Causal Preference Elicitation

Expert-in-the-loop Bayesian Causal Discovery

Edwin V. Bonilla, He Zhao & Daniel M. Steinberg

CSIRO, Australia

https://github.com/csiro-funml/CaPE

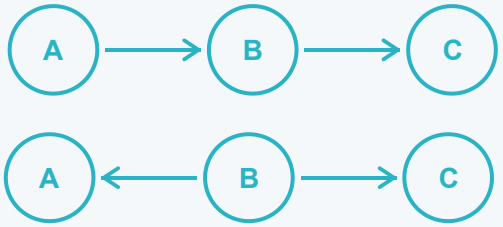
ICML 2026

CaPE

CAUSAL PREFERENCE ELICITATION

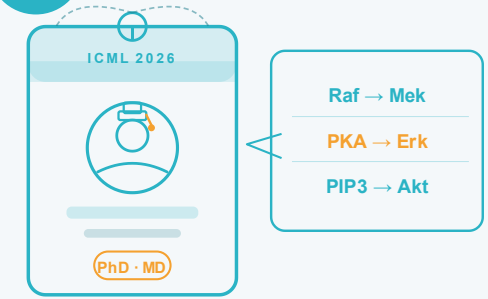
Causal discovery is hard, but experts know things

01 Markov Equivalence



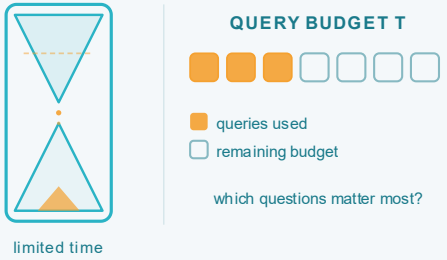
Many DAGs produce identical observational distributions. Data alone can't tell them apart.

02 Expert Knowledge Exists



Domain experts know which edges are plausible, forbidden, or directionally expected.

03 Expert Time is Scarce

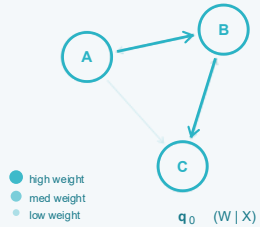


Instead of static one-shot priors we ask: which edges should we query?

CaPE -> Actively ask the right expert questions, in the right order, with Bayesian rigor

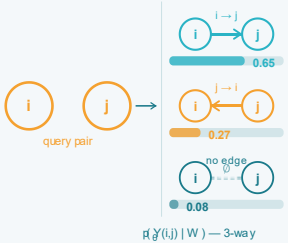
Modular Bayesian active learning over DAGs

1 Observational Posterior



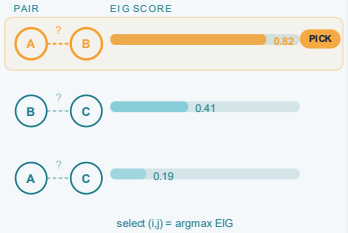
Bootstrap/MCMC/ GFlowNet
q0(W | X)

2 Expert Model



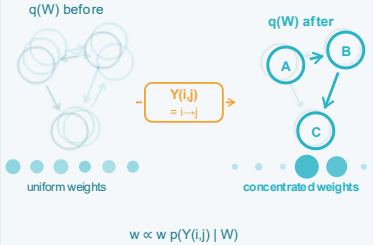
3-way likelihood:
i->j / j->i / no edge

3 Query Selection (EIG)



BALD: pick pair (i,j) that
maximises info gain

4 Bayesian Update



Importance reweighting
+ SMC rejuvenation

<- iterates T rounds until query budget exhausted

- **Black-box compatible** Any DAG sampler works: MCMC, GFlowNet, bootstrap
- **Information-efficient** EIG policy concentrates posterior much faster than random
- **Plug-and-play expert model** 3-way noisy likelihood; handles bias, abstention, noise
- **Provably consistent** Identifiability theorem under non-adversarial expert feedback

Three ingredients that make CaPE work

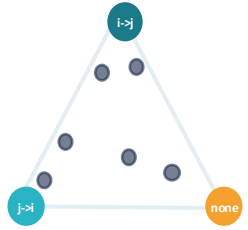
I Hierarchical Logistic Expert Model

Two logistic components per queried pair (i,j):

1) edge-existence probability p_{edge}
2) orientation probability p_{dir}

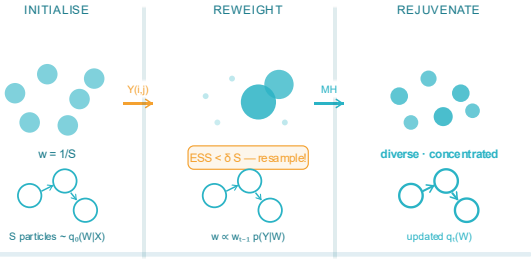
Yields a 3-class categorical likelihood over {i->j, j->i, none}.

Particle spread in simplex = high EIG



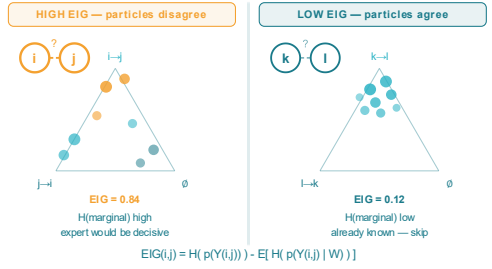
II Particle-Based Posterior (SMC)

- S weighted DAG particles updated via importance reweighting.
- ESS-triggered resampling + MH rejuvenation restores diversity while strictly enforcing acyclicity.



III BALD-Style Query Selection (EIG)

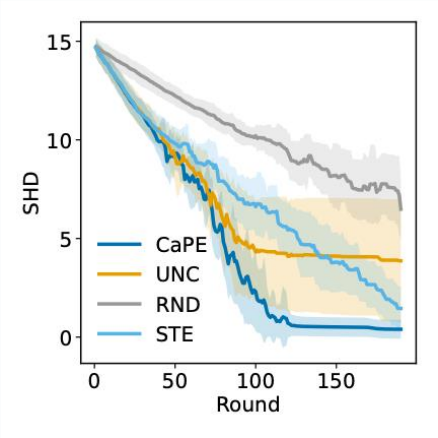
- EIG(i,j) = H(marginal predictive) - E[H(expert | W)].
- Favors pairs where particles disagree but expert is decisive.
- Equivalent to expected KL posterior contraction.



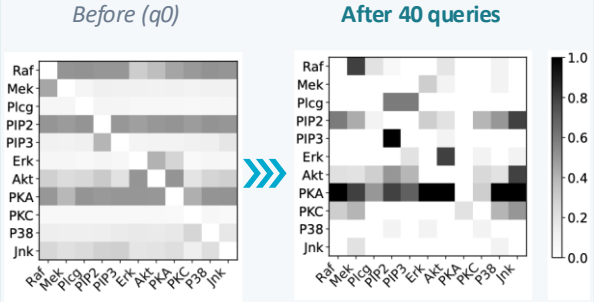
Theory: Identifiability under non-adversarial feedback . EIG = Expected KL contraction . Geometric simplex interpretation

CaPE outperforms baselines across all settings

Synthetic DAGs: SHD (↓)

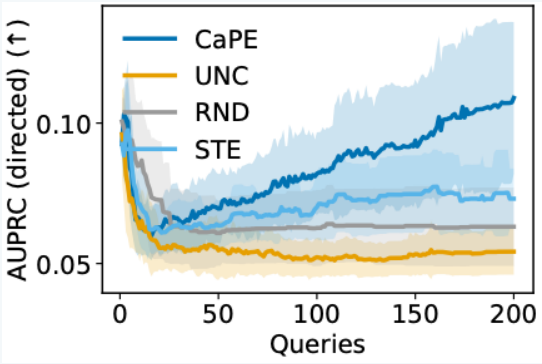


Sachs Protein: Posterior Edge Probabilities



(down) SHD by ~14 edges | Orient. F1 +0.38
observational only

CausalBench Genes (D=50): AUPRC (↑)



CaPE wins
all 3 benchmarks

-40 queries
Sachs SHD: -14 edges

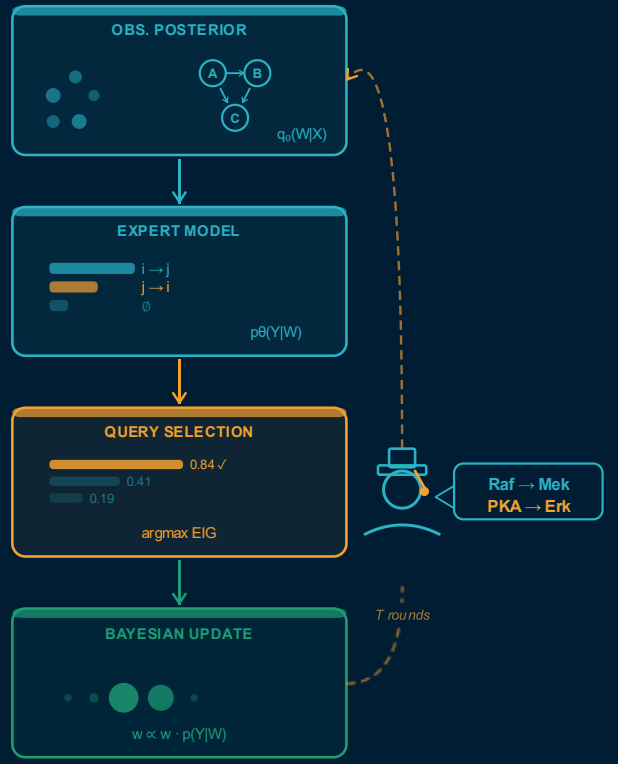
+0.38
Sachs Orientation F1

Robust
5 misspecification regimes

TAKEAWAYS

CaPE in a nutshell

- **Active querying works** EIG consistently beats baselines
- **Plug-and-play** Samples from any black-box DAG posterior as input
- **Principled** BALD/EIG = expected KL contraction; identifiability theorem
- **Validated** Synthetic graphs, Sachs protein signaling, CausalBench 50-gene
- **Robust** Stays competitive under expert bias, abstention, and heavy noise



edwin.bonilla@csiro.au
Australia's National Science Agency