Leveraging Large Language Models for Causal Discovery: a Constraint-based, Argumentation-driven Approach

Zihao Li &
Fabrizio Russo

SCAN ME

Motivation

Can we integrate semantic knowledge into data-driven causal discovery in a transparent, principled way?

Example

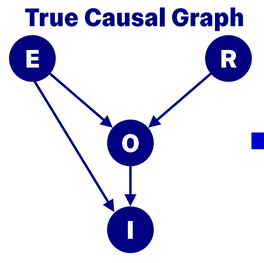
observed data

E	R	O	I
BSc	W	Fin	41
PhD	A	Ed	68
Hi	L	Of	29
BSc	B	Bu	83
MSc	A	Of	52

example CI tests

t_1: $E \not\perp\!\!\!\perp O$
t_2: $E \perp\!\!\!\perp R \mid O$
t_3: $R \perp\!\!\!\perp I \mid \{E, O\}$

True Causal Graph



more CI tests

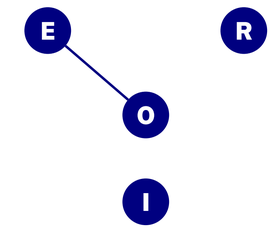
t_4: $E \perp\!\!\!\perp O \mid I$
t_5: $E \not\perp\!\!\!\perp I$
t_6: $O \perp\!\!\!\perp I$

No DAG compatible with all six CI tests

$t_1 \Rightarrow$ an active E–O path exists
$t_4 \Rightarrow$ conditioning on I blocks it, so I is a non-collider on that path
$t_1, t_4, t_5 \Rightarrow$ O and I must be dependent: $O \not\perp\!\!\!\perp I$
but t_6 asserts $O \perp\!\!\!\perp I$

Majority-PC Output

[Colombo and Maathius, 2012]



Background — Causal ABA and ABA-PC

Assumption-Based Argumentation framework $\langle L, R, A, {}^- \rangle$

Assumptions $A \in L$

$A_{arr} = \{arr_{xy} \mid x, y \in V, x \neq y\}$ $A_{dag} = A_{arr} \cup \{noe_{xy}\}$
$A_{ind} = \{(x \perp\!\!\!\perp y \mid Z) \mid Z \subseteq V, x, y \in V \setminus Z\}$ $A_{ds} = A_{dag} \cup A_{ind}$
$A_{bp} = \{bp_{p|Z} \mid p \text{ an } x - y - \text{path}\}$ $A_{csl} = A_{ds} \cup A_{bp}$

Rules R $R_{ds} = R_{dag} \cup R_{graph} \cup R_{act}$, $R_{ext} = R_{ds} \cup R_{xyZ}$

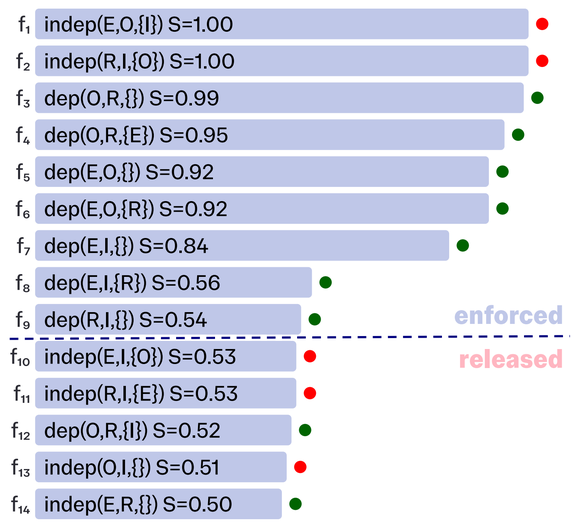
R_{graph}: $dpath_{xy} \leftarrow arr_{xy}$ $dpath_{xz} \leftarrow dpath_{xy}, arr_{yz}$
$e_{xy} \leftarrow arr_{xy}$ $e_{xy} \leftarrow arr_{yx}$
R_{xyZ}: $(x \perp\!\!\!\perp y \mid Z) \leftarrow bp_{p_1|Z}, \ldots, bp_{p_k|Z}$ $ap_{p|Z} \leftarrow p_t$
$(x \perp\!\!\!\perp y \mid Z) \leftarrow \cdot$ $\overline{(x \perp\!\!\!\perp y \mid Z)} \leftarrow \cdot$ $arr_{xy} \leftarrow$ **Facts: enter every extension**

Contraries $^- : A \to L$

$\overline{a} \leftarrow b$, $a \neq b$, $a, b \in \{arr_{xy}, arr_{yx}, noe_{xy}\}$ one edge state per pair
$\overline{arr_{x_i x_{i+1}}} \leftarrow arr_{x_1 x_2}, \ldots, arr_{x_{k-1} x_k}$ for $x_1 = x_k$ acyclicity
$\overline{noe_{xy}} \leftarrow e_{xy}$ $\overline{(x \perp\!\!\!\perp y \mid Z)} \leftarrow p_Z\ p_Z$ Z-active path $\overline{bp_{p|Z}} \leftarrow ap_{p|Z}$

Stabe Extension: set of assumptions that does not attack itself and attacks all assumptions outside the set.

14 CI tests, ranked by strength (p-value transformation)



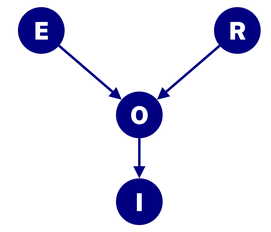
ABA-PC

PC-derived facts ranked by strength
↓
Causal ABA
↓
Stable extension ?
No: release lowest-ranked fact, rerun
Yes → stop
↓
ABAPC: 9 accepted
7 true · 2 false

stable extension ≡ a DAG + the d-separations it implies — enforced facts inconsistent ⇒ no stable extension

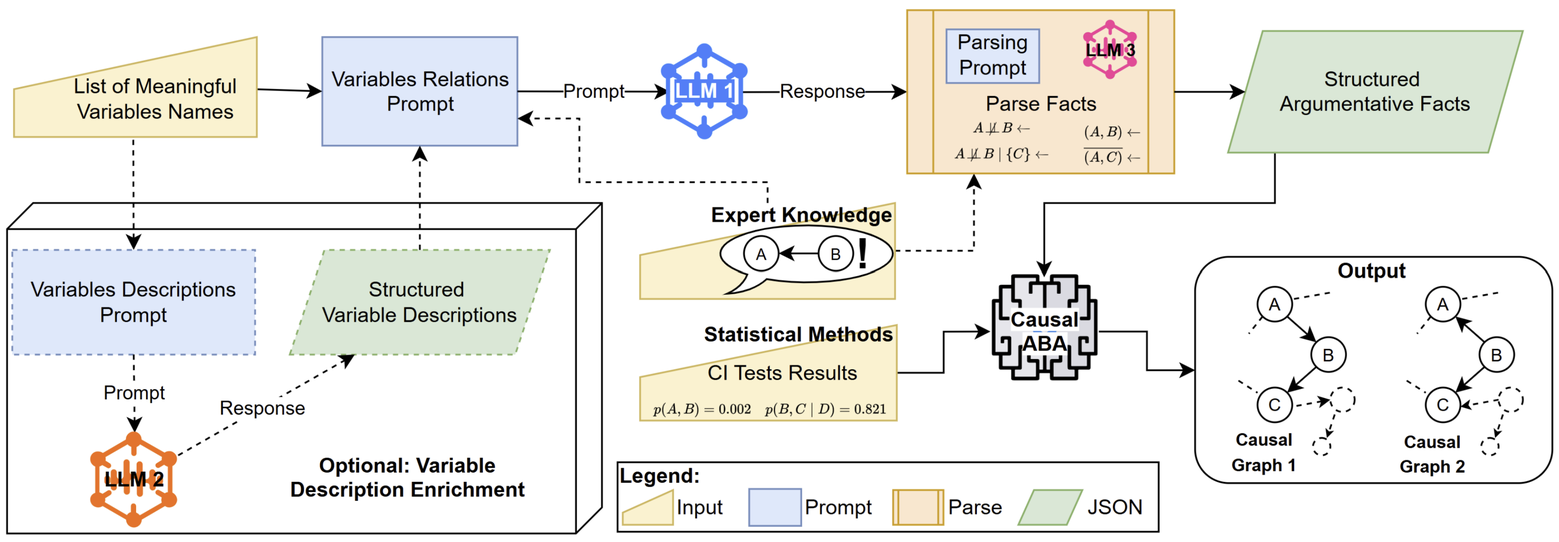
ABA-PC Output

[Russo, Rapberger & Toni, 2024]



ABAPC-LLM: robust & transparent integration of LLM knowledge

Data constrains adjacencies; consensus semantic facts constrain directions. Integration through Causal ABA.



Variables Relations Elicitation — five independent calls

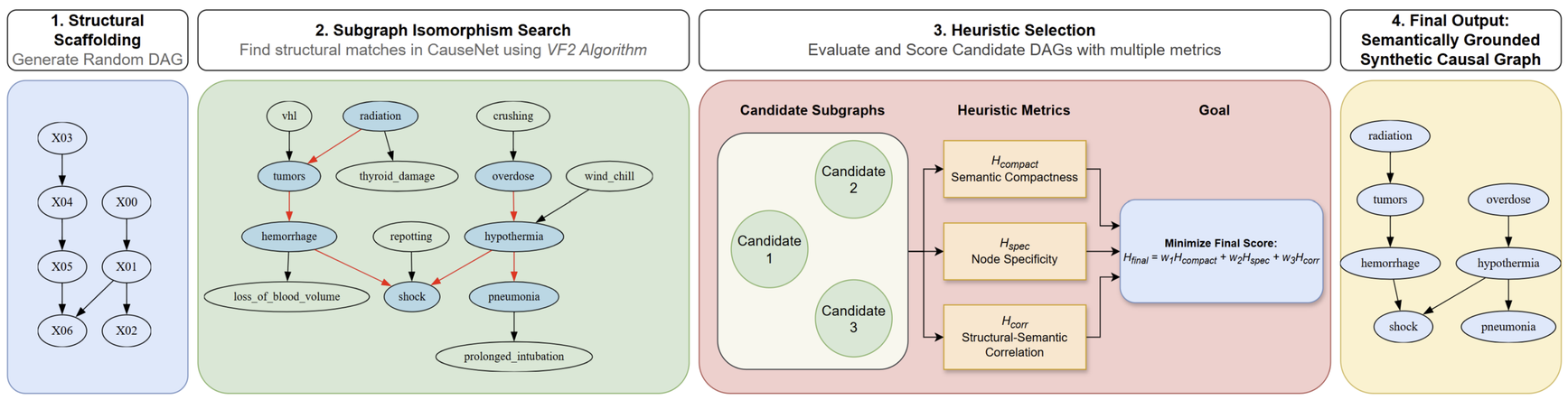
Variable names + optional descriptions → High-precision causal prompt × 5
Required and forbidden directions R_i, F_i ← Schema-guided parsing
Consensus filter $R_e = \bigcap_i R_i$, $F_e = \bigcap_i F_i$

Consensus filtering example

$arr_{OI} \leftarrow$ ●●●●● 5/5 → fact enforced (no independence found for O,I)
$\overline{arr_{IR}} \leftarrow$ ●●●●● 5/5 → fact enforced (forbidden direction)
$arr_{IR} \leftarrow$ ●●●○○ 3/5 → no consensus, discarded direction

A required arrow $arr_{xy} \leftarrow$ is enforced only if x–y survives the reduced MPC skeleton; a forbidden arrow $\overline{arr_{xy}} \leftarrow$ blocks one orientation.
LLMs orient surviving adjacencies — they can never add an edge.

A novel benchmark for LLM-aided causal discovery



Reduce memorization bias: Randomising topology makes verbatim recall of a whole benchmark DAG implausible.

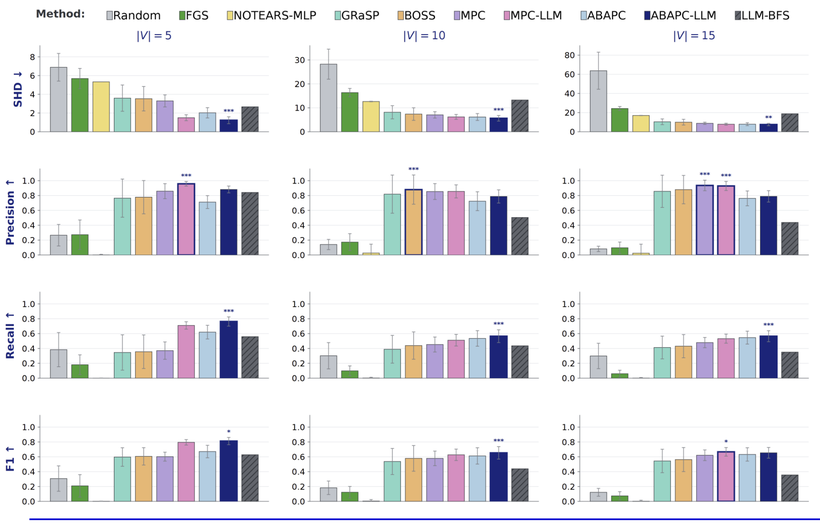
Ranking heuristic

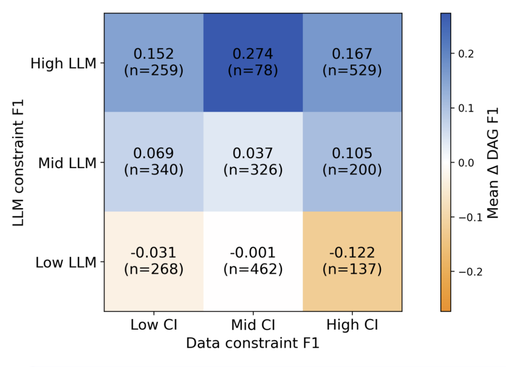
VF2 can return up to ~10^6 isomorphic sub-graphs;

$H_{compact}(S_c) = avg_{v \in S_c} [d_{cos}(emb(v), \mu_{S_c})]$ *tight semantic cluster*
$H_{spec}(S_c) = avg_{v \in S_c} [\log(\deg(v) + 1)]$ *penalise hub nodes*
$H_{corr}(S_c) = 1 - \rho(d_{graph}, d_{sem})$ *topology ↔ semantics*

Empirical Results

54 DAGs · |V| = 5/10/15 · 5,000 samples · 50 reps · Wilks G^2, α = 0.01





Key findings

Average SHD reduction

-57% vs LLM-BFS & -17.5% vs ABA-PC

1. **Lowest SHD and highest recall at every graph size (|V| = 5, 10, 15); best F1 at 5 and 10.**
2. **SHD and recall gains are significant against the next-ranked method (*** p < 0.001).**
3. **Semantic facts help most when both LLM and CI evidence are reliable.**

**LLMs supply fallible semantic evidence;
Causal ABA makes its integration constrained and traceable.**

Paper

Code + results

European Research Council
Established by the European Commission

IMPERIAL

uai2026