Introducing CausalBench: A Flexible Benchmark Framework for Causal Analysis and Machine Learning*

Ahmet Kapkiç
akapkic@asu.edu
Arizona State University
Tempe, AZ, USA

Pratanu Mandal
pmandal5@asu.edu
Arizona State University
Tempe, AZ, USA

Shu Wan
swan@asu.edu
Arizona State University
Tempe, AZ, USA

Paras Sheth
psheth5@asu.edu
Arizona State University
Tempe, AZ, USA

Abhinav Gorantla
agorant2@asu.edu
Arizona State University
Tempe, AZ, USA

Yoonhyuk Choi
ychoi139@asu.edu
Arizona State University
Tempe, AZ, USA

Huan Liu
huanliu@asu.edu
Arizona State University
Tempe, AZ, USA

K. Selçuk Candan
candan@asu.edu
Arizona State University
Tempe, AZ, USA

Abstract

While witnessing the exceptional success of machine learning (ML) technologies in many applications, users are starting to notice a critical shortcoming of ML: correlation is a poor substitute for causation. The conventional way to discover causal relationships is to use randomized controlled experiments (RCT); in many situations, however, these are impractical or sometimes unethical. Causal learning from observational data offers a promising alternative. While being relatively recent, causal learning aims to go far beyond conventional machine learning, yet several major challenges remain. Unfortunately, advances are hampered due to the lack of unified benchmark datasets, algorithms, metrics, and evaluation service interfaces for causal learning. In this paper, we introduce *CausalBench*, a transparent, fair, and easy-to-use evaluation platform, aiming to (a) enable the advancement of research in causal learning by facilitating scientific collaboration in novel algorithms, datasets, and metrics and (b) promote scientific objectivity, reproducibility, fairness, and awareness of bias in causal learning research. CausalBench provides services for benchmarking data, algorithms, models, and metrics, impacting the needs of a broad of scientific and engineering disciplines.

CCS Concepts

• **Information systems → Computing platforms**.

Keywords

Benchmark, Causality, Machine Learning, Dataset, Model, Metric

ACM Reference Format:
Ahmet Kapkiç, Pratanu Mandal, Shu Wan, Paras Sheth, Abhinav Gorantla, Yoonhyuk Choi, Huan Liu, and K. Selçuk Candan. 2024. Introducing CausalBench: A Flexible Benchmark Framework for Causal Analysis and Machine Learning. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3627673.3679218

1 Introduction

Machine learning (ML) is serving as a key pillar in scientific innovation [6] in a myriad of high-impact science domains, such as medical science [12], epidemiology [32], and environmental health [14]. Nevertheless, users are starting to notice a critical shortcoming of the traditional ML techniques, which can learn correlation-based patterns from data (Figure 1): data may contain spurious correlations and correlation is a poor substitute for causation [28].

Consequently, successfully tackling many urgent challenges in socio-economically critical domains requires a deeper understanding of causal relationships and interactions from observational data, and causal learning offers a promising alternative to correlation-based learning [1, 15, 19]. For example, developing a plan for combining, co-operating, and designing portfolios of natural and built water infrastructure requires an understanding of the causally complex interplay of entities in a multi-layer network, including physics underlying natural as well as built infrastructures for flood protection, erosion control, water storage, and purification[1] [24, 26, 27].

Standardized evaluation played a major role in ML development and contributed to the impressive impact of ML in scientific innovation. Successful early benchmarking efforts, such as UCI ML and UCI KDD repositories [9, 13, 23], not only helped guide the development of efficient and effective ML algorithms but also encouraged

*This work is supported by NSF grant 2311716, "CausalBench: A Cyberinfrastructure for Causal-Learning Benchmarking for Efficacy, Reproducibility, and Scientific Collaboration"

CIKM '24, October 21–25, 2024, Boise, ID, USA

ACM ISBN 979-8-4007-0436-9/24/10
https://doi.org/10.1145/3627673.3679218

[1]This research has been funded by a US Army Corps of Engineers Engineering With Nature Initiative through Cooperative Ecosystem Studies Unit Agreement #W912HZ-21-2-0040

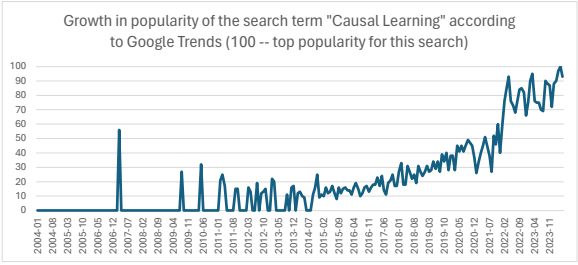

Figure 1: Causal learning has exploded in popularity in recent years

Table 1: Popular causal ML tools with the supported data, methods, and metrics [4]

		C. Effect Estim.			C. Discovery		Evaluation	
		CausalML	EconML	DoWhy	CausalNex	CausalDisc.	C-Benchm.	JustCause
Data	iid	x	x	x	x	x	x	x
	IV	x	x	x				
	Graph							
	T.Series			x				
Methods	Pro. Score		x	x				x
	Tree-based	x	x	x				
	Meta-Learn.	x	x	x				x
	Doubly ML		x	x				
	D.Robust		x	x				x
	IV	x	x	x				
	Mediation			x				
	Graph				x	x		
	Pairwise					x		
Metrics	PEHE							x
	RMSE	x					x	x
	MAE	x						x
	BIAS						x	x
	Coverage						x	
	Conf.Int.						x	
	Agg.Score						x	
	Refut.			x				
	SID					x		
	SHD					x		
	Classif.				x	x		

User Provided Metrics

collaborative research and paved the way for recent breakthroughs in deep learning. For example, to evaluate an image classifier, we have widely used metrics (e.g., accuracy, F1 score, ROC-AUC [2]), procedures (e.g., cross-validation [29]), and datasets (e.g., MNIST [8], CIFAR10 [16], ImageNet [7]). More recent frameworks, such as [30], move towards a collaborative approach, where datasets, models, and metrics are provided by the members of the community.

In this paper, we argue that the causal learning community can achieve the same by meticulously surveying the emerging field of vibrant research, systematically categorizing the existing benchmarking efforts into technically meaningful groups, and discovering the areas where further efforts are in dire need. While initial work in this area has started (Table 1), more systematic advances are required. Shared datasets and metrics for benchmarking can be extremely valuable for not only causal learning algorithm design, but also for comparison and benchmarking of available solutions. Currently, only a fraction of existing studies are replicable and with each version of a GPU driver or a Python library, performance results can vary wildly. Yet, despite the promise of advancing science and research, such data can be difficult to find and costly to annotate. Here, we argue that the recent availability of big observational data in all walks of life offers us an unprecedented opportunity to consolidate the hitherto distributed and unorganized efforts by creating a cyberinfrastructure for advancing causal learning research.

Based on this premise, here we introduce the *CausalBench* platform, a novel cyberinfrastructure for benchmarking causal learning. Aiming for a systematic, objective, and transparent evaluation of causal learning models and algorithms, CausalBench integrates publicly available benchmarks and consensus-building standards for the evaluation of causal learning models and algorithms from observational data. Consisting of a publicly accessible data and algorithm repository along with service APIs, the platform assists researchers and developers in easily applying and effectively evaluating (a) causal inference, (b) causal discovery, and (c) causal interpretability algorithms with a variety of standard metrics, procedures, and large-scale datasets.

In the rest of this paper, we first discuss the principles that are the pillars of CausalBench (Section 2). We then provide an overview of the framework and its functionalities (Section 3). In Section 4, we discuss usage scenarios of the system.

2 Key Objectives and Design Principles

As a platform to systematically and reliably benchmark causal learning models and algorithms, CausalBench aims to target the following key objectives:

- *Objective #1: Universally adopted metrics, procedures, and datasets.* This involves conducting an extensive identification of existing datasets, performance metrics, and procedures used in the evaluation of state-of-the-art causal learning algorithms, and developing an "ontology" for benchmarking to standardize the evaluation methodology, improve transparency, and promote collaboration to advance causal learning efficiently.
- *Objective #2: A standard and convenient way for the community to contribute data and models.* Different from datasets for conventional machine learning, it is often difficult to obtain the ground truth of the causal relations among observed variables, not to mention the potential existence of unobserved variables – in many cases, we have to work with datasets with incomplete causal knowledge. We need to make it easy and convenient for the community to contribute new data and models.
- *Objective #3: Trustable (transparent, reproducible) benchmarking.* In addition to making data, models, and metrics available to the researchers, the system should enable trustable, fair, reproducible, and open benchmarking of the available models and algorithms. In particular, all steps of an executed experiment, including the data, hyperparameters, as well as hardware/software configuration must be recorded and made transparently available to help support interpretation of the experiment results.
- *Objective #4: Fair and flexible comparisons of models and algorithms.*Conversely, one should be able to explore the results of recorded benchmark experiments and compare existing solutions fairly and flexibly. Fairness implies that if the models are compared, these models and the experiment settings must be compatible, and/or any differences in data, hyperparameters, and hardware/software settings that may impact the results must be highlighted. A fair system should account for biases caused by algorithms or system configurations. Flexibility means that the users of the system must be able to *slice-and-dice* the benchmark experiments in different ways, based on a different grouping or slicing criteria.

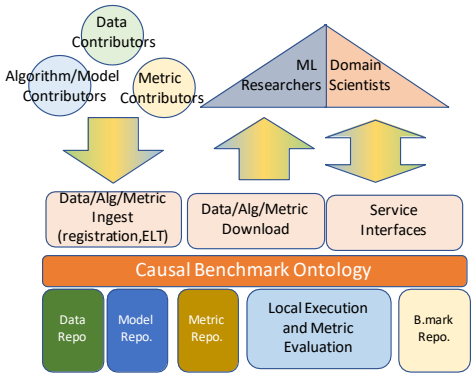

Figure 2: Overview of CausalBench

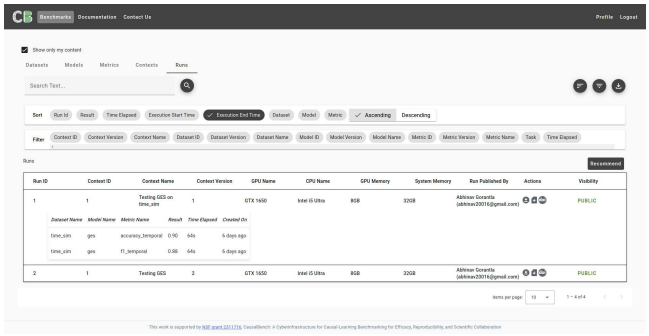

Figure 3: CausalBench runs page

3 CausalBench Framework

3.1 Overall Architecture

CausalBench is designed to enable its users to easily add new relevant datasets, models, and metrics (Figure 2). The platform boasts several key components:

- A web-based dataset, model, and metric registration module provides a guided interface through which a provider registers a dataset, a model, or a metric with CausalBench. Registration involves the systematic acquisition of metadata needed for the discovery, access, and use of data and models.
- A data, model, and metric repository manages metadata associated with all registered datasets, models, and metrics and ensures that these persist and are accessible. The repository further stores (a) benchmark contexts and experiment setups consisting of data, model, and metric components and (b) authenticated performance results of benchmark runs and the associated metadata (e.g., hyperparameters, hardware/software setups).
- A benchmark runs page (Figure 3) where performance results of runs, including results, system information, and a DOI attached to each benchmark run, is displayed. Experiment results are in a tabular format that can be sorted and filtered.
- A CausalBench console-based Python package supports the execution of causal machine learning experiments. The package enables quantitative evaluation of the models (for accuracy and efficiency) based on datasets in the repository using local CPU and GPU resources.
- A web interface supports browsing through repositories of datasets, models, metrics and benchmark contexts, exploring (slice-and-dice) experiments across the runs executed through CausalBench. In addition to providing data download links and data descriptions, the platform also offers accessible APIs of evaluation metrics and service interfaces.

3.2 Benchmarking Causal ML Models

CausalBench includes several core components. These include **datasets**, $\mathcal{D}$, which are data files and configuration files that describe the properties of the data in the data files; **models**, $\mathcal{M}$, which are algorithms written in Python that take in a dataset and execute a particular model, producing outputs based on the tasks and models; and **metrics** $\mathcal{A}$, which are Python implementation of metric calculations that take in the outputs provided by the model and output a numerical value, based on its configuration. CausalBench follows a flexible approach, where datasets, models, and metrics can be re-used for different causal machine learning tasks. The set of all causal machine learning tasks available at CausalBench is denoted as $\mathcal{T}$. Given the above, a **benchmark context**, C, includes a subset (denoted by the subscript $_P$) of datasets $\mathcal{D}$, models $\mathcal{M}$ and accuracy metrics $\mathcal{A}$, along with the appropriate parameter and hyperparameter settings:

$$C = \{(\mathcal{D}_P, \mathcal{M}_P, \mathcal{A}_P, \mathcal{H}_P), \mathcal{D}_P \subseteq \mathcal{D},\ \mathcal{M}_P \subseteq \mathcal{M},\ \mathcal{A}_P \subseteq \mathcal{A}\}.$$

Above, $\mathcal{H}_P$ denotes the set of **parameter and hyperparameter settings** applicable to the execution or training of the models. Note that the benchmark context can equivalently be seen as a set of **benchmark scenarios**:

$$C = \{(d, m, \mathcal{A}_P, h) \mid d \in \mathcal{D}_P,\ m \in \mathcal{M}_P,\ h \in \mathcal{H}_P\}.$$

An **instrumented context**, $\mathcal{I}$, is a coupling of these benchmark scenarios with a particular user hardware/software system, s:

$$\mathcal{I}(C, s) = \{(d, m, \mathcal{A}_P, h, s) \mid d \in \mathcal{D}_P,\ m \in \mathcal{M}_P, ; h \in \mathcal{H}_P\}.$$

A **benchmark run**, $\mathcal{R}(\mathcal{I}(C, s))$, then, is the recording of the outputs of the execution of the benchmark scenarios in an instrumented context, $\mathcal{I}$:

$$\{(A, T, S; d, m, h, s) \mid (d, m, \mathcal{A}_P, h, s) \in \mathcal{I}(C, s)\},$$

where A is a set of key-value pairs recording the value for each accuracy metric $a \in \mathcal{A}_P$. T is a set of key-value pairs recording the timing values for each timing metrics, such as *CPU-time*, *GPU-time*; and S is a set of key-value pairs recording the system usage values for each resource metrics, such as *CPU-memory*, *GPU-memory*. Noting that the timing metrics T and resource metrics S are measured for each benchmark scenario. CausalBench stores authenticated benchmark runs of its users in public or private repositories and allows a user to compare multiple runs (that are accessible to them) of a task, dataset, and/or model.

3.3 Reproducibility and Versioning

In order to enable reproducible research on causal machine learning, once a dataset, model, or a metric is declared as public and is included in at least one public run, it becomes permanent in the system and cannot be removed. Benchmark runs that are made public are registered with an open-access repository, Zenodo, and

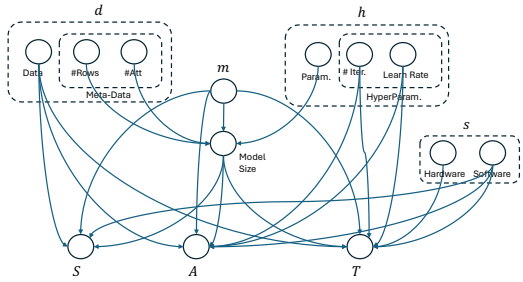

Figure 4: Outline of the causal graph enabling the causally-informed exploration and analysis of a benchmark

are associated with a unique document object identifier (DOI). Of course, over time, datasets, models, and metrics may evolve. For any public, and therefore permanent, component, this involves the creation of a new version of the component, with its own unique identifier, maintained along with the old version.

3.4 CausalBench Features

The Python package for CausalBench is written in Python 3.10, and facilitates the creation, uploading, and executing of core components (a dataset, model, or metric) and contexts. Running CausalBench requires signing up to the system through the CausalBench website and providing the user credentials in the config file of the Python package. A user has several options available on launch: downloading/uploading a component, declaring and executing a benchmark run, and exploring existing benchmarks.

Exploring Data, Model, and Metric Repositories. Users can browse the repositories of available datasets, models, and metrics created by themselves or made public by other users. Each component is visualized as a card, providing an overview of the relevant statistics of the components. Clicking on any card provides details and allows downloading the component. The cards corresponding to the versions of the same component are clustered and stacked.

Execution and Registration of Benchmark Runs. A benchmark run is essentially a benchmark scenario (a combination of datasets, models, and metrics) instrumented and executed on the user's local resources. The UI helps the user in the process of creating benchmark scenarios by filtering out incompatible components and highlighting suitable ones as the user starts declaring aspects of the benchmark scenario. This suggestion feature works based on the inputs and the outputs of each component and their task type. Executing a benchmark run includes creating an instance of the benchmark scenario with current system and environment configurations on the local machine, running configurations for each combination of the core components, and uploading the execution results, including the corresponding resource usage information back to CausalBench repositories. Once declared public, these results are registered as permanent and associated with DOIs.

Causally-Informed Exploration and Analysis of Benchmark Runs. A user can visualize and explore a benchmark run, consisting of multiple benchmark scenarios, instrumented and executed on the same hardware by the same user. This involves slicing and dicing a benchmark run based on the datasets and models and comparing the different metric results and resource consumption. The entire benchmark run or its various subsets can be downloaded by the user for external analysis and visualization (Figure 4). In addition, the user can create *virtual* benchmark runs by declaring a new benchmark context and collecting all compatible benchmark scenarios that have been instrumented, executed, and recorded in CausalBench at different times, potentially by different users. This enables the user to explore the performance of the models on different hardware/software settings.

Since accuracy, timing, and resource usage of the models may be impacted by the properties of the data, underlying parameter/hyperparameter settings, as well as hardware/software configurations, CausalBench provides services to (a) disaggregate, de-bias, and explain the various factors impacting accuracy, time, and/or resource performance of the benchmark runs, as well as (b) propose new scenarios to execute to obtain a more robust understanding of the model performance.

Figure 4 provides the outline of the causal graph that forms the basis of these causally-informed exploration and analysis services. More specifically, CausalBench leverages a priori causal knowledge, described in the form of a causal graph, to boost the representational ability and achieve better explanations and recommendations. Specifically, given a causal graph (possibly enriched by data-driven causal impact analysis [3, 17, 20–22, 25]) describing the underlying causal relationships among the various factors impacting performance, CausalBench integrates this information into the learning process to ensure that explanations and recommendations are causally-robust. The causally-informed exploration and analysis services provided by CausalBench includes the following:

- *Causal impact and sensitivity analysis:* The benchmark data are analyzed through a causal effect discovery algorithm [18, 33] to quantify the impacts of various factors on the target accuracy, time, or resource usage in a given context.
- *Causal ranking and exploration:* Given a set of potentially conflicting decision parameters, the causal graph is also used to identify a non-dominating (pareto-optimal) subset of the runs that best highlight/explain the underlying trade-offs.
- *Causal prediction (with knowledge transfer):* Given a causal model and a benchmark of runs, CausalBench provides causally-informed performance predictions under new settings [5, 31]. CausalBench tackles data sparsity through causally-informed knowledge transfer across simulation contexts, by disaggregating shareable and non-shareable information relying on the underlying causal structure.
- *Causal recommendations:* CausalBench aggregates the above impact analysis, ranking, and prediction services into a causally-informed recommendation service, which recommends additional benchmark configurations to execute.

4 Demonstration Scenarios

The demonstration scenarios include (a) dataset, model, and metric registration, (b) exploration, (c) benchmark context declaration, (d) benchmark instrumentation and execution, and (e) benchmark result exploration. Three sample scenarios are outlined next:

- Scenario 1: User registers → logs in → retrieves the API key → downloads CB → implements their own dataset/model/metric → uploads the items, creating a submission → runs the context → posts the results → makes the results public and obtains DOI.
- Scenario 2: User logs in → browses through an array of datasets, metrics, models, and contexts by sorting and filtering → creates a benchmark context by selecting several sets, models, and metrics → downloads and instruments the benchmark → executes the benchmark → uploads results to CB and obtains DOI.
- Scenario 3: User logs in → creates a virtual benchmark context by selecting several datasets and models → CB aggregates and presents matching benchmark runs → user slices-and-dices the runs and obtains causal explanations and causally-informed recommendations for additional benchmark contexts to execute.

CausalBench is accessible at [11] and a 3-minute video recording showcasing the major features of CausalBench is available at [10].

5 Conclusions

In this paper, we introduced CausalBench, a platform designed to support the benchmarking of causal learning models by facilitating scientific collaboration on novel algorithms, datasets, and metrics and promoting reproducibility in causal learning research.

References

[1] Fahim Tasneema Azad, K. Selçuk Candan, Ahmet Kapkiç, Mao-Lin Li, Huan Liu, Pratanu Mandal, Paras Sheth, Bilgehan Arslan, Gerardo Chowell-Puente, John Sabo, Rebecca Muenich, Javier Redondo Anton, and Maria Luisa Sapino. 2024. A Vision for Spatio-Causal Situation Awareness, Forecasting, and Planning. *ACM Transactions on Spatial Algorithms and Systems (TSAS)* (2024). Accepted for publication.

[2] Andrew P. Bradley. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* 30, 7 (1997), 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2

[3] Debo Cheng, Jiuyong Li, Lin Liu, Jixue Liu, and Thuc Duy Le. 2024. Data-Driven Causal Effect Estimation Based on Graphical Causal Modelling: A Survey. *ACM Comput. Surv.* 56, 5, Article 127 (jan 2024), 37 pages. https://doi.org/10.1145/3636423

[4] Lu Cheng, Ruocheng Guo, Raha Moraffah, Paras Sheth, K. Selçuk Candan, and Huan Liu. 2022. Evaluation Methods and Measures for Causal Learning Algorithms. *IEEE Transactions on Artificial Intelligence* 3, 6 (2022), 924–943. https://doi.org/10.1109/TAI.2022.3150264

[5] Yoonhyuk Choi, Jiho Choi, Taewook Ko, Hyungho Byun, and Chong-Kwon Kim. 2022. Review-Based Domain Disentanglement without Duplicate Users or Contexts for Cross-Domain Recommendation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM '22)*. Association for Computing Machinery, New York, NY, USA, 293–303. https://doi.org/10.1145/3511808.3557434

[6] Iain M Cockburn, Rebecca Henderson, and Scott Stern. 2018. *The impact of artificial intelligence on innovation*. Vol. 24449. National bureau of economic research Cambridge, MA, USA.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[8] Li Deng. 2012. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine* 29, 6 (2012), 141–142.

[9] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[10] Kapkic et. al. 2024. CausalBench Demo Video. https://drive.google.com/drive/folders/1ckfKqci1sj8u7G6QAOb02YBSA3zqShxO?usp=sharing.

[11] Kapkic et. al. 2024. CausalBench Website. www.causalbench.org.

[12] Arunim Garg and Vijay Mago. 2021. Role of machine learning in medical research: A survey. *Comput. Sci. Rev.* 40, C (may 2021), 17 pages. https://doi.org/10.1016/j.cosrev.2021.100370

[13] S Hettich and S. D Bay. 1999. The UCI KDD Archive. http://kdd.ics.uci.edu

[14] M. Hino, E. Benami, and N. Brooks. 2018. Machine learning for environmental monitoring. *Nature Sustainability* 1, 10 (01 Oct 2018), 583–588. https://doi.org/10.1038/s41893-018-0142-9

[15] Michael Höfler. 2005. The Bradford Hill considerations on causality: a counterfactual perspective. *Emerging Themes in Epidemiology* 2, 1 (03 Nov 2005), 11. https://doi.org/10.1186/1742-7622-2-11

[16] Alex Krizhevsky. 2012. Learning Multiple Layers of Features from Tiny Images. *University of Toronto* (05 2012).

[17] Mao-Lin Li, K. Selçuk Candan, and Maria Luisa Sapino. 2023. CTT: Causally Informed Tensor Train Decomposition. In *IEEE Big Data*. 1180–1187.

[18] Ana Rita Nogueira, Andrea Pugnana, Salvatore Ruggieri, Dino Pedreschi, and João Gama. 2022. Methods and tools for causal discovery and causal inference. *WIREs Data Mining and Knowledge Discovery* 12, 2 (2022), e1449. https://doi.org/10.1002/widm.1449 arXiv:https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1449

[19] Mattia Prosperi, Yi Guo, Matt Sperrin, James S. Koopman, Jae S. Min, Xing He, Shannan Rich, Mo Wang, Iain E. Buchan, and Jiang Bian. 2020. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence* 2, 7 (01 Jul 2020), 369–375. https://doi.org/10.1038/s42256-020-0197-y

[20] Amit Sharma, Jake M. Hofman, and Duncan J. Watts. 2015. Estimating the Causal Impact of Recommendation Systems from Observational Data. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation* (Portland, Oregon, USA) *(EC '15)*. Association for Computing Machinery, New York, NY, USA, 453–470. https://doi.org/10.1145/2764468.2764488

[21] Paras Sheth, Ruocheng Guo, Lu Cheng, Huan Liu, and K. Selçuk Candan. 2023. Causal Disentanglement for Implicit Recommendations with Network Information. *ACM Trans. Knowl. Discov. Data* 17, 7 (2023), 94:1–94:18.

[22] Paras Sheth, Ruocheng Guo, Kaize Ding, Lu Cheng, K. Selçuk Candan, and Huan Liu. 2022. Causal Disentanglement with Network Information for Debiased Recommendations. In *SISAP*. 265–273.

[23] Paras Sheth, Ujun Jeong, Ruocheng Guo, Huan Liu, and K. Selçuk Candan. 2021. CauseBox: A Causal Inference Toolbox for Benchmarking Treatment Effect Estimators with Machine Learning Methods. In *CIKM*. 4789–4793.

[24] Paras Sheth, Ting Liu, Durmus Doner, Qi Deng, Yuhang Wei, Rebecca Muenich, John Sabo, K. Selçuk Candan, and Huan Liu. 2022. Causal Discovery for Feature Selection in Physical Process-Based Hydrological Systems. In *2022 IEEE International Conference on Big Data (Big Data)*. 5568–5577. https://doi.org/10.1109/BigData55660.2022.10020794

[25] Paras Sheth, Raha Moraffah, Tharindu S. Kumarage, Aman Chadha, and Huan Liu. 2024. Causality Guided Disentanglement for Cross-Platform Hate Speech Detection. In *WSDM*. 626–635.

[26] Paras Sheth, Ahmadreza Mosallanezhad, Kaize Ding, Reepal Shah, John Sabo, Huan Liu, and K. Selçuk Candan. 2023. STREAMS: Towards Spatio-Temporal Causal Discovery with Reinforcement Learning for Streamflow Rate Prediction. In *CIKM*. 4815–4821.

[27] Paras Sheth, Reepal Shah, John Sabo, K. Selçuk Candan, and Huan Liu. 2022. STCD: A Spatio-Temporal Causal Discovery Framework for Hydrological Systems. In *IEEE Big Data*. 5578–5583.

[28] Herbert A. Simon. 1977. *Spurious Correlation: A Causal Interpretation*. Springer Netherlands, Dordrecht, 93–106. https://doi.org/10.1007/978-94-010-9521-1_7

[29] M. Stone. 2018. Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society: Series B (Methodological)* 36, 2 (12 2018), 111–133. https://doi.org/10.1111/j.2517-6161.1974.tb00994.x arXiv:https://academic.oup.com/jrsssb/article-pdf/36/2/111/49096683/jrsssb_36_2_111.pdf

[30] Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. 2014. OpenML: networked science in machine learning. *SIGKDD Explor. Newsl.* 15, 2 (jun 2014), 49–60. https://doi.org/10.1145/2641190.2641198

[31] Song Wei, Ronald Moore, Hanyu Zhang, Yao Xie, and Rishikesan Kamaleswaran. 2023. Transfer Causal Learning: Causal Effect Estimation with Knowledge Transfer. In *ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH)*. https://openreview.net/forum?id=V3GGYh8CKq

[32] Timothy L. Wiemken and Robert R. Kelley. 2020. Machine Learning in Epidemiology and Health Outcomes Research. *Annual Review of Public Health* 41, Volume 41, 2020 (2020), 21–36. https://doi.org/10.1146/annurev-publhealth-040119-094437

[33] Alessio Zanga, Elif Ozkirimli, and Fabio Stella. 2022. A Survey on Causal Discovery: Theory and Practice. *International Journal of Approximate Reasoning* 151 (2022), 101–129. https://doi.org/10.1016/j.ijar.2022.09.004