dfki ai

CaMoRe - Causal Models & Representations

Neuro-Mechanistic Modeling

Are our DAGs correct?

Recent Developments in Causal Discovery Evaluation

Jonas Wahl
DFKI
Causality-XAI Winter School, Paris
23.10.2025

Overview

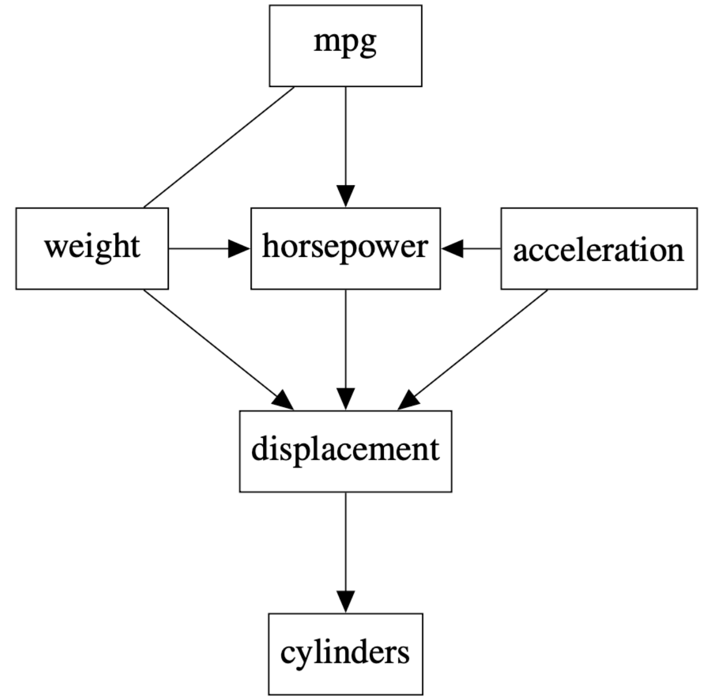
Recap

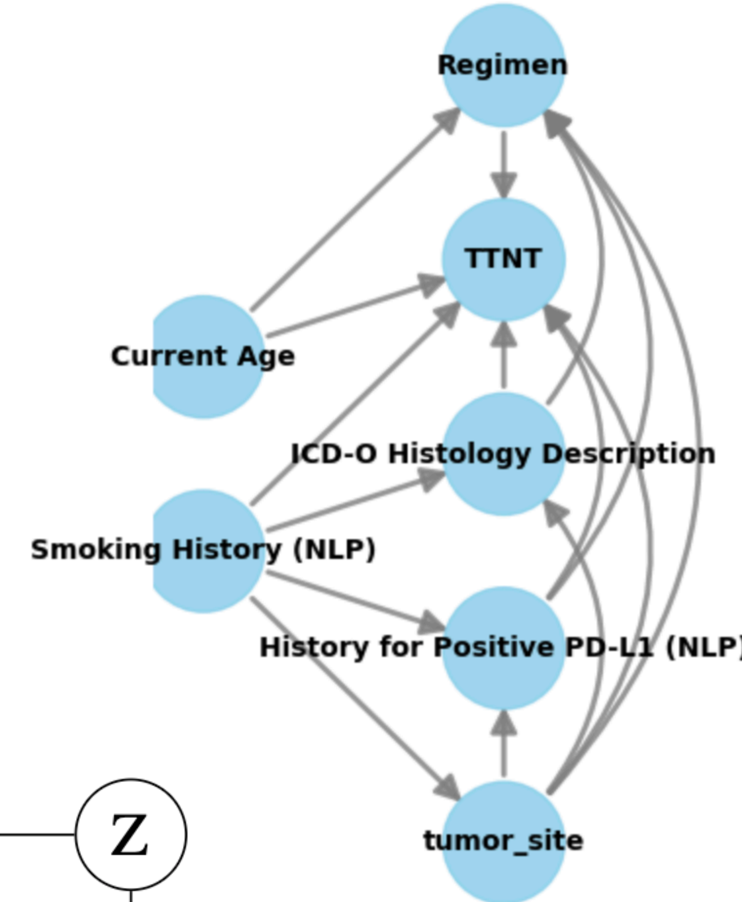
Causal Discovery Evaluation for Method Developers

Causal Discovery Evaluation for Practitioners

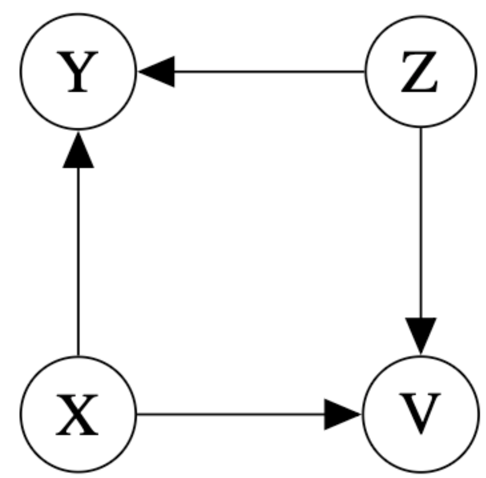
Conclusion

Recap





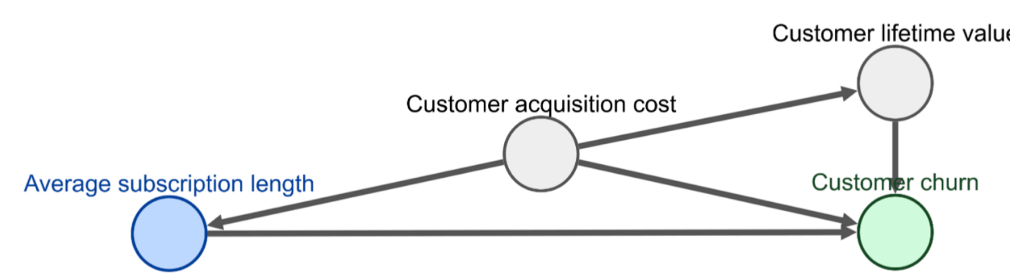
- Causal models are designed to model…



  - Observations (“non-causal”)
  - Interventions:

    how does Y respond to an external change of X?

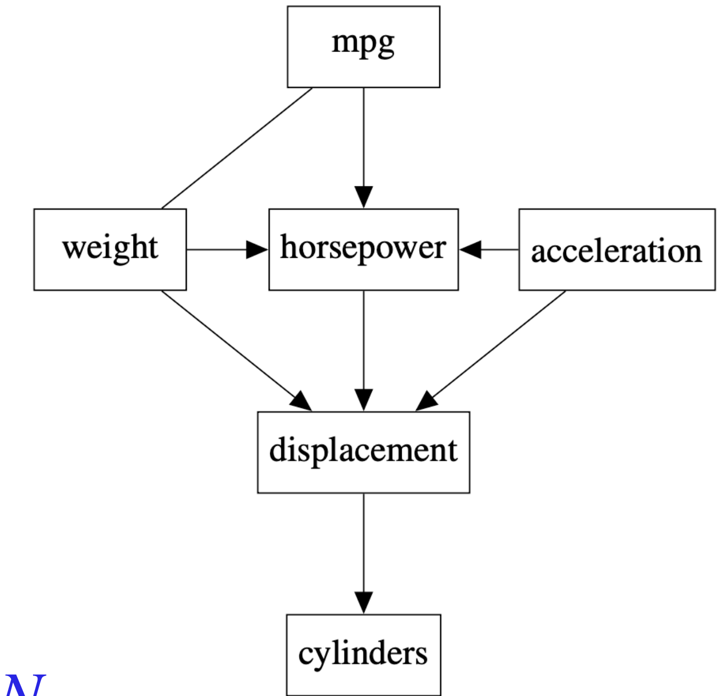
  - Counterfactuals

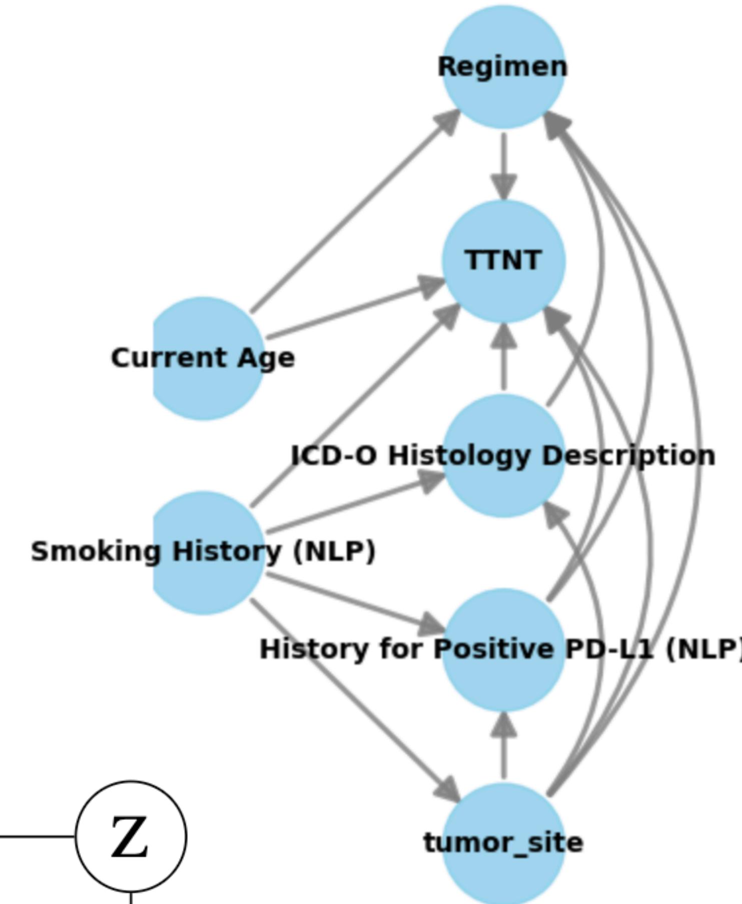


CaMoRe -
Causal Models &
Representations

dfki ai
Neuro-
Mechanistic
Modeling

Recap

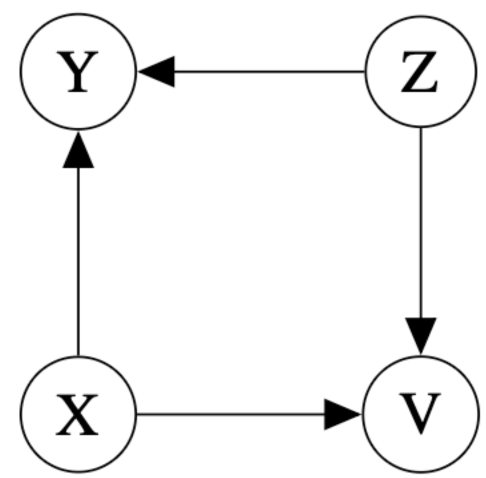




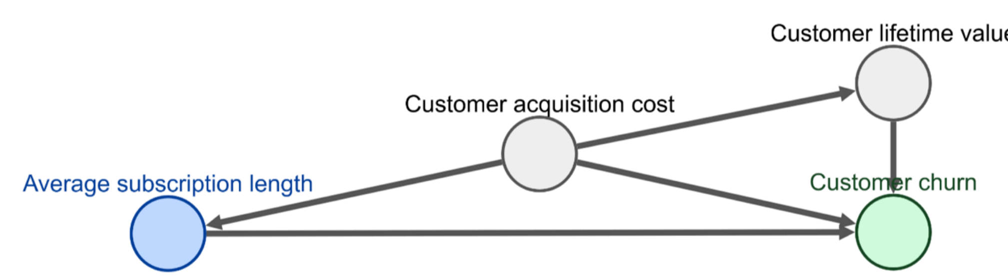
- Structural causal models (SCMs)

$$X_i := f_i(\mathrm{pa}(X_i), \eta_i) \qquad i = 1, \ldots, N$$

consist of three fundamental ingredients:



- A causal graph describing the causal parents $\mathrm{pa}(X_i)$
- Functions that describe the causal mechanisms
- Noise distributions



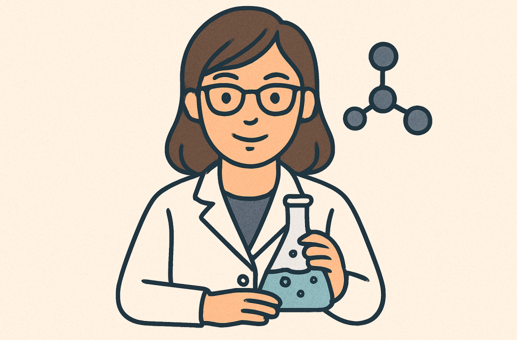
CaMoRe - Causal Models & Representations

dfki ai

Neuro-Mechanistic Modeling

Recap

- Two ways to get a causal graph:



Defined by expert

causal discovery algorithm

or combination of both

Causal Discovery Algorithms

- Long list of algorithms that learn causal graphs from data.
    - Most of them focus on observational data
    - For theoretical guarantees, this requires strong assumptions!
    - Fundamentally, it is assumed that there the is a 'ground truth' structural causal model

$$X_i := f_i(\mathrm{pa}(X_i), \eta_i) \qquad i = 1,\ldots,N$$

with causal graph $\mathcal{G}$ and observational distribution $\mathbb{P} = \mathbb{P}(X_1, \ldots, X_N)$ that accurately describes the data-generating process.

CaMoRe - Causal Models & Representations

dfki ai Neuro-Mechanistic Modeling

Causal Discovery Algorithms

- Input: $\mathbb{P}$, target: $\mathcal{G}$
- Structural assumptions on $\mathcal{G}$:
    - Causal sufficiency = no hidden confounding
    - Acyclicity

      –> directed acyclic graphs (DAGs)
    - Time series vs. 'equilibrium' model

Causal Discovery Algorithms

- Input: $\mathbb{P}$, target: $\mathscr{G}$

- Distributional assumptions on $\mathbb{P}$:

    - e.g. Gaussian vs. Non-Gaussian

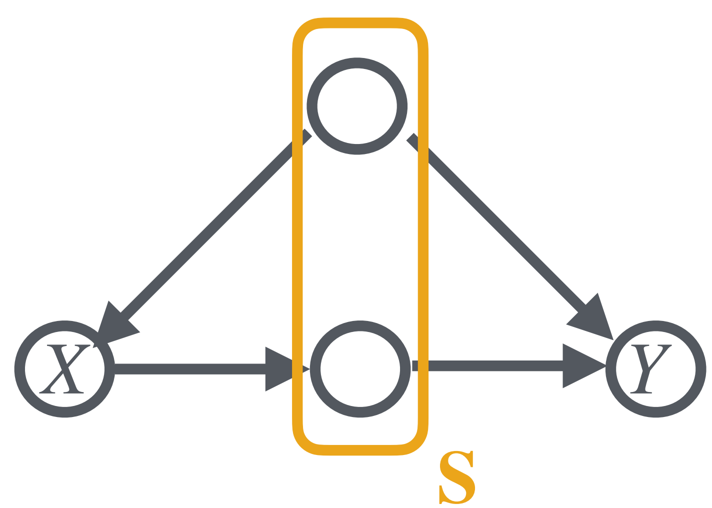
- Assumptions on mechanisms, e.g. linearity;

Causal Discovery Algorithms

- Input: $\mathbb{P}$, target: $\mathcal{G}$
- Sampling assumptions:
    - Infinite sample (for theoretical guarantees)
    - i.i.d.-ness vs. auto-correlated in finite samples
- Algorithm-specific under-the-hood design choices

Causal Discovery Algorithms

- Input: $\mathbb{P}$, target: $\mathcal{G}$
- Assumptions connecting $\mathbb{P}$ and $\mathcal{G}$:
    - Markov property: Nodes are independent of graphical non-descendants given their graphical parents
    - Can also expressed by the graphical operation of d-separation



Causal Discovery Algorithms

- Input: $\mathbb{P}$, target: $\mathscr{G}$
- Assumptions connecting $\mathbb{P}$ and $\mathscr{G}$:
  - Markov property:

    d-separation on $\mathscr{G}$ $\Rightarrow$ conditional independence in $\mathbb{P}$

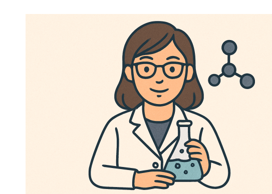
  - Faithfulness:

    d-separation on $\mathscr{G}$ $\Leftarrow$ conditional independence in $\mathbb{P}$

Causal Discovery Evaluation

- How does a **method developer** evaluate whether the a causal discovery method is working well **in general**?

    - Benchmarks? Practices? Metrics?



- How should a **practitioner** evaluate the output of a CD method on their **specific dataset**?

A typical causal discovery paper

My new Causal Discovery Method

C. D. Covery

October 2025

Abstract

In this paper, I introduce my causal discovery method FIND-CAUSAL-GRAPH.

1 Introduction

...

2 Theoretical Results

Theorem 1. *Under assumptions (1-5), FIND-CAUSAL-GRAPH identifies the ground truth causal graph* $\mathcal{G}$ *up to the following notion of equivalence in the infinite sample limit.*

3 Empirical Evaluation

3.1 Simulated Data

We evaluate FIND-CAUSAL-GRAPH on simulated data in the following setup...

3.2 Real-world example

FIND-CAUSAL-GRAPH finds the following causal graph in our real-world example which seems plausible to us.

4 Conclusion

...

CaMoRe - Causal Models & Representations

dfki ai

Neuro-Mechanistic Modeling

Method Evaluation

- Usual practice:
  - Simulate data from ground truth models that satisfy your method's assumptions
  - compare to similar methods / state-of-the-art

CaMoRe - Causal Models & Representations

dfki ai

Neuro-Mechanistic Modeling

Method Evaluation

- Criticisms:
  - assumptions are never fully satisfied in real data

    ➡ evaluate robustness to at least some degree of assumption violations

    ➡ Montagna et al. (2023)

Assumption violations in causal discovery and the robustness of score matching

Francesco Montagna
MaLGa, Università di Genova

Atalanti A. Mastakouri
AWS

Elias Eulig
German Cancer Research Center (DKFZ)

Nicoletta Noceti
MaLGa, Università di Genova

Lorenzo Rosasco
MaLGa, Università di Genova
MIT, CBMM
Istituto Italiano di Tecnologia

Dominik Janzing
AWS

Bryon Aragam
University of Chicago

Francesco Locatello
Institute of Science and Technology Austria (ISTA)

Method Evaluation

- Criticism:
  - Simulated data may have exploitable but unrealistic properties

    ➡Reisach et al. (2023, 2024) (var- and R2-sortability)

    Beware of the Simulated DAG!
    Causal Discovery Benchmarks May Be Easy To Game

    Alexander G. Reisach[1,2] **Christof Seiler**[2,3] **Sebastian Weichwald**[1]

    [1]Department of Mathematical Sciences, University of Copenhagen, Denmark
    [2]Department of Data Science and Knowledge Engineering, Maastricht University, The Netherlands
    [3]Mathematics Centre Maastricht, Maastricht University, The Netherlands

    ➡See also Lohse and Wahl (2025) for an investigation of sortability in the context of time series data

    Sortability of Time Series Data

    Christopher Lohse *lohsec@tcd.ie*
    School of Computer Science and Statistics
    University of Dublin Trinity College
    IBM Research Europe, Dublin

    Jonas Wahl *jonas.wahl@dfki.de*
    Deutsches Forschungszentrum für künstliche Intelligenz (DFKI)

    Reviewed on OpenReview: https://openreview.net/forum?id=OGvmCpcHdV

    ➡See Ormaniec et al (2024), Herman et al. (2025) for suggestions on how to avoid sortability

Method Evaluation

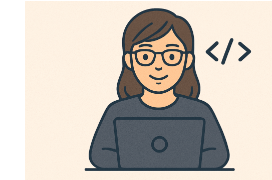
- Criticism:
  - Need for causal comparison metrics

    ➡Peters and Bühlmann (2014), Henckel et al. (2024), Wahl and Runge (2025)

CaMoRe -
Causal Models &
Representations

dfki ai
Neuro-
Mechanistic
Modeling

Method Evaluation



- Criticism:
  - Need for negative controls

    ➡Helby Petersen (2025)

Are You Doing Better Than Random Guessing? A Call for Using Negative Controls When Evaluating Causal Discovery Algorithms

Anne Helby Petersen[1]

[1]Section of Biostatistics, Department of Public Health, University of Copenhagen, Copenhagen, Denmark

CaMoRe -
Causal Models &
Representations

dfki ai

Neuro-
Mechanistic
Modeling

Method Evaluation

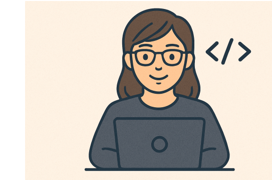
- Criticism:
  - Lack of good real-world benchmarking data sets
- Hard to find non-synthetic examples with trustworthy descriptions in terms of causal graphs

  ➡ Gamella et al. (2025)

  ➡ See also work on micro-service networks, e.g. Lohse et al. (2025)

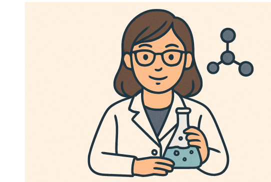
- Need more domain-specific benchmarks: data ≠ data

Method Evaluation (Summary)



- Lots of smart suggestions on how to improve the evaluation of causal discovery algorithms
- But: limited adaptation of these tools
- Need for community effort:
    - Unified software package
    - Community guidelines
    - competitions

Output Evaluation for Users



- User's don't have a 'ground truth', that's want they want to find!

➡ Cannot test whether method produces 'correct' output, only whether it is

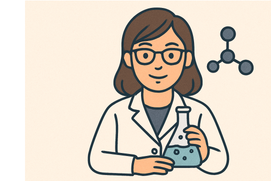
  - consistent with external knowledge
  - internally consistent
  - sensitive to changes

CaMoRe - Causal Models & Representations

dfki ai

Neuro-Mechanistic Modeling

Internal consistency



- Faller et al. (2024): Self-compatibility

  ➡ Run causal discovery on all variables and subsets of variables and check whether results are consistent.

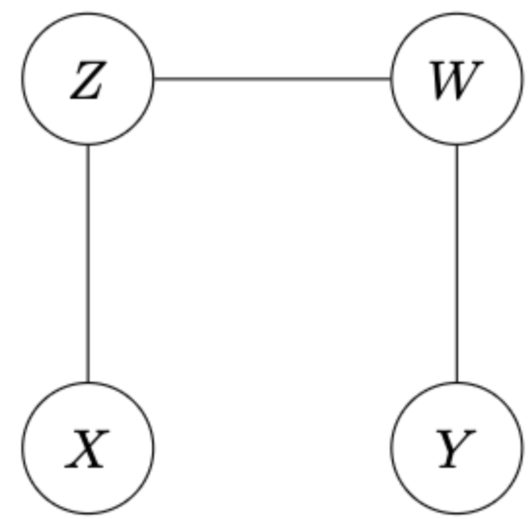
- Faltenbacher*, Wahl* et al. (2025):

  ➡ For causal discovery based on conditional independence testing

  ➡ Check whether the CD output is consistent with the tests of conditional independence it ran

How can outputs be inconsistent?

- Example:
    - X and Y test independent
    - But in the output graph, a path between them remains open implying dependence



$\mathcal{G}_{out}$

CaMoRe - Causal Models & Representations

dfki ai

Neuro-Mechanistic Modeling

Internal Inconsistency: a sanity check

- Inconsistencies are both a bug and a feature!

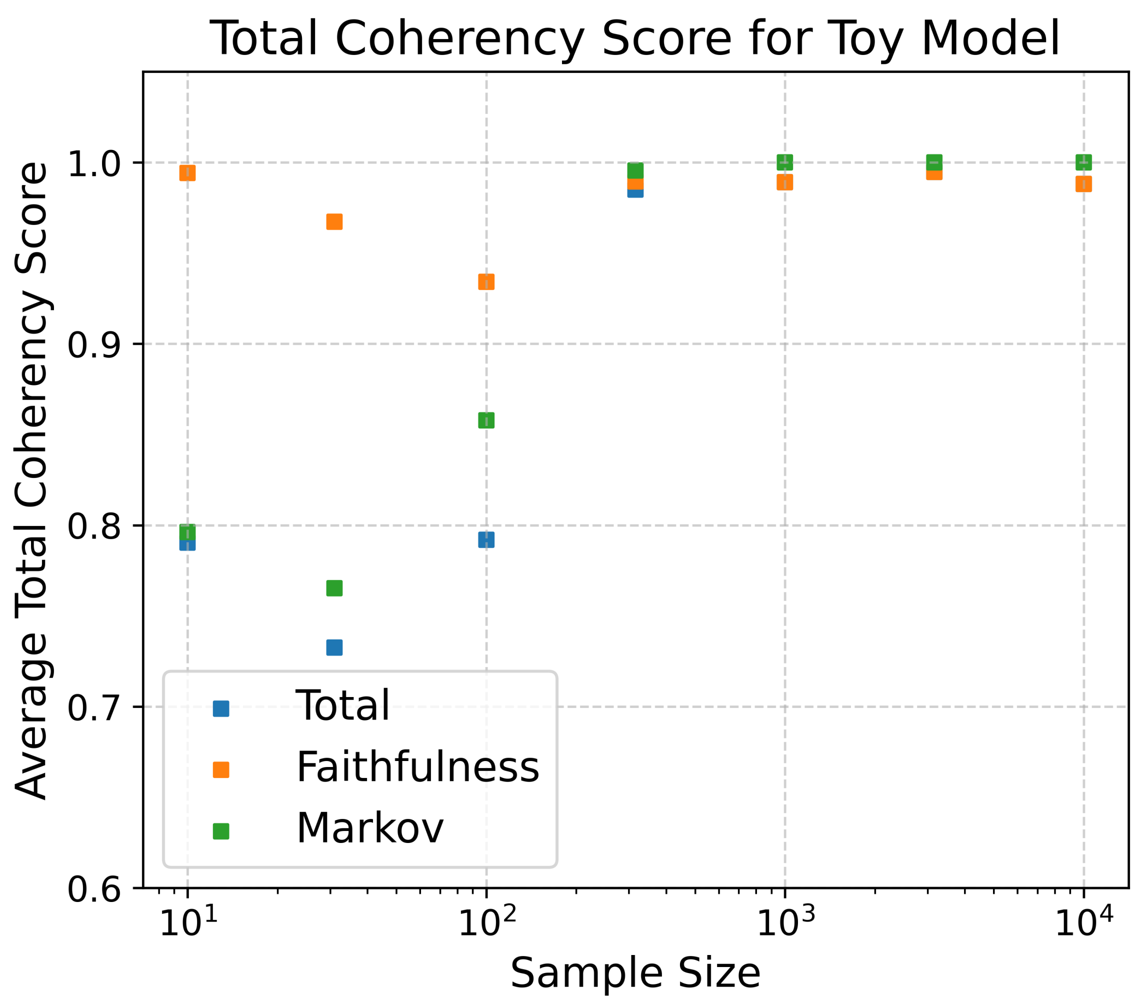
  ➡ We want the most consistent graph from our model class (e.g. DAGs)

  ➡ If the most consistent graph still has many inconsistencies this signals assumption violations!

- These scores provide a tool to
  - judge the influence of sample sizes
  - test sensitivity to hyperparameters

Internal Inconsistency: a sanity check

- Scores for PC algorithm on 'clean' linear SCM on 5 variables across different sample sizes.



CaMoRe - Causal Models & Representations

Neuro-Mechanistic Modeling

Internal Inconsistency: a sanity check

- Correlation with SHD to ground truth.

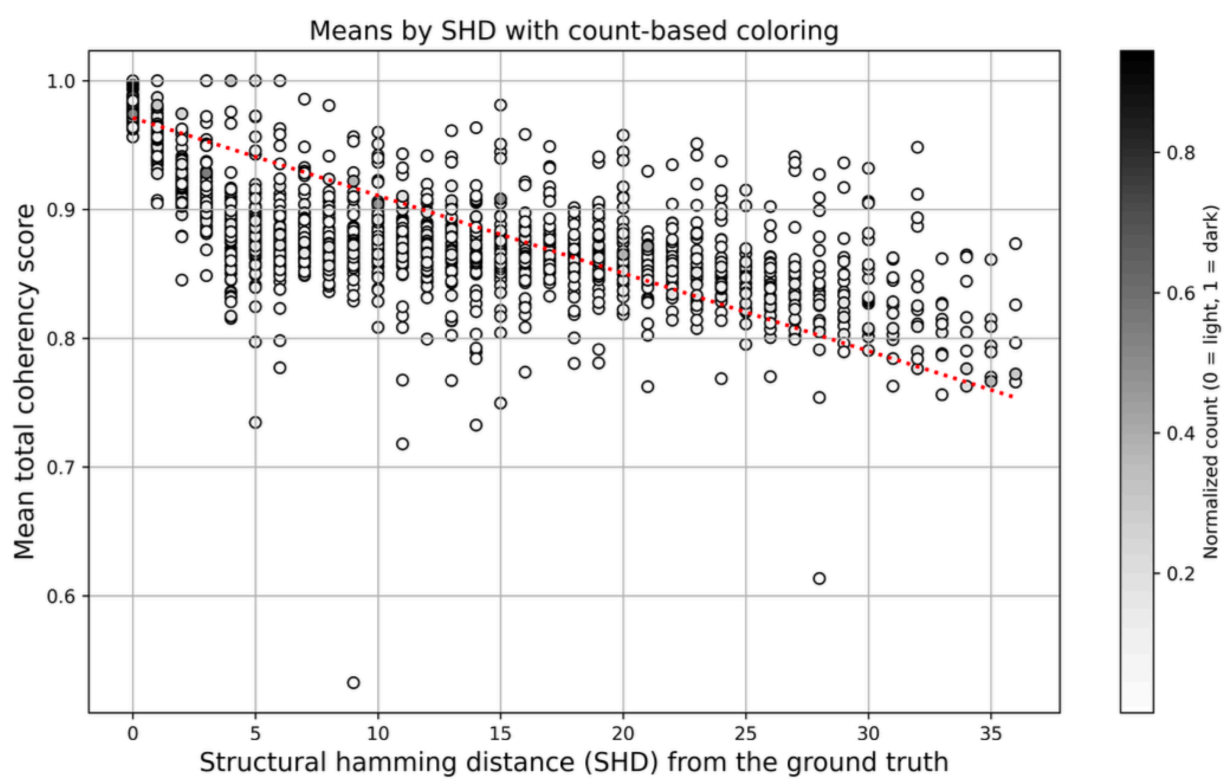


Figure 6: We generated 127000 data sets from 127 ground truth DAGs with 4 to 10 nodes from sparse to fully connected. In this plot, we show the 127 means each over 1000 DAGs with the same configuration. The red dotted line shows the weighted regression of the mean scores over 1000 random DAGs each on the SHD weighted by their counts.

CaMoRe - Causal Models & Representations

dfki ai

Neuro-Mechanistic Modeling

Concluding thoughts

- Domain-specific evaluation:

  A causal discovery method that is good for all data is too much to ask!

- Task-specific evaluation:

  Instead of ‘is our method able to find the ground truth?’, we should focus more on ‘is our method useful for task X?’

- More work on sensitivity analysis and uncertainty quantification needed: e.g. inject weak synthetic noise

CaMoRe - Causal Models & Representations

dfki ai Neuro-Mechanistic Modeling

dfki ai

CaMoRe -
Causal Models &
Representations

Neuro-
Mechanistic
Modeling

Are our DAGs useful?

Recent Developments in Causal Model Evaluation

Jonas Wahl
DFKI
Causality-XAI Winter School, Paris
23.10.2025

Thanks for listening!

References

- Eulig, Elias, et al. "Toward falsifying causal graphs using a permutation-based test." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 39. No. 25. 2025.
- Ramsey, Joseph D., Bryan Andrews, and Peter Spirtes. "Choosing dag models using markov and minimal edge count in the absence of ground truth." *arXiv preprint arXiv:2409.20187* (2024).
- Shipley, Bill. "A new inferential test for path models based on directed acyclic graphs." *Structural Equation Modeling* 7.2 (2000): 206-218.
- Montagna, Francesco, et al. "Assumption violations in causal discovery and the robustness of score matching." *Advances in Neural Information Processing Systems* 36 (2023): 47339-47378.
- Reisach, Alexander, 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 (2021): 27772-27784.
- Lohse, Christopher and Wahl, Jonas. "Sortability of Time Series Data". *Transactions on Machine Learning Research (TMLR)*, 2835-8856, 2025.
- Herman, Rebecca, et al. "Unitless Unrestricted Markov-Consistent SCM Generation: Better Benchmark Datasets for Causal Discovery". *Proceedings of the Fourth Conference on Causal Learning and Reasoning*, PMLR 275:1506-1531, 2025.
- Ormaniec, Weronika, et al. "Standardizing structural causal models." *arXiv preprint arXiv:2406.11601* (2024).
- Wahl, Jonas, and Jakob Runge. "Separation-Based Distance Measures for Causal Graphs." *International Conference on Artificial Intelligence and Statistics*. PMLR, 2025.
- Henckel, Leonard, Theo Würtzen, and Sebastian Weichwald. "Adjustment identification distance: A gadjid for causal structure learning." *arXiv preprint arXiv:2402.08616* (2024).
- Petersen, Anne Helby. "Are you doing better than random guessing? A call for using negative controls when evaluating causal discovery algorithms." *arXiv preprint arXiv:2412.10039* (2024).
- Gamella, Juan L., Jonas Peters, and Peter Bühlmann. "Causal chambers as a real-world physical testbed for AI methodology." *Nature Machine Intelligence* 7.1 (2025): 107-118.
- Faller, Philipp M., et al. "Self-compatibility: Evaluating causal discovery without ground truth." *International Conference on Artificial Intelligence and Statistics*. PMLR, 2024.
- Faltenbacher*, Sofia, Wahl*, Jonas, et al. "Internal Incoherency Scores for Constraint-based Causal Discovery Algorithms." *arXiv preprint arXiv:2502.14719* (2025).

CaMoRe -
Causal Models &
Representations

dfki ai
Neuro-
Mechanistic
Modeling