CAUSALRIVERS - SCALING UP BENCHMARKING OF CAUSAL DISCOVERY FOR REAL-WORLD TIME-SERIES

Anonymous authors
Paper under double-blind review

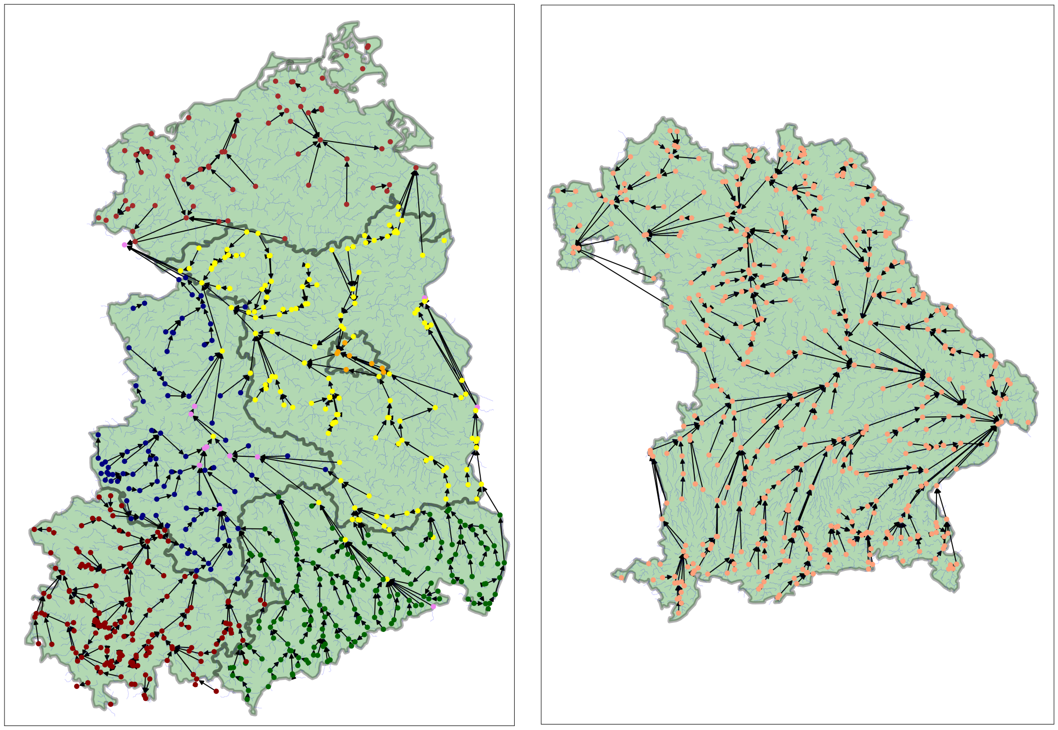
ABSTRACT

Causal discovery, or identifying causal relationships from observational data, is a notoriously challenging task, with numerous methods proposed to tackle it. Despite this, in-the-wild evaluation is still lacking, as works frequently rely on synthetic data evaluation and sparse real-world examples under critical theoretical assumptions. Real-world causal structures, however, are often complex, evolving over time, non-linear, and influenced by unobserved factors, making it hard for practitioners to select appropriate methods. To bridge this gap, we introduce **CausalRivers**, the largest in-the-wild causal discovery benchmarking kit for time series data to date. CausalRivers features an extensive dataset on river discharge that covers the complete eastern German territory (666 measurement stations) and the state of Bavaria (494 measurement stations). It spans the years 2019 to 2023 with a 15-minute temporal resolution. Further, we provide data from a recent flood around the Elbe River, as an event with a pronounced distributional shift. Leveraging multiple sources of information and time-series meta-data, we constructed two distinct causal ground truth graphs (Bavaria and eastern Germany). These graphs can be sampled to generate thousands of subgraphs to benchmark causal discovery across diverse and challenging settings. To demonstrate the utility of our benchmarking kit, we evaluate several causal discovery approaches through multiple experiments and introduce effective baselines, identifying several areas for enhancement. CausalRivers has the potential to facilitate robust evaluations and comparisons of causal discovery methods. Besides this primary purpose, we also expect that this dataset will be relevant for connected areas of research, such as time series forecasting and anomaly detection. Based on this, we hope to establish benchmark-driven method development that fosters advanced techniques for causal discovery, as is the case for many other areas of machine learning.

1 INTRODUCTION

Causal discovery, the process of identifying causal relationships from observational data, has made significant theoretical progress over the past decade (Pearl, 2009), (Peters et al., 2017). This has led to the development of various methods (Vowels et al., 2022), (Assaad et al., 2022) that especially bear potential for fields where randomized controlled trials are impractical due to restrictions concerning interventions, such as earth sciences, neuroscience, and economics. However, despite this progress, causal discovery remains a predominantly theoretically motivated area of research. We argue that one of the primary reasons for this is the challenge practitioners face in selecting appropriate causal discovery strategies, especially given the strong assumptions these methods are often required to make about the underlying data, e.g. causal sufficiency, linearity, or the absence of hidden confounders. As an example, methods based on additive noise models (ANMs, (Peters et al., 2011)) assume specific noise distributions, while constraint-based approaches like PC (Spirtes et al., 2001) and FCI (Spirtes, 2001) assume that causal relationships underlying observational data are of a faithful nature, an assumption that was criticized by Andersen (2013).

Violations of these assumptions are particularly very common in fields like neuroscience or climate science, where the data-generating process is complex, often unknown, and typically influenced by unobserved confounding factors. This, in turn, also limits the reliability of synthetic benchmarking, as data-generating processes fail to meet the complexity of real-world scenarios, leading to inflated assessments of method performance, as discussed in Reisach et al. (2021).



(a) Eastern Germany (b) Bavaria

Figure 1: The causal ground truth graphs for river discharge measurement stations are provided with this benchmarking kit. Jointly, these graphs hold over 1000 nodes. Different colors specify different data origins that we specify in appendix A.1.

Additionally, even extensive survey papers like Vowels et al. (2022) can only provide limited guidance for practitioners, as they cannot directly address which methods might provide meaningful insights when assumptions are violated. Furthermore, a large part of the causal discovery literature relies on either purely synthetic experiments (Pamfil et al., 2020) and simple real-world examples with few nodes Mooij et al. (2016), (Runge et al., 2019). This situation seems to be especially pronounced for time-series data, as even fewer datasets are available. Instead, the focus of many works lies on proving theoretical guarantees under assumptions as proof of their validity. While these insights are by no means unnecessary and provide an essential foundation for methods evaluation, they provide, again, limited help when faced with the complexity and unpredictability of the real world.

Here, we feel like it is necessary to remind ourselves of the iron rule of explanation as the cornerstone of modern science (Strevens, 2020): *"scientists [...] resolve their differences of opinion by conducting empirical tests"*. In machine learning, this is implemented through benchmark datasets, which provide standardized environments for rigorous evaluation of the performance of competing methods. These benchmarks not only facilitate fair comparisons but also reveal systematic weaknesses, and, thus, actively contribute to method development. For instance, computer vision was reshaped by the ImageNet challenge that brought the surprising performance of the AlexNet architecture to the field's attention (Alom et al., 2018). In a similar vein, we believe that a large-scale and realistic benchmark dataset for causal discovery could have a profound impact on the field. We also find that no such benchmark has been established for causal discovery from time series for which we provide evidence in the next chapter.

To bridge this gap, and inspired by a single five-node example in Muñoz-Marí et al. (2020), we introduce **CausalRivers**, the by far largest in-the-wild causal discovery benchmarking kit, specifically for time series data, to date. CausalRivers features an extensive dataset on river discharge, spanning from the year 2019 to the end of 2023, with a 15-minute resolution. It covers the entirety of the eastern German territory (666 measurement stations) and the state of Bavaria (494 measurement stations). Further, we include an additional dataset from a subset of stations, which exhibits a pronounced distributional shift through a very recent extreme precipitation event. To complement this dataset, we constructed two causal ground truth graphs (Figure 1), that include all measurement stations. For this, we leveraged multiple informational sources such as Wikipedia crawls and remote sensing. Further information on the data origins is included in appendix A.1. Importantly, as the full ground truth graphs hold over 1000 nodes, a direct application of causal discovery approaches to these time series is unfeasible. Instead, we provide sampling strategies to generate thousands of subgraphs with a flexible amount of nodes and unique graph characteristics such as single-sink nodes, root causes, hidden-confounding, or simply connected graphs. Along with the general characteristics of river discharge, which we discuss later, the dataset allows us to assess the impact of conditions such as e.g., high-dimensionality, non-linearity, non-stationarity, seasonal patterns, the presence of hidden confounding (through weather), misalignment of causal lag and sampling rate, and generally distributional shifts on method performance.

To demonstrate our benchmarking kit, we conducted three sets of experiments, providing an overview of potential benchmarking use cases. First, we provide experiments on multiple sets of subgraphs. For this, we report performances of well-known causal discovery approaches, provide naive yet effective baselines, and evaluate some recent deep learning approaches. Here, we find that simple strategies can be robust, where many causal discovery methods struggle. Second, we evaluate how the selection of specifically informative subsections of observational data can affect the performance of different methods, something that could prove helpful in real-world applications. Finally, we provide some examples of how domain adaption might be an interesting tool to cope with the complex nature of the provided data distribution. Here we find mixed results, as the impact of such a selection depends on the specific causal discovery approach. To make usage as accessible as possible, we provide a ready-to-use benchmark package with many features as a repository here: *ANONYMOUS*. With this benchmarking kit, we hope to pave the way for more benchmark-focused method development and provide the groundwork for closing the gap between causal discovery research and its potential applications. Finally, we are looking forward to seeing whether the provided data, as the amount of time-series data is extensive, might also be interesting to related disciplines such as time-series forecasting, anomaly detection, or regime change identification. To summarize, this work provides the following contributions:

- The largest real-world benchmark for causal discovery from time series to date
- A comparison of established causal discovery methods on in-the-wild data.
- An introduction and a ready-to-use implementation of the complete benchmarking kit.

2 BACKGROUND

The impact of benchmarking becomes evident in various fields where large-scale and realistic datasets have driven significant advances. As already mentioned, computer vision was reshaped by the ImageNet challenge that brought the surprising performance of the AlexNet architecture to the field's attention (Alom et al., 2018). Other examples are GLUE (Wang et al., 2019), which has become a standard for evaluating natural language processing models. Next to this, the SQuAD benchmark (Rajpurkar et al., 2016) has pushed the state-of-the-art in question-answering. Further, WMT-2014 (Bojar et al., 2014) helped with establishing Transformers (Vaswani et al., 2017) as the dominant architecture in natural language processing. Similarly, the LAION-5B dataset (Schuhmann et al., 2022) has driven the development of vision foundation models. Moreover, RESISC45 (Cheng et al., 2017) helped cement deep learning for remote-sensing scene classification. Finally, the Cityscapes benchmark (Cordts et al., 2016) has accelerated research in autonomous driving, while the CASP13 benchmark has revolutionized protein folding, via AlphaFold (AlQuraishi, 2019).

In a similar vein, we believe that a large-scale and realistic benchmark dataset for causal discovery could have a profound impact on the field. To date, however, such a benchmark is lacking. To visualize this absence, we provide an overview of existing datasets (Table 1) that either cover real-world data or attempt to imitate specific characteristics of real-world domains (semi-synthetic data). For completeness's sake, we also include datasets that only provide sample data (no temporal dimension) as well as some datasets that are considered for average treatment effect estimation, since it is possible to repurpose them for causal discovery. As can be observed from our summary, while we found almost 30 distinct datasets, few of them provide time-series data. Further, many datasets that provide authentic, real-world data have a limited number of nodes included, making it hard to rely on

Table 1: An extensive list, not only including time-series data, of works used to evaluate causal discovery approaches. A ✓ for ”Time” denotes that the data source is a time series. A ✓ for ”Real world” denotes that both observational data and ground truth causal graphs are not synthetic. Further, ⊘ denotes no theoretical limit on the number of variables as datasets have synthetic components. We emphasize that there is no comparable-sized benchmark for time-series data to date.

Topic	Origin	Time	Real world	Number of variables
Semi synthetic generation^{ATE}	Neal et al. (2021)	✗	✗	⊘
Semi synthetic generation^{ATE}	Shimoni et al. (2018)	✗	✗	⊘
Gen expressions	Dibaeinia & Sinha (2020)	✗	✗	⊘
Production line	Göbler et al. (2024)	✗	✗	⊘
Gen expressions	Van den Bulcke et al. (2006)	✗	✗	⊘
Gen networks	Pratapa et al. (2020)	✗	✗	⊘
Visual understanding	McDuff et al. (2022)	✗	✗	⊘
Mixed Challenge^{ATE}	Dorie et al. (2019)	✗	✗	⊘
Mixed Challenge^{ATE}	Hahn et al. (2019)	✗	✗	⊘
Benchmark kit (LLM)	Zhou et al. (2024b)	✗	✓	109
Single-cell perturbation	Chevalley et al. (2023)	✗	✓	622
Mixed Challenge	Guyon et al. (2008)	✗	✓	132
Cause-effect pairs	Mooij et al. (2016)	✗	✓	100×2
Congenital heart disease	Spiegelhalter et al. (1993)	✗	✓	20
Lung Cancer	Lauritzen & Spiegelhalter (1988)	✗	✓	8
Food manufacturing	Menegozzo et al. (2022)	✗	✓	34
Protein signaling	Sachs et al. (2005)	✗	✓	11
Bridges	Yoram Reich (1989)	✗	✓	12
Abalons	Warwick Nash (1994)	✗	✓	8
Arrow of time	Bauer et al. (2016)	✗	✓	⊘
Pain diagnosis	Tu et al. (2019)	✗	✓	14
Aerosols	Jesson et al. (2021)	✓	✓	14
Industrial systems	Mogensen et al. (2024)	✓	✓	233
Semi synthetic generation	Cheng et al. (2023)	✓	✗	⊘
ODE	Kuramoto (1975)	✓	✗	⊘
Gen networks	Greenfield et al. (2010)	✓	✗	⊘
FMRI	Smith et al. (2011)	✓	✗	50
Benchmark kit (CauseMe)	Muñoz-Marí et al. (2020)	✓	✓/ ✗	5 / ⊘
Benchmark kit (OCBD)	Zhou et al. (2024a)	✓	✓/ ✗	11 / ⊘
Multi-Benchmark	**CausalRivers**	✓	✓	>1000

them for benchmarking as they become susceptible to overfitting. Of course, we are not the first to recognize the difficulty of benchmarking and comparisons in the causal discovery literature. Often, this situation is attributed to the fact that causal ground truth, along with proper observational data, is notoriously hard to find (Mogensen et al., 2024), (Niu et al., 2024). Noteworthy, some works that attempt to improve on this situation through other means are Montagna et al. (2023), which tries to assess the robustness of causal discovery methods towards violations of their assumptions, or Faller et al. (2024), which attempts to score methods based on their consistency on multiple subsets of data. Further, some approaches such as Muñoz-Marí et al. (2020), Niu et al. (2024) or Zhou et al. (2024b), aim to provide benchmarking through a collection of varying synthetic and semi-synthetic data sources. While these approaches are, of course, a step in the right direction and should be considered along real-world benchmarking, they are not sufficient to fully dissect performance differences of varying causal discovery methods for in-the-wild applications. Finally, as on recent and promising attempt to benchmark causal discovery performance, we want to mention Mogensen et al. (2024) as complementing work. Here, the ground truth graph is of sufficient size (Table 1) to properly benchmark performance. Further, sufficient time series data is available. Importantly, as the domain is completely distinct from ours, we see this work as a promising additional benchmarking approach.

3 BENCHMARK DESCRIPTION

Table 2: Overview of the three provided datasets in CausalRivers.

Name	Nodes	Edges	Start date	End date	Resolution
RiversEastGermany	666	651	1.1.2019	31.12.2023	15min
RiversBavaria	495	490	1.1.2019	31.12.2023	15min
RiversElbeFlood	44	29	09.09.2024	10.10.2024	15min

Here we provide information on the origin of the data included in our benchmark kit, as well as on the causal ground truth construction. Next, we discuss unique challenges for causal discovery on in-the-wild datasets and some specific features that are native to our data domain: Hydrology. Finally, to provide a comprehensive overview, we also include a list of features that we provide next to the data in our benchmarking kit, such as sampling strategies and naive baselines.

3.1 BENCHMARK CONSTRUCTION

This benchmarking kit is concerned with river discharge, so the amount of water that flows through a river. It is measured in m^3/s. As the amount of water measured at an upstream station directly influences the amount of water measured by all downstream stations at a later point in time, we consider them as causal. Through causal discovery, these causal relationships are potentially recoverable from observational data, in this case, time series data, alone. To produce the datasets provided in our benchmarking kit, we began by collecting information on available measurement stations in our selected geographical area. Through cooperation with eight different German state agencies (each state has its own network of measurement stations that serve primarily for flood prevention), we were provided with raw time series data along with some measurement station metadata. After some initial filtering (mostly removing duplicates and broken measurements), we ended up with around 666 and 494 valid time series for the selected time intervals. To construct the causal ground truth for these measurement stations, we leveraged a mixture of meta-information provided by the state agencies, remote-sensing (Wickel et al., 2007), Wikipedia information crawls and handcrafting for a semi-automatic construction of the graph. Further, all edges were double-checked by hand in the final stage to correct for potential matching errors. For documentary purposes, we provide the full construction pipeline in *ANONYMOUS* and note that it was specifically constructed in a way that allows adding additional nodes in the future. With this, and especially as there was recently a call for less static benchmarks (Shirali et al., 2023), we leave room to extend the provided data in the future. In summary, we provide three distinct sets of time series as displayed in Table 2, along with two ground truth causal graphs (Figure 1), as the RiversElbeFlood causal ground truth is a subset of the RiversEastGermany graph. Importantly, we envision RiversEastGermany as the primary benchmarking source as it is more diverse in terms of geography and data origin than RiversBavaria. Alternatively, we suggest RiversBavaria as a tool for the exploration of domain adaptation.

3.2 BENCHMARKING KIT FEATURES

To maximize the usability of this benchmarking kit, we provide additional tools and resources along with the time series and causal ground truth graphs. These tools and resources should allow researchers to tailor the dataset to their specific needs and evaluate the performance of methods in a more targeted and streamlined manner. Specifically, we provide:

- Tools to sample from ground truth causal graphs to access subgraphs with an arbitrary number of nodes. Further, subgraphs can be restricted through specific graph characteristics such as the connectivity or the geographical reality or data source. An example of such a sample can be found in Figure 2
- Tools to assess climatic conditions, especially precipitation, around any node by building on the German weather service DWD. These tools might be helpful for dissecting confounding effects and selecting specifically interesting time-series windows.
- Preprocessing, data loaders, and display tools for all included datasets.

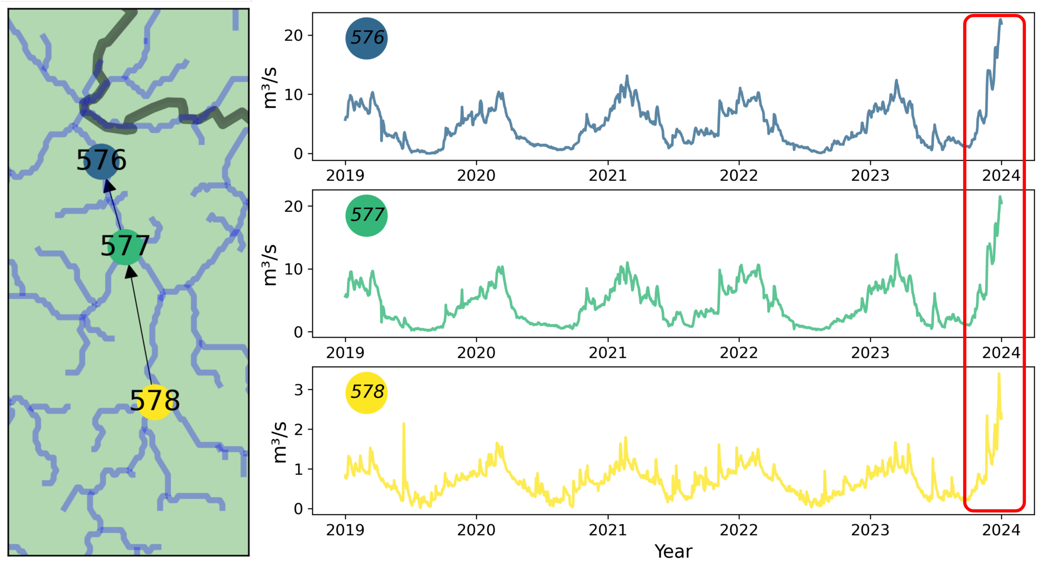


Figure 2: A single randomly sampled causal relationship along with time series data, originating from CausalRivers. A massive precipitation event is marked in *red*.

- An implementation of three naive baseline strategies that we deem necessary to evaluate performance properly (listed below).
- Tutorials on the usage of all provided tools and reproduce the results of section 4.

3.3 Baseline strategies

With our benchmarking kit, we provide three baseline causal discovery strategies. First, we determine the causal direction between two nodes (and the two corresponding time-series), here denoted as x_1 and x_2, purely based on cross-correlation between x_1 and lagged versions of x_2. For this, we look for the lag at which the cross-correlation is maximized. If this lag is negative, meaning the highest correlation is between the present of x_1 and the future of x_2, $x_1 \longrightarrow x_2$ is inferred, $x_1 \longleftarrow x_2$ otherwise. We call this strategy simply **CC** for Cross-Correlation.

Second, we rely on the actual magnitude of the time series, featuring a principle of causality that can be found in physics, where the mass of an object determines the causal direction (e.g. gravity). While in Physics, the arrow of causation typically points at the object with the lower mass, for rivers, this is reversed, as it is technically impossible that a very big river flows in a smaller river (at least without river splits). To leverage this principle, we simply asssume $x_1 \longrightarrow x_2$ if the mean of x_1 is bigger than the mean of x_2 . We call this strategy Reverse Physical, in short, **RP**. Notably, both RP and CC, decide on one direction for each potential edge. However, as it is typically the case that rivers only flow in a single location, we additionally restrict these strategies to select only one of the remaining links for each parent node. This is done either by selecting the next larger river (+N) or the biggest river (+B) as the only link or the river with the highest cross-correlation as the successor (+C). Finally, we evaluate the union between RP and CC, which we denote **Combo** and where we also test each restriction.

3.4 Unique Characteristics

Because our benchmark dataset covers a large area of Germany and is combined from multiple data sources, it exhibits a number of interesting unique features. Further, the domain of Hydrology brings, of course, its unique characteristics. In the following, we will discuss these attributes to help understand the complexity of the dataset. With this, we also hope to shine a light on the specific challenges and opportunities it provides for causal discovery.

Geographical Realities With over 1000 nodes, the datasets cover a wide range of geographical conditions, such as mountainous, coastal, and urban areas, and a wide variety of distances between stations. With this, it also covers a wide range of causal structures, lags, and strengths. Additionally, while the geographic closeness of nodes, influences the difficulty of detecting a causal relationship, other factors such as effect strength and elevation (and with this flow speed) also play a major role. The dataset includes a range of interesting geographical anomalies, such as dams, pump water storages, artificial canals, and tide effects, which can affect the causal relationships between nodes by altering the flow rate, water level, and consistency of relationships. A full list of cases, that we found particularly interesting is provided here: *ANONYMOUS*.

Weather Confounding Weather confounding plays a significant role in the innovation of all time series in the dataset. Rainfall can occur in a single node, across all nodes, or in a subset of nodes. Therefore, the impact of weather might be beneficial to determine causal direction (e.g., in the case where precipitation occurs in a single location) or be detrimental (e.g., in the case where precipitation occurs sequentially at different locations and in the reverse direction of the causal link). Further, as rainfall appears suddenly, the dataset is characterized by non-stationarity, non-linearity, and seasonal patterns. To visualize, Figure 2 displays the effect of a massive precipitation event at the end of the time series that affects all nodes.

Causal Lag Due to the varying distance and elevation between nodes, the speed of the rivers, and, in turn, the lag at which the causal effect occurs varies greatly throughout the dataset. Moreover, the causal lag of a specific relationship differs throughout the years as it depends on the amount of water that is present at a given time (the more water, the higher the velocity of the river.) We estimate this, along with weather confounding, to be a core challenge of the benchmark, as many causal discovery methods assume a static causal structure with a fixed lag.

Sampling Rate The sampling rate at which data is collected directly impacts the accuracy of inferred causal relationships (Gong et al., 2017; 2015). If the sampling rate is too low, critical causal interactions between variables are missed. Moreover, high-frequency sampling may increase the computational burden and result in models that overfit transient fluctuations rather than true causal interactions. As the dataset is provided in a 15-minute resolution, it allows researchers to explore the impact of different sampling and aggregation rates on causal discovery performance in real-world applications. Especially as the resolution far precedes the expected causal lag since stations often lie multiple kilometers apart.

Domain Biases Causal discovery methods typically integrate, besides sometimes allowing for the provision of a skeleton graph (Runge et al., 2019), little domain knowledge concerning potential causal links. Here, we want to note that depending on the domain, this might be unnecessarily agnostic. In the case of this benchmark kit, we note two specific features that, if leveraged, could be beneficial to improve performance. First, rivers typically have a single endpoint. Therefore nodes in this benchmark, with some exceptions in the form of river splits, also typically have a single child node. Secondly, the magnitude of the time series can reveal unlikely relationships as the amount of water is unlikely to reduce along the causal direction. While these specific biases here are quite specific to Hydrology, we expect that other biases in a similar manner exist in other domains and could also be utilized there. CausalRivers provides a foundation to explore such biases.

4 EXPERIMENTAL RESULTS AND DISCUSSION

To demonstrate our benchmark kit, we conducted three experiments demonstrating examples for possible use cases and gaining interesting insights into the performance of various causal discovery strategies. During these experiments, we deploy the following well-established methods from the literature: **PCMCI** with a linear conditional independence test (Runge et al., 2019), **Varlingam** (Hyvärinen et al., 2010), **Dynotears** (Pamfil et al., 2020) and a simple linear Granger causal approach (**VAR**), aiming at covering all archetypes (Assaad et al., 2022). Further, we evaluate two recent approaches featuring deep-learning techniques. First, a nonlinear Granger causal approach (CDMI, (Ahmad et al., 2022)), that analyzes residuals of deep networks under knockoff interventions to determine Granger causal relationships. Second, Causal Pretraining (**CP**, (Stein et al., 2024)), which learns a direct mapping (either a GRU or a Transformer) from multivariate time series to a causal graph from synthetic data and performs zero-shot inference for real-world samples. Notably, we specifically chose to include CP as it directly allows for domain adaption via finetuning. Finally, we always provide the performance of our proposed naive baselines along with these results.

As causal discovery methods typically come with at least some hyperparameters, we perform a rudimentary hyperparameter search per method to select proper values. For all experiments we test different resolutions (15min, 1H, 6H, 12H) and evaluate different max lags (3 and 5 for each resolution) if necessary. While we also evaluate a few method-specific parameters, we typically select default parameters. We report a full list of hyperparameter combinations evaluated in Appendix A.2. Notably, methods that require few hyperparameter configurations are more likely to be successful in practice, which should be considered when comparing methods. For all experiments, we chose to report the maximum F1 score (so the peak of the F1 score threshold relationship) of the best-performing hyperparameter combination as the final performance measure. Importantly, we ignore autoregressive links as these are always present and could potentially skew results. Here, it is important to keep in mind that this is a rather agnostic approach towards method failure. Performance is potentially overestimated when either a high variance of performance between different hyperparameter combinations exists or it is hard to determine decision thresholds. These are both complications that should be kept in mind for actual real-world applications.

4.1 Experiment set 1 - Varying Graph structures

As the first and most extensive experiment set, we perform causal discovery on subgraphs with varying graph characteristics and with the full-time series available. We take RiversEastGermany as the base graph for this experiment. For each set except the last one, we report results for graphs with three or five nodes. Notably, while we find these sub-selection criteria to be a great start for comparison, many other characteristics could be explored, such as, e.g., single-sink nodes, empty graphs, or causal pairs, to name only a few. The following graph characteristics were analyzed:

Random: We sample all possible connected subgraphs with three or five nodes. Notably this covers the entire dataset and with this, the complete diversity of conditions that the benchmarking kit offers.

Close: We sample all possible connected sub-graphs where every edge has a maximum geographic distance of five km. By excluding long distances, the causal effect should be more pronounced. Notably, all subgraphs of this selection are also included in ”Random”.

Random + 1: We sample all possible connected sub-graphs that have two or four nodes. We then add another disconnected node to the graph. To prevent confounding, we sample the random nodes from the coast and border area where we have several completely disconnected nodes.

Root cause: We sample all possible connected sub-graphs that have three or five nodes and where each has a maximum of one parent. With this, graphs are connected in the form of a single chain. We consider this useful for works on root-cause analysis (Ikram et al., 2022). Notably, all subgraphs of this selection are also included in ”Random”.

Confounding: Probably, the most interesting set, we here select sub-graphs with four or six nodes and where a single node has multiple children (while rare, these examples exist when rivers are naturally or artificially splitting). We then remove the node that has multiple children from the sample to simulate permanent hidden confounding scenarios.

Disjoint: We sample all possible connected sub-graphs that have five nodes and combine two of them into a single disjoint graph. To prevent connectivity, we choose to combine sampled with the largest possible distance between them. With this, we aim to evaluate how methods perform under a larger number of potential non-related variables.

The largest set, Random-5, holds more than 7500 subgraphs. The smallest set, Confounder-3, holds only 24 subgraphs. A full list of set sizes is reported in Appendix A.3. We report the results of this experiment in Table 3. With some exceptions, we found that our baselines are robust across the board, often achieving the highest F1 max. Additionally, they require no hyperparameter selection, a feature that we earlier noted to be beneficial for in-the-wild applications. Concerning established causal discovery approaches, we find the linear Granger causal approach (VAR) to be the most reliant. Further, we find that established causal discovery strategies often perform not better than even a null model, suggesting that they struggle with the nature of the provided data. As an explanation for these results, we propose that for some methods, the optimal decision boundary differs from sample to sample. As we calculate the F1 max once on the full graph set, as we deemed this more practical, the provided results do not account for this. We plan to further investigate this hypothesis in the future. Interestingly, some methods can improve over others for certain graph characteristics (e.g., PCMCI and CP on Random+1 with 3 variables). As the graph sets can be further split, CausalRivers should allow us to further analyze the corresponding underlying principles. We denote this as an additional future area of research. Finally, while both CP and CDMI allow for non-linearity and, to some extent, seasonality, we found no evidence for their superiority over linear approaches.

Table 3: F1 max scores for Experiment Set 1. Null model refers to predicting no causal Links which achieves the smallest possible F1 max. * CP networks are not able to process more than five variables. The datasets are ordered after their assumed difficulty. With some exceptions, baseline approaches achieve the most robust performance.

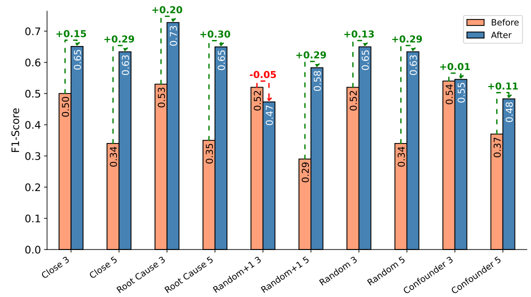
	Close		Root cause		Random +1		Random		Confounder		Disjoint
Method	3	5	3	5	3	5	3	5	3	5	10
Null model	0.50	0.34	0.50	0.33	0.29	0.26	0.50	0.34	0.54	0.36	0.16
CC	0.57	0.46	0.61	0.45	0.37	0.36	0.61	0.48	0.57	0.38	0.24
CC+C	0.52	0.41	0.61	0.52	0.45	0.43	0.59	0.47	0.54	0.37	0.47
RP	0.73	0.54	0.68	0.47	0.45	0.42	**0.71**	0.52	0.56	0.47	0.29
RP+B	**0.76**	**0.64**	0.50	0.33	0.55	**0.48**	0.68	0.49	0.54	0.42	0.23
RP+N	0.58	0.37	**0.75**	**0.63**	0.47	0.39	0.63	0.43	0.54	0.36	0.35
RPCC	0.62	0.52	0.59	0.43	0.42	0.40	0.62	0.51	0.54	0.38	0.30
RPCC+B	0.62	0.55	0.50	0.33	0.48	0.41	0.57	0.44	0.54	0.36	0.20
RPCC+N	0.54	0.45	0.64	0.55	0.44	0.39	0.58	0.44	0.54	0.36	0.36
RPCC+C	0.58	0.49	0.58	0.48	0.49	0.44	0.60	0.47	0.54	0.39	**0.50**
VAR	0.72	0.59	0.66	0.50	0.51	0.47	0.70	**0.54**	**0.58**	**0.49**	0.39
Dynotears	0.50	0.42	0.50	0.34	0.29	0.37	0.50	0.42	0.55	0.37	0.37
Varlingam	0.50	0.35	0.50	0.35	0.33	0.29	0.50	0.35	0.56	0.39	0.21
PCMCI	0.50	0.34	0.50	0.35	0.42	0.37	0.51	0.36	0.56	0.37	0.39
CDMI	0.50	0.34	0.51	0.33	0.31	0.27	0.50	0.33	0.54	0.36	0.17
CP (Gru)	0.50	0.34	0.53	0.35	0.52	0.29	0.52	0.34	0.54	0.37	-*
CP (Transf)	0.50	0.34	0.52	0.40	**0.54**	0.34	0.50	0.38	0.55	0.36	-*

4.2 Experiment set 2 - Time series subsampling

Given that the full time-series is very long (roughly 175k time steps for the original resolution), we were interested in whether selecting specific shorter, and hopefully informative, subsections might influence the performance of causal discovery algorithms. As a motivation, one might imagine that the full time-series most likely holds sections with little innovation, displays annual patterns, and includes nonstationary windows with high amounts of change (such as RiversElbeFlood). To test whether providing only a subselection can improve in-the-wild causal discovery, we restrict the causal ground truth graph to the 44 nodes included in RiversElbeFlood. We then compare the causal discovery performance on the RiversElbeFlood dataset with the performance on the full-time series and with the performance on a month with almost no recorded precipitation (Oktober 2021) in the selected region. Concerning subgraphs, we simply sample all possible graphs with five nodes, equal to the sampling strategy ”random” from Experiment Set 1. We provide the results of this comparison in Figure 3a. While our results suggest that Flood data generally decreases performance, the dataset with little precipitation shows mixed results. Notably, however, it strongly reduces the performance of Dynotears, which we take as evidence that it can affect method performance in some cases, which has implications for real-world applications. We attribute this to the fact that Dynotears is a gradient-based method that could be affected more by little innovation in the data. Next, we note that the performance on this subset of the ground truth causal graph is generally higher than in

(a) Changes in method performance depending on the provided data. We find mixed results with some pronounced differences, e.g. for Dynotears.

	Full TS	No Rain	Flood
CC	**0.56**	0.51 ↓	**0.56**
RP	**0.62**	**0.62**	0.52 ↓
RPCC	**0.61**	0.59 ↓	0.56 ↓
VAR	0.62	↑ **0.63**	0.57 ↓
Dynotears	0.61	0.40 ↓	↑ **0.62**
Varlingam	0.39	↑ **0.41**	0.35 ↓
PCMCI	0.35	↑ **0.37**	0.34 ↓
CP	**0.43**	0.38 ↓	0.41 ↓



(b) Performance increase, achieved through finetuning CP on domain samples. Such a domain adaptation strongly increases performance.

Figure 3: F1 max scores for Experiment Set 2 (a) and Experiment Set 3 (b). In (a), we mark increases and decreases in performance with ↑ and ↓, respectively. Further, the highest performance per method is marked in **bold**.

experiment set 1. We attribute this to the geographical location and the data origin of the nodes included in RiversElbeFlood. Despite clear results, focusing on such a selection strategy might be a way forward to make causal discovery methods more robust in real-world applications.

4.3 Experiment set 3 - Domain adaption

As a final evaluation, we leverage the fact that we include two distinct ground truth graphs to provide results on whether domain adaptation can be leveraged to improve causal discovery performance. As this area of research is not yet widely explored, we provide a first example of domain adaptation via Causal Pretraining (CP), a method that specifically allows for it, as causally pre-trained neural networks can be updated by finetuning in a supervised manner. We, therefore, investigate whether the previously reported performance of CP on the RiversEastGermany dataset can be improved. To execute this, we leverage RiversBavaria and sample training examples (identical to sampling strategy "random" and for five variables) from it on which we finetune a pre-trained network provided by Stein et al. (2024). We perform a small hyperparameter search, testing for different values of the learning rate, weight decay, time-series resolution, normalization, and the CP architecture. After training, we evaluate the network that achieved the highest F1 max during training (a GRU on 6H resolution and no normalization) again on all graph characteristics that were evaluated during Experiment Set 1. We report the results in Figure 3b. With the exception of one graph set, finetuning (and with that domain adaption) strongly improves the performance of CP. Further, on the graph set characteristic that CP was fine-tuned on, the final performance of CP (0.633 F1 max) clearly surpasses the previously best scoring method (VAR with an F1 max of 0.54). We take this as strong evidence that domain adaptation should be explored further by the community.

5 Conclusion

In this paper, we presented CausalRivers, the largest in-the-wild causal discovery benchmarking kit for time series data to date. After motivating the need for such a benchmark by summarizing alternative datasets, we discussed the benchmarking kit and its unique challenges and opportunities. Further, we conducted a set of experiments, aiming at an evaluation of causal discovery approaches in real-world applications and an exploration of potential beneficial strategies. As our experiments showed, many well-established causal discovery methods underperform in real-world applications and are outperformed by simple but robust baseline strategies. With this, we conclude that more research is necessary, focusing on in-the-wild robustness, potentially through selecting relevant sections of a given time series, and domain adaptation. To conclude, we hope that this work provides the foundation for a benchmark-driven method development of causal discovery methods. We also hope to inspire the development of other benchmarking approaches and are excited to see which causal discovery approaches prove to be the most successful in the end.

REFERENCES

Wasim Ahmad, Maha Shadaydeh, and Joachim Denzler. Causal Discovery using Model Invariance through Knockoff Interventions. In *ICML 2022: Workshop on Spurious Correlations, Invariance and Stability*, July 2022. URL https://openreview.net/forum?id=OcNeMVbIdCF.

Md Zahangir Alom, Tarek M. Taha, Christopher Yakopcic, Stefan Westberg, Paheding Sidike, Mst Shamima Nasrin, Brian C. Van Esesn, Abdul A. S. Awwal, and Vijayan K. Asari. The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches, September 2018. URL http://arxiv.org/abs/1803.01164. arXiv:1803.01164 [cs].

Mohammed AlQuraishi. AlphaFold at CASP13. *Bioinformatics*, 35(22):4862–4865, November 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz422. URL https://doi.org/10.1093/bioinformatics/btz422.

Holly Andersen. When to Expect Violations of Causal Faithfulness and Why It Matters. *Philosophy of Science*, 80(5):672–683, December 2013. ISSN 0031-8248, 1539-767X. doi: 10.1086/673937. URL https://www.cambridge.org/core/journals/philosophy-of-science/article/when-to-expect-violations-of-causal-faithfulness-and-why-it-matters/307D69C797503709BEB5ED34A350EBAF.

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, February 2022. ISSN 1076-9757. doi: 10.1613/jair.1.13428. URL https://www.jair.org/index.php/jair/article/view/13428.

Stefan Bauer, Bernhard Schölkopf, and Jonas Peters. The Arrow of Time in Multivariate Time Series. In *Proceedings of The 33rd International Conference on Machine Learning*, pp. 2043–2051. PMLR, June 2016. URL https://proceedings.mlr.press/v48/bauer16.html.

O. Bojar, C. Buck, C. Federmann, B. Haddow, P. Koehn, J. Leveling, C. Monz, P. Pecina, M. Post, H. Saint-Amand, R. Soricut, L. Specia, and A. Tamchyna. *Findings of the 2014 Workshop on Statistical Machine Translation*. Stroudsburg, PAAssociation for Computational Linguistics, 2014. ISBN 9781941643174. URL https://dare.uva.nl/search?identifier=9fb31ff0-f332-4fd5-939d-a7fd446a06d8.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proceedings of the IEEE*, 105(10):1865–1883, October 2017. ISSN 1558-2256. doi: 10.1109/JPROC.2017.2675998. URL https://ieeexplore.ieee.org/abstract/document/7891544.

Yuxiao Cheng, Ziqian Wang, Tingxiong Xiao, Qin Zhong, Jinli Suo, and Kunlun He. CausalTime: Realistically Generated Time-series for Benchmarking of Causal Discovery, October 2023. URL http://arxiv.org/abs/2310.01753. arXiv:2310.01753 [cs, stat].

Mathieu Chevalley, Yusuf Roohani, Arash Mehrjou, Jure Leskovec, and Patrick Schwab. CausalBench: A Large-scale Benchmark for Network Inference from Single-cell Perturbation Data, July 2023. URL http://arxiv.org/abs/2210.17283. arXiv:2210.17283 [cs].

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3213–3223, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Cordts_The_Cityscapes_Dataset_CVPR_2016_paper.html.

Payam Dibaeinia and Saurabh Sinha. SERGIO: A Single-Cell Expression Simulator Guided by Gene Regulatory Networks. *Cell Systems*, 11(3):252–271.e11, September 2020. ISSN 2405-4712. doi: 10.1016/j.cels.2020.08.003. URL https://www.sciencedirect.com/science/article/pii/S2405471220302878.

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. Automated versus Do-It-Yourself Methods for Causal Inference: Lessons Learned from a Data Analysis Competition. *Statistical Science*, 34(1):43–68, February 2019. ISSN 0883-4237, 2168-8745. doi: 10.1214/18-STS667. URL https://projecteuclid.org/journals/statistical-science/volume-34/issue-1/Automated-versus-Do-It-Yourself-Methods-for-Causal-Inference/10.1214/18-STS667.full.

Philipp M. Faller, Leena C. Vankadara, Atalanti A. Mastakouri, Francesco Locatello, and Dominik Janzing. Self-Compatibility: Evaluating Causal Discovery without Ground Truth. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, pp. 4132–4140. PMLR, April 2024. URL https://proceedings.mlr.press/v238/faller24a.html.

Mingming Gong, Kun Zhang, Bernhard Schoelkopf, Dacheng Tao, and Philipp Geiger. Discovering Temporal Causal Relations from Subsampled Data. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 1898–1906. PMLR, June 2015. URL https://proceedings.mlr.press/v37/gongb15.html.

Mingming Gong, Kun Zhang, Bernhard Schölkopf, Clark Glymour, and Dacheng Tao. Causal Discovery from Temporally Aggregated Time Series. *Uncertainty in artificial intelligence : proceedings of the ... conference. Conference on Uncertainty in Artificial Intelligence*, 2017:269, August 2017. ISSN 1525-3384. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5995575/.

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, October 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0013397. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0013397.

Isabelle Guyon, Constantin Aliferis, Greg Cooper, André Elisseeff, Jean-Philippe Pellet, Peter Spirtes, and Alexander Statnikov. Design and Analysis of the Causation and Prediction Challenge. In *Proceedings of the Workshop on the Causation and Prediction Challenge at WCCI 2008*, pp. 1–33. PMLR, December 2008. URL http://proceedings.mlr.press/v3/guyon08a.html.

Konstantin Göbler, Tobias Windisch, Mathias Drton, Tim Pychynski, Steffen Sonntag, and Martin Roth. $\texttt{causalAssembly}$: Generating Realistic Production Data for Benchmarking Causal Discovery, February 2024. URL http://arxiv.org/abs/2306.10816. arXiv:2306.10816 [cs, stat].

P. Richard Hahn, Vincent Dorie, and Jared S. Murray. Atlantic Causal Inference Conference (ACIC) Data Analysis Challenge 2017, May 2019. URL http://arxiv.org/abs/1905.09515. arXiv:1905.09515 [stat].

Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O. Hoyer. Estimation of a Structural Vector Autoregression Model Using Non-Gaussianity. *Journal of Machine Learning Research*, 11(56):1709–1731, 2010. ISSN 1533-7928. URL http://jmlr.org/papers/v11/hyvarinen10a.html.

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, December 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/c9fcd02e6445c7dfbad6986abee53d0d-Abstract-Conference.html.

Andrew Jesson, Peter Manshausen, Alyson Douglas, Duncan Watson-Parris, Yarin Gal, and Philip Stier. Using Non-Linear Causal Models to Study Aerosol-Cloud Interactions in the Southeast Pacific, November 2021. URL http://arxiv.org/abs/2110.15084. arXiv:2110.15084 [physics].

Yoshiki Kuramoto. Self-entrainment of a population of coupled non-linear oscillators. *Mathematical Problems in Theoretical Physics*, 39:420–422, January 1975. doi: 10.1007/BFb0013365. URL https://ui.adsabs.harvard.edu/abs/1975LNP....39..420K. ADS Bibcode: 1975LNP....39..420K.

S. L. Lauritzen and D. J. Spiegelhalter. Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):157–194, 1988. ISSN 2517-6161. doi: 10.1111/j.2517-6161.1988.tb01721.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1988.tb01721.x.

Daniel McDuff, Yale Song, Jiyoung Lee, Vibhav Vineet, Sai Vemprala, Nicholas Alexander Gyde, Hadi Salman, Shuang Ma, Kwanghoon Sohn, and Ashish Kapoor. CausalCity: Complex Simulations with Agency for Causal Discovery and Reasoning. In *Proceedings of the First Conference on Causal Learning and Reasoning*, pp. 559–575. PMLR, June 2022. URL https://proceedings.mlr.press/v177/mcduff22a.html.

Giovanni Menegozzo, Diego Dall'Alba, and Paolo Fiorini. CIPCaD-Bench: Continuous Industrial Process datasets for benchmarking Causal Discovery methods. In *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)*, pp. 2124–2131, August 2022. doi: 10.1109/CASE49997.2022.9926420. URL https://ieeexplore.ieee.org/abstract/document/9926420. ISSN: 2161-8089.

Søren Wengel Mogensen, Karin Rathsman, and Per Nilsson. Causal discovery in a complex industrial system: A time series benchmark. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, pp. 1218–1236. PMLR, March 2024. URL https://proceedings.mlr.press/v236/mogensen24a.html.

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*, 36:47339–47378, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/93ed74938a54a73b5e4c52bbaf42ca8e-Abstract-Conference.html.

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. ISSN 1533-7928. URL http://jmlr.org/papers/v17/14-518.html.

J. Muñoz-Marí, G. Mateo, J. Runge, and G. Camps-Valls. CauseMe: An online system for benchmarking causal discovery methods., 2020. In preparation (2020).

Brady Neal, Chin-Wei Huang, and Sunand Raghupathi. RealCause: Realistic Causal Inference Benchmarking, March 2021. URL http://arxiv.org/abs/2011.15007. arXiv:2011.15007 [cs, stat].

Wenjin Niu, Zijun Gao, Liyan Song, and Lingbo Li. Comprehensive Review and Empirical Evaluation of Causal Discovery Algorithms for Numerical Data, July 2024. URL http://arxiv.org/abs/2407.13054. arXiv:2407.13054 [cs].

Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. DYNOTEARS: Structure Learning from Time-Series Data. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pp. 1595–1605. PMLR, June 2020. URL https://proceedings.mlr.press/v108/pamfil20a.html.

Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3 (none):96–146, January 2009. ISSN 1935-7516. doi: 10.1214/09-SS057. URL https://projecteuclid.org/journals/statistics-surveys/volume-3/issue-none/Causal-inference-in-statistics-An-overview/10.1214/09-SS057.full.

Jonas Peters, Dominik Janzing, and Bernhard Scholkopf. Causal Inference on Discrete Data Using Additive Noise Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12): 2436–2450, December 2011. ISSN 1939-3539. doi: 10.1109/TPAMI.2011.71. URL `https://ieeexplore.ieee.org/abstract/document/5740928`.

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

Aditya Pratapa, Amogh P. Jalihal, Jeffrey N. Law, Aditya Bharadwaj, and T. M. Murali. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods*, 17(2):147–154, February 2020. ISSN 1548-7105. doi: 10.1038/s41592-019-0690-6.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ Questions for Machine Comprehension of Text, October 2016. URL `http://arxiv.org/abs/1606.05250`. arXiv:1606.05250 [cs].

Alexander Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the Simulated DAG! Causal Discovery Benchmarks May Be Easy to Game. In *Advances in Neural Information Processing Systems*, volume 34, pp. 27772–27784. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper/2021/hash/e987eff4a7c7b7e580d659feb6f60c1a-Abstract.html`.

Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting causal associations in large nonlinear time series datasets. *Science Advances*, 5(11):eaau4996, November 2019. ISSN 2375-2548. doi: 10.1126/sciadv.aau4996. URL `http://arxiv.org/abs/1702.07007`. arXiv:1702.07007 [physics, stat].

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, April 2005. doi: 10.1126/science.1105809. URL `https://www.science.org/doi/10.1126/science.1105809`.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, December 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/hash/a1859debfb3b59d094f3504d5ebb6c25-Abstract-Datasets_and_Benchmarks.html`.

Yishai Shimoni, Chen Yanover, Ehud Karavani, and Yaara Goldschmnidt. Benchmarking Framework for Performance-Evaluation of Causal Inference Analysis, March 2018. URL `http://arxiv.org/abs/1802.05046`. arXiv:1802.05046 [cs, stat].

Ali Shirali, Rediet Abebe, and Moritz Hardt. A Theory of Dynamic Benchmarks, March 2023. URL `http://arxiv.org/abs/2210.03165`. arXiv:2210.03165.

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, January 2011. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2010.08.063. URL `https://www.sciencedirect.com/science/article/pii/S1053811910011602`.

David J. Spiegelhalter, A. Philip Dawid, Steffen L. Lauritzen, and Robert G. Cowell. Bayesian Analysis in Expert Systems. *Statistical Science*, 8(3):219–247, 1993. ISSN 0883-4237. URL `https://www.jstor.org/stable/2245959`.

Peter Spirtes. An Anytime Algorithm for Causal Inference. In *International Workshop on Artificial Intelligence and Statistics*, pp. 278–285. PMLR, January 2001. URL `https://proceedings.mlr.press/r3/spirtes01a.html`.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, January 2001. ISBN 9780262527927. Google-Books-ID: OZ0vEAAAQBAJ.

Gideon Stein, Maha Shadaydeh, and Joachim Denzler. Embracing the black box: Heading towards foundation models for causal discovery from time series data, February 2024. URL http://arxiv.org/abs/2402.09305. arXiv:2402.09305 [cs].

Michael Strevens. *The Knowledge Machine: How Irrationality Created Modern Science*. Liveright Publishing, October 2020. ISBN 9781631491382. Google-Books-ID: ISXWDwAAQBAJ.

Ruibo Tu, Kun Zhang, Bo Bertilson, Hedvig Kjellstrom, and Cheng Zhang. Neuropathic Pain Diagnosis Simulator for Causal Discovery Algorithm Evaluation. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://papers.nips.cc/paper_files/paper/2019/hash/4fdaa19b1f22a4d926fce9bfc7c61fa5-Abstract.html.

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):43, January 2006. ISSN 1471-2105. doi: 10.1186/1471-2105-7-43. URL https://doi.org/10.1186/1471-2105-7-43.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, December 2017. URL http://arxiv.org/abs/1706.03762. arXiv:1706.03762 [cs].

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):82:1–82:36, November 2022. ISSN 0360-0300. doi: 10.1145/3527154. URL https://doi.org/10.1145/3527154.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding, February 2019. URL http://arxiv.org/abs/1804.07461. arXiv:1804.07461 [cs].

Tracy Sellers Warwick Nash. Abalone, 1994. URL https://archive.ics.uci.edu/dataset/1.

B. A. Wickel, B. Lehner, and N. Sindorf. HydroSHEDS: A global comprehensive hydrographic dataset. *American Geophysical Union, Fall Meeting*, 2007:H11H–05, December 2007. URL https://ui.adsabs.harvard.edu/abs/2007AGUFM.H11H..05W. ADS Bibcode: 2007AGUFM.H11H..05W.

Steven Fenves Yoram Reich. Pittsburgh Bridges, 1989. URL https://archive.ics.uci.edu/dataset/18.

Wei Zhou, Hong Huang, Guowen Zhang, Ruize Shi, Kehan Yin, Yuanyuan Lin, and Bang Liu. OCDB: Revisiting Causal Discovery with a Comprehensive Benchmark and Evaluation Framework, June 2024a. URL http://arxiv.org/abs/2406.04598. arXiv:2406.04598 [cs].

Yu Zhou, Xingyu Wu, Beicheng Huang, Jibin Wu, Liang Feng, and Kay Chen Tan. CausalBench: A Comprehensive Benchmark for Causal Learning Capability of Large Language Models, April 2024b. URL http://arxiv.org/abs/2404.06349. arXiv:2404.06349 [cs].