nature machine intelligence

Article

https://doi.org/10.1038/s42256-023-00744-z

Deep learning of causal structures in high dimensions under data limitations

Received: 13 April 2022

Accepted: 20 September 2023

Published online: 26 October 2023

Check for updates

Kai Lagemann [1] ✉, Christian Lagemann [2], Bernd Taschler [1,3] & Sach Mukherjee [1,4] ✉

Causal learning is a key challenge in scientific artificial intelligence as it allows researchers to go beyond purely correlative or predictive analyses towards learning underlying cause-and-effect relationships, which are important for scientific understanding as well as for a wide range of downstream tasks. Here, motivated by emerging biomedical questions, we propose a deep neural architecture for learning causal relationships between variables from a combination of high-dimensional data and prior causal knowledge. We combine convolutional and graph neural networks within a causal risk framework to provide an approach that is demonstrably effective under the conditions of high dimensionality, noise and data limitations that are characteristic of many applications, including in large-scale biology. In experiments, we find that the proposed learners can effectively identify novel causal relationships across thousands of variables. Results include extensive (linear and nonlinear) simulations (where the ground truth is known and can be directly compared against), as well as real biological examples where the models are applied to high-dimensional molecular data and their outputs compared against entirely unseen validation experiments. These results support the notion that deep learning approaches can be used to learn causal networks at large scale.

Causality remains an important open area in artificial intelligence (AI) research[1,2], and the task of identifying causal relationships between variables is key in many scientific domains[3]. The rich body of work in learning causal structures includes methods such as PC[4], LiNGAM[5], IDA[6], GIES[7], RFCI[8], ICP[9] and MRCL[10]. Scaling causal structure learning to larger problems has been facilitated by reformulation as a continuous optimization problem[11], and recent neural approaches, such as SDI[12], DCDI[13], DCD-FG[14] and ENCO[15], have demonstrated state-of-the-art performance (Supplementary section 1 provides a detailed discussion). However, learning causal structures from data remains nontrivial and continues to pose challenges, particularly under the conditions (high dimensionality, limited data sizes and hidden variables, for example) seen in many real-world problems.

In biomedicine, causal networks representing the interplay between entities such as genes or proteins play a central conceptual and practical role. Such networks are increasingly understood to be context-dependent, and are thought to underpin aspects of disease heterogeneity and the variation in therapeutic response (for example, refs. 16–19). A key bottleneck in characterizing such heterogeneity lies in the challenging nature of learning causal structures at scale, because of general methodological issues as well as aspects relevant in the biological domain such as high dimensionality, complex underlying events, the presence of hidden/unmeasured variables, limited data and noise levels.

In this Article, we propose a deep architecture for causal learning that is particularly motivated by high-dimensional biomedical

[1]Statistics and Machine Learning, German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany. [2]Institute of Aerodynamics, RWTH Aachen University, Aachen, Germany. [3]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, UK. [4]MRC Biostatistics Unit, University of Cambridge, Cambridge, UK. ✉e-mail: kai.lagemann@dzne.de; sach.mukherjee@dzne.de

problems. The approach we put forward operates within an emerging causal risk paradigm (Methods and Supplementary section 2) that allows us to leverage AI tools and scale to very high-dimensional problems involving thousands of variables. The learners proposed allow for the integration of partial knowledge concerning a subset of causal relationships and then seek to generalize beyond what is initially known to learn relationships between all variables. This corresponds to a common scientific use-case in which some prior knowledge is available at the outset—from previous experiments or scientific background knowledge—but it is desired to go beyond what is known to learn a model spanning all available variables.

A large part of the causal structure learning literature involves learning models that allow an explicit description of the relevant data-generating model (including both observational and interventional distributions) and are in that sense 'generative' (see, for example, ref. 3 and references therein). Taking a different approach, a line of recent work, including refs. 10,20–22, has considered learning indicators of causal relationships between variables (without necessarily learning full details of the underlying data-generating models), and this can be viewed as being related to notions of causal risk[23]. Such indicators may encode, for example, whether, for a pair of variables A and B, A has a causal influence on B, B on A, or neither.

The approach we propose, called 'deep discriminative causal learning' (D^{2}CL), is in the latter vein. We consider a version of the causal structure learning problem in which the desired output consists of binary indicators of causal relationships between observed variables[10,23], that is, a directed graph with nodes identified with the variables. Available multivariate data X are transformed to provide inputs to a neural network (NN), whose outputs are estimates of the causal indicators. D^{2}CL differs from classical causal structure learning approaches both in terms of the underlying framework (based on causal risk rather than generative causal models) and in leveraging NNs. The assumptions underlying the approach are also different in nature from those in classical causal structure learning and concern higher-level regularities in the data-generating processes (Methods). A number of recent papers, including refs. 12–15, also leverage neural approaches for learning causal structures and share a basis in the continuous optimization framework introduced in ref. 11 based on a directed acyclic graph (DAG) framework. D^{2}CL, in contrast, uses a risk-based approach that is not based on DAGs. Eigenmann et al. [23] studied causal risk for the assessment of existing learners; instead, we leverage the notion of causal risk to propose a new learner. In common with D^{2}CL, the recently proposed CSIvA method[24] seeks to directly map input data to a graph output. The key difference is that, while CSIvA uses a meta-learning scheme based on large-scale synthetic data, D^{2}CL is based on supervised learning using data from a specific system of interest (for example, a biological system; see Supplementary section 1 for a more detailed overview and comparison). We show that context-specific training allows D^{2}CL to successfully learn structures in a range of scenarios, including challenging real-world experimental data (as detailed in the following). Furthermore, D^{2}CL is demonstrably scalable to large numbers of variables (we show examples ranging up to $p = 50{,}000$ nodes) and applicable in regimes where very large sample data or strong simulation engines are not available.

Framework overview

We propose an end-to-end neural approach to learn causal networks from a combination of empirical data X and prior causal knowledge Π. The general D^{2}CL workflow and its application to biomolecular problems are summarized in Fig. 1. Here we provide a very brief, high-level summary of the main ideas. A detailed presentation of the methodology and associated discussion (including of causal semantics and assumptions) are provided in the Methods and Supplementary section 2.

Suppose $X_1, \ldots, X_p$ is a set of variables whose mutual causal relationships are of interest. Let G^* denote an (unknown) graph whose directed edges encode these causal relationships. D^{2}CL seeks to learn G^* from two inputs: (1) empirical data X containing measurements on each of the variables of interest and (2) prior causal knowledge Π concerning a subset of causal relationships. This corresponds to a common paradigm in real-world scientific settings, where some data are measured on variables of interest, but only limited knowledge about causal relationships is available at the outset (for example, from prior scientific knowledge or specific experiments).

We formalize the task in the following way. For each ordered pair of variables with indices (i, j) whose causal status is not known via Π, our goal is to learn an indicator of whether or not X_i has a causal influence on X_j. D^{2}CL treats these causal indicators as 'labels' in a machine learning sense, using the available inputs to learn a suitable mapping. The goal of the mapping is to minimize discrepancy with respect to the true, unknown causal status; this learning task can be viewed through the lens of causal risk[23]. In all experiments, the learner never has access to data in which the parent node of an unknown edge was intervened on. This makes learning challenging, as we require generalization to interventional regimes/distributions that are entirely unseen.

Learning is carried out using a flexible, neural model F_θ with a set of trainable parameters θ. The model is trained in a specific fashion that leverages the input information Π as a supervision/training signal to allow the model to learn representations suitable for generalization to novel causal relationships (the Methods provides details and a discussion of the assumptions). The network F_θ combines a convolutional neural network (CNN) and a graph neural network (GNN) to resolve distributional and graph structural regularities (Fig. 2). In image processing, CNNs make use of certain properties, such as spatial invariance, that exploit the notion of an image as a function on the plane. Here we leverage the CNN toolkit to capture distributional information in data X, represented as images. We create these visual representations for two-tuples of nodes. Specifically, for a variable pair (i, j) we use the $n \times 2$ submatrix $X_{(\cdot,[ij])}$, to form a bivariate kernel density estimate $f_{ij} = \mathrm{KDE}(X_{(\cdot,[ij])})$ that is treated as an image input. Note that this is in general asymmetric in the sense that $f_{ij} \neq f_{ji}$. This is important, as we want to learn ordered/directed relationships (symmetry here would imply an inability to distinguish the causal direction). The GNN is aimed at capturing graph structural regularities and to this end learns a state embedding h_j that contains the information of the neighbourhood for each node j. The GNN requires a graph as input; we provide an initial input graph $\hat{G}_0$ via computationally lightweight routines solely based on the available data, X (Methods).

Finally, following training, the model F—with parameters now fixed as a function of inputs X and Π—can be used to assign causal status to any pair via an inference step. In the experiments described in the following, the global model output is tested systematically at large scale against either the true graph G^* (in simulations) or against entirely unseen interventional experiments (for real biological examples).

Our focus is on causal learning for real-world, high-dimensional problems with thousands of nodes and limited data, motivated by large-scale biomedical problems. Within the causal risk paradigm[10,23] we use here, acyclicity (of the directed graphs to be learned) is not assumed, nor is availability of any standard factorization of the joint probability distribution. It is not required that data samples in X are drawn from a single distribution; instead, data can be drawn from, for example, a mix of observational and interventional distributions, and the causal characteristics of these regimes (for example, which node(s) or latents were intervened on) need not be known in advance. Nor is it required that we have interventional data or prior information concerning all nodes. On the contrary, in all experiments, the learner never has access to data in which the parent node of an unknown edge was intervened on nor prior information concerning the unknown edge. This is a common real-world set-up, in particular for emerging experimental designs in biology (examples are described in the following). We emphasize that the NNs used are

https://doi.org/10.1038/s42256-023-00744-z

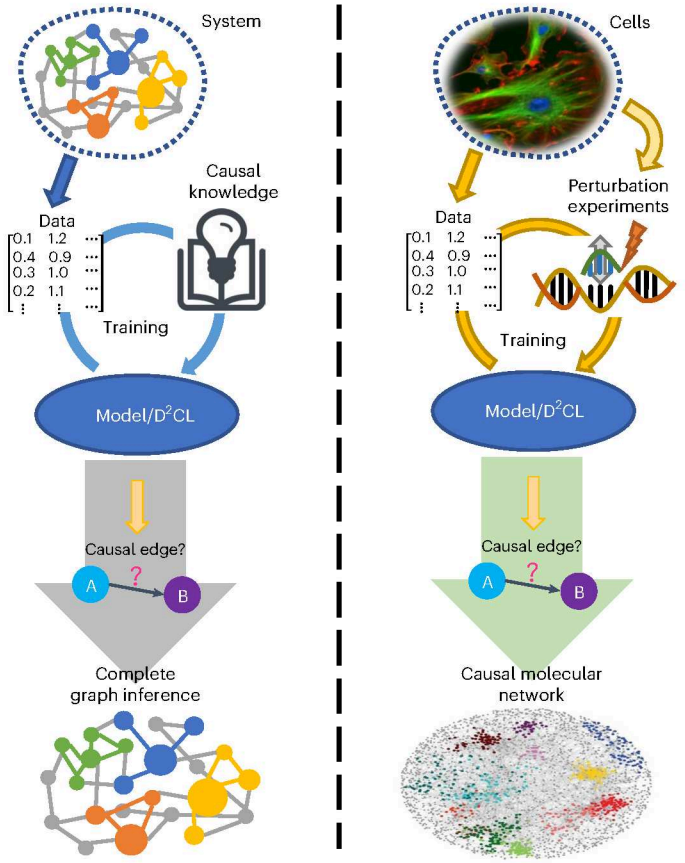

Fig. 1 | Conceptual overview of the proposed learning scheme and its application in large-scale biological experiments. The neural architecture learns causal structures by combining data and prior knowledge, resulting in a graph G intended to represent causal, not just correlational, relationships between system variables. In an abstract workflow (left), empirical data from a specific system are combined with prior causal knowledge to estimate the unknown causal structure. In the biological problem workflow (right), data are gathered from a specific biological system, and causal prior knowledge is derived from established science or interventional experiments on the system. The model seeks to generalize from the limited inputs to learn a global graph, spanning all system variables.

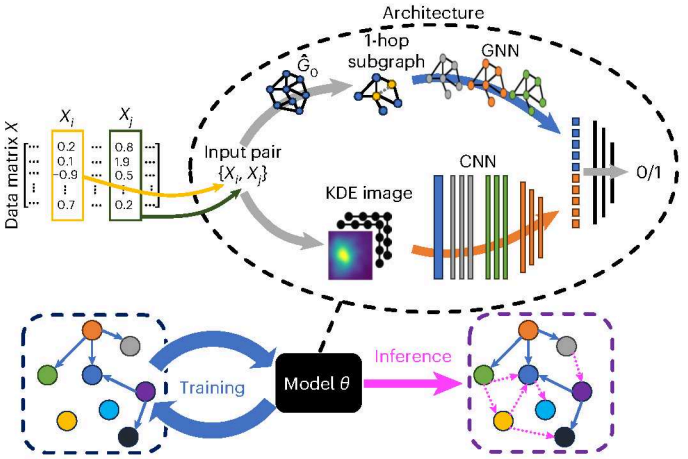

Fig. 2 | Overview of the D²CL architecture, training and inference. D²CL combines empirical data on multiple variables with prior causal knowledge to learn causal relationships between variables. For any pair of variables X_i and X_j (corresponding to two columns of the input data matrix), D²CL seeks to learn whether X_i has a causal influence on X_j, X_j on X_i, or neither. This is done using a neural architecture with two components: a CNN tower aimed at learning distributional features and a GNN tower that detects structural regularities. For an ordered pair (X_i, X_j), the CNN tower captures distributional information via a density estimate that traverses the tower to form an embedding. The GNN tower extracts a subgraph from an initial graph $\hat{G}_0$ and computes an embedding containing structural information on the neighbourhood of the nodes. The CNN and GNN embeddings are then merged through multiple layers, which finally output the probability of a directed causal relationship. The input causal information is used to provide a training signal (see main text for details). During inference, the network generalizes beyond the initial inputs to provide an estimate of the global graph spanning all variables of interest.

not rotation-invariant and hence can break symmetries and allow inference of causal direction.

Results

We use both simulated data and real biological data to assess performance. In all cases, the model output is tested with respect to causal relationships that are entirely unseen in the sense that (1) causal relationships on which the model output is tested are disjoint from those provided as inputs during training and (2) no data used to define causal relationships against which the model output is tested appear in inputs to the models. Additional results, as well as details of the experimental protocols, are provided in Supplementary sections 3 and 4.

Simulation benchmarks

We tested the methods using data generated from a (linear or nonlinear) structural equation model (SEM) with noise, based on a known underlying causal graph G^*. The protocol is outlined in Fig. 3a, with further details provided in Supplementary section 3. In brief, data were generated via structural equations of the form $X_i = f_i(\mathrm{Pa}_{G^*}(X_i), U_{X_i})$, for $i = 1, \ldots, p$, where p is the total number of variables, $\mathrm{Pa}_{G^*}(X_i)$ is the set of parents for node i in the true graph G^*, and U_{X_i} are noise variables (exogenous and jointly independent). The functions f_i are unknown to the learners. Varying the noise magnitude allows us to control the signal-to-noise ratio (SNR), and varying p allows us to understand the effect of dimensionality. The output was tested against the true, gold-standard causal structure G^* and hence assessed in causal (and not correlational or predictive) terms.

In-system, out-of-distribution evaluation. Here, model training uses (limited) prior knowledge and data from a given system, and assessment is carried out with respect to unknown edges within the same system (test and training edges are always entirely disjoint). This is out-of-distribution in the sense that the learner never has access to samples from the test interventional distributions, but in-system, because all data are from the same overall data-generating system. This corresponds to a common scientific use-case where the goal is to learn a model for a specific system of interest given available data on that system. Figure 3c shows results for a problem of dimension $p = 1{,}500$ using a nonlinear transition function (the tangent hyperbolic; other functions and configurations are shown in Supplementary Tables 2 (area under the curve, AUC) and 3 (area under the precision-recall-curve, AUPRC)) and varying SNR. (For these first results, we restricted the dimension of the problem to facilitate comparison to existing approaches that are less scalable than D²CL; higher-dimensional examples appear in the following.) Note that pairwise correlations between the variables ('Pearson') are ineffective; this is expected due to the presence of latent variables in all experiments and the fundamental difference between correlational and causal relationships. Overall, D²CL remains effective across a broad range of SNRs, as well as for a range of linear and nonlinear problems and problem sizes (Supplementary Table 1). We also compared D²CL to DCD-FG[14] and ENCO[15], two recently proposed, scalable neural-causal learners. Owing to computational considerations, we

https://doi.org/10.1038/s42256-023-00744-z

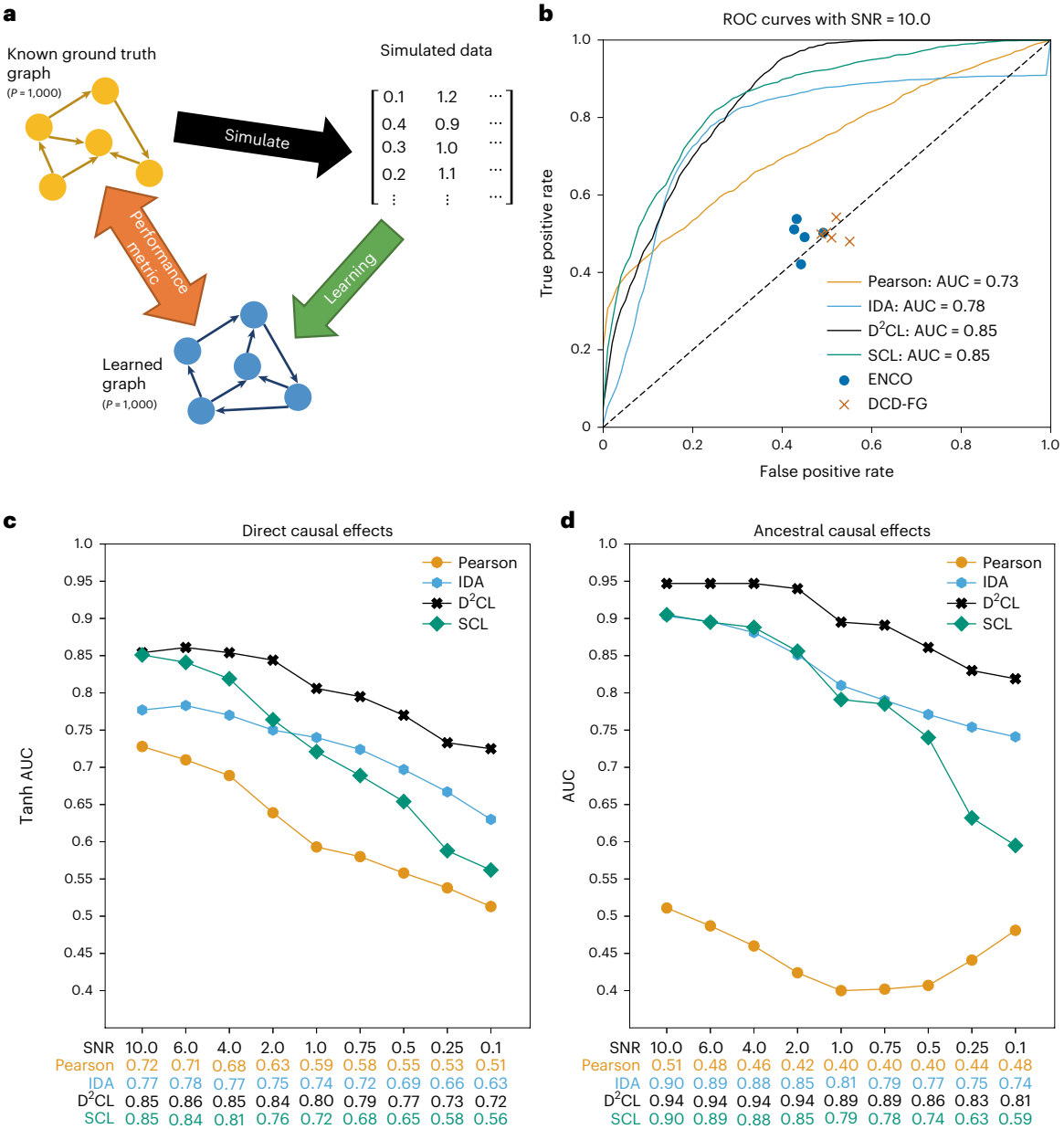

Fig. 3 | Results for large-scale simulated data. a, Overview of the experimental workflow. Data were simulated from known, gold-standard causal graphs, and the output of the learners was compared with the true, underlying graph to quantify the ability to recover the causal structure. Finite-sample empirical data were generated using a directed causal graph of specified dimension p, specifically via linear and nonlinear structural equation models with noise. **b**, ROC curves for an illustrative nonlinear case (the tangent hyperbolic), with an SNR of 10.0, for direct causal relations in a graph with $p = 1{,}500$ nodes. D^{2}CL (black) is compared against Pearson correlation coefficients (orange), IDA (cyan), SCL (green), ENCO (blue) and DCD-FG (brown). The ROC curve and the area under the ROC curve (AUC) are given for algorithms providing a continuous output (Pearson, IDA, SCL and D^{2}CL). The binary graph estimates of ENCO and DCD-FG are represented by single markers for five different runs. **c**, Results for an illustrative nonlinear case (the tangent hyperbolic), at varying noise levels, for direct causal relationships. The causal area under the ROC curve (AUC; with respect to the causal ground truth graph, see Methods and Supplementary section 3 for details) is shown as a function of SNR for an experiment with $p = 1{,}500$ variables and a sample size of $n = 1{,}024$. Results for other linear and nonlinear functions are provided in Supplementary section 4. D^{2}CL (blue) is compared with Pearson correlations (orange; this is a non-causal baseline), IDA (cyan) and SCL (green). **d**, Results for indirect causal relationships, with other settings as in **c**. Here, causal AUC is shown with respect to a graph encoding causal, but potentially indirect, relationships. (Results shown are averages over five datasets at each specified SNR).

restricted this comparison to a subset of the simulations. Illustrative results are provided in Fig. 3b. We find that neither approach is effective in this case, possibly due to the limited data and the presence of latent variables.

In addition, we tested the effectiveness of D^{2}CL for additive and multiplicative Gaussian noise with varying SNRs under settings with hard deterministic and stochastic interventions. We refer the interested reader to Supplementary section 3 for a definition of an intervention and the types used. The test results (AUC and AUPRC values) are summarized in Supplementary Tables 8 and 9 and support the notion that D^{2}CL is robust to different types of noise.

The graph G^* in the above examples encodes direct causal relationships as there is an edge from one node to another if the former appears in the equation for the latter. However, in many real-world examples, interest focuses also on indirect effects, which may be mediated by other nodes. For example, if node A has a direct effect

https://doi.org/10.1038/s42256-023-00744-z

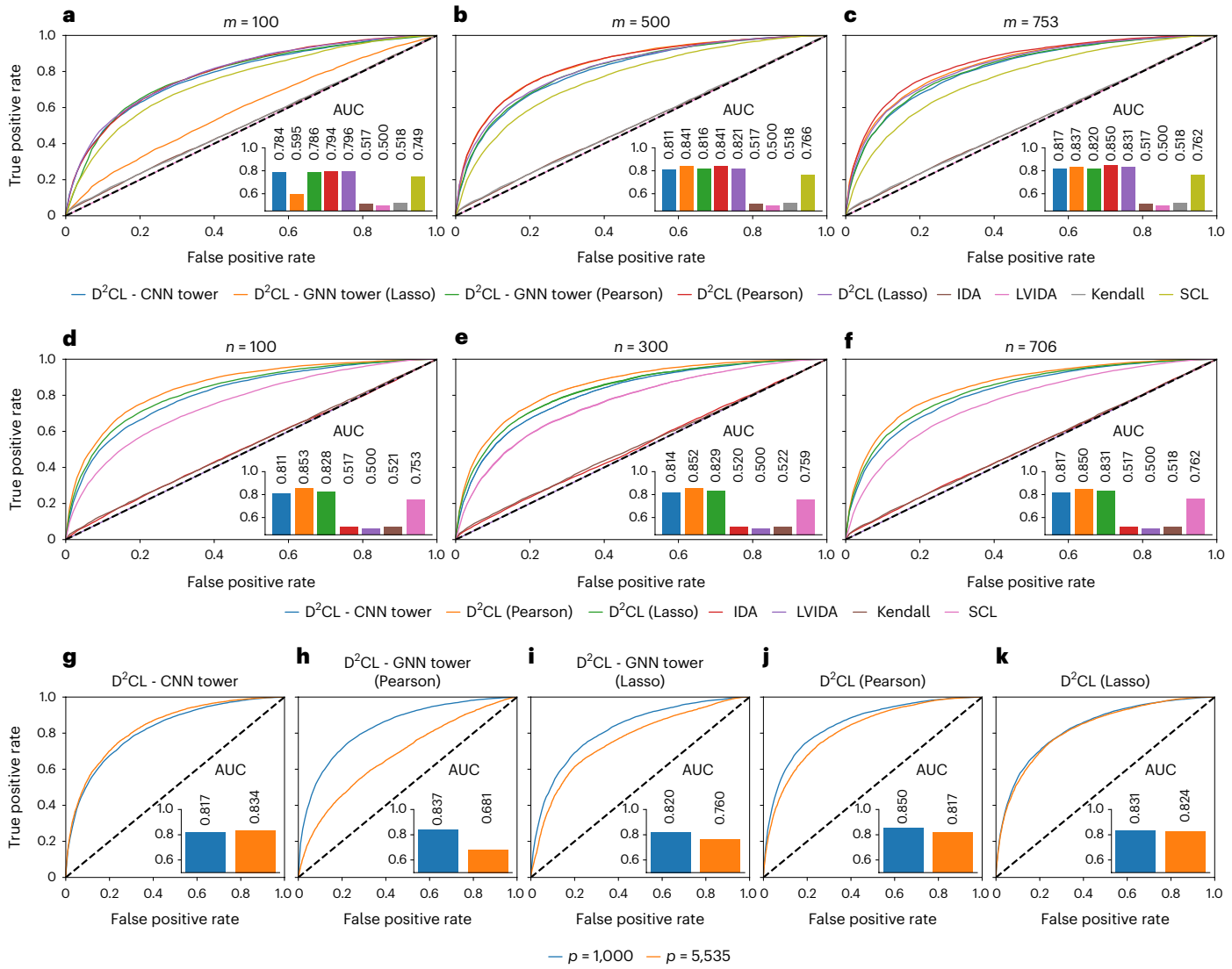

Fig. 4 | Results for the yeast gene deletion experiments. Causal learning methods, including D^{2}CL, were applied to gene expression measurements from yeast cells. Performance was quantified using causal ROC curves (and AUCs) computed with respect to a causal ground truth obtained from entirely unseen interventional experiments (see main text for details). **a**–**c**, The number of interventions m whose effects are available to the learner was varied (with the problem dimension fixed to $p = 1{,}000$ and the sample size to $n = 706$): $m = 100$ (**a**), 500 (**b**) and 753 (**c**). **d**–**f**, The sample size n of the data matrix X was varied (with the problem dimension fixed to $p = 1{,}000$ and the number of available interventions fixed to $m = 753$): $n = 100$ (**d**), 300 (**e**) and 706 (**f**). **g**–**k**, Analogous results for a higher-dimensional setting covering all available genes (roughly the full yeast genome) with $p = 5{,}535$ (with $n = 706$ and $m = 753$) for the indicated arrangements. Here, only D^{2}CL variants are shown, as the other methods could not be run due to the computational burden in this higher-dimensional case. Comparison with the corresponding $p = 1{,}000$ case demonstrates the scalability of D^{2}CL, with performance broadly maintained in the higher-dimensional setting. The D^{2}CL variants shown include a CNN tower alone (**g**), GNN tower alone (**h**,**i**) and the entire D^{2}CL architecture (**j**,**k**); methods compared against include IDA, LVIDA, Kendall correlations (as a non-causal baseline) and SCL (see main text and Supplementary sections 1 and 3 for details and references). For D^{2}CL and its variants, two different initial graph estimates were used based respectively on Pearson correlation coefficients ('Pearson') and on a lightweight regression ('Lasso'); details are provided in the main text.

on B, and B on C, intervention on A may change C, even though A does not itself appear in the equation for C. To test the ability to learn indirect edges, we proceeded as above, but with the inputs Π being indirect edges and the output tested against the true indirect graph. Results are presented in Fig. 3d. D^{2}CL outperforms existing methods across a range of SNRs and also in other linear/nonlinear problem configurations (Supplementary Tables 4 and 5). IDA performs well in the case of a linear SEM, but not for functions based on nonlinear multilayer perceptrons. D^{2}CL appears to be the most noise-robust of the methods tested. These results show that D^{2}CL can learn indirect causal edges over many variables under conditions of noise and nonlinearity.

Out-of-system, out-of-distribution evaluation. D^{2}CL is trainable using (limited) data from a specific system (for example, a specific biological system, such as cells of a particular kind, or a disease state). However, it is interesting to see whether it is possible to generalize to different systems. To this end, we trained D^{2}CL on a dataset from a certain system and cross-evaluated the trained model on data from another system (a different simulation regime). The results are provided in Supplementary Tables 10 and 11. Some generalization appears possible, suggesting that D^{2}CL can find signals that are causally informative in a cross-system sense, although performance is always worse relative to in-system training (this is expected in our framework, and we emphasize that we do not claim any general ability to achieve out-of-system generalization).

https://doi.org/10.1038/s42256-023-00744-z

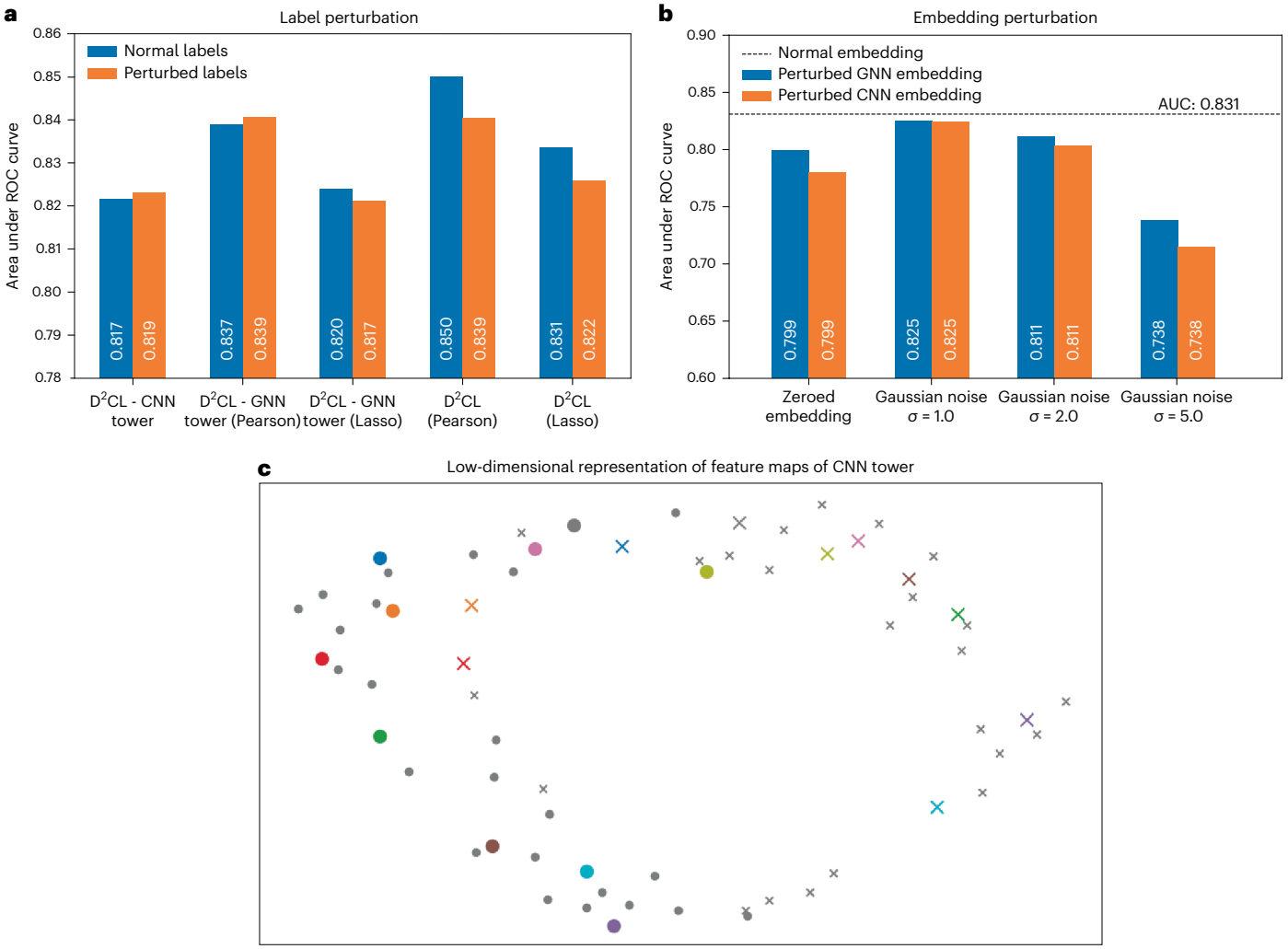

Fig. 5 | Sensitivity to incorrect causal inputs and additional results on causal direction. a, Robustness to incorrect causal inputs. The sensitivity of D^{2}CL to errors in prior/input causal knowledge Π was studied by artificially introducing errors into Π, with 10% of inputs corrupted (experiments used the yeast gene deletion data; see main text for details). Results quantified via causal AUC (as in the main results, computed with respect to an experimentally defined ground truth), shown for several D^{2}CL variants. **b**, An ablation-like study in which failures of either the CNN (orange) or the GNN (blue) tower within D^{2}CL were artificially introduced. The relevant embedding was either set to zero or to zero-mean Gaussian noise (with scale as shown). The unaffected case is given as a dashed black line. **c**, Causal direction analysis (see main text for details). Low-dimensional representations of latent feature maps of the converged CNN tower at two different layer depths. Edges $A \to B$ are shown as filled circles and reverse edges $B \to A$ as x-shaped markers. An edge and its corresponding reverse are shown in the same colour. For improved readability, only ten random pairs are highlighted in colours and bigger markers. We see that the embedding is not invariant with respect to causal direction and is able to effectively identify the correct direction (as shown also in an additional experiment, see main text). The different D^{2}CL variants include a CNN tower alone, a GNN tower for two different initial graph estimates, and the complete network for the same two initial graph estimates. Initial graph estimates for the GNN and combined models were either based on Pearson correlation coefficients ('Pearson') or a lightweight regression ('Lasso'; see main text for details).

Nevertheless, these results broadly support the notion of large-scale meta-learning for causal structures[24].

Large-scale evaluation. Finally, to test the scalability of D^{2}CL to high-dimensional problems, we considered a problem with $p = 50{,}000$ variables (that is, $p = 50{,}000$ nodes in the ground-truth graph; note that none of the compared methods can practically scale to this setting). We considered learning of direct causal relationships; the results are shown in Supplementary Table 6 and support the notion that D^{2}CL can scale to problems spanning many thousands of variables.

Large-scale biological data

To study performance in the context of real biological data, we leveraged a large set of gene deletion experiments in yeast[25], which have previously been used for causal learning[9,10,26]. The experiments involve measuring gene expression in yeast cells under each of a large number of interventions (gene deletions; Supplementary section 3 provides further details).

In biological experiments, causal effects may be indirect, and we sought to learn a directed graph with nodes corresponding to p observed genes and edges representing (possibly indirect) causal influences. Such edges are scientifically interesting and amenable to experimental verification, as noted in refs. 22,27. Cycles can arise in systems biology (see, for example, ref. 28) and we do not enforce acyclicity (see ref. 29 and references therein for a discussion of cyclic causality). A fuller discussion of the causal interpretation of laboratory experiments is beyond the scope of this Article, but relevant work includes refs. 29–31, and we direct the interested reader to these references for further discussion.

Because causal background knowledge is an input for our approach, it is relevant to consider performance as a function of the

https://doi.org/10.1038/s42256-023-00744-z

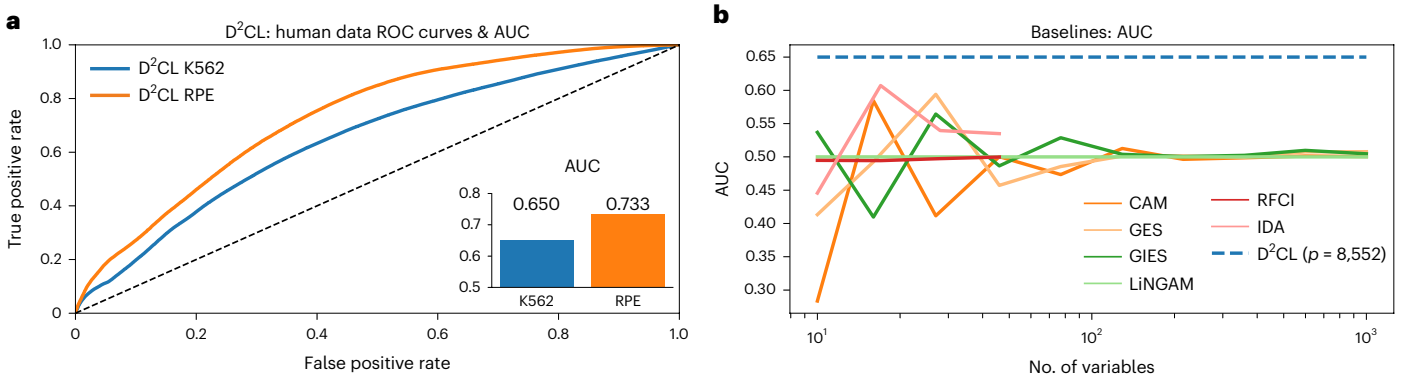

Fig. 6 | Results for high-dimensional human data. Single-cell CRISPR-based experiments (due to ref. 32) were used to illustrate the use of the proposed approaches in a high-dimensional human cell setting. Performance was quantified using causal ROC curves (and AUC) computed with respect to a causal ground truth obtained from entirely unseen interventional experiments (see main text for details). **a**, Results from D^{2}CL applied to data obtained from RPE cells and a cancer cell line (K562) in problems spanning more than 8,000 variables (other methods could not be practically run in this case due to the computational burden). **b**, Performance of existing causal learning approaches (on K562 data) as a function of problem dimension. The dashed line indicates D^{2}CL performance on the full problem (p = 8,552 variables).

amount of such input. To this end, we fixed the problem size to p = 1,000 and varied the number of interventions m whose effects were available to the learner (Supplementary section 3 provides details). As each experiment involves only a subset of the entire yeast genome, latent variables are present by design. The input prior knowledge Π is derived from the causal status, but, as in all experiments, is strictly disjoint with respect to any test edges.

Results are presented in Fig. 4a–c, including the area under the receiver operating characteristic (ROC) curve (AUC; computed with respect to an experimentally determined gold standard; Supplementary section 3). Overall, the proposed methods perform well, achieving good results in this high-dimensional, limited-data regime. Next, to shed light on data efficiency, we varied the sample size n of the data matrix X (Fig. 4d–f).

Finally, we tested the performance in a higher-dimensional example spanning all p = 5,535 available genes (Fig. 4g–k) and found that D^{2}CL remains effective at the genome scale. Interestingly, although the CNN tower performs particularly well, the GNN tower degrades more. This may be because larger p leads to a larger number of variable pairs (which is helpful for the CNN), but also to a (rapid) increase in the number of nodes and edges in the GNN subgraphs and hence a harder GNN learning task in practice.

D^{2}CL leverages prior causal knowledge. However, in practice, the available causal inputs Π may be incorrect, for example, due to flawed initial experiments or errors in the known science. To study sensitivity to flawed causal inputs, we introduced errors into Π. This was done by perturbing 10% of the inputs (that is, labelling causal pairs as non-causal and vice versa) at the outset. The results are shown in Fig. 5a and demonstrate a level of robustness to such perturbation. We also see a benefit of the dual network variants; this is investigated further in Fig. 5b. For the latter, in general, the embedding of either tower is modified immediately before the fusion layer. We considered several different modifications: setting the embedding of one tower to zero and hence effectively removing all information from this tower, or applying Gaussian noise with magnitude σ = 1.0, σ = 2.0 and σ = 5.0.

Causal relations are in general directed and asymmetric, so it is interesting to explore model behaviour with respect to causal direction. Given an image representation, the CNN tower is designed to extract feature maps that are unique for ordered node pairs, that is, such that in general features differ depending on edge direction. To empirically study learning of causal direction, we constructed additional test data as follows: for each truly causal edge $k \to l$ in the test set, we also included the reverse direction $l \to k$. This means that any learner estimating undirected links would have an AUC score of 0.5 (because the output $k \to l$ entails also $l \to k$, one of which is a false positive). Supplementary Table 4 shows that D^{2}CL is indeed capable of accurately identifying causal direction. In addition, Fig. 5c shows a low-dimensional representation of the feature maps of the converged CNN tower. These feature maps differ by causal direction ($k \to l$ versus $l \to k$) throughout the forward pass, supporting the foregoing arguments.

High-dimensional CRISPR-based perturbations

Finally, we used recent, single-cell clustered regularly interspaced short palindromic repeats (CRISPR)-based interventional experiments[32] to illustrate the use of the proposed approaches in very high-dimensional data from human cells. The experimental protocol (see ref. 32 for full details) includes a large number of interventions in a leukaemia cell line (K562) and in retinal pigment epithelial (RPE) cells. The K562 and RPE experiments include gene-expression levels for a total of, respectively, p = 8,552 and p = 8,833 genes (Supplementary section 3 provides details). This is a challenging setting due to the known complexity of regulatory events in human cells and high levels of variability and noise in single-cell protocols. The results are presented in Fig. 6 and demonstrate good performance for RPE, and slightly worse performance, but still nontrivial consistency with the experimental gold standard, for K562. Additional plots in Fig. 6 and Supplementary Fig. 3 show the performance and runtime for a set of baseline algorithms. These results demonstrate two key points. First, the runtime for many available algorithms grows so rapidly with increasing number of variables as to render them unsuitable for problems at this scale. Second, for existing methods that are at all able to scale to larger problems, performance is considerably less effective than D^{2}CL in this setting.

Conclusions

Emerging experimental protocols, involving combinations of perturbations and high-dimensional readouts, are allowing for new, scalable ways to query molecular networks in a context-specific fashion. Combined with scalable causal learning tools, these approaches have the potential to strongly impact disease biology by allowing global networks, spanning thousands or tens of thousands of variables, to be investigated across many different contexts.

Networks learned in this way could, in the future, be leveraged to allow for the prediction of disease phenotypes or drug response under novel perturbations (this is a different task from standard supervised learning, because the test case involves an unseen perturbation to the system). Furthermore, in the area of personalized medicine, such an

approach could even allow for rational optimization over potential therapeutic strategies, because the latter are often interventions targeted at molecular nodes.

Our model leverages deep learning tools to learn causal relationships between variables at large scale. However, and in contrast to well-established approaches based on causal graphical models, it provides only a structural output rather than a probability model of the underlying system. It is also interesting to contrast D^{2}CL with the recently proposed CSIvA[24]. Both approaches pursue, in a sense, a 'direct' mapping of data inputs to graph outputs, with a key difference being that CSIvA uses meta-learning and seeks to generalize across systems, whereas D^{2}CL uses supervised learning to generalize to new interventions on a given system (for example, a biological system of interest). An interesting direction for future work may be to combine both approaches, for example by using CSIvA to provide the initial input to D^{2}CL; this would combine general, simulation-based learning and data-efficient, system-specific training.

At present, rigorous theory and an understanding of the theoretical properties of the kind of approach studied here remain lacking. A key direction for future theoretical work will be to understand the precise conditions for the underlying system that are needed to ensure that direct mapping approaches can guarantee the recovery of specific causal structures. An interesting observation is that the proposed approach may benefit from a 'blessing of dimensionality', because the learning problem will typically enjoy a larger number of examples as dimension p grows. Conversely, and in contrast to established statistical causal models, our approach (at the current stage) cannot be used in the small-p regime, because the number of examples will be too small for deep learning.

Methods

In this section, we provide information on the causal interpretation of our learning scheme, as well as a more detailed presentation of the architecture and associated implementation.

Notation

Observed variables with index set $V = \{1, \ldots, p\}$ are denoted $X_1, \ldots, X_p$. The variables will be identified with vertices in a directed graph G whose vertex and edge sets are denoted $V(G)$ and $E(G)$, respectively. We occasionally overload G to refer also to the corresponding binary adjacency matrix, using G_{ij} to refer to the entry (i,j) of the adjacency matrix, as will be clear from context. We use linear indexing of variable pairs to aid formulation as a machine learning problem. Specifically, an ordered pair $(i,j) \in V \times V$ has an associated linear index $k \in \mathcal{K} = \{1, \ldots, K\}$, where K is the total number of variable pairs of interest. Where useful, we make the mapping explicit, denoting the linear index corresponding to a pair (i,j) as $k(i,j)$ and the variable pair corresponding to a linear index k as $(i(k), j(k))$. The linear indices of pairs whose causal relationships are unknown and of interest are $\mathcal{U} \subset \mathcal{K}$, and those pairs known in advance via input knowledge Π are $\mathcal{T}(\Pi) \subset \mathcal{K}$. In all experiments, $\mathcal{T}(\Pi)$ and $\mathcal{U}$ are disjoint; that is, no prior causal information is available on the pairs $\mathcal{U}$ of interest.

Problem statement

We focus on the setting in which the available inputs are

- (I1) Empirical data: an $n \times p$ data matrix X whose columns correspond to variables $X_1, \ldots, X_p$.
- (I2) Causal background knowledge Π providing information on a subset $\mathcal{T}(\Pi) \subset \mathcal{K}$ of causal relationships.

For (I2), we assume that information is available concerning the causal status of a subset of variable pairs. That is, for some variable pairs (X_i, X_j) the correct binary indicator G^*_{ij}, representing the presence/absence of an edge in the target graphical object, is provided as an input. In terms of linear indexing, these can be viewed as available 'labels' of causal status for the pairs $\mathcal{T}(\Pi) \subset \mathcal{K}$. No specific assumption is made on the data X, but, in line with our focus on generalizing to unseen causal relationships, it is assumed that it does not contain interventional data corresponding to the pairs in $\mathcal{U}$. Furthermore, in all experiments, not only are the sets $\mathcal{T}$ and $\mathcal{U}$ disjoint, but we enforce the stronger requirement that $u \in \mathcal{U} \Rightarrow \nexists j : k(i(u), j) \in \mathcal{T}$. This means that all interventions on which models are tested are entirely novel, that is, unrepresented in the inputs to the learner, either as data or prior input. This also means that the learner has no access whatsoever to samples from the test interventional distributions, and all experiments are out-of-distribution in this sense.

The learning task can thus be formulated as follows: given inputs (I1) and (I2), the goal is to estimate, for each ordered pair of variables (X_i, X_j) with unknown causal relationship, whether or not X_i has a causal influence on X_j.

Summary of the learning scheme

With the notation above, our goal is to learn a graph whose nodes correspond to the variables $X_1, \ldots, X_p$ and whose edges represent causal relationships. To this end, we train a parameterized network F_θ, that is, a nonlinear function F with a set of unknown, trainable parameters θ. This is possible, because we know for each pair $k \in \mathcal{T}$ the causal status G^*_{ij} based on input information Π. The architecture we use as F_θ is detailed below, but for now assume this has been specified. Then, given the data X and the training labels $Y_k = G^*_{i(k),j(k)}$ for all pairs $k \in \mathcal{T}(\Pi)$, we train the set of parameters $\hat{\theta}(X, \Pi)$ under a loss that is supervised by the (causal) labels Y_k.

At this stage, the trained network $F_{\hat{\theta}(X,\Pi)}$ allows assignment of causal status to any pair, because it gives an estimate of the entire graph including those pairs whose causal status was unknown. The output is given by

$$\hat{G}_{ij}(X, \Pi) = \begin{cases} F_{\hat{\theta}(X,\Pi)}(i, j; X) & \text{if } k(i,j) \notin \mathcal{T}(\Pi) \\ Y_{k(i,j)}(\Pi) & \text{otherwise} \end{cases} \quad (1)$$

where (i,j) are ordered variable pairs. Note that the overall estimate depends solely on the data X and causal information Π. By default, no change is made for pairs $\mathcal{T}$ whose status was known at the outset. Reference [23] studied causal notions of risk based on loss functions of the form that compare a graph estimate $\hat{G}$ with ground truth G^*. In our setting, we consider a classification-type loss on the variable pairs k, where the causal status of known pairs $\mathcal{T}(\Pi)$ provides the training 'labels'. We therefore use the corresponding binary cross-entropy loss, augmented by additional terms that, for example, prevent exploding weights.

Causal interpretation of the learning scheme

D^{2}CL outputs a directed graph. The discriminative nature of D^{2}CL means that the notion of causal influence encoded by the edges is rooted in the application setting and input information Π, because causal semantics are inherited via the problem setting rather than specified by a generative model (see ref. 10 for related discussions). Indeed, in the experiments we showed that D^{2}CL could be used to successfully learn either direct or indirect/ancestral causal relationships.

Here we provide some intuition as to why discriminative learning can be effective in this setting. However, we note that the following arguments are not intended to constitute a rigorous theory at this stage, but rather to help gain an understanding of the conditions under which discriminative causal structure learning may be expected to be effective.

We start with a general causal framework and then introduce assumptions for D^{2}CL (the meta-generator assumption (MGA) and the dominant cause under single intervention (DCSI), described in the following sections). Following refs. 1,33, we assume decomposition of the underlying system into modular and independent mechanisms:

Independent causal mechanisms (ICMs). The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other.

For variables X_i assume a structural causal model with equations $X_i = f_i(\mathrm{Pa}_{G^*}(X_i), U_{X_i}),\ i = 1, \ldots, p$, where $\mathrm{Pa}_{G^*}(X_i)$ denotes the set of parents in the ground-truth graph G^* for node i, and f_i is a node-specific function. Exogeneous noise terms U_{X_i} are assumed jointly independent and distributed as $U_{X_i} \sim p_i$, where p_i is a node-specific density.

Our approach treats the f_i and p_i as unknown, but assumes they are related at a higher level. This can be formalized as a meta-generator assumption as follows.

Meta-generator assumption (MGA). For a specific system W, the functions f_i and noise distributions p_i are (independently) generated as $f_i \sim \mathcal{F}_W$ and $p_i \sim \mathcal{P}_W$, where $\mathcal{F}_W$ denotes a function generator and $\mathcal{P}_W$ a stochastic generator, that are specific to the applied problem setting W.

MGA is motivated by the notion that in any particular real-world system, underlying (biological, physical, social and so on) processes tend to share some functional and stochastic aspects, which impart some higher-level regularity. That is, MGA states that, in a given applied context, functions f_i and (independent causal mechanism-consistent) noise terms U_{X_i}, while unknown, varied and potentially complex, are nonetheless related at a 'meta'-level. The generators $\mathcal{F}_W$, $\mathcal{P}_W$ are random processes, representing, respectively, a 'distribution over functions' and a 'distribution over distributions', whose role here is to capture the notion of relatedness among f_i functions (respectively p_i) in a given setting W. Note that $\mathcal{F}_W, \mathcal{P}_W$ are treated as unknown and never directly estimated.

As mentioned in the problem statement, we focus on the causal status of variable pairs (X_i, X_j) (rather than general tuples), which denotes the simplest possible case under MGA. Furthermore, in both our work and the majority of interventional studies in applications such as biology, single interventions (rather than joint interventions on multiple nodes) are the norm. This requires the additional assumption, DCSI.

Dominant cause under single interventions (DCSI). A sufficiently large change in one of potentially multiple causes leads to a change with respect to the effect. Therefore, single interventions are sufficient to drive variation in the child distribution.

From MGA and DCSI to discriminative causal structure learning. Consider an applied problem W with underlying causal graph G^*_W, treated as fixed but unknown. The associated functions and noise terms are also unknown but assumed to follow MGA. Then, under DCSI, we have that all pairs of the form (X_i, X_j) have underlying relationships of the form $X_j = f_j(X_i, U_{X_j})$ with components following the MGA (that is, drawn from generators $\mathcal{F}_W$, $\mathcal{P}_W$). This in turn suggests that within the setting W, identification of causal pairs can be treated as a classification problem, as all pairs share the same generators. In other words, MGA restricts the distribution over relations of variables and noise terms to system-specific generators.

Note that no particular assumption is made on the individual functions f_j, only that they are mutually related on a higher level. Furthermore, the generators themselves need not be known nor directly estimated; rather, it is only important that they are shared across the applied setting W. Note that a model learned for setting W will not in general be able to classify pairs in an entirely different applied setting W' (because the generators may then differ strongly); that is, we do not seek to learn 'universal' patterns that apply to all causal relations in any system whatsoever. The classification task of D^{2}CL aims to tell apart causal relationships (assumed drawn from the system-specific generators) from non-causal ones. We note that, in real systems, f_i functions may be coupled via constraints on global functionality, and are thus non-independent; however, the good performance seen in the results empirically justifies the approach. Despite the initial theoretical ideas described above, rigorous theory and the theoretical properties of the kind of approach studied here remain to be understood, in particular the precise conditions for the underlying system needed to ensure that the classification-type approach can guarantee recovery of specific causal structures. We emphasize also that in contrast to classical causal learning schemes, for example, based on causal DAGs, we cannot at this stage make theoretical statements concerning underlying multivariate distributions and their link to edges estimated by our models. Our goal is good performance in an edge-wise sense (as detailed above), and the core assumptions (formalized above) concern a limited notion of classifiability. We note also that our models at present learn edges separately and do not impose any particular wider/global constraints (such as acyclicity or path constraints), although this could in principle be done within the causal risk framework.

Architecture details

CNN tower. To capture distributional information from empirical data X, a preprocessing step is required. In principle, this could be done via a variety of multidimensional transformations of X. We consider the simplest possible case, namely for a pair (i, j) to consider only the corresponding columns i and j in the data matrix X. Specifically, we use the $n \times 2$ submatrix $X_{(:,[ij])}$ to form a bivariate kernel density estimate $f_{ij} = \mathrm{KDE}(X_{(:,[ij])})$. Note that this is, in general, asymmetric in the sense that $f_{ij} \neq f_{ji}$, which is important as we want to learn ordered/directed relationships. In other words, this ensures that, in general, the CNN tower can output different probabilities for edges $A \to B$ and $B \to A$ (for any two nodes A and B). Evaluations of the KDE at equally spaced grid points on the plane (that is, numerical values from the induced density function) are treated as the input to the CNN. The KDE itself is a standard bivariate approach using automated bandwidth selection following refs. 34,35. This provides an 'image' of the data and allows us to leverage existing image analysis ideas. Furthermore, we concatenate channelwise the numerical KDE values on the regularly spaced grid with a positional encoding of the grid points.

The concrete network architecture of our CNN tower is inspired by a ResNet-54 architecture[36]. From a high-level perspective, it consists of a stem, five stages with [3, 4, 6, 3, 3] ResNet blocks and multiple fully connected layers that transform the high-level feature maps into a latent space that is merged with the output of the GNN tower. The first ResNet block at each stage downsamples the spatial dimensions of the output of the previous stage by a factor of two. To enhance the computational efficiency of the bottleneck layers in each ResBlock, channel down- and upsampling exploiting 1×1 convolutions is performed before and after each feature-extraction CNN layer[37]. We replaced ReLU activations by the parametric counterpart PReLU[38], allowing us to learn the slope of the negative part at negligible additional computational costs, which addresses the problem of dying neurons. Following ref. 39, we chose a full pre-activation of the convolutional layers, normalization–activation–convolution.

GNN tower. Our GNN tower builds on the SEAL architecture of ref. 40 and the resulting graph convolutional neural network (GCNN) for link prediction. The underlying notion is that a heuristic function predicts scores for the existence of a link. However, instead of employing predefined heuristics (such as the Katz coefficient or PageRank), an adaptive function is learned in an end-to-end fashion, which is formulated as a graph classification problem on enclosing subgraphs. Reference [40] showed that a γ-decaying heuristic can be approximated by an h-hop neighbourhood while the approximation error is at least decreasing exponentially. These findings suggest that it is possible to learn high-order graph structure features from local enclosing subgraphs instead of the entire graph, which can be exploited for link prediction. Consider the pair of nodes of interest (i, j); the GNN tower is intended to infer causally interesting node features and state embeddings based on a local 1-hop enclosing subgraph extracted from the approximated

input graph $\hat{G}_0$. For node pair (i,j), we first extract a set of nodes $\mathcal{N}$ with all nodes that are connected to either node i or node j based on the adjacency matrix of the approximated input graph $\hat{G}_0$. The edge structure within the subgraph $G_{i,j}$ is then reconstructed by pulling out all edges from $\hat{G}_0$ for which the parent and child node are in $\mathcal{N}$. The order of the nodes is shuffled for each subgraph. The node features in every input subgraph consist of structural node labels that are assigned by a double-radius node labelling (DRNL) heuristic[40] and the individual data features. In a first step, the distances between node i and all other nodes of the local subgraph except node j are computed. The same is repeated for node j. A hashing function then transforms the two distance labels into a DRNL label that assigns the same label to nodes that are on the same 'orbit' around the centre nodes i and j. During the training process, the DRNL label is transformed into a one-hot encoded vector and passed to the first graph convolutional layer. In contrast to traditional CNNs, GCNNs do not benefit strongly from very deep architecture design[41,42]. Therefore, our GNN tower consists only of four sequentially stacked graph convolutional layers. The activation function is defined as the hyperbolic tangent. Because the number of nodes in the enclosing subgraph for each pair of variables (i,j) is different, a SortPooling layer[43] is applied to select the top k nodes according to their structural role within the graph. Afterwards, one-dimensional convolutions extract features from the selected state embeddings.

Embedding fusion. Each tower outputs a high-dimensional embedding of the individual features found. These embeddings are concatenated and further processed by multiple fully connected layers. Finally, the last layers output the log-likelihood of a directed edge from node i to node j.

Implementation details. All network architectures are implemented in the open-source framework PyTorch[44]. The GNN is coded based on the Deep Graph Library[45]. All modules are initialized from scratch using random weights. During training, we apply an Adam-Optimizer[46] starting at an initial learning rate of $\epsilon_0 = 0.0001$. The learning rate is reduced by a factor of five once the evaluation metrics stop improving for 15 consecutive epochs. The minimum learning rate is set to $\epsilon_{\min} = 10^{-8}$. The training predictions are supervised on the binary cross-entropy loss between estimated and ground-truth edge labels. The evaluation metric is the (held-out) area under the ROC curve. Every network architecture is trained for 100 epochs. All computations are run on multiple graphics processing unit (GPU) nodes simultaneously, each equipped with eight Nvidia Tesla V100 GPUs.

Data availability

Data files are publicly available as follows. Yeast gene deletion data are from ref. 25. CRISPR perturbation data are from ref. 32. The pseudocode for data simulation is provided in Supplementary section 5.

Code availability

A Code Ocean compute capsule, which contains a pre-built compute environment and the source code of D^{2}CL, is available at https://codeocean.com/capsule/4465854/tree/v1 ref. 47.

References

1. Peters, J., Janzing, D. & Schölkopf, B. *Elements of Causal Inference: Foundations and Learning Algorithms* (MIT Press, 2017).
2. Arjovsky, M., Bottou, L., Gulrajani, I. & Lopez-Paz, D. Invariant risk minimization. Preprint at https://arxiv.org/abs/1907.02893 (2019).
3. Heinze-Deml, C., Maathuis, M. H. & Meinshausen, N. Causal structure learning. *Annu. Rev. Stat. Appl.* **5**, 371–391 (2018).
4. Spirtes, P., Glymour, C. & Scheines, R. *Causation, Prediction and Search* (MIT Press, 2000).
5. Shimizu, S., Hoyer, P. O., Hyvärinen, A. & Kerminen, A. A linear non-Gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.* **7**, 2003–2030 (2006).
6. Maathuis, M. H., Kalisch, M. & Bühlmann, P. Estimating high-dimensional intervention effects from observational data. *Ann. Stat.* **37**, 3133–3164 (2009).
7. Hauser, A. & Bühlmann, P. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *J. Mach. Learn. Res.* **13**, 2409–2464 (2012).
8. Colombo, D., Maathuis, M. H., Kalisch, M. & Richardson, T. S. Learning high-dimensional directed acyclic graphs with latent and selection variables. *Ann. Stat.* **40**, 294–321 (2012).
9. Peters, J., Bühlmann, P. & Meinshausen, N. Causal inference using invariant prediction: identification and confidence intervals. *J. R. Stat. Soc.* **78**, 947–1012 (2016).
10. Hill, S. M., Oates, C. J., Blythe, D. A. & Mukherjee, S. Causal learning via manifold regularization. *J. Mach. Learn. Res.* **20**, 127 (2019).
11. Zheng, X., Aragam, B., Ravikumar, P. K. & Xing, E. P. DAGs with no tears: continuous optimization for structure learning. In *Proc. Advance in Neural Information Processing Systems* Vol. 31, 9472–9483, (eds Bengio, S. et al.) (Curran Associates, 2018).
12. Ke, N. R. et al. Learning neural causal models from unknown interventions. Preprint at https://arxiv.org/abs/1910.01075 (2019).
13. Brouillard, P., Lachapelle, S., Lacoste, A., Lacoste-Julien, S. & Drouin, A. Differentiable causal discovery from interventional data. *Adv. Neural Inf. Process. Syst.* **33**, 21865–21877 (2020).
14. Lopez, R., Hütter, J.-C., Pritchard, J. & Regev, A. Large-scale differentiable causal discovery of factor graphs. *Adv. Neural Inf. Process. Syst.* **35**, 19290–19303 (2022).
15. Lippe, P., Cohen, T. & Gavves, E. Efficient neural causal discovery without acyclicity constraints. In *International Conference on Learning Representations* (2022).
16. Ideker, T. & Krogan, N. J. Differential network biology. *Mol. Syst. Biol.* **8**, 565 (2012).
17. Hill, S. M. et al. Inferring causal molecular networks: empirical assessment through a community-based effort. *Nat. Methods* **13**, 310–318 (2016).
18. Hill, S. M. et al. Context specificity in causal signaling networks revealed by phosphoprotein profiling. *Cell Syst.* **4**, 73–83 (2017).
19. Kuenzi, B. M. & Ideker, T. A census of pathway maps in cancer systems biology. *Nat. Rev. Cancer* **20**, 233–246 (2020).
20. Lopez-Paz, D., Muandet, K., Schölkopf, B. & Tolstikhin, I. Towards a learning theory of cause-effect inference. In *Proc. 32nd International Conference on Machine Learning* Vol. 37, 1452–1461 (eds Bach, F. et al.) (PMLR, 2015).
21. Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J. & Schölkopf, B. Distinguishing cause from effect using observational data: methods and benchmarks. *J. Mach. Learn. Res.* **17**, 1–102 (2016).
22. Noè, U., Taschler, B., Täger, J., Heutink, P. & Mukherjee, S. Ancestral causal learning in high dimensions with a human genome-wide application. Preprint at https://arxiv.org/abs/1905.11506 (2019).
23. Eigenmann, M., Mukherjee, S. & Maathuis, M. Evaluation of causal structure learning algorithms via risk estimation. In *Proc. 36th Conference of Uncertainty in Artificial Intelligence 2020, UAI 2020* Vol. 124, 151–160 (eds Peters, J. et al.) (PMLR, 2020).
24. Ke, N. R. et al. Learning to induce causal structure. Preprint at https://arxiv.org/abs/2204.04875 (2022).
25. Kemmeren, P. et al. Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell* **157**, 740–752 (2014).
26. Meinshausen, N. et al. Methods for causal inference from gene perturbation experiments and validation. *Proc. Natl Acad. Sci. USA* **113**, 7361–7368 (2016).

27. Zhang, J. Causal reasoning with ancestral graphs. *J. Mach. Learn. Res.* **9**, 1437–1474 (2008).
28. Alon, U. *An Introduction to Systems Biology: Design Principles of Biological Circuits* (CRC Press, 2019).
29. Hyttinen, A., Eberhardt, F. & Hoyer, P. O. Learning linear cyclic causal models with latent variables. *J. Mach. Learn. Res.* **13**, 3387–3439 (2012).
30. Eberhardt, F. & Scheines, R. Interventions and causal inference. *Philos. Sci.* **74**, 981–995 (2007).
31. Kocaoglu, M., Shanmugam, K. & Bareinboim, E. Experimental design for learning causal graphs with latent variables. In *Proc. Advance in Neural Information Processing Systems* Vol. 30, 7018–7028, (eds Guyon, I. et al.) (Curran Associates, 2017).
32. Replogle, J. M. et al. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell* **185**, 2559–2575 (2022).
33. Schölkopf, B. et al. On causal and anticausal learning. In *Proc. 29th International Conference on Machine Learning, ICML 2012* 459–466 (eds Langford, J. et al.) (icml.cc/Omnipress, 2012).
34. Silverman, B. W. *Density Estimation for Statistics and Data Analysis* (Chapman & Hall, 1986).
35. Turlach, B. Bandwidth selection in kernel density estimation: a review. Technical Report (1999).
36. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*) 770–778 (IEEE, 2016).
37. Szegedy, C. et al. Going deeper with convolutions. In *Proc. 2015 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*) 1–9 (IEEE, 2015).
38. He, K., Zhang, X., Ren, S. & Sun, J. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In *Proc. 2015 IEEE International Conference on Computer Vision* (*ICCV*) 1026–1034 (IEEE, 2015).
39. Xie, S., Girshick, R., Dollár, P., Tu, Z. & He, K. Aggregated residual transformations for deep neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*) 5998–5995 (IEEE, 2017).
40. Zhang, M. & Chen, Y. Link prediction based on graph neural networks. In *Proc. Advances in Neural Information Processing Systems 2018* Vol. 31, 5165–5175 (eds Bengio, S. et al.) (Curran Associates, 2018).
41. Chen, D. et al. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. *Computing Research Repository* (*CoRR*) https://doi.org/10.1609/aaai.v34i04.5747 (2019).
42. Li, Q., Han, Z. & Wu, X.-M. Deeper insights into graph convolutional networks for semi-supervised learning. In *Proc. 32nd AAAI Conference on Artificial Intelligence* 3538–3545 (eds McIlraith, S. et al.) (AAAI, 2018).
43. Zhang, M., Cui, Z., Neumann, M. & Chen, Y. An end-to-end deep learning architecture for graph classification. In *Proc. 32nd AAAI Conference on Artificial Intelligence* 4438–4445 (eds McIlraith, S. et al.) (AAAI, 2018).
44. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. In *Proc. Advances in Neural Information Processing Systems* Vol. 32, 8026–8037 (eds Wallach, H. et al.) (Curran Associates, 2019).
45. Wang, M. et al. Deep Graph Library: a graph-centric, highly-performant package for graph neural networks. Preprint at https://arxiv.org/abs/1909.01315 (2019).
46. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. In *3rd International Conference on Learning Representations* (2015).
47. Lagemann, K., Lagemann, C., Taschler, B. & Mukherjee, S. Deep learning of causal structures in high dimensions under data limitations https://codeocean.com/capsule/4465854/tree/v1*CodeOcean* (2023).

Acknowledgements

This work was partly supported by the German Federal Ministry of Education and Research (BMBF) project 'LODE', the UK Medical Research Council (MC-UU-00002/17) and the National Institute for Health Research (Cambridge Biomedical Research Centre at the Cambridge University Hospitals NHS Foundation Trust).

Author contributions

Methods were developed by K.L. and S.M. Implementation and experiments were performed by K.L., supported by C.L. B.T. contributed to the design and implementation of experiments using the baseline algorithms. The manuscript was written by K.L. and S.M., with input from C.L. and B.T. The research was supervised by S.M.

Funding

Open access funding provided by Deutsches Zentrum für Neurodegenerative Erkrankungen e.V. (DZNE) in der Helmholtz-Gemeinschaft.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-023-00744-z.

Correspondence and requests for materials should be addressed to Kai Lagemann or Sach Mukherjee.

Peer review information *Nature Machine Intelligence* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editor: Liesbeth Venema, in collaboration with the *Nature Machine Intelligence* team.