,k'}\). Define \(\tilde{L} = D - \frac{W}{2}\) and \(L = D - \frac{1}{2} \tilde{L} D^{-1/2}\) (i.e. the normalised graph Laplacian; all matrices with \(O(m^2)\) entries are denoted as bold capitals to emphasise the potential bottleneck that is associated with storage and manipulation of these matrices). Let
\[
f = \begin{bmatrix} f^L \\ f^U \end{bmatrix} \in \mathbb{R}^{m \times 2}
\]
be a matrix corresponding to a classification function \(f : S \rightarrow \mathbb{R}^2\) evaluated at the \(m\) variable pairs \(K\), with the superscripts indicating correspondence with the labelled and unlabelled pairs. We would like to an estimate \(\hat{f}\) to agree with the known labels \(y^L\) and also to take account of the manifold structure encoded in \(L\). Following [Belkin et al. (2006)] we consider an objective function
\[
J(f, y) = \frac{1}{m_L} \text{tr}\{f^L y^L\} + \frac{\lambda}{m^2} \text{tr}\{f^T L f\}
\]
where the first term relates the known labels to the classification \(f\) and the second term imposes “smoothness” on the label assignment in the sense of encouraging solutions where the labels do not change quickly with respect to the distance metric (see [Belkin et al. (2006)] for full details). The tuning parameter \(\lambda\) may be set by cross-validation. We considered a coarse grid \(\{0.001, 0.01, 0.1, 1.0\}\) but found results to be insensitive to the choice of \(\lambda\), hence for all results below we simply set \(\lambda = 0.001\).

Given training labels \(y^L\), label estimates are obtained by minimising the objective function described above, as described in [Belkin et al. (2006)]. This gives
\[
(\hat{f}, \hat{y}^U) = \arg\min_{f, y^U} J \left( f, \begin{bmatrix} y^L \\ y^U \end{bmatrix} \right)
\]
where as above \(f \in \mathbb{R}^{n \times 2}\) and \(y^U \in \{0, 1\}^{m_U \times 2}\). Here \(\hat{y}^U\) provides a point estimate for the unknown labels while \(\hat{f}^U\) is real-valued and can be used to rank candidate pairs. The optimization problem in Eqn. 4 admits a closed-form solution (see Sec 4.2 of [Belkin et al. (2006)] for details) and this was used here. The computational cost is \(O(m^3)\). Computation for large-scale semi-supervised learning has been studied in the literature (see e.g. [Fergus et al. 2009]) and a number of approaches could be used to scale up to larger problems.

### 3 Empirical Results

We tested our approach using three datasets with different characteristics. The key features of each dataset are outlined below, with a full description of each dataset appearing in the respective subsection. In all cases the goal was to predict the effect of an unseen intervention.
• **D1: Yeast knockout data.** Here, we used a dataset due to Kemmeren et al. (2014), previously considered for causal discovery in Peters et al. (2016); Meinshausen et al. (2016). The data consist of a large number of gene deletion experiments with corresponding gene expression measurements.

• **D2: Kinase intervention data from human cancer cell lines.** These data, due to Hill et al. (2017), involve a small number of interventions on human cells, with corresponding protein measurements over time.

• **D3: Protein data from cancer patient samples.** These data arise from The Cancer Genome Atlas (TCGA) and are presented in Akbani et al. (2014). There are no interventional data, but the data pertain to relatively well-understood biological processes.

An appealing feature of our approach is the simplicity with which it can be applied to diverse problems. In each case below, we simply concatenate available data to form the matrix $X$ and available knowledge/interventions to form $\Phi$, then directly apply the methods as described.

### 3.1 General Problem Set-Up

The basic idea in all three problems was the same: given data on a set of variables, for each (ordered) pair $(i,j)$ of variables we sought to determine whether or not $i$ has a causal effect on $j$. In the case of datasets D1 and D2 the results were directly assessed against the outcome of experiments involving explicit interventions. The availability of a large number of interventions in D1 allowed a wider range of experiments, whereas D2 is a much smaller dataset (but from human cells), allowing only a relatively limited assessment. In the case of D3, where direct interventional data (i.e. interventions on the same biological material that give rise to the training data) were not available but the relevant biological mechanisms are relatively well-understood, we compared results to a reference causal graph derived from the biochemical literature (the literature itself is in effect an encoding of extensive interventional experiments combined with biochemical knowledge).

Within the semi-supervised set-up, a subset of pairs were labelled at the outset and the remaining pairs were unlabelled. All empirical results below are for unlabelled pairs, i.e. in all cases the prediction of a causal effect or otherwise was unseen in the sense that the relevant interventional data on the pair of interest (or relevant entry in the reference causal graph) was not available to the estimator.

Note that where interventional data were available (D1 and D2), although the data allow us to read off causal effects, any such effects may be indirect: if variable $j$ changes under intervention on variable $i$ we can conclude that $i$ has a causal effect on $j$. However the effect might be mediated by other variables that may or may not be measured. For D3, the reference causal graph encodes causal effects that may be indirect, but are at least direct relative to the set of measured variables: an edge from variable $i$ to variable $j$ in the reference causal graph indicates that $i$ has a causal effect on $j$, and this effect is not mediated by other measured variables, but may be mediated by unmeasured variables.

### 3.2 Dataset D1: Yeast Gene Expression

**Data.** The data consisted of gene expression measurements for a total of $p_{\text{total}} = 6170$ genes. Some of the data samples were measurements after knocking out a specific gene (interventional data) and
the other samples were without any such intervention (observational data), with sample sizes of \(n_{\text{int}} = 1479\) and \(n_{\text{obs}} = 160\) respectively. Each of the genes intervened on was one of the \(p_{\text{total}}\) genes. Let \(t(l)\) be the index of the gene targeted by the \(l\)th intervention. That is, the \(l\)th interventional sample was an experiment in which gene \(t(l)\) was knocked out. Let \(T = \{t(1), \ldots, t(n_{\text{int}})\}\) be the subset of genes that were the target of an interventional experiment.

**Problem set-up.** Our problem set-up was as follows. We sampled a subset \(C \subset T\) of the genes that were intervened upon, with \(|C| = 50\), and treated this as the vertex set of interest (i.e. setting \(V = C\) and \(p = |C| = 50\)). The goal was to uncover the mutual causal relationships between these \(p\) variables.

Since by design interventional data were available for all variables \(j \in C\), we used these data to define an interventional “gold standard”. To this end we used a robust z-score that considered the change in a variable of interest under intervention, relative to its observational variation. Let \(X_{ij}^{\text{int}}\) denote the measured expression level of gene \(j\) following intervention on gene \(i\). For any pair of genes \(i, j \in C\) we say that \(i\) has a causal effect on \(j\) if and only if \(Z_{ij} = |X_{ij}^{\text{int}} - M_j^{\text{obs}}|/IQR_j^{\text{obs}} > \tau\), where \(M_j^{\text{obs}}\) is the median level of gene \(j\) (calculated using half of the observational data samples; the remaining samples were used as training data, see below), \(IQR_j^{\text{obs}}\) the corresponding interquartile range and \(\tau = 5\) was a fixed threshold. In other words, if \(G\) is the true graph of indirect causal effects between the genes in \(C\), and \(A_G\) the corresponding binary adjacency matrix, we have \(A_G(i,j) = 1 \iff Z_{ij} > \tau\). To rule out absence of causal effects we required that \(A_G(i,j) = 1\) for at least 2.5% of gene pairs \(i, j \in C, i \neq j\).

Given data on genes \(C\) and the corresponding ground-truth matrix \(A_G\), we set up the learning problem as follows. We treated a fraction \(\rho\) of the entries in \(A_G\) as the available labels \(\Phi\) and then, using these labels and data on the variables \(C\), we performed semi-supervised learning as described. This gave estimates for the remaining (unseen) entries in \(A_G\), which we compared against the corresponding true values. The data matrix \(X\) comprised expression measurements for \(n_{\text{obs}} = 80\) observational data samples (those samples not used to calculate the robust z-scores), plus \(n_{\text{int}}\) interventional data samples where genes outside the set of interest were intervened upon, i.e. a subset of the 1429 genes in \(T \setminus C\). This set-up ensured that \(X\) include neither any of the interventional nor observational data that was used to obtain the ground-truth matrix \(A_G\). The total amount of training data is denoted by \(n_{\text{train}} = n_{\text{obs}} + n_{\text{int}}\). We considered \(n_{\text{train}} = 200, 500\) and \(1000\) (corresponding to \(n_{\text{int}} = 120, 420\) and \(920\) respectively, sampled at random).

**Results.** We compared the proposed semi-supervised learning-based approach (SSCD) with the following methods to provide a baseline comparison (we note that none of the benchmark methods exploited the background knowledge \(\Phi\) and in that sense should not be regarded as direct competitors):

- **Pearson** and **Kendall** correlation coefficients. A correlation coefficient was computed for each pair of variables \(i, j \in C, i \neq j\) and subsequently thresholded.

- Penalized regression with an \(L_1\) penalty ([**Lasso**](https://en.wikipedia.org/wiki/Lasso); [Tibshirani, 1996](https://doi.org/10.1007/s11222-007-9058-7)). Each variable \(j \in C\) was regressed on all other variables \(i \in C, i \neq j\) to obtain regression coefficients.

- Intervention-calculus when the DAG is absent ([**IDA**](https://en.wikipedia.org/wiki/Intervention-calculus); [Maathuis et al., 2009, 2010](https://www.nature.com/articles/nature07739)). A lower bound for the total causal effect of variable \(i\) on variable \(j\) was estimated for each pair \(i, j \in C, i \neq j\).
Figure 1: Results for dataset D1 (yeast data), random sampling of causal labels. Area under the ROC curve (AUC), with respect to held-out causal influences (determined from interventional data), as a function of the fraction $\rho$ of labels available in the semi-supervised set-up (labels were sampled at random). Results are shown for three training data sample sizes $n_{\text{train}}$ and for five methods: the proposed semi-supervised approach, SSCD; Pearson and Kendall correlations; Lasso regression; and intervention-calculus when the DAG is absent (IDA). Results are mean values over 25 iterations and error bars indicate standard error of the mean. See text for details.

All the methods result in a score $s_{ij}$ for all pairs $i, j \in C, i \neq j$ (i.e. correlation or regression coefficients, total causal effects, or, for SSCD, the real-valued $\hat{f}$ in (4)). For each method, these scores were thresholded and pairs $(i, j)$ whose absolute values of the score fell above the threshold were labelled as “causal”. Varying the threshold and calculating true positives and false positives with respect to the binary unseen entries in the matrix $A_C(G)$ resulted in an ROC curve.

Figure 1 shows the area under the ROC curve (AUC) as a function of the proportion $\rho$ of entries in $A_C(G)$ that were observed, for the three sample sizes. Results were averaged over 25 iterations. SSCD had a marked improvement in performance over the benchmark methods for all considered values of $n_{\text{train}}$ and $\rho$. As expected, SSCD performed better as a larger proportion of entries in $A_C(G)$ were observed (i.e. for larger $\rho$). For the benchmark methods, any variation in performance with $\rho$ was solely due to the changing test set (these methods do not use the background knowledge $\Phi$). The error bars tended to become larger with $\rho$ due to the decreasing size of the test set. All methods had a similar performance for the three sample sizes, with some modest increases in AUC with increasing $n_{\text{train}}$.

In the above results the pairs whose causal relationship was to be predicted were chosen at random (i.e. the set of unlabelled pairs was a random subset of the set of all pairs). In contrast, in some settings it may be relevant to predict the effect of intervening on variable $i$, without knowing the effect of intervening on $i$ on any other variable. For this setting, the unlabelled set should
Figure 2: Results for dataset D1 (yeast data), row-wise sampling of causal labels. As Figure 1, except the subset of labels available in the semi-supervised set-up were obtained by sampling entire rows of the causal adjacency matrix. As before, a proportion \( \rho \) were sampled. The remaining rows were then used as test data.

comprise entire rows of the causal adjacency matrix \( A_C(G) \). Figure 2 considers this case. To ensure a sufficient number of rows were non-empty, we imposed the additional restriction on the gene subset \( C \) that at least half of the rows had at least one causal effect. This row-wise case is a harder problem, reflected in the lower AUC values for SSCD compared with Figure 1, but we still found that SSCD outperformed the benchmark methods.

## 3.3 Dataset D2: Protein Time-Course Data

**Data.** The data consisted of protein measurements for \( p = 35 \) proteins measured at seven time points in four different “cell lines” (BT20, BT549, MCF7 and UACC812; these are laboratory models of human cancer) and under eight growth conditions. The proteins under study act as kinases (i.e. catalysts for a biochemical process known as phosphorylation) and interventions were carried out using kinase inhibitors that block the kinase activity of specific proteins. A total of four intervention regimes were considered, plus a control regime with no interventions. The data used here were a subset of the complete dataset reported in detail in Hill et al. (2017) and were also previously used in a Dialogue for Reverse Engineering Assessments and Methods (DREAM) challenge on learning causal networks (Hill et al., 2016).

**Problem set-up.** Treating each cell line as a separate, independent problem, the intervention regimes were used to define an interventional “gold standard”, in a similar vein as for Dataset D1. This followed the procedure described in detail in Hill et al. (2016) with an additional step of taking a majority vote across growth conditions to give a causal gold standard for each cell line.
Figure 3: Results for dataset D2 (protein time course data). Each panel is a different cell line, with its own training and (interventional) test data. AUC is with respect to an entirely held-out intervention. See text for details.

c. For each cell line $c$, we formed a data matrix $X_c$ consisting of all available data for the $p = 35$ proteins except for one of the intervention regimes. The intervention regime not included was a kinase inhibitor targeting the protein mTOR. This intervention was entirely held out and used to provide the test labels. As background knowledge $\Phi_c$ we took as training labels causal effects under the other interventions. With this set-up, the task was to determine the (indirect) causal effects of the entirely unseen intervention. Note that each cell line $c$ was treated as an entirely different dataset and task, with its own data matrix, background knowledge and interventional test data.

Results. Figure 3 shows AUCs (with respect to changes seen under the test intervention) for each of the four cell lines and each of the methods. There was no single method that outperformed all others across all four cell lines. SSCD performed particularly well relative to the other methods for cell lines BT549 and MCF7, was competitive for cell line UACC812, but performed less well for cell line BT20. We note also that, for cell lines BT549 and MCF7, the performance of SSCD was competitive with the best performers in the DREAM challenge and with an analysis reported in Hill et al. (2017). The latter involved a complex Bayesian model specifically designed for such data. In contrast, SSCD was applied directly to a data matrix comprising all training samples simply collected together.

3.4 Dataset D3: Human Cancer Data

Data. The data consisted of protein measurements for $p = 35$ proteins measured in $n = 820$ human breast cancer samples (from biopsies). The data originate from The Cancer Genome Atlas (TCGA) Project, are described in Akbani et al. (2014) and were retrieved from The Cancer Proteome Atlas (TCPA) data portal (Li et al., 2013; http://tcpaportal.org; data release version 4.0; Pan-Can 19
Figure 4: Results for dataset D3 (human cancer data). Data are protein measurements from breast cancer patient samples from The Cancer Genome Atlas (TCGA). AUC is with respect to a reference causal graph based on the (causal) biochemical literature. Results are mean values over 25 iterations and error bars indicate standard error of the mean. See text for details.

Level 4 data). Data for many cancer types are available, but here we focus on a single type (breast cancer) to minimize the potential for confounding by cancer type. It is at present difficult to carry out interventions in biopsy data of this kind. However, we focused on the same 35 proteins as in Dataset D2, whose mutual causal relationships are relatively well-understood, and used a reference causal graph for these proteins based on the biochemical literature (as reported in Hill et al., 2017).

Problem set-up. We formed data matrices $X$ consisting of measurements for the $p = 35$ proteins for three different sample sizes: (i) $n_{\text{train}} = 200$, (ii) $n_{\text{train}} = 500$ or (iii) all $n_{\text{train}} = 820$ patient samples. For (i) and (ii) patient samples were selected at random. We then used a random fraction $\rho$ of the reference graph as background knowledge, testing output on the (unseen) remainder.

Results. Figure 4 shows AUCs (with respect to the held-out causal labels) as a function of the proportion $\rho$ of causal edges that were observed, for each of the methods and for the three sample sizes. Results were averaged over 25 iterations. SSCD outperformed the other methods, with the magnitude of the gain in performance increasing with $\rho$, and wither performance still competitive for $\rho = 0.2$. Results were qualitatively similar for the three sample sizes, with increases in AUC for $n_{\text{train}} = 820$ and $n_{\text{train}} = 500$ relative to $n_{\text{train}} = 250$, particularly for SSCD. As in Figures 1 and 2, the error bars become larger as $\rho$ increases due to the size of the test set decreasing.

4 Discussion

In this paper, we proposed a way in which causal discovery can be framed as a task in semi-supervised learning. The approach we propose stands in contrast to more explicitly causal methods
in that it learns potentially relevant regularities directly from data. The main advantage of this approach is that it allows regularities in the data to emerge via learning, rather than having to be encoded via an explicit causal or mechanistic model.

In the examples we considered, the gold-standard causal relationships could be indirect. We conjecture that with suitable inputs our framework could be used to learn direct relationships. However, exploring the learning of direct causal relationships using real experimental data is challenging, since such data typically contain readouts after interventions on individual vertices (as in the yeast data). Such data are not in themselves sufficient to provide clear evidence of direct effects, since it may be that the true causal path includes an unknown and unmeasured variable.

We do not expect our approach – or indeed any causal discovery method – to be everywhere and always the best performer. Causal discovery remains an open area and at this stage we think it may be fruitful to revisit some causal inference problems within a machine learning context.

Appendix A

In this appendix we provide proofs for the theoretical results in the main text.

**Proposition 1.** Given \( S, S' \in \mathcal{S}_n \), we obtain from the pseudo-metric properties of \( d_\mathcal{P} \) each of (i) positivity \( d_\mathcal{S}(S, S') = d_\mathcal{P}(\kappa(S), \kappa(S')) \geq 0 \), (ii) pseudo-identity \( d_\mathcal{S}(S, S) = d_\mathcal{P}(\kappa(S), \kappa(S)) = 0 \), (iii) symmetry \( d_\mathcal{S}(S, S') = d_\mathcal{P}(\kappa(S), \kappa(S')) = d_\mathcal{P}(\kappa(S'), \kappa(S)) = d_\mathcal{S}(S', S) \) and (iv) the triangle inequality \( d_\mathcal{S}(S, S'') = d_\mathcal{P}(\kappa(S), \kappa(S'')) \leq d_\mathcal{P}(\kappa(S), \kappa(S')) + d_\mathcal{P}(\kappa(S'), \kappa(S'')) = d_\mathcal{S}(S, S') + d_\mathcal{S}(S', S'') \), for all \( S, S', S'' \in \mathcal{S}_n \).

Suppose now that \( \kappa \) is injective and \( d_\mathcal{P} \) is a metric on \( \mathcal{P} \). Then if \( d_\mathcal{S}(S, S') = d_\mathcal{P}(\kappa(S), \kappa(S')) \) = 0, it follows that \( \kappa(S) = \kappa(S') \) which (from assumption on \( \kappa \)) implies \( S = S' \) in \( \mathcal{S}_n \). Thus under these additional assumptions, \( d_\mathcal{S} \) is a metric on \( \mathcal{S}_n \).

**Proposition 2.** The non-negativity, symmetry and sub-additivity properties of \( d_\mathcal{P} \) are clear, so all that remains is to establish that \( d_\mathcal{P}(\Pi, \Pi') = 0 \) implies \( \Pi = \Pi' \). Let \( \Pi, \Pi' \) have densities \( \pi, \pi' \). From the definition of \( \mathcal{P} \), both \( \pi \) and \( \pi' \) are continuous on \( \mathcal{X}_2 \). The result is then immediate from the fact that, since \( \pi \) and \( \pi' \) are continuous and \( \mathcal{X}_2 \) is compact, then \( \int_{\mathcal{X}_2} |\pi(x) - \pi'(x)|^2 \, d\Lambda_2(x) = 0 \) implies \( \pi \) and \( \pi' \) must be identical as functions on \( \mathcal{X}_2 \). Thus \( \Pi = \Pi' \) as required.

**Proposition 3.** Observe that, using Prop. 2 for sub-additivity of the metric \( d_\mathcal{P} \),

\[
d_\mathcal{S}(S^{(n)}, \tilde{S}^{(n)}) = d_\mathcal{P}(\kappa(S^{(n)}), \kappa(\tilde{S}^{(n)})) \\
\leq d_\mathcal{P}(\kappa(S^{(n)}), \Pi) + d_\mathcal{P}(\Pi, \Pi') + d_\mathcal{P}(\Pi', \kappa(\tilde{S}^{(n)})) \\
= d_\mathcal{P}(\Pi, \Pi') + \|\pi - \kappa(S^{(n)})\|_2 + \|\tilde{\pi} - \kappa(\tilde{S}^{(n)})\|_2
\]

Since \( \kappa \) is consistent we have \( \|\pi - \kappa(S^{(n)})\|_2 = o_P(1) \) and \( \|\tilde{\pi} - \kappa(\tilde{S}^{(n)})\|_2 = o_P(1) \). This completes the proof.

**Proposition 4.** This proof extends the simpler proof given for the univariate case in Theorem 6.11 of Wasserman (2006). For convenience, and without loss of generality, we suppose that \( \mathcal{X}_2 = [0, 1]^2 \).
Let
\[ p_{i,j} = \int_{B_{i,j}} \pi \, d\Lambda \]
be the probability mass assigned to \( B_{i,j} \), so that, from binomial properties, the mean and variance of the histogram estimator \( \kappa(S^{(n)}(x)) \) at the point \( x \in \mathcal{X}^2 \) are
\[
m(x) = \frac{p_{i,j}}{h^2} \]
\[
v(x) = \frac{p_{i,j}(1 - p_{i,j})}{nh^4}.
\]
Let \( b(x) = m(x) - \pi(x) \) denote the bias of the histogram estimator. The mean square of the error \( \pi(x) - \kappa(S^{(n)}(x)) \) at a point \( x \in \mathcal{X}^2 \) can be bias-variance decomposed:
\[
\mathbb{E}\{[\pi(x) - \kappa(S^{(n)}(x)) ]^2\} = b(x)^2 + v(x)
\]
The aim is to obtain independent bounds on both the bias and variance terms next.

To bound the bias term, Taylor’s theorem gives that, for \( x, x' \in B_{i,j} \),
\[
\pi(x') = \pi(x) + (x' - x)^\top \cdot \nabla \pi(x) + \frac{1}{2} (x' - x)^\top \, R_{i,j}(x) \, (x' - x)
\]  
(5)
where the remainder term satisfies
\[
\| \| R_{i,j}(x) \|_{\text{max}} \leq \sup_{x'' \in B_{i,j}} \| \nabla^2 \nabla^\top \pi(x'') \|_{\text{max}} (\text{Taylor})
\]
\[
\leq \sup_{x'' \in \mathcal{X}^2} \| \nabla^2 \nabla^\top \pi(x'') \|_{\text{max}} < \infty \quad \text{(continuous on compact domain)}.
\]

Here \( \| M \|_{\text{max}} = \max\{M_{i,j}\} \) and \( \nabla^2 \nabla^\top \pi \) denotes the Hessian, which exists since \( \pi \) is twice continuously differentiable in \( \mathcal{X}^2 \). Thus for \( x \in B_{i,j} \), integrating Eqn. 5,
\[
\int_{B_{i,j}} \pi(x') \, d\Lambda^2(x') = h^2 \pi(x) + h^2 \left( \frac{h}{2} \left[ \frac{2i - 1}{2j - 1} \right] - x \right) \cdot \nabla \pi(x) + E_{i,j}(x)
\]
where the new remainder term can be bounded:
\[
| E_{i,j}(x) | = \left| \frac{1}{2} \int_{B_{i,j}} (x' - x)^\top \, R_{i,j}(x) \, (x' - x) \, d\Lambda_2(x') \right|
\]
\[
\leq \frac{1}{2} \int_{B_{i,j}} \| x' - x \|^2 \, d\Lambda_2(x') \times \sup_{x'' \in \mathcal{X}^2} \| \nabla^2 \nabla^\top \pi(x'') \|_{\text{max}}
\]
\[
\leq 8h^4 \sup_{x'' \in \mathcal{X}^2} \| \nabla^2 \nabla^\top \pi(x'') \|_{\text{max}} =: C h^4
\]  
(6)
where the constant \( C \) is independent of \( x \) and \( i, j \). The number 8 (which is not sharp) is obtained from trivial but tedious computation of the integral in Eqn. 6 and bounding each term in the result. Now, for \( x \in B_{i,j} \), the bias is expressed using Eqn. 5 as
\[
b(x) = \frac{1}{h^2} \int_{B_{i,j}} \pi(x') \, d\Lambda_2(x') - \pi(x)
\]
\[
= \left( \frac{h}{2} \left[ \frac{2i - 1}{2j - 1} \right] - x \right) \cdot \nabla \pi(x) + \frac{1}{h^2} E_{i,j}(x).
\]  
14
Now we integrate this expression over \( x \in B_{i,j} \):

\[
\int_{B_{i,j}} b^2 d\Lambda_2 = \int_{B_{i,j}} \left\{ \left( \frac{h}{2} \left\lfloor \frac{2i-1}{2j-1} \right\rfloor - x \right) \cdot \nabla \pi(x) + \frac{1}{h^2} E_{i,j}(x) \right\}^2 d\Lambda_2(x)
\]

\[
\leq \int_{B_{i,j}} \left\{ \left( \frac{h}{2} \left\lfloor \frac{2i-1}{2j-1} \right\rfloor - x \right) \cdot \nabla \pi(x) \right\}^2 d\Lambda_2(x)
\]

\[
+ 2\int_{B_{i,j}} \left| \left( \frac{h}{2} \left\lfloor \frac{2i-1}{2j-1} \right\rfloor - x \right) \cdot \nabla \pi(x) \right| \frac{1}{h^2} |E_{i,j}(x)| d\Lambda_2(x)
\]

\[
+ \int_{B_{i,j}} \frac{1}{h^4} E_{i,j}(x)^2 d\Lambda_2(x)
\]

\[
\leq \int_{B_{i,j}} \left\{ \left( \frac{h}{2} \left\lfloor \frac{2i-1}{2j-1} \right\rfloor - x \right) \cdot \nabla \pi(x) \right\}^2 d\Lambda_2(x)
\]

\[
+ 2Ch^2 \int_{B_{i,j}} \left| \left( \frac{h}{2} \left\lfloor \frac{2i-1}{2j-1} \right\rfloor - x \right) \cdot \nabla \pi(x) \right| d\Lambda_2(x) + C^2 h^2
\]

To bound these integrals we use Cauchy-Schwarz:

\[
\int_{B_{i,j}} \left\{ \left( \frac{h}{2} \left\lfloor \frac{2i-1}{2j-1} \right\rfloor - x \right) \cdot \nabla \pi(x) \right\}^2 d\Lambda_2(x) \leq \int_{B_{i,j}} \left\| \frac{h}{2} \left\lfloor \frac{2i-1}{2j-1} \right\rfloor - x \right\|^2_2 \| \nabla \pi(x) \|^2_2 d\Lambda_2(x)
\]

\[
\leq \frac{h^2}{2} \int_{B_{i,j}} \| \nabla \pi(x) \|^2_2 d\Lambda_2(x)
\]  \( (7) \)

and

\[
\int_{B_{i,j}} \left| \left( \frac{h}{2} \left\lfloor \frac{2i-1}{2j-1} \right\rfloor - x \right) \cdot \nabla \pi(x) \right| d\Lambda_2(x) \leq \int_{B_{i,j}} \left\| \frac{h}{2} \left\lfloor \frac{2i-1}{2j-1} \right\rfloor - x \right\|_2 \| \nabla \pi(x) \|_2 d\Lambda_2(x)
\]

\[
\leq \frac{h}{\sqrt{2}} \int_{B_{i,j}} \| \nabla \pi(x) \|_2 d\Lambda_2(x).
\]  \( (8) \)

Both expressions in Eqns. \( 7 \) and \( 8 \) are finite since the integrand is continuous and the domain is compact. The total integrated bias is thus bounded as

\[
\int_{\mathcal{X}_2} b^2 d\Lambda_2 \leq \frac{h^2}{2} \int_{\mathcal{X}_2} \| \nabla \pi(x) \|^2_2 d\Lambda_2(x) + C^2 h^2 + O(h^3)
\]

To bound the variance term, from the integral form of the mean value theorem we have that, for some \( x_{i,j} \in B_{i,j} \),

\[
p_{i,j} = \int_{B_{i,j}} \pi d\Lambda_2 = h^2 \pi(x_{i,j}).
\]

The application of the integral form of the mean value theorem is valid since \( \pi \) is continuous on
\[ \mathcal{X}_2 \text{. Then:} \]

\[
\int_{\mathcal{X}_2} v^2 d\Lambda_2 = \sum_{i,j=1}^m \int_{B_{i,j}} v d\Lambda_2 \\
= \sum_{i,j=1}^m \int_{B_{i,j}} p_{i,j} \left(\frac{1 - p_{i,j}}{nh^4}\right) d\Lambda_2 \\
= \frac{1}{nh^2} - \frac{1}{nh^2} \sum_{i,j=1}^m p_{i,j}^2 \\
= \frac{1}{nh^2} - \frac{1}{n} \left( \int_{\mathcal{X}_2} \pi^2 d\Lambda_2 + o(1) \right) = \frac{1}{nh^2} + O \left( \frac{1}{n} \right)
\]

Putting this all together to obtain a bound:

\[
\mathbb{E} \| \pi - \kappa(S^{(n)}) \|_2^2 = \int_{\mathcal{X}_2} b^2 d\Lambda_2 + \int_{\mathcal{X}_2} v d\Lambda_2 \quad \text{(Fubini)} \\
\leq \frac{h^2}{2} \int_{\mathcal{X}_2} \|
abla \pi(x)\|_2^2 d\Lambda_2(x) + C^2 h^2 + O(h^3) + \frac{1}{nh^2} + O \left( \frac{1}{n} \right) \tag{9}
\]

where \( \mathbb{E} \) denotes expectation with respect to sampling of the data \( S^{(n)} \sim \Pi \). From inspection of Eqn. \[9\] the estimator error vanishes provided that \( h \) is chosen such that \( nh^2 \to \infty \). Since convergence in expectation implies convergence in probability, we have established that \( \| \pi - \kappa(S^{(n)}) \|_2 = o_P(1) \).

The bandwidth \( h^* \), which minimises the upper bound in Eqn. \[9\] is

\[
h^* = \frac{1}{n^{1/4}} \left( \int_{\mathcal{X}_2} \|
abla \pi(x)\|_2^2 d\Lambda_2(x) + 2C^2 \right)^{1/4}
\]

and with this choice we have that \( \mathbb{E} \| \pi - \kappa(S^{(n)}) \|_2^2 = O_P(n^{-1/2}) \). For \( h = h^* \) we have thus established that \( \| \pi - \kappa(S^{(n)}) \|_2 = O_P(n^{-1/4}) \).

\[ \Box \]

**Acknowledgements:** CJO was supported by the ARC Centre of Excellence for Mathematics and Statistics, Australia, and the Lloyd’s Register Foundation programme on data-centric engineering at the Alan Turing Institute, UK.

**References**

J. Pearl. *Causality*. Cambridge University Press, 2009.

M.H. Maathuis, M. Kalisch, and P. Bühlmann. Estimating high-dimensional intervention effects from observational data. *Annals of Statistics*, 37(6A):3133–3164, 2009.

S.E.F. Spencer, S.M. Hill, and S. Mukherjee. Inferring network structure from interventional time-course experiments. *Annals of Applied Statistics*, 9(1):507–524, 2015.
D. Lopez-Paz, K. Muandet, B. Schölkopf, and I. Tolstikhin. Towards a learning theory of causation. In Proceedings of the 32nd International Conference on Machine Learning, 2015.

J.M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. Distinguishing cause from effect using observational data: Methods and benchmarks. Journal of Machine Learning Research, 17 (32):1–102, 2016.

C.J. Oates, F. Dondelinger, N. Bayani, J. Korkola, J.W. Gray, and S. Mukherjee. Causal network inference using biochemical kinetics. Bioinformatics, 30(17):i468–i474, 2014.

M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. Journal of Machine Learning Research, 7:2399–2434, 2006.

M.P. Wand and M.C. Jones. Kernel Smoothing. CRC Press, 1994.

R. Fergus, Y. Weiss, and A. Torralba. Semi-supervised learning in gigantic image collections. In Proceedings of the 23rd Annual Conference on Neural Information Processing Systems, pages 522–530, 2009.

P. Kemmeren, K. Sameith, L.A. van de Pasch, J.J. Benschop, T.L. Lenstra, T. Margaritis, E. O’Duibhir, E. Apweiler, S. van Wageningen, C.W. Ko, S. van Heesch, M.M. Kashani, G. Ampatziadis-Michailidis, M.O. Brok, N.A.C.H. Brabers, A.J. Miles, D. Bouwmeester, S.R. van Hooff, H. Bakel, E. Sluiters, L.V. Bakker, B. Snel, P. Lijnzaad, D. van Leenen, M.J.A. Groot Koerkamp, and F.C.P. Holstege. Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. Cell, 157(3):740–752, 2014.

J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference using invariant prediction: Identification and confidence intervals. Journal of the Royal Statistical Society: Series B, 78(5):947–1012, 2016.

N. Meinshausen, A. Hauser, J.M. Mooij, J. Peters, P. Versteeg, and P. Bühlmann. Methods for causal inference from gene perturbation experiments and validation. Proceedings of the National Academy of Sciences, 113(27):7361–7368, 2016.

S.M. Hill, N.K. Nesser, K. Johnson-Camacho, M. Jeffress, A. Johnson, C. Boniface, S.E.F. Spencer, Y. Lu, L.M. Heiser, Y. Lawrence, N.T. Pande, J.E. Korkola, J.W. Gray, G.B. Mills, S. Mukherjee, and P.T. Spellman. Context specificity in causal signaling networks revealed by phosphoprotein profiling. Cell Systems, 4(1):73–83, 2017.

R. Akbani, P.K.S. Ng, H.M. Werner, M. Shahmoradgoli, F. Zhang, Z. Ju, W. Liu, J.Y. Yang, K. Yoshihara, J. Li, S. Ling, E.G. Seviour, P.T. Ram, J.D. Minna, L. Diao, P. Tong, J.V. Heymach, S.M. Hill, F. Dondelinger, N. Städler, L.A. Byers, F. Meric-Bernstam, J.N. Weinstein, B.M. Broom, R.G.W. Verhaak, H. Liang, S. Mukherjee, Y. Lu, and G.B. Mills. A pan-cancer proteomic perspective on The Cancer Genome Atlas. Nature Communications, 5:3887, 2014.

R. Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B, 58(1):267–288, 1996.
M.H. Maathuis, D. Colombo, M. Kalisch, and P. Bühlmann. Predicting causal effects in large-scale systems from observational data. *Nature Methods*, 7(4):247–248, 2010.

S.M. Hill, L.M. Heiser, T. Cokelaer, M. Unger, N.K. Nesser, D.E. Carlin, Y. Zhang, A. Sokolov, E.O. Paull, C.K. Wong, K. Graim, A. Bivol, H. Wang, F. Zhu, B. Afsari, L.V. Danilova, A.V. Favorov, W.S. Lee, D. Taylor, C.W. Hu, B.L. Long, D.P. Noren, A.J. Bisberg, The HPN-DREAM Consortium, G.B. Mills, J.W. Gray, M. Kellen, T. Norman, S. Friend, A.A. Qutub, Y. Fertig, E.J. Guan, M. Song, J.M. Stuart, P.T. Spellman, H. Koeppel, G. Stolovitzky, J. Saez-Rodriguez, and S. Mukherjee. Inferring causal molecular networks: Empirical assessment through a community-based effort. *Nature Methods*, 13(4):310–318, 2016.

J. Li, Y. Lu, R. Akbani, Z. Ju, P. L. Roebuck, W. Liu, J-Y. Yang, B.M. Broom, R.G.W. Verhaak, D.W. Kane, C. Wakefield, J.N. Weinstein, G.B. Mills, and H. Liang. TCPA: a resource for cancer functional proteomics data. *Nature Methods*, 10(11):1046–1047, 2013.

L. Wasserman. *All of Nonparametric Statistics*. Springer, 2006.