Proceedings of Machine Learning Research 275:1–40, 2025 4th Conference on Causal Learning and Reasoning

The Landscape of Causal Discovery Data: Grounding Causal Discovery in Real-World Applications

Philippe Brouillard PHILIPPEBROUILLARD@GMAIL.COM
Mila-Québec, Université de Montréal

Chandler Squires CSQUIRES@ANDREW.CMU.EDU
Carnegie Mellon University

Jonas Wahl JONAS.WAHL@DFKI.DE
Deutsches Forschungszentrum für künstliche Intelligenz (DFKI)

Konrad P. Körding KORDING@UPENN.EDU
University of Pennsylvania

Karen Sachs SACHSKAREN@GMAIL.COM
Next Generation Analytics and Modulo Bio

Alexandre Drouin[†] ALEXANDRE.DROUIN@SERVICENOW.COM
ServiceNow Research, Mila-Québec, Université Laval

Dhanya Sridhar[†] DHANYA.SRIDHAR@MILA.QUEBEC
Mila-Québec, Université de Montréal

[†] *Equal supervision*

Editors: Biwei Huang and Mathias Drton

Abstract

Causal discovery aims to automatically uncover causal relationships from data, a capability with significant potential across many scientific disciplines. However, its real-world applications remain limited. Current methods often rely on unrealistic assumptions and are evaluated only on simple synthetic toy datasets, often with inadequate evaluation metrics. In this paper, we substantiate these claims by performing a systematic review of the recent causal discovery literature. We present applications in biology, neuroscience, and Earth sciences—fields where causal discovery holds promise for addressing key challenges. We highlight available simulated and real-world datasets from these domains and discuss common assumption violations that have spurred the development of new methods. Our goal is to encourage the community to adopt better evaluation practices by utilizing realistic datasets and more adequate metrics.

Keywords: Causal discovery, evaluation metrics, real-world applications

1. Introduction

In many scientific endeavors, researchers are not merely interested in identifying statistical patterns, but in understanding the underlying causal relationships that govern complex systems. They want to answer causal questions such as "What would be the impact of changing a specific variable on this system?". This kind of question cannot be answered by purely statistical models. For instance, in healthcare, understanding causal relationships is essential to determine the efficacy of treatments leading to better patient outcomes and more efficient resource allocation. If a purely statistical model is used instead, the model might rely on spurious correlations, leading to erroneous conclusions. Causal

© 2025 P. Brouillard, C. Squires, J. Wahl, K.P. Körding, K. Sachs, A. Drouin[†] & D. Sridhar[†].

discovery aims at recovering causal relations directly from data, allowing us to answer causal queries. While causal inference is challenging, most scientific fields could benefit from that capability.

That being said, the field of causal discovery is predominantly method-driven rather than application-driven (Rolnick et al., 2024): the community produces new methods and algorithms at high speed but still relies on toy datasets and simple metrics for their evaluation (Gentzel et al., 2019), impeding its development and applicability to real-world problems. Recently, a plethora of surveys of causal discovery have covered existing causal discovery methods (Wang et al., 2024; Hasan et al., 2023; Zanga et al., 2022; Assaad et al., 2022; Vowels et al., 2022; Zhou and Chen, 2022; Nogueira et al., 2021; Guo et al., 2020; Glymour and Zhang, 2019; Malinsky and Danks, 2018; Singh et al., 2017), but none focused on the datasets and real-world applications to which these methods were applied. However, using good datasets and benchmarks is just as crucial as having good algorithms. For example, this has been pivotal in the recent deep learning boom with datasets such as ImageNet (Deng et al., 2009) and its associated challenge (Russakovsky et al., 2015). Beyond the choice of datasets, there is also a need for deeper consideration of the types of problems to which causal discovery can and should be applied. Over-reliance on simple settings makes the field disconnected from real-world challenges, and without practical applications, causal discovery risks becoming merely theoretical storytelling.

The goal of this review is to incite the community to be more application-driven: we do that by surveying the recent literature and highlighting key methodological shortcomings to be improved, as well as identifying fields that seem ripe to benefit the application of causal discovery. First, by performing a systematic review, we show in Section 4 that the field of causal discovery still relies on synthetic datasets and a low diversity of real-world datasets. Also, in most studies, inadequate metrics are used for evaluation. Second, in Section 5, we show that many alternatives exist to simple synthetic datasets, both pseudo-real and real-world datasets, and we provide a list of some common datasets that are used to assess new causal discovery methods (see our github repo). Finally, we highlight a few key scientific fields — biology, neuroscience, and Earth sciences — where a significant amount of real-world data is generated and causal discovery should be a short-term target. Overall, the resulting overview reveals that real-world applications frequently challenge established causal discovery assumptions and may serve as catalysts for innovation, underscoring the importance of grounding research in practical scenarios and utilizing real-world datasets over purely synthetic ones.

2. Background

Causal models make formal predictions about the effects of intervention, i.e., external manipulations that set a variable to some specific value or distribution. While many approaches exist to this end, this section briefly presents one specific class of causal models that is popular in the field. We detail the entailed assumptions and introduce causal discovery (for details, see Peters et al. (2017); Pearl (2009)).

Causal Bayesian Network. A *Causal Bayesian Network* (CBN) consists of a *directed acyclic graph* (DAG) $G=(V,E)$ with $|V|=d$ and a random vector $X=(X_1,\ldots,X_d)\sim P_X$ whose entries correspond to the nodes of G. The distribution P_X is connected to the graph G by the Markov property which asserts that P_X factorizes as

$$P_X = \prod_{i=1}^{d} P^{(0)}(X_i \,|\, pa_i^G), \tag{1}$$

where pa_i^G are the parents of X_i in the graph G. Up to this point, this model is a standard Bayesian network. The causal semantics stem from the interventional interpretation of the edge directions

and the fact that interventions on variables can also be considered. Let $(I_1,...,I_k)$ be a collection of interventional targets. Each interventional target $I_j \subseteq [d]$ represents a set of variables that have been intervened upon during intervention j. The distribution induced by the j-th intervention is given by

$$P_X^{(j)} = \prod_{i \notin I_j} P^{(0)}(X_i \,|\, pa_i^G) \prod_{i \in I_j} P^{(j)}(X_i \,|\, pa_i^G), \tag{2}$$

where $P^{(0)}$ are the observational conditionals that stay invariants (i.e., same as in Eq 1) and $P^{(j)}$ are conditionals intervened upon which are specific to the interventional distribution j. This is a general formulation, with perfect interventions being a notable specific case that corresponds to a setting where the conditional $P^{(j)}(X_i)$ does not depend on its parents pa_i^G.

Causal discovery. The task of recovering the graph G from a dataset $\mathcal{D}$ (possibly containing interventional data) is called causal discovery. Constraint-based methods such as the PC algorithm (Spirtes et al., 2001) perform conditional independence tests to recover G, while score-based methods achieve this by finding the graph that maximizes a score, such as the Bayesian Information Criterion (BIC) (Chickering, 2002). Some hybrid methods combine aspects of both approaches. Other methods make parametric assumptions on the functional form of the causal mechanisms or the variable distributions and orient edges based on detected asymmetries. For a more complete presentation, see for example Glymour and Zhang (2019). All of these methods are only guaranteed to recover the correct "ground-truth" graph in the infinite sample limit if the data-generating mechanism satisfies specific assumptions.

Common assumptions. Causal discovery relies on many assumptions, some directly induced by the CBN approach: 1) acyclicity of the graph over variables, 2) *causal sufficiency* which refers to the fact that there are no unobserved confounders, i.e., variables that are parents to more than one X_i, 3) the *faithfulness* assumption that stipulates that conditional independencies in the distribution P_X implies the corresponding d-separation in the graph G, and 4) the random variables provide an appropriate representation to reason about the problem of interest (Spirtes, 2009; Eberhardt, 2016). As already mentioned above, many methods also assume a particular functional form of the causal mechanisms (e.g., linearity). In practice, as we will explain in more detail in Section 5.4, most of these assumptions are violated in real-world problems.

Even when all these assumptions are satisfied, causal discovery is a hard task, both combinatorically and statistically. The space of DAGs scales super-exponentially with respect to the number of variables, and the assumptions above only guarantee correctness in the infinite sample limit, while in practice, one also has to deal with finite sample errors. Moreover, when the data is purely observational, without further assumptions one can at best identify an equivalence class of graphs, called the Markov Equivalence Class. Utilizing interventions represents the optimal strategy for overcoming obstacles in causal identifiability since it can greatly shrink the size of the equivalence class. If single-target interventions are performed on every node except one, the ground-truth graph is identifiable and, in general, fewer interventions are required (Eberhardt et al., 2005).

Evaluation. To evaluate the performance of causal discovery algorithms, there are, broadly speaking, four classes of metrics: structural, qualitative, observational, and interventional. *Structural* metrics consist of comparing the learned graph to the ground-truth graph using distances, such as the *structural Hamming distance* (SHD) which counts the total number of edges that are missing, superfluous, and reversed. *Qualitative* assessments consist of experts in the field who will discuss, based on their domain knowledge, the plausibility of some causal relations. This is similar to structural measure, but it is used when the ground-truth graph is not known. *Observational* and *interventional* metrics correspond to

evaluating how well the learned model predicts held-out observational data and data from an unseen intervention, respectively. The latter is arguably closest to what most practitioners care about: the ability to predict the effect of unseen interventions. Appendix D provides further details on these metrics.

3. From Purely Synthetic Datasets to Real-World Datasets

We first describe different families of datasets that are available for evaluating causal discovery methods before analyzing their use in recent papers.

3.1. Synthetic Datasets

We need to address the elephant in the room: many causal discovery methods are evaluated only on simple synthetic datasets that do not reflect any real-world phenomenon (Gentzel et al., 2019) (a claim that we will also demonstrate in Section 4). Moreover, many of these synthetic datasets even use exactly the same generator as the model fitted. Still, synthetic datasets are used since they offer many advantages: the ground-truth causal graph is known, a large sample size can be used and different properties of the causal model can be precisely controlled (e.g. density of the graph, number of vertices, functional form, etc) to assess a method. By design, the generated data will perfectly respect many stringent assumptions such as causal sufficiency, faithfulness, a particular functional form, etc. Moreover, a motivation for using synthetic datasets might come from the misconception that real-world datasets are scarce or impossible to evaluate quantitatively, a notion we aim to refute.

Synthetic datasets are usually generated by following these steps: first, the causal graph is sampled, then the causal mechanisms parameters, and finally, the data is sampled using ancestral sampling (i.e., by sampling the variables following their topological ordering). Common approaches to sample a DAG include the Erdős-Rényi scheme, which samples directed edges with a fixed probability, and scale-free networks (Barabási and Albert, 1999), which samples edges following a preferential attachment process (Barabási, 2009). The causal mechanisms parameters often follow a particular functional form assumed by the causal discovery method. For instance, one of the most common causal mechanisms is linear relations with Gaussian noise. Alternatives include nonlinear (Peters et al., 2014) and post-nonlinear additive noise models (Zhang and Hyvarinen, 2012) which are often used since they lead to identifiability results.

However, as highlighted in Reisach et al. (2021) and Reisach et al. (2024), the way these datasets are generated can sometimes be problematic as it introduces artifacts that some causal discovery methods may exploit. Namely, Reisach et al. (2021) showed that one can recover the causal ordering of some synthetic datasets simply by sorting the variables according to their variances. Even when these pitfalls are avoided by changing the data generating process (Andrews and Kummerfeld, 2024; Ormaniec et al., 2024), synthetic datasets are much simpler than their real-world counterpart and thus, the performance of proposed causal discovery methods is overestimated (Eigenmann et al., 2020). As we will elaborate in Section 5.4, real-world problems rarely conform to many of the assumptions built into synthetic datasets. We finish by noting that synthetic datasets can still be used for more realistic evaluation by benchmarking causal discovery methods in scenarios where these stringent assumptions are violated, as in Montagna et al. (2024).

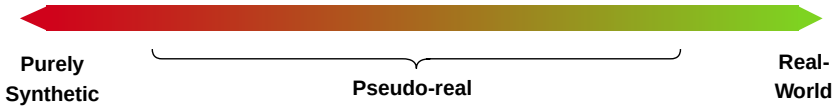

Figure 1: The realism of datasets spans a spectrum, from purely synthetic to real-world datasets.

3.2. Real-World Datasets

In the end, what we really care about is the application of causal discovery to real-world problems. Real-world datasets are particularly interesting since they often break common assumptions and inform the causal discovery community of what are remaining and interesting challenges to overcome. The primary limitation of real-world datasets, compared to synthetic and pseudo-real datasets, is the difficulty in evaluating the quality of discovered structures due to the absence of a known ground truth, making most assessments qualitative. However, when interventional data are available, we stress the fact that quantitative evaluation is possible via interventional metrics.

3.3. Pseudo-Real Datasets

Pseudo-real datasets are designed to be similar to real-world data while retaining the benefits of synthetic datasets: a known ground-truth graph, adherence to common assumptions, and control over generation parameters. As a result, some strongly advocate to use this type of data (Glymour et al., 2019).

Many pseudo-real datasets rely on a data generation process inspired by mathematical models, such as ordinary or stochastic differential equations, used in their respective fields. In biology, many simulators that generate synthetic gene expression data have been proposed: *SynTReN* (Van den Bulcke et al., 2006), *GeneNetWeaver* (Schaffter et al., 2011), *BEELINE* (Pratapa et al., 2020), *SERGIO* (Dibaeinia and Sinha, 2020). Similarly, in neuroscience, various simulators have been proposed such as simulated spiking interactions between neurons from hippocampus (Bezaire, 2015), simulated network dynamics between network areas approximated by mean-field dynamics along with fMRI signal generation and calibrated against some brain data (Smith et al., 2011). Additionally, several datasets derive from variants of the virtual brain project (Sanz Leon et al., 2013). In Earth sciences, Ebert-Uphoff and Deng (2017) simulate data that reflects typical advection and diffusion processes in the planet's atmosphere to investigate unexplained connections found by causal discovery algorithms on real-world data.

Alternatively, some pseudo-real datasets are generated by directly learning a model from real-world datasets. Once fitted, the model can produce examples similar to the original data under different conditions, with the model's graph serving as the ground truth. This often leads to datasets that, by design, respect most common causal discovery assumptions. The most popular resource of that type is the bnlearn repository (Scutari, 2009; Friedman et al., 1997) that contains several datasets from a wide range of fields. More recently, in the medical setting, Tu et al. (2019) created a simulator for neuropathic pain diagnosis. In the field of manufacturing (Vuković and Thalmann, 2022), Göbler et al. (2023) proposed a Benchmark called *Causalassembly* where a model has been fitted to real production line data. Runge et al. (2019) propose the platform *CauseMe* that contains several time-series datasets, some pseudo-real and some real with a consensus graph. Finally, Lawrence et al. (2021); Cheng et al. (2024) propose more general frameworks where several datasets from different fields can be combined for generating realistic time-series data. Simply put, their methods can be applied to any real-world dataset and yield a new simulator.

The realism of datasets can be viewed as a spectrum, with purely synthetic datasets at one end and real-world datasets at the other (see Fig. 1). Pseudo-real datasets fall in between and they can greatly vary in their realism: on one end, they can resemble purely synthetic datasets by respecting all common assumptions and integrating little information from real-world data (e.g., only the graph), at the other end, they can represent a significant improvement over synthetic datasets as they resemble real-world problems and can violate common causal assumptions. As an example of the latter, the simulator of Smith et al. (2011) generates cycles and the simulator of Tu et al. (2019) can generate data with unknown confounders, selection bias, and missing data. In short, a good simulator should faithfully replicate real-world datasets in all their complexity and, hopefully, causal discovery methods that perform well on the simulator should also transfer to real-world datasets. To do so, considering real-world problems and datasets is essential when designing pseudo-real datasets to ensure their realism and practical relevance.

4. Systematic Review of the Causal Discovery Literature

To better understand the trends in the causal discovery community regarding dataset use and evaluation metrics, we conducted a systematic literature review similar to the one of Gentzel et al. (2019). Using the Semantic Scholar API (Kinney et al., 2023), we collected scientific articles on causal discovery published between 2019 and 2024 at major machine learning conferences (NeurIPS, ICLR, ICML, AISTATS, UAI, AAAI, and CLeaR). We collected a total of 221 papers and, after manually filtering them, we retained 167 papers. A detailed presentation of our methodology, along with some additional results, can be found in Appendix A. The list of selected papers and the analysis code are available at our github repo.

Fig 2 shows the distribution of dataset types used in the selected studies. We observe that 20% of these only make use of purely synthetic datasets, while 62% rely on real-world datasets. Most real-world datasets are small, with 80% containing 20 or fewer variables. Fig 3 shows the field of provenance of pseudo-real and real-world datasets. Biology is by far the most prevalent field. This is partly explained by the ubiquitous use of the flow cytometry dataset, often simply named *Sachs* (Sachs et al., 2005). It is the only real-world dataset considered for 35% of all the papers relying on real-world datasets (we redo our analysis excluding it in Appendix A.4). For the pseudo-real datasets, the most commonly used datasets come from the bnlearn repository (Scutari, 2009; Friedman et al., 1997).

These two widely used datasets have some notable limitations. For the Sachs datasets, the consensus network used is not fully consistent with the one given in Sachs et al. (2005), in particular, the cycles are often omitted; the ground truth provided is not definitive (see Ramsey and Andrews (2018); Mooij et al. (2020)), and varies between studies. Unfortunately, the existence of a ground truth network leads to an over-reliance on structural metrics (less than 5% of studies use interventional metrics) - even though the dataset includes several interventions. Also, some studies rely on a pseudo-real version of the dataset generated by fitting a model to a consensus network (without cycles) (Scutari, 2009). Finally, the dataset is often not discriminatory for causal discovery methods: the reported performance has peaked at an SHD of around 12. For the pseudo-real datasets from bnlearn, most studies rely only on structural metrics and we note that the way the datasets are generated, all common assumptions are respected.

Table 1 summarizes the types of metrics used to evaluate performance on simulated (synthetic and pseudo-real) versus real-world datasets. Structural metrics dominate for simulated datasets, where the ground-truth graph is known (100% of studies use them). In contrast, evaluations on real-world datasets depend more heavily on qualitative assessments (36% vs. 3%). Structural metrics are also widely used for real-world datasets (67%) which can be explained by the overreliance on real-world

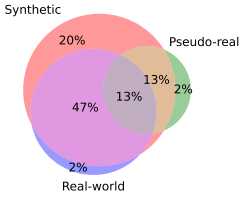

Figure 2: Distribution of papers based on used dataset types.

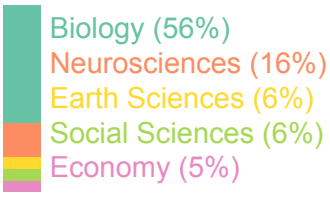

Figure 3: Common fields of the pseudo-real and real-world datasets.

	Simulated	Real-world
Structural	100.0%	67.3%
Qualitative	3.0%	36.4%
Observational	5.5%	5.6%
Interventional	9.1%	7.5%

Table 1: Percentage of studies using evaluation metrics.

datasets that contain a ground-truth graph. Both observational and interventional metrics are rare across dataset types, used in fewer than 10% of studies. Overall, most studies rely solely on structural metrics: 86% for the simulated data and 54% for the real-world data. We also note that for real-world datasets that do contain interventions, interventional metrics were not used 89% of the time.

In summary, we observed that 1) the choice of datasets could be improved to be closer to realistic settings and 2) most studies rely only on structural or qualitative evaluations. We want to emphasize that the community can readily improve its approach. First, by incorporating a broader range of real-world tasks such as some suggested in the following section. Second, by using interventional metrics as they assess the outcomes we truly care about — namely, the effects of unseen interventions.

5. Real-World Datasets: Examples and Unique Challenges

This section highlights three scientific fields—biology, neuroscience, and Earth sciences—that offer numerous real-world datasets for the causal discovery community. We outline key datasets where new methods have been applied and suggest others as promising candidates (see Appendix C for lists and links to datasets). For each field, we explain the nature of the datasets, their challenges, and some potential opportunities. Finally, we also present some works that expand causal discovery methods by tackling unique challenges of real-world datasets where most common assumptions do not hold.

5.1. Biology: Biomolecular Networks

The field of cellular biology stands out as a key area for applying causal discovery (Lagani et al., 2016; Uhler, 2024): first since biologists already analyze cellular processes like metabolism and DNA repair through the lens of networks and pathways (Pavlopoulos et al., 2011; Alberts et al., 2022), but also because it is driven by recent advances in technology that have enabled the creation of large-scale datasets, often including samples generated under interventional conditions. An enhanced understanding of these networks can shed light on development, disease, and other biological processes (Emmert-Streib et al., 2014). Many technologies now exist for inferring cellular activity, with some focusing on *messenger RNA* (mRNA) levels, others on protein levels, and some targeting entire cell populations (*bulk* methods), while newer methods allow observation of individual cells. One prevalent approach is *single-cell RNA sequencing* (scRNA-seq). Gene-editing techniques like CRISPR (Qi et al., 2013) are also of great interest, as they can naturally be framed as interventions. Technologies such as Perturb-seq (Dixit et al., 2016) combine gene perturbations with single-cell RNA-sequencing, making them ideal for generating datasets well-suited to causal discovery. We discuss in more depth the field-specific characteristics of such data in Appendix B.

Table 2: Commonly used biomolecular network datasets. *Int*, *n*, and *d* represent respectively the number of interventions, data points, and features commonly used in causal discovery papers.

Dataset	Description	Int.	n	d
Wille et al. (2004)	Gene expression microarray (*A. thaliana*)	-	118	39
Dixit et al. (2016)	Perturb-seq (bone marrow-derived dendritic cells)	8	14427	24
Replogle et al. (2022)	Perturb-seq (cell line K562)	1158	310385	8552
Replogle et al. (2022)	Perturb-seq (cell line RPE1)	651	247914	8833
Frangieh et al. (2021)	Perturb-CITE-seq (melanoma cells)	249	218331	1000
Sachs et al. (2005)	Flow cytometry (CD4+ T cells)	6	5846	11

Biomolecular datasets. In Table 2, we summarize several biological datasets that have been studied in causal discovery, with an emphasis on gene expression datasets. Prior to the development of scRNA-seq, the gene expression microarray dataset of Wille et al. (2004) was a common testbed for causal discovery and other graphical structure learning algorithms (Drton and Perlman, 2007; Bühlmann et al., 2014). Following the development of Perturb-seq, several papers have applied causal discovery to such perturbational gene expression datasets. Often, the datasets have to be preprocessed and only a small subset of the genes are used to perform causal discovery. For instance, the complete dataset of bone marrow-derived dendritic cells (BMDCs) from Dixit et al. (2016) contains over $\sim$30,000 measurements of 32,777 genes under CRISPR/Cas9 gene deletion perturbations. Following the authors' practice, researchers in causal discovery focus on 24 genes that code for highly influential transcription factors and use a shortened version of the datasets that passed a quality control (Wang et al., 2017; Yang et al., 2018; Varici et al., 2021).

More recently, Replogle et al. (2022) introduced three much more comprehensive Perturb-seq datasets, including a dataset of 2.5 million K562 cells under thousands of interventions, and two smaller datasets focused on more putatively important genes. These datasets have primarily been used for directly predicting the effects of genetic perturbations (Lopez et al., 2023; Roohani et al., 2024), but have also been considered in causal discovery (Xue et al., 2023; Lagemann et al., 2023). Indeed, Chevalley et al. (2022) use these two datasets as the basis of *CausalBench*, a benchmarking suite for causal discovery. Lastly, the Perturb-CITE-seq datasets from Frangieh et al. (2021) have been widely used in causal discovery (Lopez et al., 2022; Sethuraman et al., 2023; Rohbeck et al., 2024); these datasets also include protein expression data, which is typically ignored.

As mentioned in Section 4, the *Sachs* dataset (Sachs et al., 2005) is by far one of the most commonly used real-world datasets in the causal discovery community. Sachs et al. (2005) pioneered the use of individual cells as the smallest observational unit of intact biological systems, catapulting the available dataset size from dozens or hundreds (of mice, patients, etc) to thousands of cells. This is a flow cytometry dataset that includes abundance measurements for 11 proteins and phospholipids over 7466 CD4+ T cells exposed to nine perturbation conditions. Causal discovery algorithms are often applied to a limited version of the dataset that includes only the 5846 measurements from these seven conditions, see e.g., Wang et al. (2017), Yang et al. (2018), and Squires et al. (2020). The popularity of this dataset is partially accounted for by the fact that, in contrast with the papers above, Sachs et al. (2005) introduced a consensus network based on existing biological literature. In the absence of such a network, one often resorts to comparing against partial ground truth, e.g., as done by Frot et al. (2019), who compare against the reference database TRRUST (Han et al., 2015).

Caveats, challenges, and opportunities. We note that while graphs are prevalent in biology, the "textbook" examples are significantly different from the kinds of networks learned through causal discovery (Tejada-Lapuerta et al., 2023). In a way, causal sufficiency never holds since biologists typically conceive of networks that involve several different types of molecules, such as membrane channel proteins, enzymes, various other kinds of proteins, and RNA. Meanwhile, causal discovery methods are typically applied to datasets that contain measurements of only a single type of molecule, e.g. gene expression datasets. Thus, the networks returned by typical causal discovery algorithms only explicitly involve genes, though some recent methods are also designed to include other latent factors (Squires et al., 2022; Lopez et al., 2023). There is also an opportunity to use datasets involving more direct quantification than latent correlates such as mRNA (see Fig 6). Technologies such as CRISPR are amazing as they yield many interventional data. We note however that these interventions are often imperfect; in particular, a knocked-out gene may persist in the system, even when it was theoretically completely removed. It is not common practice in the literature to check the perturbations for efficacy, leading to potential issues with both training and validation.

5.2. Neuroscience

Neuroscience is often concerned with understanding mechanisms, which ultimately is about causality (Ross and Bassett, 2024). It distinguishes the connectome, which describes the wires – the observable physical connections between neurons or brain areas – from the effectome, which describes the causal influences between brain regions (Pospisil et al., 2024). And when it comes to causality, there is a wide spectrum of approaches, including those that assume that correlation is causation and those that ask for perturbations (Siddiqi et al., 2022). There has been growing interest in how we can uncover genuine causal relationships from neuronal recordings, establishing causal inference as a central paradigm in neuroscience research (Reid et al., 2019).

Observational data. Most causal discovery studies in neuroscience are almost entirely focused on observational data where there is no known ground truth. Most branches of neuroscience produce datasets that are used to obtain insights into causal relations. This includes spiking data (Stevenson et al., 2008), signals typically recorded at milliseconds resolution of which we currently record about 3000 simultaneously from many brain regions and that is high signal to noise (Stevenson and Körding, 2011). This includes fMRI datasets that are typically recording either about 10^4 voxels or roughly 10^2 brain areas at roughly 1Hz resolution (Smith et al., 2011). There are many other modalities including Ca2+ imaging, EEG, MEG, and fNIRS. The key is that there are plenty of datasets available and they are generally either purely observational (and without ground truth causal labels) or come from simulations. We give a list of some frequently used datasets in the field of causal discovery in Appendix C.1. We note that while there are no ground-truth graphs for most datasets, for some, we can rely on the known anatomical connectivity. For example, if there are no anatomical connections, there can not be a direct causal connection (Monti et al., 2020; Bird and Burgess, 2008).

Challenges and opportunities. There are major problems for causal inference from brain data. To start with, none of the recording methods obtains data from more than a vanishing subset of underlying variables (e.g., thousands out of many billions of neurons). As such, all observational datasets have dramatically more confounders than observed variables (Mehler and Körding, 2018). Many causal inference techniques popular in neuroscience also assume an absence of cycles (Friston et al., 2011; Zeki and Shipp, 1988) however, the existence of feedback loops is arguably a key principle of brain connectivity (Braitenberg, 1985). For a more complete list of challenges, see Ramsey et al. (2010);

Mehler and Körding (2018); Stevenson and Körding (2010); Ocker et al. (2017); Das and Fiete (2020), who list specific problems of applying causal discovery to brain signals.

Recent advancements are also paving the way for performing targeted interventions. Recently, concurrent electrical stimulation with fMRI (es-fMRI) has been proposed (Oya et al., 2017) and causal discovery, namely fGES, has been applied on such dataset (Dubois et al., 2020). Combining large-scale perturbations with transcranial magnetic stimulation (TMS) with brain imaging is an interesting avenue to acquire interventional data (Oathes et al., 2021). Electrical and optogenetic stimulation, which uses light to stimulate genetically modified neurons, is also a promising way to obtain interventional data on animal models (Stroh and Diester, 2012; Lepperød et al., 2023; Lu et al., 2024). All these studies produce interventional data allowing for a more reliable evaluation of causal discovery methods by verifying if they correctly predict the effects of perturbations.

5.3. Earth sciences

In the Earth sciences, a field in which controlled experimentation is virtually impossible, researchers rely on a mixture of observational data and physics-based simulations of varying degrees of complexity. Most data is time series or spatio-temporal data and as a consequence, time series causal discovery methods dominate the field, see Runge et al. (2019, 2023).

Reanalysis and observational data. Due to the intricacies of measuring atmospheric and surface variables across large spatial and temporal scales (e.g., irregular measurement locations or measurement times, meteorological conditions affecting remote sensing capabilities), most studies involving causal discovery in the Earth sciences do not use purely observational data. Instead, the most commonly used type of data, in particular for atmospheric variables, is *reanalysis data*. Reanalysis data is imputed by fitting observations to numerical meteorological prediction models and is thus pseudo-real in the sense of Section 3.3. There are several large reanalysis projects led by national research institutes that make reanalysis data available to the public, including the NCEP/NCAR 40-year reanalysis project and the ERA reanalysis project (Hersbach et al., 2020). These databases contain a wide range of atmospheric parameters such as temperature, humidity, pressure, and wind speed direction (Kalnay et al., 2018). Runge et al. (2019) discuss some of the general challenges of these datasets: strong autocorrelation, time delays, time aggregation, unobserved variables, and more. Examples of causal discovery applications on reanalysis data include Kretschmer et al. (2017); Saranya Ganesh et al. (2023); Iglesias-Suarez et al. (2024) in which causal discovery is used as a feature selection pre-processing step for downstream prediction tasks and neural network parameter selection. Kretschmer et al. (2018) investigate interactions between global modes of climate variability in the Earth system, so-called teleconnections, using ERA reanalysis data. Di Capua et al. (2020b) combine reanalysis data with climate indices available in the KNMI Climate Explorer to investigate teleconnections in boreal summer; see also Saggioro et al. (2020) for another causal discovery application to teleconnections using climate indices. Di Capua et al. (2019, 2020a) use causal discovery on observational data from the Climate Prediction Center (CPC) global rainfall dataset as well as ERA reanalysis surface temperature data to examine causal drivers of Indian summer monsoon rainfall. Engelke and Hitz (2020); Améndola et al. (2021); Tran et al. (2024) apply causal discovery methods targeting extreme events to a river flow network dataset.

In Environmental Science, causal discovery has been applied in Krich et al. (2021, 2022) to atmospheric flux data from the FLUXNET dataset (Pastorello et al., 2020). FLUXNET contains measurements of carbon, water vapor and energy exchange in different regions of the planet. Guo et al. (2024)

investigate the influence of ozone levels on influenza with three causal discovery approaches, using data from the Tropospheric Ozone Assessment Report (TOAR) database and the CDC Influenza report.

Physics-based model data. In addition to direct observations and reanalysis data, climate scientists employ large-scale global or regional climate models to simulate interventions, most notably to investigate global warming under different carbon emission scenarios. Global climate models are coordinated within the Climate Model Intercomparison Project (CMIP), currently in its 6th phase (Eyring et al., 2016). However, these simulators are so computationally demanding that it can take months to run a single simulation (Balaji et al., 2017), making it hard to simulate an abundance of interventional data. Additionally, while there is a huge amount of data that has been produced by climate model runs, different datasets are often inconsistent (e.g., due to a different space or time resolution) and may be hard to retrieve. Recently, an effort has been made to make curated versions of these datasets available (Watson-Parris et al., 2022; Kaltenborn et al., 2024). Applications of causal discovery to CMIP data include Karmouche et al. (2024) who compare the output of the causal discovery method PCMCI+ across different climate models and Nowack et al. (2020) who investigate whether CMIP6 models whose causal discovery output graphs are similar to the graph found on reanalysis data exhibit better performance on a downstream prediction task. Simpler data simulators for climate-specific causal discovery that are faster to run but far less detailed have been developed in Ebert-Uphoff and Deng (2017) and Tibau et al. (2022).

Evaluation of causal discovery output graphs. Due to the unfeasibility of interventions, it is usually impossible to directly validate the output of a causal discovery method. In addition, as in almost all real-world applications causal discovery assumptions are almost certainly violated, and the degree of violation is often difficult to estimate. Therefore, Earth scientists resort to softer plausibility criteria, for instance by asking whether the returned network is consistent with physical laws. Sometimes more than one causal discovery algorithm is applied to verify whether conclusions are consistent across methods, e.g. in Guo et al. (2024). As Earth scientists are well aware that such validations need to be handled with care due to the danger of confirmation bias, causal discovery is predominantly used for feature selection (Kretschmer et al., 2017; Saranya Ganesh et al., 2023; Iglesias-Suarez et al., 2024) or model comparison (Nowack et al., 2020; Karmouche et al., 2024).

5.4. Challenges of Real-World Datasets

In this section, we highlight several causal discovery works that have been designed specifically to answer challenges arising in the field of biology and neuroscience. By exploring these works, we aim to illustrate how the violation of standard assumptions can drive innovation, offering insights that purely synthetic or pseudo-real datasets alone might not provide.

High-dimensionality. Real-world datasets often present a high number of features. For instance, brain imaging datasets can contain tens of thousands of features corresponding to individual voxels. Ramsey et al. (2017) proposed fGES, a modification of the popular score-based method GES that assumes linearity, that can scale to a million variables. Gene regulatory network (GRN) datasets, which frequently encompass the entire human genome with around 20,000 features, pose a similar computational challenge, especially for nonlinear causal discovery methods. This is why most applications usually focus on a much smaller subset of genes (often less than a hundred). Recently, a few works have focused on adapting existing nonlinear methods to scale to a much higher number of features (of the order of thousands) (Lee et al., 2019; Lopez et al., 2022). One specificity of the data that can be leveraged is its modularity. Gene regulatory networks form modules or programs of genes that act together. Segal et al. (2005); Lopez et al. (2022) have used this prior to learning more efficiently causal structures.

Heterogeneity. The heterogeneity of biological data often necessitates the integration of multiple datasets to achieve a comprehensive understanding of the underlying biological processes. To address this, Triantafillou and Tsamardinos (2015) and Huang et al. (2020) introduced methods for combining datasets that share a subset of variables, allowing for the leveraging of complementary information across datasets. The heterogeneity can also arise from datasets generated under different populations, such as cell types or disease states. Recognizing this, researchers have proposed methods to model biological data as a mixture of DAGs, each representing a distinct causal structure corresponding to a specific population (Saeed et al., 2020b). Finally, brain imaging datasets are often collected from a cohort of subjects. Although there are strong shared connectivities across the subjects (Damoiseaux et al., 2006), each subject also exhibits unique brain connectivity patterns. Exploring methods to conduct multisubject analyses presents a compelling research challenge that has been explored in Oates et al. (2014, 2016), Monti and Hyvärinen (2018), and Huang et al. (2019).

Cyclic models. While GRNs and brain connectivity networks contain undoubtedly feedback loops (Ferrell Jr, 2013), most causal discovery methods assume acyclicity. Recent works motivated by GRNs (Rohbeck et al., 2024; Sethuraman et al., 2023; Sethuraman and Fekri, 2024) and by the brain examples (Sanchez-Romero et al., 2019) have continued the exploration of cyclic causal models.

Off-target interventions. While a gene knockout is usually considered as an intervention targeting a specific gene, in reality, gene knockouts exhibit off-target effects (Fu et al., 2013). In causality terms, this phenomenon is called fat-hand interventions and has been investigated in different biological contexts (Eaton and Murphy, 2007; Choo et al., 2024).

Measurement error. Technologies such as scRNA-seq can fail to detect some RNA at low levels and will report mistakenly many expression levels at zeros (a phenomenon called dropout) (Hicks et al., 2018). Saeed et al. (2020a); Ke et al. (2023); Dai et al. (2024) have proposed causal discovery methods that take into account this type of measurement error.

6. Conclusion

We systematically surveyed recent work in causal discovery research, focusing on datasets and evaluations used in these studies. Our findings reveal that not much has changed since the study of Gentzel et al. (2019), indicating that the time is well overdue for a critical change in the field. Most studies still only use structural metrics instead of interventional ones. Several studies only include synthetic datasets and while several do include real-world datasets, they often rely on the same ones which have some major limitations. Furthermore, most causal discovery methods rely on strong assumptions that real-world datasets rarely satisfy. Overall, causal discovery still has considerable progress to make before it can be directly applied; practitioners tend to be aware of its limitations and they employ it pragmatically, for instance as an exploratory tool, rather than as a means to derive an irrevocable causal truth. Finally, although we focused on causal discovery, in Appendix E we discuss how similar problems are also present in the emerging field of causal representation learning where simple toy datasets are mostly used and where the common assumptions of the field probably don't hold in real-world settings. We offer recommendations and urge researchers in this field to also use more realistic datasets.

We also explored in more detail the real-world datasets used in causal discovery. A key observation from our exploration is the increased availability of these kinds of datasets, alongside a trend towards larger and more detailed real-world datasets in recent years. In the field of biology, biomolecular network datasets contain even more interventions than before thanks to new technological advances. These datasets present an invaluable opportunity for the advancement of causal discovery and could

also be used in tandem with optimal experimental design as explored in Cho et al. (2016); Ness et al. (2017); Agrawal et al. (2019); Tigas et al. (2022); Zhang et al. (2023). Additionally, we showed that real-world domains provide a fertile ground for pushing the boundaries of causal discovery methods since they challenge existing assumptions.

Our conclusion in recommending the use of empirical datasets echoes the one from Gentzel et al. (2019). To be clear, synthetic datasets are useful, but they should be complemented by more realistic evaluations on pseudo-real and real-world datasets. When interventional data are present, good quantitative evaluation on real-world datasets exists. However, in many fields besides biology, interventional data are hard to come by and thus pseudo-real datasets might be more adequate. They conserve most of the synthetic datasets' advantages while being more realistic. Still, the creation of pseudo-real datasets should always remain grounded by considering real-world datasets and the assumptions they violate. Through this review, which compiles an extensive list of both simulators and empirical datasets, we aim to motivate researchers to diversify their dataset usage, moving beyond the confines of synthetic data to embrace the complexity and richness of the real world in their causal discovery endeavors. We hope this effort will also foster the collaboration between causal discovery researchers and practitioners, leading to even more relevant real-world datasets and better integration of domain knowledge.

Acknowledgments

PB acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) and acknowledges Assya Trofimov for helpful discussions. JW received funding from the European Research Council (ERC) Starting Grant CausalEarth (Grant Agreement No. 948112) and from the German Federal Ministry of Education and Research (BMBF) as part of the project MAC-MERLin (Grant Agreement No. 01IW24007). CS received funding from Valence Labs and acknowledges Jason Hartford for helpful discussions. AD acknowledges Sara Magliacane for helpful discussions. KS was funded in part by NIMH grant 1R44MH135465. DS acknowledges support from NSERC Discovery Grant RGPIN-2023-04869, and a Canada-CIFAR AI Chair.

References

Silvia Acid and Luis M de Campos. Searching for bayesian network structures in the space of restricted acyclic partially directed graphs. *Journal of artificial intelligence research*, 18:445–490, 2003.

Raj Agrawal, Chandler Squires, Karren Yang, Karthikeyan Shanmugam, and Caroline Uhler. Abcd-strategy: Budgeted experimental design for targeted causal structure discovery. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3400–3409. PMLR, 2019.

Ossama Ahmed, Frederik Träuble, Anirudh Goyal, Alexander Neitz, Manuel Wuthrich, Yoshua Bengio, Bernhard Schölkopf, and Stefan Bauer. Causalworld: A robotic manipulation benchmark for causal structure and transfer learning. In *International Conference on Learning Representations*, 2021.

Kartik Ahuja, Jason Hartford, and Yoshua Bengio. Weakly supervised representation learning with sparse perturbations. In *Advances in Neural Information Processing Systems*, 2022.

B. Alberts, R. Heald, A. Johnson, D. Morgan, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell (Seventh Edition)*. W. W. Norton, Incorporated, 2022. ISBN 9780393884647.

Carlos Améndola, Benjamin Hollering, Seth Sullivant, and Ngoc Tran. Markov equivalence of max-linear Bayesian networks. In Cassio de Campos and Marloes H. Maathuis, editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 1746–1755. PMLR, 27–30 Jul 2021.

Tara V Anand, Adele H Ribeiro, Jin Tian, and Elias Bareinboim. Causal effect identification in cluster dags. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12172–12179, 2023.

Bryan Andrews and Erich Kummerfeld. Better simulations for validating causal discovery with the dag-adaptation of the onion method. *arXiv preprint arXiv:2405.13100*, 2024.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Charles K Assaad, Emilie Devijver, and Eric Gaussier. Survey and evaluation of causal discovery methods for time series. *Journal of Artificial Intelligence Research*, 73:767–819, 2022.

K Bach and M Lichman. Uci machine learning repository (2013) university of california. *School of Information and Computer Science*.

Venkatramani Balaji, Eric Maisonnave, Niki Zadeh, Bryan N Lawrence, Joachim Biercamp, Uwe Fladrich, Giovanni Aloisio, Rusty Benson, Arnaud Caubel, Jeffrey Durachta, et al. Cpmip: measurements of real computational performance of earth system models in cmip6. *Geoscientific Model Development*, 10(1):19–34, 2017.

Albert-László Barabási. Scale-free networks: a decade and beyond. *science*, 325(5939):412–413, 2009.

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286 (5439):509–512, 1999.

Sander Beckers and Joseph Y Halpern. Abstracting causal models. In *Proceedings of the aaai conference on artificial intelligence*, volume 33, pages 2678–2685, 2019.

Sander Beckers, Frederick Eberhardt, and Joseph Y Halpern. Approximate causal abstractions. In *Uncertainty in artificial intelligence*, pages 606–615. PMLR, 2020.

Marianne Bezaire. *Modeling physiological oscillations in a biologically constrained CA1 network from two perspectives: full-scale parallel network and rationally reduced Network Clamp*. University of California, Irvine, 2015.

Chris M Bird and Neil Burgess. The hippocampus and memory: insights from spatial processing. *Nature reviews neuroscience*, 9(3):182–194, 2008.

Valentino Braitenberg. Charting the visual cortex. In *Cerebral Cortex 3: Visual Cortex*, pages 379–414. Plenum Press, 1985.

Jessica Brandi, Roberta Noberini, Tiziana Bonaldi, and Daniela Cecconi. Advances in enrichment methods for mass spectrometry-based proteomics analysis of post-translational modifications. *Journal of Chromatography A*, 1678:463352, 2022. ISSN 0021-9673.

Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco S Cohen. Weakly supervised causal representation learning. *Advances in Neural Information Processing Systems*, 35:38319–38331, 2022.

Philippe Brouillard, Sébastien Lachapelle, Julia Kaltenborn, Yaniv Gurwicz, Dhanya Sridhar, Alexandre Drouin, Peer Nowack, Jakob Runge, and David Rolnick. Causal representation learning in temporal data via single-parent decoding. *arXiv preprint arXiv:2410.07013*, 2024.

Peter Bühlmann, Jonas Peters, and Jan Ernest. Cam: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014.

Barbara M Byrne. *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. routledge, 2013.

Riccardo Cadei, Lukas Lindorfer, Sylvia Cremer, Cordelia Schmid, and Francesco Locatello. Smoke and mirrors in causal downstream tasks. *Advances in Neural Information Processing Systems*, 37: 26082–26112, 2025.

Krzysztof Chalupka, Pietro Perona, and Frederick Eberhardt. Visual causal feature learning. *arXiv preprint arXiv:1412.2309*, 2014.

Krzysztof Chalupka, Frederick Eberhardt, and Pietro Perona. Causal feature learning: an overview. *Behaviormetrika*, 44:137–164, 2017.

Yuxiao Cheng, Ziqian Wang, Tingxiong Xiao, Qin Zhong, Jinli Suo, and Kunlun He. Causaltime: Realistically generated time-series for benchmarking of causal discovery. In *The Twelfth International Conference on Learning Representations*, 2024.

Mathieu Chevalley, Yusuf Roohani, Arash Mehrjou, Jure Leskovec, and Patrick Schwab. Causalbench: A large-scale benchmark for network inference from single-cell perturbation data. *arXiv preprint arXiv:2210.17283*, 2022.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

Hyunghoon Cho, Bonnie Berger, and Jian Peng. Reconstructing causal biological networks through active learning. *PloS one*, 11(3):e0150611, 2016.

Davin Choo, Kirankumar Shiragur, and Caroline Uhler. Causal discovery under off-target interventions. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 1621–1629. PMLR, 02–04 May 2024.

Haoyue Dai, Ignavier Ng, Gongxu Luo, Peter Spirtes, Petar Stojanov, and Kun Zhang. Gene regulatory network inference in the presence of dropouts: a causal view. In *The Twelfth International Conference on Learning Representations*, 2024.

Jessica S Damoiseaux, Serge ARB Rombouts, Frederik Barkhof, Philip Scheltens, Cornelis J Stam, Stephen M Smith, and Christian F Beckmann. Consistent resting-state networks across healthy subjects. *Proceedings of the national academy of sciences*, 103(37):13848–13853, 2006.

Abhranil Das and Ila R Fiete. Systematic errors in connectivity inferred from activity in strongly recurrent networks. *Nature Neuroscience*, 23(10):1286–1296, 2020.

Christian Schroeder de Witt, Catherine Tong, Valentina Zantedeschi, Daniele De Martini, Alfredo Kalaitzis, Matthew Chantry, Duncan Watson-Parris, and Piotr Bilinski. Rainbench: Towards data-driven global precipitation forecasting from satellite imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14902–14910, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

G Di Capua, M Kretschmer, J Runge, A Alessandri, RV Donner, B van den Hurk, R Vellore, R Krishnan, and D Coumou. Long-lead statistical forecasts of the indian summer monsoon rainfall based on causal precursors. *Weather and Forecasting*, 34(5):1377, 2019.

G. Di Capua, M. Kretschmer, R. V. Donner, B. van den Hurk, R. Vellore, R. Krishnan, and D. Coumou. Tropical and mid-latitude teleconnections interacting with the indian summer monsoon rainfall: a theory-guided causal effect network approach. *Earth System Dynamics*, 11(1):17–34, 2020a. doi: 10.5194/esd-11-17-2020.

Giorgia Di Capua, Jakob Runge, Reik V Donner, Bart van den Hurk, Andrew G Turner, Ramesh Vellore, Raghavan Krishnan, and Dim Coumou. Dominant patterns of interaction between the tropics and mid-latitudes in boreal summer: Causal relationships and the role of time-scales. *Weather and Climate Dynamics Discussions*, 2020:1–28, 2020b.

Adriana Di Martino, Chao-Gan Yan, Qingyang Li, Erin Denio, Francisco X Castellanos, Kaat Alaerts, Jeffrey S Anderson, Michal Assaf, Susan Y Bookheimer, Mirella Dapretto, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular psychiatry*, 19(6):659–667, 2014.

Payam Dibaeinia and Saurabh Sinha. Sergio: a single-cell expression simulator guided by gene regulatory networks. *Cell systems*, 11(3):252–271, 2020.

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P Fulco, Livnat Jerby-Arnon, Nemanja D Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, et al. Perturb-seq: dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *cell*, 167(7): 1853–1866, 2016.

Sven Dorkenwald, Claire E McKellar, Thomas Macrina, Nico Kemnitz, Kisuk Lee, Ran Lu, Jingpeng Wu, Sergiy Popovych, Eric Mitchell, Barak Nehoran, et al. Flywire: online community for whole-brain connectomics. *Nature methods*, 19(1):119–128, 2022.

Mathias Drton and Michael D Perlman. Multiple testing and error control in gaussian graphical model selection. *Statistical Science*, 22(3):430–449, 2007.

Julien Dubois, Hiroyuki Oya, J Michael Tyszka, Matthew Howard III, Frederick Eberhardt, and Ralph Adolphs. Causal mapping of emotion networks in the human brain: Framework and initial findings. *Neuropsychologia*, 145:106571, 2020.

Daniel Eaton and Kevin Murphy. Exact bayesian structure learning from uncertain interventions. In *Artificial intelligence and statistics*, pages 107–114. PMLR, 2007.

Frederick Eberhardt. Green and grue causal variables. *Synthese*, 193:1029–1046, 2016.

Frederick Eberhardt, Clark Glymour, and Richard Scheines. On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, UAI'05, page 178–184. AUAI Press, 2005.

Imme Ebert-Uphoff and Yi Deng. Causal discovery in the geosciences—using synthetic data to learn how to interpret results. *Computers & geosciences*, 99:50–60, 2017.

Marco Eigenmann, Sach Mukherjee, and Marloes Maathuis. Evaluation of causal structure learning algorithms via risk estimation. In *Conference on Uncertainty in Artificial Intelligence*, pages 151–160. PMLR, 2020.

Frank Emmert-Streib, Matthias Dehmer, and Benjamin Haibe-Kains. Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks. *Frontiers in cell and developmental biology*, 2:38, 2014.

Sebastian Engelke and Adrien S. Hitz. Graphical models for extremes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4):871–932, 2020. doi: https://doi.org/10.1111/rssb.12355.

Elias Eulig, Atalanti A Mastakouri, Patrick Blöbaum, Michaela Hardt, and Dominik Janzing. Toward falsifying causal graphs using a permutation-based test. *arXiv preprint arXiv:2305.09565*, 2023.

Veronika Eyring, Sandrine Bony, Gerald A Meehl, Catherine A Senior, Bjorn Stevens, Ronald J Stouffer, and Karl E Taylor. Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Development*, 9(5):1937–1958, 2016.

James E Ferrell Jr. Feedback loops and reciprocal regulation: recurring motifs in the systems biology of the cell cycle. *Current opinion in cell biology*, 25(6):676–686, 2013.

Chris J Frangieh, Johannes C Melms, Pratiksha I Thakore, Kathryn R Geiger-Schuller, Patricia Ho, Adrienne M Luoma, Brian Cleary, Livnat Jerby-Arnon, Shruti Malu, Michael S Cuoco, et al. Multimodal pooled perturb-cite-seq screens in patient models define mechanisms of cancer immune evasion. *Nature genetics*, 53(3):332–341, 2021.

Nir Friedman, Moises Goldszmidt, David Heckerman, and Stuart Russell. Where is the impact of bayesian networks in learning. In *International Joint Conference on Artificial Intelligence*. Citeseer, 1997.

Karl J Friston, Baojuan Li, Jean Daunizeau, and Klaas E Stephan. Network discovery with dcm. *Neuroimage*, 56(3):1202–1221, 2011.

Benjamin Frot, Preetam Nandy, and Marloes H Maathuis. Robust causal structure learning with some hidden variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81 (3):459–487, 2019.

Yanfang Fu, Jennifer A Foden, Cyd Khayter, Morgan L Maeder, Deepak Reyon, J Keith Joung, and Jeffry D Sander. High-frequency off-target mutagenesis induced by crispr-cas nucleases in human cells. *Nature biotechnology*, 31(9):822–826, 2013.

Juan L Gamella, Jonas Peters, and Peter Bühlmann. Causal chambers as a real-world physical testbed for ai methodology. *Nature Machine Intelligence 7, 107–118*, 2025.

Dan Garant and David Jensen. Evaluating causal models by comparing interventional distributions. *arXiv preprint arXiv:1608.04698*, 2016.

Josh Gardner, Zoran Popovic, and Ludwig Schmidt. Benchmarking distribution shift in tabular data with tableshift. *Advances in Neural Information Processing Systems*, 36, 2024.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34:9574–9586, 2021.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks. In *International Conference on Machine Learning*, pages 7324–7338. PMLR, 2022.

Amanda Gentzel, Dan Garant, and David Jensen. The case for evaluating causal models using interventional measures and empirical data. *Advances in Neural Information Processing Systems*, 32, 2019.

Clark Glymour and Kun Zhang. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:418407, 2019.

Clark Glymour, Joseph D Ramsey, and Kun Zhang. The evaluation of discovery: Models, simulation and search through "big data". *Open Philosophy*, 2(1):39–48, 2019.

Konstantin Göbler, Tobias Windisch, Tim Pychynski, Steffen Sonntag, Martin Roth, and Mathias Drton. causalassembly: Generating realistic production data for benchmarking causal discovery. *CoRR*, 2023.

Alex Greenfield, Aviv Madar, Harry Ostrer, and Richard Bonneau. Dream4: Combining genetic and dynamic information to identify biological networks and dynamical models. *PloS one*, 5(10): e13397, 2010.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

Fang Guo, Pei Zhang, Vivian Do, Jakob Runge, Kun Zhang, Zheshen Han, Shenxi Deng, Hongli Lin, Sheikh Taslim Ali, Ruchong Chen, et al. Ozone as an environmental driver of influenza. *Nature Communications*, 15(1):3763, 2024.

Ruocheng Guo, Lu Cheng, Jundong Li, P Richard Hahn, and Huan Liu. A survey of learning causality with data: Problems and methods. *ACM Computing Surveys (CSUR)*, 53(4):1–37, 2020.

Alexander Hägele, Jonas Rothfuss, Lars Lorch, Vignesh Ram Somnath, Bernhard Schölkopf, and Andreas Krause. Bacadi: Bayesian causal discovery with unknown interventions. In *International Conference on Artificial Intelligence and Statistics*, pages 1411–1436. PMLR, 2023.

Heonjong Han, Hongseok Shim, Donghyun Shin, Jung Eun Shim, Yunhee Ko, Junha Shin, Hanhae Kim, Ara Cho, Eiru Kim, Tak Lee, et al. Trrust: a reference database of human transcriptional regulatory interactions. *Scientific reports*, 5(1):11432, 2015.

Uzma Hasan, Emam Hossain, and Md Osman Gani. A survey on causal discovery methods for iid and time series data. *Transactions on Machine Learning Research*, 2023.

Leonard Henckel, Theo Würtzen, and Sebastian Weichwald. Adjustment identification distance: A gadjid for causal structure learning. In Negar Kiyavash and Joris M. Mooij, editors, *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pages 1569–1598. PMLR, 15–19 Jul 2024.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

Stephanie C Hicks, F William Townes, Mingxiang Teng, and Rafael A Irizarry. Missing data and technical variability in single-cell rna-sequencing experiments. *Biostatistics*, 19(4):562–578, 2018.

Biwei Huang, Kun Zhang, Yizhu Lin, Bernhard Schölkopf, and Clark Glymour. Generalized score functions for causal discovery. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1551–1560, 2018.

Biwei Huang, Kun Zhang, Pengtao Xie, Mingming Gong, Eric P Xing, and Clark Glymour. Specific and shared causal relation modeling and mechanism-based clustering. *Advances in Neural Information Processing Systems*, 32, 2019.

Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21(89):1–53, 2020.

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999.

Fernando Iglesias-Suarez, Pierre Gentine, Breixo Solino-Fernandez, Tom Beucler, Michael Pritchard, Jakob Runge, and Veronika Eyring. Causally-informed deep learning to improve climate models and projections. *Journal of Geophysical Research: Atmospheres*, 129(4):e2023JD039202, 2024.

Azam Ikram, Sarthak Chakraborty, Subrata Mitra, Shiv Saini, Saurabh Bagchi, and Murat Kocaoglu. Root cause analysis of failures in microservices through causal discovery. *Advances in Neural Information Processing Systems*, 35:31158–31170, 2022.

Eric Jonas and Konrad Paul Körding. Could a neuroscientist understand a microprocessor? *PLoS computational biology*, 13(1):e1005268, 2017.

Eugenia Kalnay, Masao Kanamitsu, Robert Kistler, William Collins, Dennis Deaven, Lev Gandin, Mark Iredell, Suranjana Saha, Glenn White, John Woollen, et al. The ncep/ncar 40-year reanalysis project. In *Renewable energy*, pages Vol1_146–Vol1_194. Routledge, 2018.

Julia Kaltenborn, Charlotte Lange, Venkatesh Ramesh, Philippe Brouillard, Yaniv Gurwicz, Chandni Nagda, Jakob Runge, Peer Nowack, and David Rolnick. Climateset: A large-scale climate model dataset for machine learning. *Advances in Neural Information Processing Systems*, 36, 2024.

S. Karmouche, E. Galytska, G. A. Meehl, J. Runge, K. Weigel, and V. Eyring. Changing effects of external forcing on atlantic–pacific interactions. *Earth System Dynamics*, 15(3):689–715, 2024. doi: 10.5194/esd-15-689-2024.

Nan Rosemary Ke, Aniket Rajiv Didolkar, Sarthak Mittal, Anirudh Goyal, Guillaume Lajoie, Stefan Bauer, Danilo Jimenez Rezende, Yoshua Bengio, Christopher Pal, and Michael Curtis Mozer. Systematic evaluation of causal discovery in visual model based reinforcement learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Nan Rosemary Ke, Sara-Jane Dunn, Jorg Bornschein, Silvia Chiappa, Melanie Rey, Jean-Baptiste Lespiau, Albin Cassirer, Jane Wang, Theophane Weber, David Barrett, et al. Discogen: Learning to discover gene regulatory networks. *arXiv preprint arXiv:2304.05823*, 2023.

R. Kinney, C. Anastasiades, R. Authur, I. Beltagy, J. Bragg, A. Buraczynski, I. Cachola, S. Candra, Y. Chandrasekhar, A. Cohan, and M. Crawford. The semantic scholar open data platform. 2023.

Marlene Kretschmer, Jakob Runge, and Dim Coumou. Early prediction of extreme stratospheric polar vortex states based on causal precursors. *Geophysical research letters*, 44(16):8592–8600, 2017.

Marlene Kretschmer, Judah Cohen, Vivien Matthias, Jakob Runge, and Dim Coumou. The different stratospheric influence on cold-extremes in eurasia and north america. *npj Climate and Atmospheric Science*, 1(1):44, 2018.

Christopher Krich, Mirco Migliavacca, Diego G Miralles, Guido Kraemer, Tarek S El-Madany, Markus Reichstein, Jakob Runge, and Miguel D Mahecha. Functional convergence of biosphere–atmosphere interactions in response to meteorological conditions. *Biogeosciences*, 18(7):2379–2404, 2021.

Christopher Krich, Miguel D Mahecha, Mirco Migliavacca, Martin G De Kauwe, Anne Griebel, Jakob Runge, and Diego G Miralles. Decoupling between ecosystem photosynthesis and transpiration: a last resort against overheating. *Environmental Research Letters*, 17(4):044013, 2022.

Sébastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ica. In *First Conference on Causal Learning and Reasoning*, 2022.

Sébastien Lachapelle, Tristan Deleu, Divyat Mahajan, Ioannis Mitliagkas, Yoshua Bengio, Simon Lacoste-Julien, and Quentin Bertrand. Synergies between disentanglement and sparsity: Generalization and identifiability in multi-task learning. *International Conference on Machine Learning*, 2023.

Vincenzo Lagani, Sofia Triantafillou, Gordon Ball, Jesper Tegnér, and Ioannis Tsamardinos. Probabilistic computational causal discovery for systems biology. *Uncertainty in biology: a computational modeling approach*, pages 33–73, 2016.

Kai Lagemann, Christian Lagemann, Bernd Taschler, and Sach Mukherjee. Deep learning of causal structures in high dimensions under data limitations. *Nature Machine Intelligence*, 5(11): 1306–1316, 2023.

Andrew R Lawrence, Marcus Kaiser, Rui Sampaio, and Maksim Sipos. Data generating process to evaluate causal discovery techniques for time series data. *arXiv preprint arXiv:2104.08043*, 2021.

Hao-Chih Lee, Matteo Danieletto, Riccardo Miotto, Sarah T Cherng, and Joel T Dudley. Scaling structural learning with no-bears to infer causal transcriptome networks. In *Pacific Symposium on Biocomputing 2020*, pages 391–402. World Scientific, 2019.

Mikkel Elle Lepperød, Tristan Stöber, Torkel Hafting, Marianne Fyhn, and Konrad Paul Körding. Inferring causal connectivity from pairwise recordings and optogenetics. *PLoS Computational Biology*, 19(11):e1011574, 2023.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Efstratios Gavves. Causal representation learning for instantaneous and temporal effects in interactive systems. In *The Eleventh International Conference on Learning Representations*, 2022a.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Stratis Gavves. Citris: Causal identifiability from temporal intervened sequences. In *International Conference on Machine Learning*, pages 13557–13603. PMLR, 2022b.

Chenxi Liu, Yongqiang Chen, Tongliang Liu, Mingming Gong, James Cheng, Bo Han, and Kun Zhang. Discovery of the hidden world with large language models. *arXiv preprint arXiv:2402.03941*, 2024.

Yuejiang Liu, Alexandre Alahi, Chris Russell, Max Horn, Dominik Zietlow, Bernhard Schölkopf, and Francesco Locatello. Causal triplet: An open challenge for intervention-centric causal representation learning. *arXiv preprint arXiv:2301.05169*, 2023.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning*, 2020.

Christopher Lohse and Jonas Wahl. Sortability of time series data. *arXiv preprint arXiv:2407.13313*, 2024.

Romain Lopez, Jan-Christian Hütter, Jonathan Pritchard, and Aviv Regev. Large-scale differentiable causal discovery of factor graphs. *Advances in Neural Information Processing Systems*, 35: 19290–19303, 2022.

Romain Lopez, Natasa Tagasovska, Stephen Ra, Kyunghyun Cho, Jonathan Pritchard, and Aviv Regev. Learning causal representations of single cells via sparse mechanism shift modeling. In *Conference on Causal Learning and Reasoning*, pages 662–691. PMLR, 2023.

Dian Lu, James Stieger, Zoe Lusk, Vivek Buch, and Josef Parvizi. Causal cortical and thalamic connections in the human brain. *bioRxiv*, pages 2024–06, 2024.

Daniel Malinsky and David Danks. Causal discovery algorithms: A practical guide. *Philosophy Compass*, 13(1):e12470, 2018.

Daniel Marbach, Robert J Prill, Thomas Schaffter, Claudio Mattiussi, Dario Floreano, and Gustavo Stolovitzky. Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the national academy of sciences*, 107(14):6286–6291, 2010.

Riccardo Massidda, Atticus Geiger, Thomas Icard, and Davide Bacciu. Causal abstraction with soft interventions. In *Conference on Causal Learning and Reasoning*, pages 68–87. PMLR, 2023.

David Marc Anton Mehler and Konrad Paul Körding. The lure of misleading causal statements in functional connectivity research. *arXiv preprint arXiv:1812.03363*, 2018.

Francesco Montagna, Atalanti Mastakouri, Elias Eulig, Nicoletta Noceti, Lorenzo Rosasco, Dominik Janzing, Bryon Aragam, and Francesco Locatello. Assumption violations in causal discovery and the robustness of score matching. *Advances in Neural Information Processing Systems*, 36, 2024.

Ricardo Pio Monti and Aapo Hyvärinen. A unified probabilistic model for learning latent factors and their connectivities from high-dimensional data. *arXiv preprint arXiv:1805.09567*, 2018.

Ricardo Pio Monti, Kun Zhang, and Aapo Hyvärinen. Causal discovery with general non-linear relationships using non-linear ica. In *Uncertainty in artificial intelligence*, pages 186–195. PMLR, 2020.

Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *Journal of Machine Learning Research*, 17(32):1–102, 2016.

Joris M Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 21(99):1–108, 2020.

Gemma E Moran, Dhanya Sridhar, Yixin Wang, and David M Blei. Identifiable deep generative models via sparse decoding. *arXiv preprint arXiv:2110.10804*, 2021.

Vivian Yvonne Nastl and Moritz Hardt. Do causal predictors generalize better to new domains? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Robert Osazuwa Ness, Karen Sachs, Parag Mallick, and Olga Vitek. A bayesian active learning experimental design for inferring signaling networks. In *Research in Computational Molecular Biology: 21st Annual International Conference, RECOMB 2017, Hong Kong, China, May 3-7, 2017, Proceedings 21*, pages 134–156. Springer, 2017.

Tung Nguyen, Jason Jewik, Hritik Bansal, Prakhar Sharma, and Aditya Grover. Climatelearn: Benchmarking machine learning for weather and climate modeling. *Advances in Neural Information Processing Systems*, 36, 2024.

Ana Rita Nogueira, João Gama, and Carlos Abreu Ferreira. Causal discovery in machine learning: Theories and applications. *Journal of Dynamics & Games*, 8(3), 2021.

Peer Nowack, Jakob Runge, Veronika Eyring, and Joanna D Haigh. Causal networks for climate model evaluation and constrained projections. *Nature communications*, 11(1):1415, 2020.

Chris J Oates, L Costa, and Tom E Nichols. Toward a multisubject analysis of neural connectivity. *Neural Computation*, 27(1):151–170, 2014.

Chris J Oates, Jim Q Smith, Sach Mukherjee, and James Cussens. Exact estimation of multiple directed acyclic graphs. *Statistics and Computing*, 26:797–811, 2016.

Desmond J Oathes, Nicholas L Balderston, Konrad P Körding, Joseph A DeLuisi, Gianna M Perez, John D Medaglia, Yong Fan, Romain J Duprat, Theodore D Satterthwaite, Yvette I Sheline, et al. Combining transcranial magnetic stimulation with functional magnetic resonance imaging for probing and modulating neural circuits relevant to affective disorders. *Wiley Interdisciplinary Reviews: Cognitive Science*, 12(4):e1553, 2021.

Gabriel Koch Ocker, Yu Hu, Michael A Buice, Brent Doiron, Krešimir Josić, Robert Rosenbaum, and Eric Shea-Brown. From the statistics of connectivity to the statistics of spike times in neuronal networks. *Current opinion in neurobiology*, 46:109–119, 2017.

Weronika Ormaniec, Scott Sussex, Lars Lorch, Bernhard Schölkopf, and Andreas Krause. Standardizing structural causal models. *arXiv preprint arXiv:2406.11601*, 2024.

Hiroyuki Oya, Matthew A Howard, Vincent A Magnotta, Anton Kruger, Timothy D Griffiths, Louis Lemieux, David W Carmichael, Christopher I Petkov, Hiroto Kawasaki, Christopher K Kovach, et al. Mapping effective connectivity in the human brain with concurrent intracranial electrical stimulation and bold-fmri. *Journal of neuroscience methods*, 277:101–112, 2017.

Pekka Parviainen and Samuel Kaski. Learning structures of bayesian networks for variable groups. *International Journal of Approximate Reasoning*, 88:110–127, 2017.

Gilberto Pastorello, Carlo Trotta, Eleonora Canfora, Housen Chu, Danielle Christianson, You-Wei Cheah, Cristina Poindexter, Jiquan Chen, Abdelrahman Elbashandy, Marty Humphrey, et al. The fluxnet2015 dataset and the oneflux processing pipeline for eddy covariance data. *Scientific data*, 7(1):225, 2020.

Georgios A Pavlopoulos, Maria Secrier, Charalampos N Moschopoulos, Theodoros G Soldatos, Sophia Kossida, Jan Aerts, Reinhard Schneider, and Pantelis G Bagos. Using graph theory to analyze biological networks. *BioData mining*, 4:1–27, 2011.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Jonas Peters and Peter Bühlmann. Structural intervention distance for evaluating causal graphs. *Neural computation*, 27(3):771–799, 2015.

Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(58):2009–2053, 2014.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Russell A Poldrack, Timothy O Laumann, Oluwasanmi Koyejo, Brenda Gregory, Ashleigh Hover, Mei-Yen Chen, Krzysztof J Gorgolewski, Jeffrey Luci, Sung Jun Joo, Ryan L Boyd, et al. Long-term neural and physiological phenotyping of a single human. *Nature communications*, 6(1):8885, 2015.

Dean A Pospisil, Max J Aragon, Sven Dorkenwald, Arie Matsliah, Amy R Sterling, Philipp Schlegel, Szi-chieh Yu, Claire E McKellar, Marta Costa, Katharina Eichler, et al. The fly connectome reveals a path to the effectome. *Nature*, 634(8032):201–209, 2024.

Aditya Pratapa, Amogh P Jalihal, Jeffrey N Law, Aditya Bharadwaj, and TM Murali. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature methods*, 17(2):147–154, 2020.

Lei S Qi, Matthew H Larson, Luke A Gilbert, Jennifer A Doudna, Jonathan S Weissman, Adam P Arkin, and Wendell A Lim. Repurposing crispr as an rna-guided platform for sequence-specific control of gene expression. *Cell*, 152(5):1173–1183, 2013.

Joseph Ramsey and Bryan Andrews. Fask with interventional knowledge recovers edges from the sachs model. *arXiv preprint arXiv:1805.03108*, 2018.

Joseph Ramsey, Madelyn Glymour, Ruben Sanchez-Romero, and Clark Glymour. A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International journal of data science and analytics*, 3:121–129, 2017.

Joseph D Ramsey, Stephen José Hanson, Catherine Hanson, Yaroslav O Halchenko, Russell A Poldrack, and Clark Glymour. Six problems for causal inference from fmri. *neuroimage*, 49(2): 1545–1558, 2010.

Francesco Randi, Anuj K Sharma, Sophie Dvali, and Andrew M Leifer. Neural signal propagation atlas of caenorhabditis elegans. *Nature*, 623(7986):406–414, 2023.

Stephan Rasp, Peter D Dueben, Sebastian Scher, Jonathan A Weyn, Soukayna Mouatadid, and Nils Thuerey. Weatherbench: a benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11):e2020MS002203, 2020.

Andrew T Reid, Drew B Headley, Ravi D Mill, Ruben Sanchez-Romero, Lucina Q Uddin, Daniele Marinazzo, Daniel J Lurie, Pedro A Valdés-Sosa, Stephen José Hanson, Bharat B Biswal, et al. Advancing functional connectivity research from association to causation. *Nature neuroscience*, 22(11):1751–1760, 2019.

Alexander Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated dag! causal discovery benchmarks may be easy to game. *Advances in Neural Information Processing Systems*, 34:27772–27784, 2021.

Alexander Reisach, Myriam Tami, Christof Seiler, Antoine Chambaz, and Sebastian Weichwald. A scale-invariant sorting criterion to find a causal order in additive noise models. *Advances in Neural Information Processing Systems*, 36, 2024.

Joseph M Replogle, Reuben A Saunders, Angela N Pogson, Jeffrey A Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J Wagner, Karen Adelman, Gila Lithwick-Yanai, et al. Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. *Cell*, 185(14):2559–2575, 2022.

Travers Rhodes and Daniel Lee. Local disentanglement in variational auto-encoders using jacobian l_1 regularization. *Advances in Neural Information Processing Systems*, 34:22708–22719, 2021.

Martin Rohbeck, Brian Clarke, Katharina Mikulik, Alexandra Pettet, Oliver Stegle, and Kai Ueltzhöffer. Bicycle: Intervention-based causal discovery with cycles. In *Causal Learning and Reasoning*, pages 209–242. PMLR, 2024.

Paul Rolland, Volkan Cevher, Matthäus Kleindessner, Chris Russell, Dominik Janzing, Bernhard Schölkopf, and Francesco Locatello. Score matching enables causal discovery of nonlinear additive noise models. In *International Conference on Machine Learning*, pages 18741–18753. PMLR, 2022.

David Rolnick, Alan Aspuru-Guzik, Sara Beery, Bistra Dilkina, Priya L Donti, Marzyeh Ghassemi, Hannah Kerner, Claire Monteleoni, Esther Rolf, Milind Tambe, et al. Application-driven innovation in machine learning. *arXiv preprint arXiv:2403.17381*, 2024.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with gears. *Nature Biotechnology*, 42(6):927–935, 2024.

Lauren N Ross and Dani S Bassett. Causation in neuroscience: Keeping mechanism meaningful. *Nature Reviews Neuroscience*, 25(2):81–90, 2024.

Paul K Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Causal consistency of structural equation models. *arXiv preprint arXiv:1707.00819*, 2017.

Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel D Mahecha, Jordi Muñoz-Marí, et al. Inferring causation from time series in earth system sciences. *Nature communications*, 10(1):2553, 2019.

Jakob Runge, Andreas Gerhardus, Gherardo Varando, Veronika Eyring, and Gustau Camps-Valls. Causal inference for time series. *Nature Reviews Earth & Environment*, 4(7):487–505, 2023.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721): 523–529, 2005.

Basil Saeed, Anastasiya Belyaeva, Yuhao Wang, and Caroline Uhler. Anchored causal inference in the presence of measurement error. In *Conference on uncertainty in artificial intelligence*, pages 619–628. PMLR, 2020a.

Basil Saeed, Snigdha Panigrahi, and Caroline Uhler. Causal structure discovery from distributions arising from mixtures of dags. In *International Conference on Machine Learning*, pages 8336–8345. PMLR, 2020b.

Elena Saggioro, Jana de Wiljes, Marlene Kretschmer, and Jakob Runge. Reconstructing regime-dependent causal relationships from observational time series. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30(11), 2020.

Ruben Sanchez-Romero, Joseph D Ramsey, Kun Zhang, Madelyn RK Glymour, Biwei Huang, and Clark Glymour. Estimating feedforward and feedback effective connections from fmri time series: Assessments of statistical methods. *Network Neuroscience*, 3(2):274–306, 2019.

Paula Sanz Leon, Stuart A Knock, M Marmaduke Woodman, Lia Domide, Jochen Mersmann, Anthony R McIntosh, and Viktor Jirsa. The virtual brain: a simulator of primate brain network dynamics. *Frontiers in neuroinformatics*, 7:10, 2013.

S. Saranya Ganesh, Tom Beucler, Frederick Iat-Hin Tam, Milton S. Gomez, Jakob Runge, and Andreas Gerhardus. Selecting robust features for machine-learning applications using multidata causal discovery. *Environmental Data Science*, 2:e27, 2023. doi: 10.1017/eds.2023.21.

Thomas Schaffter, Daniel Marbach, and Dario Floreano. Genenetweaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16): 2263–2270, 2011.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Marco Scutari. Learning bayesian networks with the bnlearn r package. *arXiv preprint arXiv:0908.3817*, 2009.

Eran Segal, Dana Pe'er, Aviv Regev, Daphne Koller, Nir Friedman, and Tommi Jaakkola. Learning module networks. *Journal of Machine Learning Research*, 6(4), 2005.

Muralikrishnna G Sethuraman and Faramarz Fekri. Learning cyclic causal models from incomplete data. *arXiv preprint arXiv:2402.15625*, 2024.

Muralikrishnna G Sethuraman, Romain Lopez, Rahul Mohan, Faramarz Fekri, Tommaso Biancalani, and Jan-Christian Hütter. Nodags-flow: Nonlinear cyclic causal structure learning. In *International Conference on Artificial Intelligence and Statistics*, pages 6371–6387. PMLR, 2023.

Preya Shah, Danielle S Bassett, Laura EM Wisse, John A Detre, Joel M Stein, Paul A Yushkevich, Russell T Shinohara, John B Pluta, Elijah Valenciano, Molly Daffner, et al. Mapping the structural and functional network architecture of the medial temporal lobe using 7t mri. *Human Brain Mapping*, 39(2):851–865, 2018.

Luis Vence Sheng Wu, Lei Jin and Laszlo G Radvanyi. Development and application of 'phosphoflow' as a tool for immunomonitoring. *Expert Review of Vaccines*, 9(6):631–643, 2010.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvarinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, Kenneth Bollen, and Patrik Hoyer. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research-JMLR*, 12(Apr):1225–1248, 2011.

Shan H Siddiqi, Konrad P Körding, Josef Parvizi, and Michael D Fox. Causal mapping of human brain function. *Nature reviews neuroscience*, 23(6):361–375, 2022.

Meromit Singer, Chao Wang, Le Cong, Nemanja D Marjanovic, Monika S Kowalczyk, Huiyuan Zhang, Jackson Nyman, Kaori Sakuishi, Sema Kurtulus, David Gennert, et al. A distinct gene module for dysfunction uncoupled from activation in tumor-infiltrating t cells. *Cell*, 166(6):1500–1511, 2016.

Karamjit Singh, Garima Gupta, Vartika Tewari, and Gautam Shroff. Comparative benchmarking of causal discovery techniques. *arXiv preprint arXiv:1708.06246*, 2017.

Stephen M Smith, Karla L Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F Beckmann, Thomas E Nichols, Joseph D Ramsey, and Mark W Woolrich. Network modelling methods for fmri. *Neuroimage*, 54(2):875–891, 2011.

Peter Spirtes. Variable definition and causal inference. 2009.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2001.

Chandler Squires, Yuhao Wang, and Caroline Uhler. Permutation-based causal structure learning with unknown intervention targets. In *Conference on Uncertainty in Artificial Intelligence*, pages 1039–1048. PMLR, 2020.

Chandler Squires, Annie Yun, Eshaan Nichani, Raj Agrawal, and Caroline Uhler. Causal structure discovery between clusters of nodes induced by latent factors. In *Conference on Causal Learning and Reasoning*, pages 669–687. PMLR, 2022.

Ian H Stevenson and Konrad P Körding. On the similarity of functional connectivity between neurons estimated across timescales. *PloS one*, 5(2):e9206, 2010.

Ian H Stevenson and Konrad P Körding. How advances in neural recording affect data analysis. *Nature neuroscience*, 14(2):139–142, 2011.

Ian H Stevenson, James M Rebesco, Lee E Miller, and Konrad P Körding. Inferring functional connections between neurons. *Current opinion in neurobiology*, 18(6):582–588, 2008.

Albrecht Stroh and Ilka Diester. Optogenetics: a new method for the causal analysis of neuronal networks in vivo. *e-Neuroforum*, 3(4):81–88, 2012.

Akshay Subramaniam, Sungduk Yu, Zeyuan Hu, Walter M Hannah, Liran Peng, Jerry Lin, Mohamed Aziz Bhouri, Ritwik Gupta, Björn Lütjens, Justus Will, et al. Climsim-online: A large multi-scale dataset and framework for hybrid ml-physics climate emulation. *AGU24*, 2024.

Jeff L Teeters and Friedrich T Sommer. Crcns. org: a repository of high-quality data sets and tools for computational neuroscience. *BMC Neuroscience*, 10(Suppl 1):S6, 2009.

Alejandro Tejada-Lapuerta, Paul Bertin, Stefan Bauer, Hananeh Aliee, Yoshua Bengio, and Fabian J Theis. Causal machine learning for single-cell genomics. *arXiv preprint arXiv:2310.14935*, 2023.

Johannes Textor, Benito Van der Zander, Mark S Gilthorpe, Maciej Liśkiewicz, and George TH Ellison. Robust causal inference using directed acyclic graphs: the r package 'dagitty'. *International journal of epidemiology*, 45(6):1887–1894, 2016.

WH Thompson, R Nair, H Oya, O Esteban, JM Shine, CI Petkov, RA Poldrack, M Howard, and R Adolphs. A data resource from concurrent intracranial stimulation and functional mri of the human brain. *Scientific data*, 7(1):258, 2020.

Xavier-Andoni Tibau, Christian Reimers, Andreas Gerhardus, Joachim Denzler, Veronika Eyring, and Jakob Runge. A spatiotemporal stochastic climate model for benchmarking causal discovery methods for teleconnections. *Environmental Data Science*, 1:e12, 2022.

Panagiotis Tigas, Yashas Annadani, Andrew Jesson, Bernhard Schölkopf, Yarin Gal, and Stefan Bauer. Interventions, where and how? experimental design for causal models at scale. *Advances in neural information processing systems*, 35:24130–24143, 2022.

Ngoc Mai Tran, Johannes Buck, and Claudia Klüppelberg. Estimating a directed tree for extremes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkad165, 2024.

Sofia Triantafillou and Ioannis Tsamardinos. Constraint-based causal discovery from multiple interventions over overlapping variable sets. *The Journal of Machine Learning Research*, 16(1): 2147–2205, 2015.

Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65:31–78, 2006.

Ruibo Tu, Kun Zhang, Bo Bertilson, Hedvig Kjellstrom, and Cheng Zhang. Neuropathic pain diagnosis simulator for causal discovery algorithm evaluation. *Advances in Neural Information Processing Systems*, 32, 2019.

Caroline Uhler. Building a two-way street between cell biology and machine learning. *Nature Cell Biology*, 26(1):13–14, 2024.

Tim Van den Bulcke, Koenraad Van Leemput, Bart Naudts, Piet van Remortel, Hongwu Ma, Alain Verschoren, Bart De Moor, and Kathleen Marchal. Syntren: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC bioinformatics*, 7:1–12, 2006.

David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, et al. The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013.

Burak Varici, Karthikeyan Shanmugam, Prasanna Sattigeri, and Ali Tajer. Scalable intervention target estimation in linear models. *Advances in Neural Information Processing Systems*, 34:1494–1505, 2021.

J. Von Kugelgen, Y. Sharma, L. Gresele, W. Brendel, B. Scholkopf, M. Besserve, and F. Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *Advances in Neural Information Processing Systems*, 2021.

Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34:16451–16467, 2021.

Matthew J Vowels, Necati Cihan Camgoz, and Richard Bowden. D'ya like dags? a survey on structure learning and causal discovery. *ACM Computing Surveys*, 55(4):1–36, 2022.

Matej Vuković and Stefan Thalmann. Causal discovery in manufacturing: A structured literature review. *Journal of Manufacturing and Materials Processing*, 6(1):10, 2022.

Jonas Wahl and Jakob Runge. Separation-based distance measures for causal graphs. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025.

Jonas Wahl, Urmi Ninad, and Jakob Runge. Vector causal inference between two groups of variables. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12305–12312, 2023.

Jonas Wahl, Urmi Ninad, and Jakob Runge. Foundations of causal discovery on groups of variables. *Journal of Causal Inference*, 12(1):20230041, 2024.

Lei Wang, Shanshan Huang, Shu Wang, Jun Liao, Tingpeng Li, and Li Liu. A survey of causal discovery based on functional causal model. *Engineering Applications of Artificial Intelligence*, 133:108258, 2024.

Xinyue Wang and Konrad Paul Körding. Learning domain-specific causal discovery from time series. *arXiv preprint arXiv:2209.05598*, 2022.

Xuerui Wang, Rebecca Hutchinson, and Tom M Mitchell. Training fmri classifiers to detect cognitive states across multiple human subjects. *Advances in neural information processing systems*, 16, 2003.

Yuhao Wang, Liam Solus, Karren Yang, and Caroline Uhler. Permutation-based causal inference algorithms with interventions. *Advances in Neural Information Processing Systems*, 30, 2017.

Duncan Watson-Parris, Yuhan Rao, Dirk Olivié, Øyvind Seland, Peer Nowack, Gustau Camps-Valls, Philip Stier, Shahine Bouabid, Maura Dewey, Emilie Fons, et al. Climatebench v1. 0: A benchmark for data-driven climate projections. *Journal of Advances in Modeling Earth Systems*, 14(10): e2021MS002954, 2022.

Anja Wille, Philip Zimmermann, Eva Vranová, Andreas Fürholz, Oliver Laule, Stefan Bleuler, Lars Hennig, Amela Prelić, Peter von Rohr, Lothar Thiele, et al. Sparse graphical gaussian modeling of the isoprenoid gene network in arabidopsis thaliana. *Genome biology*, 5:1–13, 2004.

Albert Xue, Jingyou Rao, Sriram Sankararaman, and Harold Pimentel. dotears: Scalable, consistent dag estimation using observational and interventional data. *arXiv preprint arXiv:2305.19215*, 2023.

Karren Yang, Abigail Katcoff, and Caroline Uhler. Characterizing and learning equivalence classes of causal dags under interventions. In *International Conference on Machine Learning*, pages 5541–5550. PMLR, 2018.

Dingling Yao, Caroline Muller, and Francesco Locatello. Marrying causal representation learning with dynamical systems for science. *arXiv preprint arXiv:2405.13888*, 2024.

Alessio Zanga, Elif Ozkirimli, and Fabio Stella. A survey on causal discovery: theory and practice. *International Journal of Approximate Reasoning*, 151:101–129, 2022.

Semir Zeki and Stewart Shipp. The functional logic of cortical connections. *Nature*, 335(6188): 311–317, 1988.

Jiaqi Zhang, Louis Cammarata, Chandler Squires, Themistoklis P Sapsis, and Caroline Uhler. Active learning for optimal intervention design in causal models. *Nature Machine Intelligence*, 5(10): 1066–1075, 2023.

Jiaqi Zhang, Kristjan Greenewald, Chandler Squires, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions. *Advances in Neural Information Processing Systems*, 36, 2024.

Kun Zhang and Aapo Hyvarinen. On the identifiability of the post-nonlinear causal model. *arXiv preprint arXiv:1205.2599*, 2012.

Y. Zheng, I. Ng, and K. Zhang. On the identifiability of nonlinear ICA sparsity and beyond. In *Advances in Neural Information Processing Systems*, 2022.

Wenxiu Zhou and QingCai Chen. A survey on causal discovery. In *China Conference on Knowledge Graph and Semantic Computing*, pages 123–135. Springer, 2022.

Appendix

Appendix A. Systematic Review

A.1. Methodology

As explained in the main text we used the Semantic Scholar API to collect papers (Kinney et al., 2023). Specifically, we used the *bulk* method to find scientific articles containing the keywords "Causal discovery", "Causal structure learning", "DAG learning", and "DAG structure learning" in their title or abstract. The list of papers was retrieved on September 19, 2024. We did not include articles from workshops at the selected conferences. We manually verified the relevance of each of the 221 papers. We removed articles that were not doing causal discovery or that did not contain any experiments. After this filtering, we kept 167 papers. The whole list of articles and their properties are accessible at our github repo as a CSV file (`curated_papers.csv`).

A.2. Description of each field

In this section, we describe briefly each field of the CSV file `curated_papers.csv`. The *title*, *year*, and *conf* represent the title of the article, the year it was made accessible (note that this is not necessarily the date of publication if, for example, it was put on an open-access repository such as arXiv), and the name of the conference where the article was published. For the *field* column, we report the field of the provenance of the pseudo-real and real-world datasets. We used the following fields: biology (bio), neuroscience (neuro), Earth Sciences (earth), economy (econ), computational systems (comp), social sciences (socio), health sciences (health), and others. For some common datasets, we noted their use in the *pseudo_datasets* and *real_datasets* columns, respectively for pseudo-real and real-world datasets. We noted in *time_series* and *interv_setting* if the proposed causal discovery method operated respectively in a time series and/or interventional setting. In *synthetic*, *pseudo_real*, and *real*, we noted if any experiments were performed on these types of datasets. In *interventions*, we report if the real-world datasets used contained interventional data. In the columns *small*, *medium*, and *big*, we reported the biggest real-world datasets used in each study. Small means 20 variables or less, medium is between 20 and 100 variables, and big is more than 100 variables. The columns *synth_structural*, *synth_observational*, *synth_interventional*, and *synth_qualitative* correspond to the type of evaluation that was used on synthetic and pseudo-real datasets. By structural, we refer to measures comparing the learned graph to the ground-truth graph such as SHD, SID, AUROC, F1-score, etc. By qualitative, we refer to any qualitative judgment that was done to assess the performance of the algorithm. Most of the time, it was about some edges of the learning graph based on some domain knowledge. By observational and interventional, we refer to metrics such as the negative log-likelihood that evaluate the learned model respectively on held-out data in the observational setting and data in an unseen interventional setting. We give more details of our classification of metrics in Appendix D. The four following columns (*real_structural*, *real_observational*, *real_ interventional*, and *real_qualitative*) are similar, but refer to the evaluation on real-world datasets. Finally, *included* denotes whether the article was included or not in our analysis.

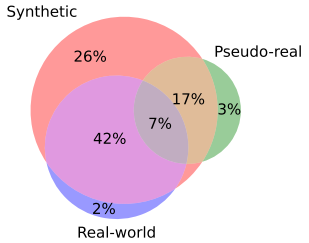

Figure 4: Distribution of papers based on used dataset types.

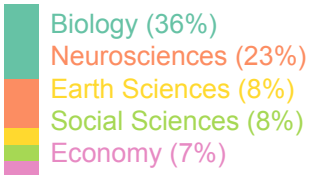

Figure 5: Most common fields of the pseudo-real and real-world datasets.

	Simulated	Real-world
Structural	100.0%	54.3%
Qualitative	3.1%	47.1%
Observational	5.5%	7.1%
Interventional	9.4%	11.4%

Table 3: Percentage of studies using evaluation metrics.

A.3. Scope and limits of the systematic review

We limited our review to papers published at major machine-learning conferences. Of course, this is not necessarily representative of what practitioners do in their respective fields. The *bulk* method of Semantic Scholar seems adequate for our use as it leads to only a few false positives, but, on the other hand, we might have missed some relevant articles. The choice of categories for the type of datasets was subjectively created on the prevalence of some datasets. Finally, the review was performed by two different reviewers. To alleviate possible bias, the reviewers reviewed a similar subsample to make sure their judgment were similar.

A.4. Additional results

Excluding papers with only Sachs. We perform the same analysis as in the main text but we exclude the papers containing only Sachs as their real-world datasets (a total of 37 papers). Overall, the results are similar, but we can notice a few interesting differences. The proportion of papers using only synthetic datasets is higher at 26% (see Fig. 4). The field of biology is still the most popular, but it is now more closely followed by the field of neuroscience (see Fig. 5). Finally, for the real-world evaluations, the use of structural metrics is lower leading to an almost equal use of qualitative and structural metrics (see Table 3). This can be explained by the frequent use of Sachs where the structural metrics are used based on the consensus network.

Most popular datasets. In this section, we briefly discuss the most used datasets for real-world and pseudo-real datasets. For the real-world datasets, besides Sachs, the Tübingen pairs (Mooij et al., 2016) is the most frequent real-world datasets. We describe in more detail this dataset in Appendix C.2. This contains only pairs of variables and the ground truth is assumed to be known, driving up the number of the use structural metrics for real-world datasets. The third most used dataset is the resting state fMRI data from Poldrack et al. (2015). The recording comes from a single subject over 84 successive days. This is a small graph (6 nodes) representing regions of the hippocampus. We note that in some studies, the different days are considered as different experimental conditions. The ground-truth graph is unknown and qualitative metrics are mostly used. Finally, in fourth position is the perturb-CITE-seq data from Frangieh et al. (2021) coming from three different cell populations, which contains approximatively 20000 genes (for all studies, only a subset is used $\leq$1000), and interventions under the form of gene knockdowns. The most common metric used is the interventional one.

For the pseudo-real datasets, the bnlearn repository (Scutari, 2009) is followed by the simulated fMRI data from Smith et al. (2011), the DREAM datasets (Marbach et al., 2010; Greenfield et al., 2010), and SERGIO that generates single-cell expression data of gene regulatory networks from

Dibaeinia and Sinha (2020). For all of them, since they are simulated, the ground-truth graph is known and structural metrics are mostly used. Note that the simulated fMRI data violates the acyclicity assumption by relying on differential equations model.

Appendix B. Biological data

B.1. Gene expression and transcriptomics.

Although a single organism is composed of vastly different types of cells (e.g., skin cells, neurons, immune cells), all of these cells have the same genetic code (DNA). Within an organism, variation in cell state is not driven by variation in genetic profiles. Rather, such variation depends heavily upon the process of *transcription*, in which (protein-coding) genes from the DNA are transcribed into messenger RNA (mRNA) molecules, which are then used as a template to synthesize the cell's proteins. One of the most important determining factors of a cell's state is its *transcriptome* (also called its *gene expression profile*), i.e., the total number of mRNA molecules transcribed from each gene. Thus, the field of *transcriptomics* is a key part of understanding questions about development, disease, and other processes (Emmert-Streib et al., 2014).

In causal discovery, one might say, as a shortcut, that Gene A regulates Gene B if changing the expression of Gene A results in a change in the expression of Gene B. Physically speaking, this relationship is mediated by other, unmeasured molecules, e.g. Gene A might code for a transcription factor (i.e., a protein) which in turn binds to a promoter region for Gene B, increasing the expression of Gene B. Thus, in causal discovery, an edge Gene $A \rightarrow$ Gene B represents the existence of such a mediated causal relationship.

Transcriptomic technologies exist both for measuring and for experimentally manipulating gene expression. Two common approaches to measuring gene expression are *microarrays*, which measure a fixed panel of genes, and the much more comprehensive *RNA sequencing (RNA-seq)*, which sequences all mRNA transcripts in an untargeted manner. Molecular measurements are either done in bulk, wherein a population of cells are lysed and an average is measured, or can be performed in single cells. *Single-cell RNA sequencing* (scRNA-seq) has obvious advantages with respect to the number of observational units; however, it should be noted that the data can be extremely sparse. Low abundance genes - including crucial regulators like transcription factors - may fall below the level of detection in individual cells, but are readily detectable in bulk. Single-cell experiments are also far more expensive.

B.2. Beyond transcriptomics.

As aforementioned, mRNA readout of gene expression is highly informative of cell state. However, one has to keep in mind that the relation between genes is always mediated: genes themselves do not execute functions in the cells, but rather the proteins which are created based on the information encoded in the genes. Roughly speaking, mRNA (from Gene X) translates into protein (to Protein X). However, factors such as RNA stability and degradation strongly affect this relationship. Hence, the *Gene Regulatory Networks* (GRN) being modeled via transcriptomics are actually carried out not by the quantified gene mRNAs, but by latent variables: the gene-encoded proteins. One way to model biomolecular networks more directly is via proteomic datasets, especially ones in which the abundance of the *activated* proteins is quantified (Sheng Wu and Radvanyi, 2010; Sachs et al., 2005; Brandi et al., 2022), though these tend to be more challenging experimentally.

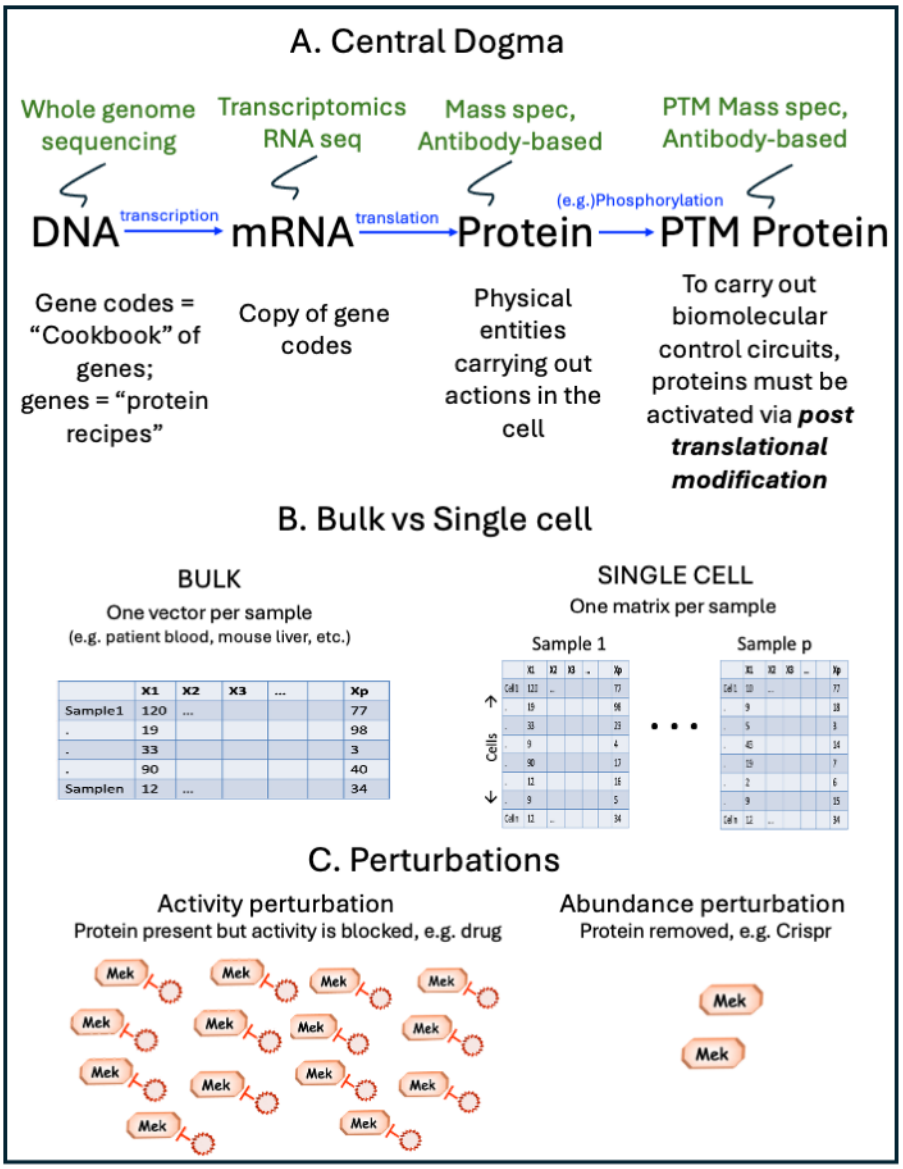

Figure 6: Details of biomolecular datasets. A. Central Dogma of biology. The unique code for each organism is encoded in the genome, consisting of a sequence of genes encoded in DNA. The uniqueness is due to variations in genes. Genes are codes or "recipes" for proteins. This code is copied into gene-specific mRNA molecules, via a process called transcription. The information from mRNA molecules is used to create unique proteins via a process called translation. Proteins that comprise the nodes of biomolecular regulatory networks must be activated via processes catalyzed by other (upstream) proteins, in a process called post-translational modification. Processes are shown in blue, measurement technologies are shown in green. Antibody-based modalities include flow cytometry and microscopy-based technologies. B. Depiction of lysate-based (bulk) measurements vs. single cell. Bulk technologies are easier and cheaper, but yield just one vector per sample; single-cell data is sparse in the context of transcriptomics, or of limited dimensionality in proteomics. C. Activity perturbations such as small molecule inhibitors (drugs), and abundance perturbations, typically carried out by genetic means such as CRISPR or ASOs. Activity perturbations leave the protein intact, but block its activity, while abundance perturbations, typically remove the protein by removing the gene from the DNA (genome), or by removing the gene's transcript.

Table 4: Links to pseudo-real datasets

Article	Name	Link
(Cheng et al., 2024)	CausalTime	link to dataset
(Göbler et al., 2023)	CausalAssembly	link to dataset
(Lawrence et al., 2021)	-	link to dataset
(Pratapa et al., 2020)	BEELINE	link to dataset
(Dibaeinia and Sinha, 2020)	SERGIO	link to dataset
(Runge et al., 2019)	CauseMe	link to dataset
(Sanchez-Romero et al., 2019)	-	link to dataset
(Tu et al., 2019)	Neuropathic Pain Diagnosis Simulator	link to dataset
(Schaffter et al., 2011)	GeneNetWeaver	link to dataset
(Smith et al., 2011)	Netsim	link to dataset
(Marbach et al., 2010)	DREAM4	link to dataset
(Van den Bulcke et al., 2006)	SynTReN	link to dataset

Lysed vs. single cells. Starting with cells in a dish or test tube, cells are either lysed and measured in bulk, or measured as individual cells. In transcriptomics, single-cell data may be very sparse, as genes may not need to be expressed all the time (if the coded proteins are stable), or may fall below the limit of detection. Bulk data is also far cheaper. In proteomics, mass spectrometry-based modalities measure the entire proteome, but are still in the very early days of single-cell capabilities, and are expensive even in bulk measurements. Most such datasets focus on proteome abundance, some also include *post-translational modification* (PTM). Antibody or label-based modalities for single-cell proteomics such as flow cytometry have been around the longest (>50 years) and are neither sparse nor prohibitively expensive, and readily report an abundance of PTM proteins. However, they must focus on far more limited sets of proteins, in the tens rather than thousands.

Activity vs. abundance perturbations. It is useful to distinguish between activity perturbations such as small molecule inhibitors (drugs), and abundance perturbations, typically carried out by genetic means such as CRISPR or ASOs. Activity perturbations leave the protein intact, but block its activity, such that it cannot activate further proteins in the signaling cascade or biomolecular network, while abundance perturbations typically remove the protein by removing the gene from the DNA (genome), or by removing the gene's transcript. Details of data modalities and intervention technologies are summarized in Fig. 6.

Appendix C. Datasets

C.1. Links to datasets

In the following tables, we provide links to the different datasets we discussed in the main text. In Table 4, 5, 6 and 7 we report the links for pseudo-real, biology, neuroscience and Earth science datasets. We also keep an updated version of these lists of datasets at our github repo. The lists contain mainly datasets that have been used in causal discovery studies. Several are particularly interesting since they contain interventional data and can violate some common assumptions. Still, we advise thoughtful use in line with our recommendations, rather than applying them blindly. Also, the lists are by no means exhaustive and we do encourage researchers to explore new applications that are not listed here. While this section provides links to pseudo-real and real-world datasets, we would also like to share the synthetic generator Causally (Montagna et al., 2024) where it is natively possible to control the modelling assumptions (e.g., presence of latent confounders, unfaithful path, etc).

Table 5: Links to biology datasets

Dataset	Description	Link
(Replogle et al., 2022)	Perturb-seq (cell line K562)	link to dataset
(Replogle et al., 2022)	Perturb-seq (cell line RPE1)	link to dataset
(Frangieh et al., 2021)	Perturb-CITE-seq data from melanoma cells	link to dataset
(Frot et al., 2019)	RNA-seq of ovarian cancer	link to dataset
(Dixit et al., 2016)	Perturb-Seq of bone marrow-derived dendritic cells	link to dataset
(Singer et al., 2016)	Naive and activated T cells (Drop-seq)	link to dataset
(Sachs et al., 2005)	Flow cytometry dataset of immune cells	link to dataset
(Wille et al., 2004)	Microarray of *A. thaliana* gene expression	link to dataset

Table 6: Links to neural datasets

Dataset	Description	Link raw data	Preprocessed
DANDI	Large collection of modern large neuroscience datasets, including optogenetics (DANDI)	raw dataset	link to database
(Dorkenwald et al., 2022)	Drosophila connectome (Flywire)	raw dataset (connectome)	simulated Ca2+ activities
(Randi et al., 2023)	C. elegans simultaneously record Ca2+ most neurons while stimulating	raw data	link to dataset
(Teeters and Sommer, 2009)	Spiking data from various sources (CRCNS)	database	–
(Thompson et al., 2020)	es-fMRI data (intracranial electrodes)	raw dataset	same link
(Shah et al., 2018)	rs-fMRI data from the medial temporal lobe	raw dataset	–
(Poldrack et al., 2015)	Hippocampal rs-fMRI (MyConnectome project)	raw dataset	–
(Di Martino et al., 2014)	rs-fMRI (ABIDE Consortium)	raw dataset	link to dataset
(Van Essen et al., 2013)	rs-fMRI (Human Connectome Project)	raw dataset	–
(Ramsey et al., 2010)	Task fMRI (Rhyme judgment)	raw dataset	link to dataset
(Wang et al., 2003)	Task fMRI (star/plus experiment)	–	–

We provide the links pointing to the original datasets. However, several of these datasets require some data preprocessing. In this paragraph, we elaborate on some datasets giving references to preprocessed versions that have been in causal discovery settings.

The two datasets from Replogle et al. (2022) have been processed and curated by Chevalley et al. (2022), and made accessible on *CausalBench*. The dataset from Frangieh et al. (2021) has been preprocessed by Lopez et al. (2022). The code is available online.

For the simulator SERGIO (Dibaeinia and Sinha, 2020), a modified version has been proposed by Hägele et al. (2023), making it possible to use custom graphs and generate (perfect) interventional data.

C.2. Miscellaneous Datasets

In this section, we elaborate on a few other source of datasets that are particularly interesting.

Computational systems. Datasets generated by existing computational systems have been recently proposed and used to evaluate causal discovery methods. They are of considerable interest since they are deterministic yet complex systems composed of many components. Also, compared to pseudo-real datasets, they are generated from real-world environments. For example, the C++ simulator of the MOS 6502 microprocessor (Jonas and Körding, 2017; Wang and Körding, 2022) is composed of 3510 transistors and 1904 connection elements where the variables of interest are the voltage of the different transistors. While the physical connections are known, it is not sufficient to know the causal graph, instead, it was determined from the perturbation of single transistors. Similarly,

Table 7: Links to Earth science datasets

Article	Name	Link
(Kaltenborn et al., 2024)	ClimateSet	link to dataset
(Nguyen et al., 2024)	ClimateLearn	contained in ClimateSet
(Subramaniam et al., 2024)	ClimSim	contained in ClimateSet
(Watson-Parris et al., 2022)	ClimateBench	contained in ClimateSet
(Rasp et al., 2020)	WeatherBench	contained in ClimateSet
(de Witt et al., 2021)	RainBench	contained in ClimateSet
–	KNMI Climate Explorer	link to database
–	NCEP/NCAR 40-year reanalysis project	link to database
(Hersbach et al., 2020)	ERA reanalysis project	link to database
–	Climate Prediction Center (CPC) global rainfall dataset	link to dataset
(Pastorello et al., 2020)	FLUXNET	link to database
–	Tropospheric Ozone Assessment Report (TOAR)	link to database
–	CDC Influenza report	link to dataset
(Eyring et al., 2016)	Climate Model Intercomparison Project (CMIP)	link to database

Table 8: Links to Miscellaneous Real-World Datasets

Article	Name	Link
(Gamella et al., 2025)	Causal Chambers	link to simulator
(Wang and Körding, 2022)	MOS 6502 microprocessor	link to dataset
(Gentzel et al., 2019)	Software Systems (PostgreSQL, JDK, Networking)	link to datasets
(Huang et al., 2018)	Archaeology data set	–
(Mooij et al., 2016)	Cause-effect pairs (Tübingen pairs)	link to dataset
(Byrne, 2013)	Teacher's burnout study	link to dataset
(Shimizu et al., 2011)	General Social Survey	link to dataset

data from the analysis of microservice-based applications (Ikram et al., 2022) (from a sock-shop demo and a production-based microservice system hosted on AWS cloud-native system) has been used to evaluate root-cause analysis methods. The data can be different metrics such as CPU and memory utilization, while the interventions can be failures such as CPU hog and memory leak. Gentzel et al. (2019) also proposed several datasets from such systems: Java Development Kit, PostgreSQL, and a web server infrastructure. Recently, datasets generated by the activation of neural networks have also been used in order to interpret their learned representations (Geiger et al., 2021, 2022).

The Causal Chambers. Recently, Gamella et al. (2025) have proposed an interesting new type of dataset where the data is generated from real-world experiments but the ground-truth graph is known. Gamella et al. (2025) designed two computer-controlled physical simulators, that can generate observational and interventional data (i.i.d. as well as time series) with experimentally verified ground-truth graphs. Due to the recency of its development, as far as we know, no causal discovery method has been evaluated on Causal Chamber data yet (see Lohse and Wahl (2024) for an investigation on var-/R^2-sortability of Causal Chambers time series data).

Tübingen pairs. The Tübingen pairs (or CauseEffectPairs benchmark) (Mooij et al., 2016) is a repository regrouping 108 datasets composed of pairs of cause and effect coming from many different domains. The range of domains is vast: climate data, biology (e.g., growth of abalone), healthcare (e.g., arrhythmia and diabetes), economy (e.g., census income), stock market, etc. Note that many of these pairs are adapted from the UCI Machine Learning Repository (Bach and Lichman) where the complete datasets are available. While being real-world data, the ground-truth causal direction is given by the authors since in many cases, it is obvious from common sense (e.g., the altitude causes the temperature).

TableShift. Recently Gardner et al. (2024) proposed TableShift: a benchmark comprising many tabular datasets chosen to assess the robustness of machine learning methods to distribution shift. Although not originally intended for causal discovery, it has been explored in Nastl and Hardt (2024), where causal discovery methods are used with the aim of finding causal parents of variables of interest.

Appendix D. Metrics to evaluate causal models

Structural metrics. The most commonly used structural metric is clearly the structural Hamming distance (SHD) (Acid and de Campos, 2003; Tsamardinos et al., 2006). The distance $\text{SHD}(\hat{G},G)$ is defined as the number of edges that should be added, removed or reversed in order to modify an estimated graph $\hat{G}$ to a target graph G. Besides SHD, other similar metrics are also often reported: precision-recall, false discovery rate, F_1 score, AUROC, etc. They can be more useful since the SHD alone can be misleading (e.g., for a really sparse graph, an empty graph can be better in terms of SHD than denser graphs that contain many ground-truth edges).

A major limitation of these metrics is that they are purely based on the graph without any notion of causality. Other structural metrics assess the distance in terms of topological ordering (e.g., Rolland et al. (2022)) conditional independencies (Textor et al., 2016), d-separation statements (Wahl and Runge, 2025), node-permutation tests (Eulig et al., 2023), etc. More focused on the effect of interventions, Peters and Bühlmann (2015) proposed the *Structural Interventional Distance* (SID) which counts the number of interventional distributions that would be wrongly computed using the parents from the learned structure as its adjustment sets. While considering interventions, this measure is still about the graph and correlates strongly with SHD (Gentzel et al., 2019). A generalization of this measure that considers other adjustment sets has also been proposed by Henckel et al. (2024). It also has the advantage of being directly applicable to CPDAGs as it returns a scalar instead of bounds and is computationally less demanding.

Interventional metrics. As previously mentioned, interventional metrics do not necessitate a known ground-truth graph as it evaluates directly how well a causal model can predict data coming from an unseen interventional distribution (Garant and Jensen, 2016). A common interventional metric is the *interventional negative log-likelihood* (I-NLL) (Lopez et al., 2022):

$$\text{I-NLL} = -\mathbb{E}_{x\sim P^{(j)}}[\log P_\theta^{(j)}(X)], \tag{3}$$

where data from $P^{(j)}$ were not part of the training set and $P_\theta^{(j)}$ is the learned model. Usually, the average is taken over multiple interventional distributions. We also note that this metric does not even require the learned model P_θ to use a graph and can thus be used with a more general class of methods. While it often takes the form of a likelihood, it can also be any distance between the learned distribution and the ground-truth one: some have used the mean absolute error (Lopez et al., 2022), total variation distance (Garant and Jensen, 2016), the KL divergence, etc. It can also take the form of the strength of a causal relation or the average/conditional treatment effect.

Appendix E. Causal Representation Learning

In this review, we focused on causal discovery where it is assumed that we have access to structured data. However, many datasets generated by real-world phenomena are unstructured data (e.g., images, videos, texts, etc). The question of how to deal with such datasets has been central in causal abstraction (Rubenstein et al., 2017; Beckers and Halpern, 2019; Beckers et al., 2020; Massidda et al., 2023),

causal feature learning (Chalupka et al., 2014, 2017), causal grouping (Parviainen and Kaski, 2017; Anand et al., 2023; Wahl et al., 2023, 2024), and causal representation learning (Schölkopf et al., 2021). In this section, we will briefly present causal representation learning where the main task is identifying latent causal variables usually from an unstructured input. This recent development opens the doors to many new practical applications where datasets are unstructured. So far, the field has focused on proving identifiability results, showing that it is possible to recover the right representation (up to some minor transformations) under some assumptions.

Formally, we have the observed variable $X=(X_1,...,X_n)$ that are generated by applying a function g to the latent variable $Z=(Z_1,...,Z_d)$ that is Markov to a graph G. The data-generating process is as follows:

$$X=g(Z), \quad P_Z=\prod_{i=1}^{d} P_i(Z_i \mid pa_i^G), \tag{4}$$

where $g:\mathbb{R}^d \to \mathbb{R}^n$ is an injective function called decoder or mixing function. The goal is to find from X a representation of Z that is causally disentangled. Disentangled representations allow interpretability and can also be useful for many downstream tasks. However, without assumptions, it is impossible to learn such a representation (Hyvärinen and Pajunen, 1999; Locatello et al., 2019). Thus, data-generating processes considered follow additional assumptions: assumptions are made on the distribution of the latent variable Z and its support, additional assumptions, such as sparsity, are made about g (Moran et al., 2021; Rhodes and Lee, 2021; Zheng et al., 2022; Brouillard et al., 2024) or the latent graph (Lachapelle et al., 2022) and often the presence of auxiliary information is assumed (e.g., Locatello et al. (2020); Von Kugelgen et al. (2021); Brehmer et al. (2022); Ahuja et al. (2022); Lachapelle et al. (2023)). Given these assumptions, many identification results have been discovered showing that the disentangled representation is unique up to some minor transformation (such as affine transformations).

However, so far, the field has compared new methods almost exclusively on simple synthetic datasets. We present in Table 9 a list of commonly used datasets in causal representation learning and show visual examples in Figure 7. These synthetic datasets are, in many respects, really not representative of real-world tasks: they focus only on problems where images are the observable input, the latent variables are always simple properties of objects (e.g., position, color, etc), and the latent variables are only a few (i.e., less than 10). Liu et al. (2023) also highlight that images coming from synthetic datasets are too simple: most have plain textures, contain only a small number of objects, and do not contain object occlusion (see Figure 7, except the image to the right). As for the evaluation metrics, the situation is similar to causal discovery. Most rely on the Mean Correlation Coefficient (MCC) that finds the best permutation to evaluate how well the learned latent variables correlate with the ground-truth latent variables. However, the MCC is, at least in some instances, not directly related to the models' performance on downstream tasks. As for the causal discovery evaluation, we recommend evaluating models on downstream tasks such as predicting the effect of interventions.

A few recent works have used real-world datasets, such as Lopez et al. (2023); Zhang et al. (2024) for gene regulatory networks, Yao et al. (2024); Brouillard et al. (2024) in the Earth science domain, and Cadei et al. (2025) in ecology. However, no realistic simulators like the one proposed for causal discovery exist. We also observe that many real-world datasets reported in Table 5, 6 and 7 are good candidates for causal representation learning methods since they are high-dimensional unstructured data before their preprocessing. These datasets offer a more diversified and challenging repertoire than what is presently used in the field. Furthermore, the common practice in causal discovery applied

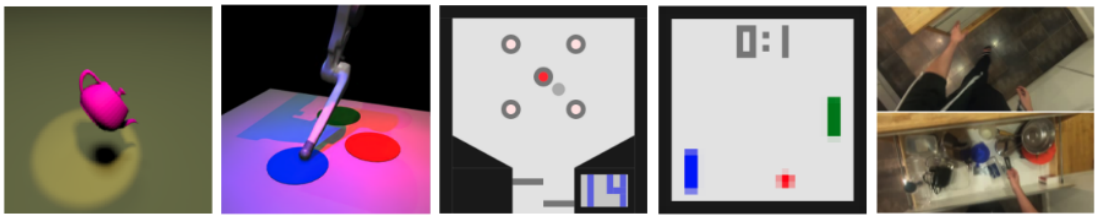

Figure 7: From left to right: Causal3DIdent, CausalCircuit, Causal Pinball, Interventional Pong, Causal triplet (see Table 9 for the references).

Table 9: List of the most common CRL datasets. d_z is the dimensionality of the latent variables.

Dataset	Description	d_z
Von Kügelgen et al. (2021)	Causal3DIdent: a 3D object under various conditions	7
Brehmer et al. (2022)	CausalCircuit: A robot arm can interact with lights	4
Lippe et al. (2022a)	Causal Pinball	5
Lippe et al. (2022b)	Interventional Pong	6
Liu et al. (2023)	Causal triplet	-

to unstructured problems is to use dimensionality reduction methods or to drop features with less variation. This constitutes an opportunity for causal representation learning since these common practices probably lead to an incorrect choice of variables.

We conclude by stating that we only focused on causal representation learning, but the realm of domains where causal methods are applied has grown abundantly yielding many other possible applications. While deviating from classical causal discovery, they often use more realistic datasets. Some new methods relying on LLM can directly rely on text metadata to learn relevant causal variables (Liu et al., 2024). Causally-inspired algorithms have also been proposed to tackle the problem of multidomain data (e.g., Arjovsky et al. (2019)), where different domains are interpreted as different interventional environments (for examples of datasets, see Gulrajani and Lopez-Paz (2021)). Finally, we could also mention the active field of causal reinforcement learning (e.g., Ahmed et al. (2021); Ke et al. (2021)).