PRIM: Meta-Learned Bayesian Root Cause Analysis

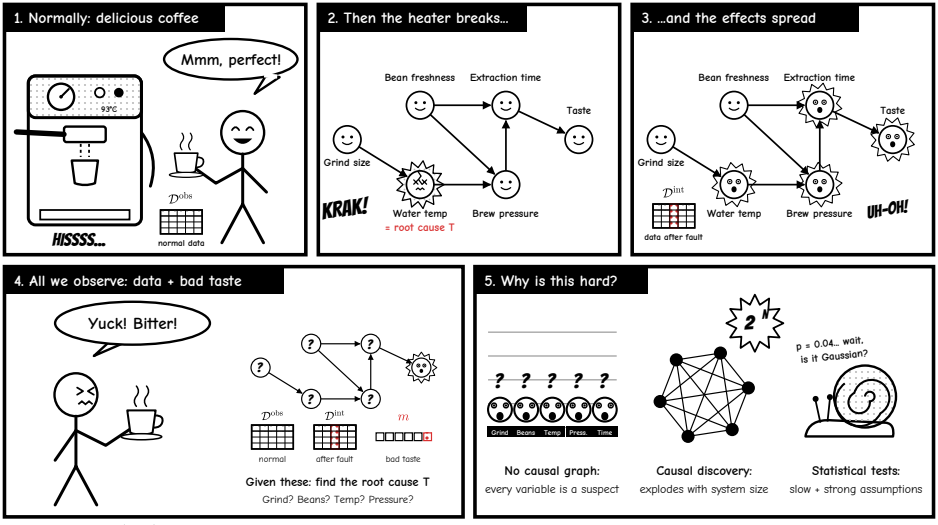
Christopher Lohse[1,2] Anish Dhir[3] Amadou Ba[2]
Bradley Eck[2] Marco Ruffini[1] Jonas Wahl[4,5]

[1]University of Dublin, Trinity College, Dublin, Ireland, [2]IBM, [3]Gatsby Computational Neuroscience Unit, University College London, [4]Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Saarbrücken, Germany, [5]Department of Philosophy, University of Bergen, Norway

Accepted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026)

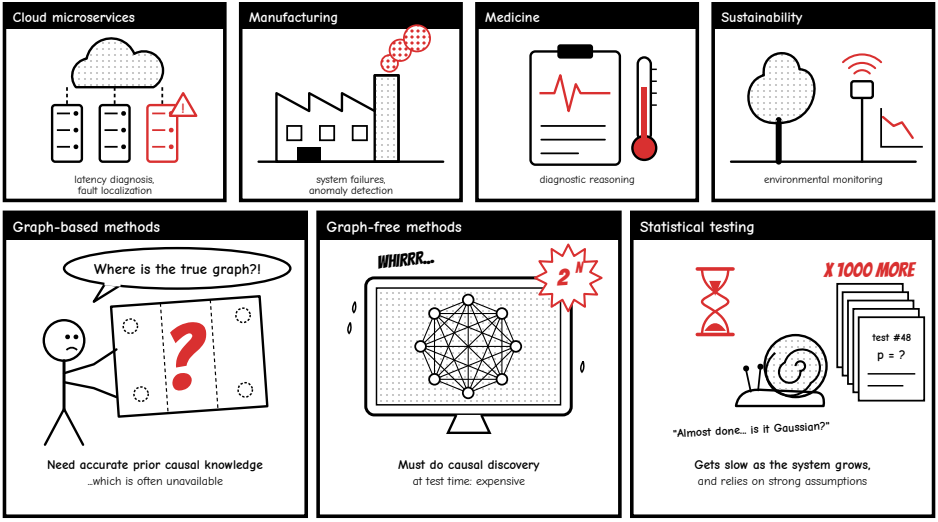
C.L. was partilaly funded by the EcoTwin project, which has received funding from the European Commission's Horizon Europe programme under grant agreement No 101157369. JW received support from the German Federal Ministry of Research, Technology and Space (BMFTR) as part of the project MAC-MERLin (Grant Agreement No. 01IW24007

The Root Cause Analysis Problem



Modelling root causes as (soft)-interventions following e.g. Budhathoki, Janzing, et al. 2021; Budhathoki, Minorics, et al. 2022

Applications & Existing Challenges



Bayesian RCA Formulation

$$\underbrace{p(T \mid \mathcal{D}^{\text{obs}}, \mathcal{D}^{\text{int}}, m)}_{\text{root-cause posterior}} = \overbrace{\sum_{\mathcal{G}} \int \underbrace{p(T \mid \mathbf{f}, \mathcal{G}, m)}_{\text{which root cause}} \underbrace{p(\mathbf{f} \mid \mathcal{G}, \mathcal{D})}_{\text{how plausible are these mechanisms?}} \underbrace{p(\mathcal{G} \mid \mathcal{D})}_{\text{how plausible}} d\mathbf{f}}^{\text{average over every possible graph and mechanism}}$$

root-cause posterior — what we want

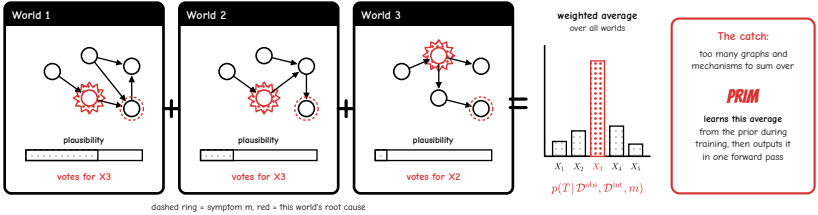
which root cause — explains m in this world?

how plausible are these mechanisms?

how plausible — is this graph?

average over every possible graph and mechanism — (all the worlds that could have produced the data)

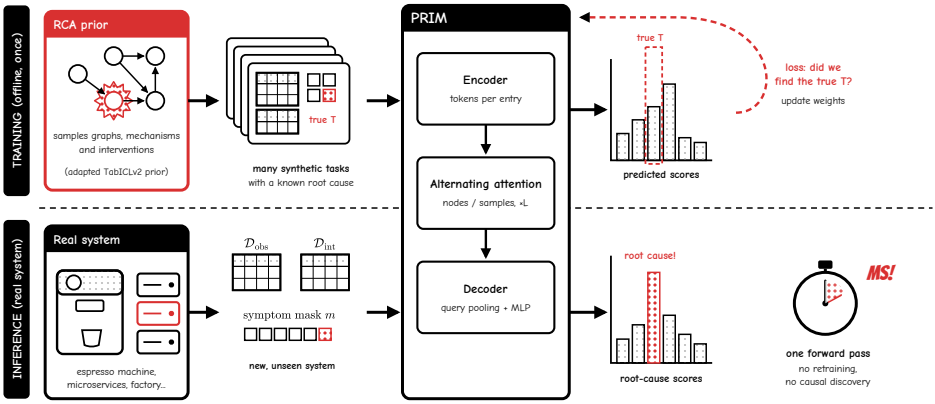
In pictures: many possible worlds



dashed ring = symptom m, red = this world's root cause

This posterior is identifiable; under standard assumptions on the mass concentration of the prior, posterior updates concentrate asymptotically on the true root cause set as the sample size goes to infinity (Ghosal and Vaart 2017).

PRIM PFN Training



Following standard PFN training procedure Hollmann, Müller, Eggensperger, et al. 2023; Hollmann, Müller, Purucker, et al. 2025

Results

Convergence: 3-node graphs



Mediator Confounder

0.9 – 1.0

Recall@1 with enough data
(chance level: 0.33)

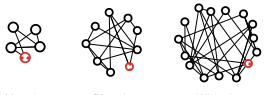
More interventional samples:

$n_{\text{int}} = 1$ → $n_{\text{int}} = 50$

0.7 → ~1.0

PRIM converges to the true root cause on both topologies

Scalability: 20-100 nodes (in-distribution)



20 nodes 50 nodes 100 nodes

PRIM stays stable as systems grow:

NN functions	0.75 – 0.84
GP functions	0.63 – 0.76

best baselines: 0.57 or less, and dropping



7 ms

inference at every size

Real-world (out-of-distribution)



PetShop (top-3 recall)

PRIM 0.79 **BEST**
Corr 0.78

CausRCA full (MAP@3)

PRIM 0.27 **BEST**
TimeRank 0.24

CausRCA sub (MAP@3)

PRIM 0.61
Baro 0.75

No causal graph given.
Best on 2 of 3 settings, competitive on the third

Thank you!

References i

- Budhathoki, Kailash, Dominik Janzing, et al. (2021). **"Why did the distribution change?"** In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 1666–1674.
- Budhathoki, Kailash, Lenon Minorics, et al. (17–23 Jul 2022). **"Causal structure-based root cause analysis of outliers".** In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri et al. Vol. 162. Proceedings of Machine Learning Research. PMLR, pp. 2357–2369. URL: https://proceedings.mlr.press/v162/budhathoki22a.html.

References ii

- Ghosal, Subhashis and Aad van der Vaart (2017). **“Consistency: General Theory”.** In: *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. Chap. 6, pp. 123–164.
- Hardt, Michaela et al. (Jan. 2024). **“The PetShop Dataset — Finding Causes of Performance Issues across Microservices”.** In: *Proceedings of the Third Conference on Causal Learning and Reasoning*. Ed. by Francesco Locatello and Vanessa Didelez. Vol. 236. Proceedings of Machine Learning Research. PMLR, pp. 957–978. URL: https://proceedings.mlr.press/v236/hardt24a.html.

References iii

- Hollmann, Noah, Samuel Müller, Katharina Eggensperger, et al. (2023). **“TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second”.** In: *The Eleventh International Conference on Learning Representations*.
- Hollmann, Noah, Samuel Müller, Lennart Purucker, et al. (Jan. 2025). **“Accurate predictions on small data with a tabular foundation model”.** en. In: *Nature* 637.8045, pp. 319–326. ISSN: 1476-4687. DOI: 10.1038/s41586-024-08328-6. URL: https://www.nature.com/articles/s41586-024-08328-6 (visited on 01/16/2026).

References iv

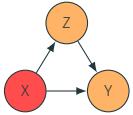
- Mehling, Carl Willy et al. (Jan. 2026). **“Enabling Joint Benchmarking of Automated Root Cause Analysis and Causal Discovery in Manufacturing Using the *causRCA* Dataset”.** In: *Procedia CIRP*. 13th CIRP Global Web Conference 139, pp. 114–120. ISSN: 2212-8271. DOI: 10.1016/j.procir.2025.09.010. URL: https://www.sciencedirect.com/science/article/pii/S221282712500976X (visited on 03/13/2026).
- Pham, Luan et al. (2025). **“Rcaeval: A benchmark for root cause analysis of microservice systems with telemetry data”.** In: *Companion Proceedings of the ACM on Web Conference 2025*, pp. 777–780.

Appendix

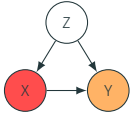
Synthetic Validation: 3-Node Structures

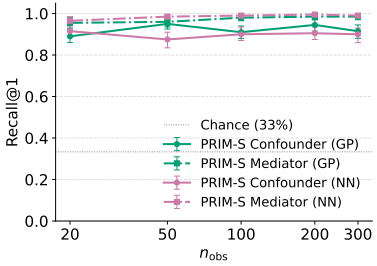
Test Topologies:

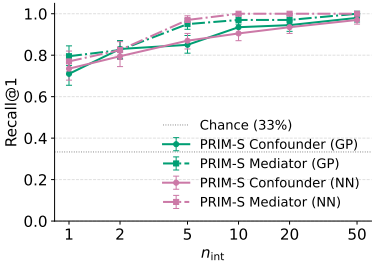
Mediator



Confounder





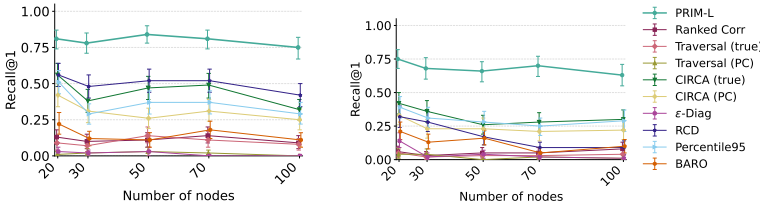


Left: Varying n_{obs}, keeping $n_{int} = 10$ Right: Varying n_{int}, keeping $n_{obs} = 100$

Legend: Root cause (X) Descendants Unaffected.

Result: PRIM competitive with graph-not-given baselines; accurately identifies root cause with sufficient data across both topologies.

Synthetic Validation: Scalability



Recall@1 vs. number of nodes (4–100 nodes). Left: Neural Network Functions, Right: Gaussian Process Functions

Key findings:

1. **PRIM outperforms all baselines** (Traversal, CIRCA, RCD, ε-diagnosis) as systems grow
2. Constant 17ms inference (independent of size)
3. RCD loses ground with scale (causal discovery is expensive) (Pham et al. 2025)

Real World Benchmarks

PetShop (microservices)

- 44 services, 48 fault injections
- 4 traffic scenarios, latency & availability metrics

CausRCA (manufacturing)

- 100 runs, 19 fault experiments
- 3 subsystems: coolant, hydraulics, probe

		ε-diag	RCD	TimeRank	Baro	Corr	CausTR+PC	PRIM†
PetShop		0.10*	0.39*	0.15	0.08	0.78*	–	**0.79**
CausRCA	Full	0.02	0.04	0.24	0.09	0.05	–	**0.27**
	Sub	0.11	0.24	0.60	**0.75**	0.33	0.57*	0.61

Table 1: Average RCA accuracy without a given graph. PetShop: top-3 recall; CausRCA: MAP@3. * from original papers Hardt et al. 2024; Mehling et al. 2026; – not reported; **bold** best per row; † ours.