WHERE THE ACCURACY GOES IN SINGLE-EDIT CAUSAL SEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Score-based structure learning returns the graph a greedy search reaches by repeatedly taking the single edit that most improves a score such as BIC, so its output is settled by how the score orders the candidate edits at each step, an order that evaluation by the final graph never observes. We measure it by replay on five discrete benchmark networks: at every step of searches that were actually run, every legal edit is rescored with the shipped implementation and against the ground-truth graph. On the three smaller networks the edit the search takes beats most accepted alternatives in oriented-edge F1, yet it is the best of them at only 45% of steps, and against its closest competitors in score it is truer 58% of the time. Reordering only the edits the score already accepts recovers 32.6% of the distance from a uniform draw to the single-edit ceiling, against 10.1% for the score's own order. Half of that gap is exact ties between Markov-equivalent successors, which the data cannot orient and the implementation breaks by enumeration order; a coin flip in its place moves the result significantly, in opposite directions on different networks. On the two larger networks the gap narrows to 4.3 points. Measured against what reordering can reach, a frontier language model in the reordering role recovers 76.2% on child and insurance, averaged over three runs, against 43.5% for the score's own order; its lead disappears when variable names are replaced by pseudonyms and survives plain-language descriptions. No open-weight model we tested beats the score significantly.

1 INTRODUCTION

Score-based structure learning recovers a causal graph from observational data by greedy search: from a starting graph it enumerates the legal single edits, scores each, and takes the one that most improves a score such as BIC, until no edit improves it (Chickering, 2002; Ramsey, 2015). Its failures are studied through the graph it ends on (Nazaret & Blei, 2024; Ejaz & Bareinboim, 2025). Methods are compared by the graph they return. That graph is the product of the order the score imposes on the candidate edits at every step, and the endpoint mixes this order with the starting graph, the move set, the stopping rule and the sample. To our knowledge, no study has measured how faithfully the score orders the candidates a search actually weighs.

We measure it at the point of choice (Figure 1). Replaying searches that were actually run, we score every legal edit at every step twice, with the shipped implementation and against the ground-truth graph, and compare the two orders.

The score's order carries information and still decides badly where it matters. On the smaller networks the edit a search takes is truer than an accepted alternative 75.1% of the time by oriented-edge F1 and 79.3% by the structural Hamming distance between CPDAGs, yet it is the best accepted edit at 44.9% and 36.5% of steps respectively: against its closest competitors in score it is truer only 57.7% of the time, and those are the edits it chooses between.

The accuracy this costs is large and has a specific structure. Holding the start, the legality rules, the in-degree cap and the budget fixed, and keeping every edit the score rejects out of reach (the score's *veto*), reordering the edits it accepts recovers three times the accuracy of its own order. Half of the gap comes from exact ties between Markov-equivalent successor graphs, which no likelihood separates and which the implementation breaks by enumeration order; that tie-break moves reported accuracy significantly, and in opposite directions on different networks, a new instance of the order

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

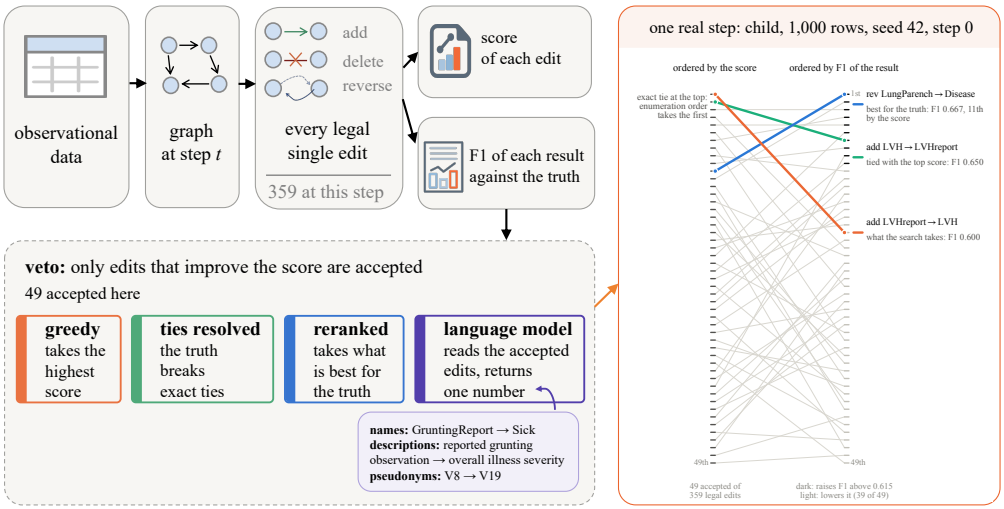

Figure 1: **What is measured.** At every step of a search that was run, every legal edit is scored twice, by the shipped score and by the F1 of the graph it produces; the score accepts only edits that improve it, and one policy picks among them. Right: step 0 of child at 1,000 rows, seed 42. The search takes an edit that lowers F1 from 0.615 to 0.600, while its reverse has exactly the same score and would raise it to 0.650.

effects known for constraint-based search and simulated benchmarks (Colombo & Maathuis, 2014; Reisach et al., 2021). Most of the remaining gap is one preference: the score takes the sparser of two candidates far more often than the sparser one is right, and its complexity term produces that preference.

Language models can supply what the data cannot (Kıcıman et al., 2024; Darvariu et al., 2024; Zhang et al., 2026). Placed in the reordering role inside the same veto, a frontier model recovers most of the reordering headroom on child and insurance, and at the ties alone it leads the enumeration order in each of three runs, short of significance. Both leads come from what the variables mean: they disappear under pseudonyms and survive under plain-language descriptions. None of the fourteen open-weight models that pass our audit beats the score's own order significantly.

Contributions.

- A replay instrument that reproduces the shipped scorer on all 2,400 recorded score deltas and measures fidelity at the point of choice under oriented-edge F1 and an equivalence-class metric, finding the score weakest against the competitors a search chooses between (Sections 2 and 3).
- A decomposition under a preserved veto that divides the ordering loss into exact indifference and misordering, and traces the misordering to a sparsity preference produced by the complexity term, a diagnosis that does not say how to set the term (Sections 4 to 6).
- Evidence that an unreported tie-break changes reported accuracy by a large margin whose sign depends on the network (Section 4).
- A measurement of what language models contribute in the one role the decomposition leaves open, with presentations that separate a variable's meaning from its spelling (Section 7).

2 What is measured, and on what

The search. The learner holds a DAG, enumerates the legal single edits from it (adding, deleting or reversing one arrow, with at most three parents per variable), scores each with BIC plus a penalty of 2 per edge, and takes the edit that improves the score most, for at most $3|V|$ steps or until no legal edit improves it. We call this policy greedy, call an edit *accepted* when it improves the score, and call the rule that no other edit is ever taken the score's *veto*. Everything the search returns is

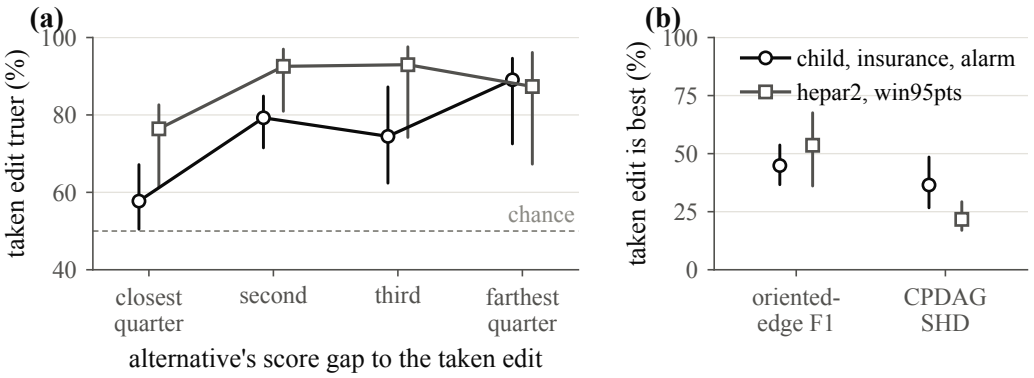

Figure 2: **The score is weakest against the edits a search chooses between. (a)** How often the edit a greedy search takes is truer, by oriented-edge F1, than an accepted alternative, with the alternatives split into quartiles of their score gap to the taken edit. **(b)** How often the taken edit is the best accepted edit, by oriented-edge F1 and by CPDAG SHD. Every step on 27 cases of the smaller networks and 12 of hepar2 and win95pts; pairs the score or the metric cannot order are dropped, and bars are 95% case-bootstrap intervals. Exact values are in Table 7.

decided by how the score orders the accepted edits at each step; the reported result shows only the final graph.

Replay. For searches that were actually run, we take each step, enumerate every legal edit from the graph held at that step, and score each candidate twice: with the *shipped* implementation of the score, the scorer released with the search and used unchanged, and by the *oriented-edge F1* of the resulting graph against the ground truth, the harmonic mean of precision and recall over directed edges. Replaying the edits the searches took reproduces 2,400 of 2,400 recorded score deltas, so every measurement below uses the shipped scorer itself. Where orientation matters we also judge graphs by the structural Hamming distance between their CPDAGs (*CPDAG SHD*), which is the same for any two Markov-equivalent graphs.

Networks and cases. Five discrete BNLEARN networks: child (20 variables), insurance (27) and alarm (37), the *smaller networks*, and hepar2 (70) and win95pts (76). A case is one network at one sample size with one data seed. Table 4 lists the cases each measurement uses.

Ties and reproducibility. Candidates whose scores differ by less than 10^{-6} are tied, and the implementation takes whichever it enumerated first. Enumeration order depends on Python's hash seed, so every run fixes `PYTHONHASHSEED`. Under four hash seeds, greedy on child at 1,000 rows, seed 42, ends at 0.750 twice and 0.792 twice; four other cases do not move.

Shares and intervals. Accuracy is reported as a share of the distance from a uniform draw to a named ceiling, pooled over cases as a ratio of sums, with 95% intervals from a bootstrap that resamples cases (Appendix B). We call a difference significant when its interval excludes zero.

3 How the score orders the candidates a search weighs

At every step of the greedy trajectories we compare the edit the search takes with each other edit the score accepted (Figure 2; exact values in Table 7). On the smaller networks the taken edit is truer than an alternative 75.1% of the time by oriented-edge F1 and 79.3% by CPDAG SHD, so the score's order carries information under either metric. Its top choice is still the best accepted edit at only 44.9% of steps by oriented-edge F1 and 36.5% by CPDAG SHD. The two readings fit together because a step offers a median of 14 accepted alternatives: beating most of them still leaves the best one untaken at most steps.

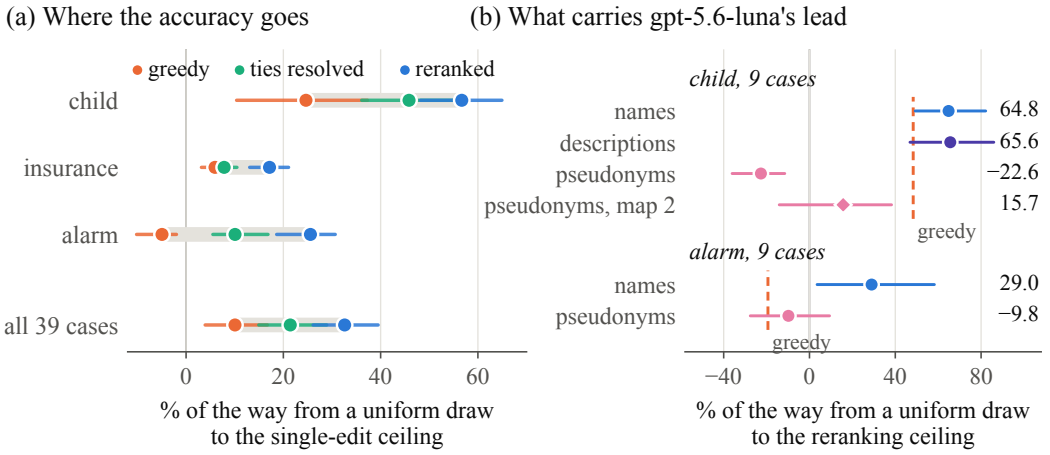

Figure 3: **Where the accuracy goes, and what carries a frontier model's lead.** Points are ratios of sums pooled over cases, with 95% case-bootstrap intervals; the colours match Figure 1. **(a)** From the same start under the same constraints and veto, greedy, greedy with only its exact ties settled by the truth, and every accepted edit reordered by the truth, on each smaller network and pooled over their 39 cases. The gap differs by network and is smallest on insurance; settling ties alone closes half of the pooled gap. **(b)** The frontier model gpt-5.6-luna in the reordering role on child, where a plain-language description of every variable is available, and on alarm, under each presentation of the same variables; the dashed line is greedy on the same cases. Replacing names with descriptions leaves the share unchanged; replacing them with pseudonyms puts it below greedy under either map. Exact values are in Tables 10 and 3.

The score is weakest against its closest competitors. Against the quarter of alternatives closest to it in score, the taken edit is truer 57.7% of the time, and against the farther half 81.8%; these close competitors are the edits a search actually chooses between. The alternatives are close in the truth as well: 86.4% of the 6,336 accepted alternatives lie within 0.05 oriented-edge F1 of the taken edit, the *near band*, and none lies beyond 0.15. Within the near band the taken edit is still truer 71.2% of the time, so what marks the score's failures is closeness in score. On hepar2 and win95pts the score is steadier: the taken edit is truer than an alternative 87.3% of the time and 76.4% against its closest quarter, though it is the best accepted edit at 53.6% of steps by oriented-edge F1 and 21.7% by CPDAG SHD.

Pairs of finished graphs behave differently. On the 1,742 pairs that PC, GES and NOTEARS produce on the same rows, BIC agrees with the truth on 82% of pairs more than 0.15 apart and 54% within the near band (Appendix D). Agreement rates measured on finished graphs therefore do not transfer to the one-edge comparisons a search makes.

4 Where the accuracy goes

A search that takes the best accepted edit at 44.9% of its steps leaves accuracy behind. To measure how much is at stake we hold everything except the order fixed. From one starting graph, under the same legality rules, in-degree cap and step budget, five policies walk:

- uniform, an accepted edit drawn uniformly, averaged over ten draws: the floor;
- greedy, the highest-scoring accepted edit: what the search does;
- prior, the accepted edit that agrees with a graph a language model proposed before seeing data;
- reranked, the accepted edit that raises oriented-edge F1 most: the *reranking ceiling*;
- oracle, the legal edit that raises F1 most, accepted or not: the *single-edit ceiling*.

reranked keeps the score's veto. It changes only the order in which accepted edits are taken, so its gap to greedy is the accuracy available to a better ordering of the same candidates.

Over 39 cases on the smaller networks (Figure 3a; by network in Table 11, every case in Table 12), as a share of the distance from uniform to the single-edit ceiling, greedy recovers 10.1%, prior 17.4% and reranked 32.6%. The gap between reranked and greedy, the *ordering loss*, is 22.5 points, with a 95% case-bootstrap interval of $[18.4, 27.0]$. reranked leads in 38 of the 39 cases, and in 13 of them greedy falls below uniform. The edits that would have helped were enumerated, legal and accepted; the score ranked them below other edits it also accepted.

Half of the ordering loss is exact indifference. A score can misorder two candidates or fail to separate them at all, and one more policy tells the two apart. greedy, ties resolved is greedy except where several accepted edits score within 10^{-6} of the best, and there it takes the one that raises F1 most. It recovers 21.4%, so exact indifference accounts for 11.4 of the 22.5 points of ordering loss: half, with an interval of $[37\%, 65\%]$. Settling ties never costs a case: it is better in 26 of the 39 cases, worse in none and neutral in 13, a mean gain of 0.040 oriented-edge F1.

Adding $A \to B$ and adding $B \to A$ usually yield Markov-equivalent graphs, which no likelihood separates and oriented-edge F1 does. On the 27 cases of the smaller networks the top-scoring accepted edit is tied with another at 99 of 285 steps, 35%; every such tie is a pair of Markov-equivalent successor graphs, and in 93 of the 99 the two differ in F1. The score is right to be indifferent here: the data cannot orient a Markov-equivalent pair, and a search over DAGs must pick one member. The pick follows an unreported rule and moves the oriented-edge accuracy that benchmarks report, a difference CPDAG SHD does not register because it scores both members alike. The finding is therefore specific to single-edit search over DAGs, the move set of the language-model proposal systems in Section 8; GES searches over equivalence classes and avoids these ties by construction, and our results do not bound what it loses.

The enumeration tie-break changes the result, with a sign that depends on the network. Candidates are generated by scanning variables in the order the network file lists them, and a tie goes to the member generated first. Against a coin flip averaged over ten draws, this file order helps on two networks and hurts on the third. As a share of the distance to the reranking ceiling it reaches 48.3% against the coin's 16.2% on child, better in 8 of 9, and 31.7% against -4.3% on insurance, better in all 9; on alarm it reaches -19.4% against 2.5%, worse in 8 of 9. Each difference is significant, also at the 98.3% level that corrects for three networks.

The sign follows the edges the search ties on. On all three networks the file order as a whole runs with the truth: between 56% and 65% of true edges point from an earlier-listed variable to a later one, and hepar2 and win95pts are listed in exact topological order. The tied edges are a different sample. Among ties where one candidate is truly better, the first-generated one is better in 24 of 37 on child, 20 of 29 on insurance and 11 of 27 on alarm. Each rate is close to a coin flip and has the sign of its network's trajectory-level effect; over the roughly ten ties a trajectory passes through, a small per-tie bias compounds. On alarm, 63% of true edges run forward in the file, against 41% of the ones its search ties on. Averaging over networks cannot remove this unreported effect, and on alarm the shipped tie-break costs accuracy that a coin flip keeps.

The result holds from every starting graph. From each of four starts (the pooled graph the score prefers and the graphs of GES under BDeu, NOTEARS and PC), reranked leads greedy with an interval that excludes zero, and greedy takes between 0.38 and 0.65 of the reachable ordering, below half from three of the four (Table 5).

On larger networks each tie is worth less. On hepar2 and win95pts greedy recovers 13.8% and reranked 18.2% over 12 cases, a gap of 4.3 points against 22.5 on the smaller networks; reranked still leads in 10 of 12 cases, and greedy never falls below uniform there, against 13 of 39 on the smaller networks. The tie mechanism is nearly unchanged. At 1,000 rows the top score is tied at 34 of 102 steps (33%) over the six cases of hepar2 and win95pts, against 32 of 83 (39%) over the nine cases of the smaller networks, and every tie is a pair of Markov-equivalent successors at both scales. What changes is the F1 at stake, 0.0124 per tie against 0.0341, because one arrow is a smaller share of a graph with more than a hundred true edges than of one with twenty-five; per case, ties carry 0.062 F1 at the larger scale against 0.110 at the smaller. The score makes the same mistake at nearly the same rate on a metric where each mistake costs less.

Table 1: **The complexity term moves the sparsity preference.** The 1,691 comparisons of the smaller networks and the 365 of hepar2 and win95pts that differ in edge count, rescored with the complexity term multiplied by γ: how often the score takes the sparser graph, how often it agrees with the truth, and the share of its errors that are taking the sparser graph when it is the worse one. Percentages; $\gamma = 1$ is the shipped score. The grey row is the shipped score; bold is the highest agreement in each group.

	child, insurance, alarm			hepar2, win95pts	
γ	takes sparser	agrees	errors from sparsity	takes sparser	agrees
1	57.7	69.3	85.9	36.7	78.1
0.75	55.0	70.8	83.2	32.1	76.7
0.5	51.3	73.2	79.1	25.8	**78.6**
0.25	45.7	**76.6**	71.4	20.8	78.1
sparser is truer		35.7		37.3	

5 THE MISORDERING IS A SPARSITY PREFERENCE

The finished-graph pairs show which way the score errs, because each pair differs in many edges at once; Section 6 then tests the mechanism inside the search. Of the 1,742 comparisons, 1,691 are between graphs that differ in edge count. The sparser graph is the truer one in 35.7% of them, $[32.4, 39.1]$, and the score prefers it in 57.7%, $[52.8, 62.9]$. Preferring the sparser graph when it is the worse one accounts for 85.9% of the score's errors, $[79.9, 91.3]$. On hepar2 and win95pts the two rates are 37.3% and 36.7% over 365 comparisons: there the score prefers the sparser graph about as often as it is right to, and these are the networks on which the pairwise readings of Section 3 are highest. Like every rate in this section, the preference is read with oriented-edge F1 as the truth; Appendix D repeats it on equivalence classes.

The configurable edge penalty does not produce the preference: adding 2 per edge moves agreement by at most one point in any band, while changing the score from BIC to BDeu moves it by 5 to 14 points (Table 8). The remaining candidate is the complexity term of the score itself, which Section 6 varies.

6 THE COMPLEXITY TERM PRODUCES THE PREFERENCE

We test a mechanism: whether the complexity term produces the preference of Section 5, and whether changing it changes which edges a search takes. Scaling this term is established practice. It is `penaltyDiscount` in Tetrad (Ramsey, 2015); its discrete counterpart is BDeu's equivalent sample size, whose effect on the recovered structure Silander et al. (2007) characterised; and Banerjee et al. (2025) sweep it against accuracy.

We rescored the same 1,691 comparisons (Table 1) with the shipped scorer and its complexity term multiplied by γ. As γ falls from 1 to 0.25 the score prefers the sparser graph monotonically less often, from 57.7% to 45.7%, a drop of 12.1 points, $[-15.1, -9.2]$, and agrees with the truth 7.3 points more often, $[4.6, 10.2]$. On hepar2 and win95pts, where the sparser graph is taken about as often as it is truer, the same scaling moves how often it is taken and leaves agreement unchanged, at $+0.0$ points, $[-7.3, 8.4]$. The term sets the preference on every network, and the preference costs accuracy where it is wrong.

We then ran the shipped search with the term multiplied by $\gamma \in \{0.25, 0.5, 0.75, 1\}$ on the smaller networks at three sample sizes and three seeds, 27 units per setting, a unit being one network, sample size and seed (Table 15); at $\gamma = 1$ the implementation is byte-identical to the shipped one. The shipped setting is best on none of the nine dataset cells and tied for best on two. At $\gamma = 0.25$ one unit moves by $+0.394$ oriented-edge F1 and another by -0.251, while across the four settings the number of steps on a cell differs by at most two and the final edge count by at most four. The term changes which edges are taken much more than how many.

Table 2: **What a model contributes in the ordering role, with real names and with pseudonyms.** Share of the distance from a uniform draw among accepted edits to the reranking ceiling, pooled over 18 cases as a ratio of sums. Pseudonyms replace each name with a label whose order is unrelated to the enumeration order. With names, only the frontier model beats greedy significantly; with pseudonyms, no model does. The frontier model's names entry is its first run, the one its pseudonym run is paired with; its three-run mean is 76.2. The grey row is the reference; bold is the one share whose lead over greedy is significant; grey type fails the position gate under pseudonyms (Appendix C).

	names	pseud.		names	pseud.
greedy, the score's own order: 43.5			prior, matched: 46.9		
gpt-5.6-luna	**72.5**	-8.0	qwen3-14b	13.4	-17.1
qwen3.8-27b	46.5	$-5.9^{\dagger}$	qwen3-30b-a3b	12.5	-13.3
qwen3-32b-fp8	40.6	-19.3	qwen3-8b	10.3	13.0
qwen3-4b-2507	31.2	5.3	glm-4.7-flash	0.4	—
qwen2.5-7b	26.4	-12.2	olmo2-7b	-2.5	—
qwen3-coder-30b	26.3	—	llama3.1-8b	-12.6	-11.9
qwen3-1.7b	19.4	-5.7	deepseek-v2-lite	-18.2	—
qwen3-4b	13.4	-10.0			

A diagnosis, without a recipe. Paired against the shipped setting and bootstrapped over the nine datasets, $\gamma = 0.25$ moves 9 units up, 9 down and leaves 9 unchanged, a mean of $+0.031$ with a 95% interval of $[-0.031, +0.112]$ and a signed-rank p of 0.27. Halving the term has an interval that excludes zero, $[+0.005, +0.063]$, with a signed-rank p of 0.11 and 16 of 27 units unchanged. The mean at $\gamma = 0.25$ is carried by the three units of alarm at 250 rows, and its largest drop is one unit of child at 250 rows. The best factor differs by network and sample size, and choosing it requires the graph the search is trying to recover.

7 What language models contribute

Half of the ordering loss is the score being exactly indifferent between Markov-equivalent successors, which no amount of observational data separates, so an outside source is the only thing that can supply the direction. We measure what language models contribute there and at every other step, in the role the decomposition defines. At each step the model sees the edits the score has accepted, in a shuffled order, and returns the number of the edit it would take, decoded at temperature 0 (Appendix C gives the prompt, the decoding and the parser). It never sees the data, the scores or the truth, and it cannot take an edit the score rejected, so its result lies on the same axis as every other policy. Except where alarm is named, shares in this section are ratios of sums over the 18 cases of child and insurance against the reranking ceiling, on which greedy reaches 43.5%; Section 4 divides by the larger single-edit ceiling over 39 cases, where the same policy reads 10.1%.

The open-weight models are served locally. The frontier model, gpt-5.6-luna, is reached through an API. It is run three times with real names in both roles, and the interval on a three-run mean resamples runs as well as cases; every other condition is run once.

No open-weight model beats the score significantly; the frontier model does. A model is admitted only if at least 90% of its answers parse and no presented position takes more than half of its choices among five or more candidates, since a model that answers a position is not ranking edits. None of the fourteen open-weight models that clear this gate beats the score's own order significantly, and eight fall significantly below it. The closest is qwen3.8-27b at 46.5%, $+3.0$ points from the score with an interval of $[-15.0, +23.5]$ (Table 14 splits every model by network). The three largest open-weight entries of our roster were not run (Appendix C), so the result covers the models we ran. The frontier model recovers most of the reordering headroom from inside the same veto: its three runs give 72.5%, 67.3% and 88.7%, a spread of 21 points, and their mean, 76.2%, is 32.7 points above greedy, $[+15.2, +54.0]$.

Table 3: **The frontier model in every condition.** Share of the distance from a uniform draw among accepted edits to the reranking ceiling, and the difference from greedy on the same cases (for the ties-only role, from the enumeration tie-break) with a 95% interval; for the three-run means the interval resamples runs as well as cases. Grey rows are the three-run means; bold differences have an interval that excludes zero.

condition	cases	runs	share	− greedy
every accepted edit				
names, mean of three runs	18	3	76.2	**+32.7** [+15.2, +54.0]
names, first run	18	1	72.5	**+29.1** [+9.9, +54.6]
pseudonyms	18	1	−8.0	**−51.5** [−68.1, −28.2]
names, child	9	1	64.8	+16.5 [−4.1, +47.2]
descriptions, child	9	1	65.6	+17.3 [−10.0, +53.9]
pseudonyms, child	9	1	−22.6	**−71.0** [−83.3, −53.1]
pseudonyms, second map, child	9	1	15.7	**−32.6** [−58.5, −5.3]
names, alarm	9	1	29.0	**+48.3** [+20.5, +97.8]
pseudonyms, alarm	9	1	−9.8	+9.5 [−10.1, +45.8]
exact ties only				
names, mean of three runs	18	3	57.3	+13.8 [−1.6, +33.8]
pseudonyms	18	1	−8.4	**−51.8** [−72.1, −31.4]
names, alarm	9	1	−5.5	+13.9 [−5.8, +58.9]

The lead comes from what the variables mean. Pseudonyms take the whole lead away: gpt-5.6-luna falls from its first run's 72.5% to -8.0%, by 80.6 points, $[67.3, 94.1]$, and no model beats the score without names. Under a second map on the nine child cases the model reaches 15.7% against the first map's -22.6%, still 32.6 points below greedy, $[5.4, 58.5]$; the level depends on the map, and the lead is gone under both.

Pseudonyms remove a variable's spelling and its meaning together. On the nine child cases a third presentation keeps the meaning and drops the spelling: each name becomes a plain-language description, written before any experiment here and released with the code. Descriptions match the names: the frontier model reaches 65.6% with descriptions, 64.8% with the names and -22.6% with pseudonyms, against 48.3% for greedy on those cases; descriptions differ from names by $+0.8$ points, $[-12.6, +15.7]$, and from pseudonyms by $+88.3$, $[64.6, 115.4]$. The spelling in the benchmark files adds nothing beyond what the meaning already carries, which is evidence against recall of these particular files; knowledge of the domain learned in another form remains possible.

At the ties, the frontier model leads and the open-weight models do not. Placed only on the exact ties, which make up a fifth of the queries, with every other step left to greedy, a model can no longer override the data, and six of the nine open-weight models we ran this way do better than when they reorder everything. None of the nine beats the enumeration tie-break significantly: eight fall below 43.5% in point estimate, at 12.8% to 37.6%, two significantly; qwen3.8-27b sits level with it at 45.1%; and six of the eight run under both presentations lose ground under pseudonyms (Table 13). The frontier model leads in each of its three runs, at 58.8%, 56.6% and 56.4%; their mean, 57.3%, is 13.8 points above the enumeration tie-break, with an interval of $[-1.6, +33.8]$ that resamples runs as well as cases, a lead consistent across runs and short of significance on these 18 cases. Under pseudonyms the same model on the same ties falls to -8.4%, 67.2 points below its first run with names, $[48.7, 91.8]$, so this lead too comes from the variables' meaning.

Table 3 collects every condition. The runs above use child and insurance. Alarm matters for two reasons: its enumeration tie-break points the wrong way, and language models are reported to recall it (Babakov et al., 2024). On its nine cases the frontier model reaches 29.0% reordering every accepted edit, 48.3 points above greedy at -19.4%, $[20.5, 97.8]$, and pseudonyms take 38.8 points of that lead away, $[14.0, 66.5]$. At alarm's ties alone it sits level with a coin flip, at -5.5% against 2.5%. Over all 27 cases both roles lead greedy, the full role by 33.7 points, $[16.2, 55.3]$, and the tie role by 15.0, $[2.3, 31.5]$, each from the first run of every case.

8 Related work

Guarantees and diagnoses of greedy score search. GES with a locally consistent score identifies the true equivalence class in the large-sample limit (Chickering, 2002). The guarantee concerns the limit point; it says nothing about how two non-equivalent graphs are ranked at a finite sample, which is the quantity we measure. Ejaz & Bareinboim (2025) give the closest existing observation: in their Example 2, over 100 runs from one true DAG, a score-improving insertion selects the truth-worse of two non-equivalent candidates in 4% of runs, as the motivation for a less greedy search. We measure the same phenomenon at every step of searches that were actually run, and separate it from the candidate set. Nazaret & Blei (2024) find GES ending below what the score can reach and attribute it to early over-insertion, measured against the true graph's score; their question is whether the walk reaches the score's optimum; we ask how well the score orders the candidates on the way.

Order and benchmark artifacts. The output of PC depends on the order in which variables are given, which Colombo & Maathuis (2014) show and remove. Reisach et al. (2021) show that simulated benchmarks leak the causal order through marginal variance and that methods exploit it. The enumeration tie-break of Section 4 is a case of the same kind in discrete single-edit search: ties inherit the variable order of the network file, which is topological for two of our five networks.

Search beyond single edits. GRaSP (Lam et al., 2022) and BOSS (Andrews et al., 2023) search over permutations, and amortized approaches (Lorch et al., 2022) replace search altogether. Permutation moves change many edges at once and fall outside the near band, so the decomposition does not apply to them.

Language models and causal structure. Kıcıman et al. (2024) evaluate language models on causal discovery benchmarks and test whether their answers reflect memorisation of the benchmark. Darvariu et al. (2024) elicit structural priors that a data-driven algorithm then weighs, the setting closest to our prior policy; Ban et al. (2025) balance fit to the data against model-derived priors, and Kampani et al. (2024) use a model's graph to initialise a differentiable objective. ReActBN (Zhang et al., 2026) is the closest design to our ordering role: at each step a model sees the ten best-scoring neighbouring graphs with their scores, may choose any of them, and a tabu list prevents revisits; its ablations, run on the data-free variant, vary whether names or descriptions are given, both of which carry meaning. Our model sees only accepted edits and no scores, cannot leave the veto, and is compared against a truth-reordering ceiling; pseudonyms are the presentation that removes meaning, and descriptions the one that keeps it without the benchmark's spelling.

Contamination and benchmarks. Babakov et al. (2024) exclude alarm and insurance from language-model evaluation because models recall them, reporting GPT-4 recall of 0.86 on alarm and 0.52 on insurance. We keep both and control for recall on the same cases, with pseudonyms throughout and descriptions on child. Gaps between small synthetic evaluations and larger or real settings are documented (Zhou et al., 2024; Chevalley et al., 2025); we report where our result weakens with scale.

9 Conclusion

A greedy score-based learner is decided by how its score orders the candidate edits. On the smaller networks the edit it takes beats most accepted alternatives, yet it is the best of them at 44.9% of steps by oriented-edge F1 and 36.5% by CPDAG SHD, and it is weakest against its closest competitors. Half of what a better order recovers in oriented-edge F1 lies in Markov-equivalent ties that an unreported enumeration order breaks, with a sign that changes between networks, and much of the rest in a sparsity preference that the complexity term produces. On child and insurance, one frontier model reading what the variables mean recovers most of it. Three practices follow: state the tie-break rule and report results over random tie-breaks, report an equivalence-class metric beside oriented-edge ones, and report accuracy over several scales of the complexity term. Appendix E states the limits of these claims.

AI USE STATEMENT

Language models are an object of study here: the ordering role of Section 7 is filled by the seventeen models listed in Appendix C, and every one of their answers is released as a per-call trace. Separately, we used a generative AI coding and research assistant for implementation (analysis scripts, experiment runners and the audit gates), data analysis, drafting and revising text, and producing figures, and as an aid in designing controls and interpreting results. We did not use generative AI to generate data: every sample is drawn from the BNLEARN networks and released as a file. The paper makes no theoretical claims or proofs. Every number in the text is recomputed from the released result files by one script and checked against the manuscript, every model arm passes a stated audit gate before it is read, and AI-written code was reviewed and tested against the shipped scorer, which the replay reproduces on 2,400 of 2,400 recorded score deltas. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The supplementary material contains the shipped search and scorer, the frozen data files (the generator does not reproduce them under current library versions; Appendix E), every result file the paper reads, the per-call trace of every model run, and the scripts that recompute every reported number and interval from them with the Python standard library, except the CPDAG conversions, which use causal-learn. Appendix A lists the cases and parameters, Appendix B defines the pooling and the intervals, and Appendix C gives the prompt, the decoding, the parser and the audit gate.

REFERENCES

Bryan Andrews, Joseph Ramsey, Ruben Sanchez-Romero, Jazmin Camchong, and Erich Kummerfeld. Fast scalable and accurate discovery of DAGs using the best order score search and grow-shrink trees. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, pp. 63945–63956, 2023.

Nikolay Babakov, Ehud Reiter, and Alberto Bugarín. Scalability of Bayesian network structure elicitation with large language models: a novel methodology and comparative analysis. *arXiv preprint arXiv:2407.09311*, 2024.

Taiyu Ban, Lyuzhou Chen, Derui Lyu, Xiangyu Wang, Qinrui Zhu, Qiang Tu, and Huanhuan Chen. Integrating large language model for improved causal discovery. *IEEE Transactions on Artificial Intelligence*, 6:3030–3042, 2025.

Ritwick Banerjee, Bryan Andrews, and Erich Kummerfeld. An extensive simulation study evaluating the interaction of resampling techniques across multiple causal discovery contexts. *arXiv preprint arXiv:2503.15436*, 2025.

Mathieu Chevalley, Yusuf H. Roohani, Arash Mehrjou, Jure Leskovec, and Patrick Schwab. A large-scale benchmark for network inference from single-cell perturbation data. *Communications Biology*, 8:412, 2025.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

Diego Colombo and Marloes H. Maathuis. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 15(116):3921–3962, 2014.

Victor-Alexandru Darvariu, Stephen Hailes, and Mirco Musolesi. Large language models are effective priors for causal graph discovery. *arXiv preprint arXiv:2405.13551*, 2024.

Adiba Ejaz and Elias Bareinboim. Less greedy equivalence search. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, 2025.

Shiv Kampani, David Hidary, Constantijn van der Poel, Martin Ganahl, and Brenda Miao. LLM-initialized differentiable causal discovery. *arXiv preprint arXiv:2410.21141*, 2024.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*, 2024.

Wai-Yin Lam, Bryan Andrews, and Joseph Ramsey. Greedy relaxations of the sparsest permutation algorithm. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 180 of *PMLR*, pp. 1052–1062, 2022.

Lars Lorch, Scott Sussex, Jonas Rothfuss, Andreas Krause, and Bernhard Schölkopf. Amortized inference for causal structure learning. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, pp. 13104–13118, 2022.

Achille Nazaret and David Blei. Extremely greedy equivalence search. In *Proceedings of the 40th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 244 of *PMLR*, pp. 2716–2745, 2024.

Joseph D. Ramsey. Scaling up greedy causal search for continuous variables. *arXiv preprint arXiv:1507.07749*, 2015.

Alexander G. Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated DAG! causal discovery benchmarks may be easy to game. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, 2021.

Tomi Silander, Petri Kontkanen, and Petri Myllymäki. On sensitivity of the MAP Bayesian network structure to the equivalent sample size parameter. In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 360–367, 2007.

Yinghuan Zhang, Yufei Zhang, Parisa Kordjamshidi, and Zijun Cui. Bayesian network structure discovery using large language models. *Transactions on Machine Learning Research*, 2026.

Wei Zhou, Hong Huang, Guowen Zhang, Ruize Shi, Kehan Yin, Yuanyuan Lin, and Bang Liu. OCDB: Revisiting causal discovery with a comprehensive benchmark and evaluation framework. *arXiv preprint arXiv:2406.04598*, 2024.

A SETUP

Table 4: **Networks and the cases each measurement uses.** A case is one network at one sample size with one data seed; every measurement uses seeds 42, 43 and 44. "Forward" is the share of true edges that point from an earlier-listed variable to a later one in the network file.

network	variables	edges	forward	sample sizes
child	20	25	56%	250, 500, 1,000, 2,000, 5,000
insurance	27	52	65%	250, 500, 1,000, 2,000, 5,000
alarm	37	46	63%	250, 1,000, 5,000
hepar2	70	123	100%	250, 1,000, 5,000
win95pts	76	112	100%	250, 1,000, 5,000

measurement	networks and sample sizes	cases
finished-graph pairs, γ (App. D, Sections 5, 6)	smaller; 250, 1,000, 5,000	27
the same	hepar2, win95pts; 250, 1,000, 5,000	18
fidelity at the point of choice (Section 3)	smaller; 250, 1,000, 5,000	27
the same	hepar2, win95pts; 250, 1,000	12
decomposition (Section 4)	smaller; every size above	39
the same	hepar2, win95pts; 250, 1,000	12
γ in the search (Section 6)	smaller; 250, 1,000, 5,000	27 per γ
language models (Section 7)	child, insurance; 250, 1,000, 5,000	18
alarm	alarm; 250, 1,000, 5,000	9

Table 5: **The decomposition from four starting graphs** (Section 4). Shares of the distance from uniform to the single-edit ceiling as ratios of sums; the gap carries a 95% case-bootstrap interval. Case counts differ because not every algorithm produced a graph on every case. Bold marks a start from which greedy takes less than half of the reachable ordering.

start	cases	greedy	reranked	gap	greedy/reranked
the pooled graph the score prefers	30	14.2	34.5	20.4 [15.6, 25.9]	**0.41**
GES under BDeu	17	30.4	47.0	16.6 [12.9, 20.3]	0.65
NOTEARS, $\lambda = 0.1$	18	23.2	48.8	25.6 [18.2, 33.3]	**0.48**
PC, χ^2, $\alpha = 0.05$	16	16.6	43.7	27.1 [21.0, 36.0]	**0.38**

Search. Moves add, delete or reverse one arrow; an edit is legal when the result is acyclic and no variable has more than three parents. The score is BIC plus 2 per edge on the discrete data, and an edit is accepted when it improves this penalised score. The budget is $3|V|$ steps, and a walk also stops when no legal edit is accepted. All policies of Section 4 start from the same graph, the pooled graph the score prefers unless a start is named, and uniform and the coin-flip tie-break are averaged over ten draws with seeds fixed per case.

Candidate graphs. PC (stable) with the χ^2 test at $\alpha \in \{0.01, 0.05, 0.10, 0.20\}$ and the G^2 test at $\alpha \in \{0.01, 0.05\}$, GES under BDeu and BIC, and NOTEARS at $\lambda \in \{0.01, 0.05, 0.1, 0.3\}$ with 100 iterations. Comparisons with equal oriented-edge F1 are skipped, since neither graph is closer to the truth.

Fidelity at the point of choice. Section 3 replays the shipped greedy walk on 39 cases, the 27 of the smaller networks at 250, 1,000 and 5,000 rows and the 12 of hepar2 and win95pts at 250 and 1,000, with the same start, legal moves, scorer, enumeration tie-break and hash seed as every other run; every replay ends at the final F1 of the frozen run. At each step every accepted edit is scored by oriented-edge F1, skeleton F1 and CPDAG SHD, with CPDAGs computed by the `dag2cpdag` routine of causal-learn. A pair is the taken edit against one other accepted edit; pairs within 10^{-6} in score, or tied on the metric, are dropped, and rates are ratios of sums over cases with the case bootstrap of Appendix B. Shuffling the scores within each step, which removes any information they carry, gives 53.1% by oriented-edge F1 and 51.1% by CPDAG SHD on the smaller networks, and 59.9% and 52.6% on hepar2 and win95pts.

B Pooling and intervals

Every share is a ratio of sums over cases, $\sum_c (a_c - u_c) / \sum_c (h_c - u_c)$, where a_c is the oriented-edge F1 a policy reaches on case c, u_c the uniform floor and h_c the ceiling named in the text. We do not average per-case ratios. Where $h_c - u_c$ is small, one edge moves that ratio by a large amount: on insurance at 5,000 rows two cases have a reranking headroom of 0.011 and 0.013 oriented-edge F1, so one flipped edge moves their share by about 100 points, and in a mean of ratios those two cases alone set the sign of four models' results. The ratio of sums weights each case by the headroom it had and needs no threshold. Under a mean of ratios the decomposition of Figure 3a reads 7.6%, 24.1% and 33.0% where the ratio of sums reads 10.1%, 21.4% and 32.6%: the order of the three policies and the reranking ceiling survive the choice, the score's own share moves by a quarter of itself, and the model tables of Section 7 move more. The choice is therefore fixed across the paper on the grounds above.

Intervals come from a bootstrap with 10,000 replicates that resamples whole cases, since the steps of one trajectory are not independent, with a fixed seed per interval so that a quantity has the same interval wherever it appears. Every contrast is computed inside a replicate, so its interval is on the difference itself. The three per-network tie-break contrasts are also read at the Bonferroni level of 98.3%, where each still excludes zero.

C MODEL PROTOCOL

Prompt. At each step the accepted edits are listed in an order shuffled with a seed fixed per case, so the presented position carries no information about the score or the enumeration. The model receives one user message and no system message. The excerpt below is the real prompt for step 0 of child at 1,000 rows, seed 42, the step of Figure 1; the full prompt lists all 49 accepted edits.

```
A search over directed acyclic graphs of these variables is choosing its
next single change. Every option below is already allowed and already
improves the statistical fit; the only question is which one is most
likely to be a true causal relation among these variables.

1. add an arrow GruntingReport -> Sick
2. add an arrow LungParench -> LungFlow
3. add an arrow LungParench -> Sick
4. add an arrow Grunting -> CO2Report
5. add an arrow Disease -> Sick
6. add an arrow Age -> LungFlow
   [ ... options 7 to 48 ... ]
49. add an arrow LVHreport -> Disease

Answer with the number of the single best option and nothing else.
```

The three presentations change only how a variable is written. The first three options of the same prompt, in the same order:

	names	descriptions	pseudonyms
1	GruntingReport → Sick	reported grunting observation → overall illness severity	V8 → V19
2	LungParench → LungFlow	lung parenchyma condition → pulmonary blood flow	V5 → V17
3	LungParench → Sick	lung parenchyma condition → overall illness severity	V5 → V19

The descriptions predate every experiment in this paper and are fixed; they are released with the code. Pseudonyms are assigned by a permutation seeded from the variable set, so they carry neither the name nor the file order.

Decoding and parsing. Temperature 0 with a fixed seed. The first request allows 24 output tokens with the model's reasoning mode disabled where the chat template supports it; an answer cut off at that length, or a template that rejects the flag, is asked again with room to reason. The answer is read from its final verdict: an explicit "answer: n", else a single number on the last non-empty line, else a single in-range number anywhere in the text. An answer naming two or more options without a verdict is refused. A refused answer receives one follow-up in the same conversation,

```
You did not give a number. Reply with the number of the single best
option from the list above, and nothing else.
```

and a step whose answer still does not parse falls back to greedy's choice and is counted against the model.

Audit gate. Before any share is read, an arm must parse at least 90% of its answers, no presented position may take more than half of its choices among five or more candidates, and its trace must cover every case with as many calls as its counters record. A server failure stops the run after five consecutive errors, so an arm cannot silently become greedy. Table 6 gives every model's audit.

Three kinds of absence. Two models fail the gate for reasons that belong to the model: llama3.2-1b answers presented position 3 on 65% of choices among five or more candidates, and gpt-oss-20b enters degenerate repetition and reaches no verdict on 31% of its steps. They are reported, and their shares are not read. Two models, gemma-4-31b and qwen3-next-80b, do not start on our serving

Table 6: **Every model in the ordering role, real names, 18 cases.** Calls are model queries; "top position" is the share of choices on the most chosen presented position among five or more candidates; "= score" is the share of choices that match greedy's, against the chance rate for the number of candidates. Checkpoint names are as stored; the frontier model is reached through an API. Grey type marks the two models that fail the gate.

model	checkpoint	calls	parsed	top position	= score	chance	gate
gpt-5.6-luna	API	204	100%	11%	18%	12%	pass
deepseek-v2-lite	DeepSeek-V2-Lite-Chat	293	100%	12%	13%	13%	pass
glm-4.7-flash	GLM-4.7-Flash	320	100%	18%	12%	13%	pass
gpt-oss-20b	gpt-oss-20b	219	69%	20%	19%	16%	fail
llama3.1-8b	Meta-Llama-3.1-8B-Instruct	308	100%	13%	12%	14%	pass
llama3.2-1b	Llama-3.2-1B-Instruct	317	100%	65%	11%	13%	fail
olmo2-7b	OLMo-2-1124-7B-Instruct	287	100%	24%	14%	13%	pass
qwen2.5-7b	Qwen2.5-7B-Instruct	335	100%	27%	13%	13%	pass
qwen3-1.7b	Qwen3-1.7B	309	96%	12%	19%	13%	pass
qwen3-14b	Qwen3-14B	272	100%	19%	15%	13%	pass
qwen3-30b-a3b	Qwen3-30B-A3B	291	100%	14%	15%	13%	pass
qwen3-32b-fp8	Qwen3-32B-FP8	267	100%	11%	16%	13%	pass
qwen3-4b	Qwen3-4B	279	100%	14%	14%	14%	pass
qwen3-4b-2507	Qwen3-4B-Instruct-2507	276	100%	18%	15%	12%	pass
qwen3-8b	Qwen3-8B	292	100%	11%	13%	13%	pass
qwen3-coder-30b	Qwen3-Coder-30B-A3B-Instruct	305	100%	19%	11%	13%	pass
qwen3.8-27b	Qwen3.8-27B	242	100%	35%	15%	13%	pass

stack; the errors are in the supplementary material. Three larger entries of the roster, deepseek-v4-flash, llama3.1-70b and llama-4-scout, were not run.

D EXACT VALUES

Every table in this appendix is written by a script from the released result files; none is typed by hand. Each subsection states which part of the body it supports.

FIDELITY AT THE POINT OF CHOICE

The step-level readings of Section 3; Appendix A gives the protocol.

Table 7: **How the edit a search takes compares with the other edits the score accepted, the values behind Figure 2.** Every greedy step on 27 cases of the smaller networks and 12 of hepar2 and win95pts. A pair is the taken edit against one other accepted edit; pairs the score or the metric cannot order are dropped. The closest quarter and the farther half split the alternatives by their score gap to the taken edit, at its first quartile and its median. Percentages with 95% case-bootstrap intervals; shuffling the scores within each step gives 51% to 60% on the two every-alternative rows.

	child, insurance, alarm	hepar2, win95pts
steps	285	243
taken edit is the best accepted edit, share of steps		
by oriented-edge F1	44.9 [36.6, 53.7]	53.6 [36.1, 67.6]
by CPDAG SHD	36.5 [26.7, 48.5]	21.7 [17.0, 29.3]
taken edit is truer than an accepted alternative		
every alternative, oriented-edge F1	75.1 [66.9, 83.5]	87.3 [75.1, 91.8]
within 0.05 oriented-edge F1	71.2 [64.5, 79.5]	87.3 [75.1, 91.8]
closest quarter in score	57.7 [50.5, 67.2]	76.4 [61.4, 82.6]
farther half in score	81.8 [69.4, 90.2]	90.2 [75.5, 96.1]
every alternative, CPDAG SHD	79.3 [74.4, 84.2]	76.2 [67.0, 79.9]

FIDELITY ON PAIRS OF FINISHED GRAPHS

The finished-graph comparisons cited in Sections 3 and 5, in full, with the sparsity preference judged by a second truth metric.

The finished graphs are those a practitioner might have produced on the same rows: PC with the χ^2 test at four significance levels and the G^2 test at two, GES under BIC and BDeu, and NOTEARS at four sparsity penalties. This gives twelve graphs per case, 324 over 27 cases on the smaller networks and 216 over 18 on hepar2 and win95pts. Pairs of Markov-equivalent graphs are dropped, since no likelihood-based score can separate them. None of the 1,742 comparisons on the smaller networks is between equivalent graphs, because twelve different settings produce graphs that differ in many edges at once. The equivalence test itself classifies six hand-built pairs correctly, including a collider against a fork on the same skeleton.

For each pair of candidate graphs on the same rows we ask whether the score prefers the one closer to the truth, and group pairs by their difference in oriented-edge F1 (Figure 4a; Table 8). BIC agrees with the truth on 82% of pairs more than 0.15 apart, 60% of pairs between 0.05 and 0.15, and 54% of pairs within 0.05, the near band of Section 3. The near-band interval runs from 47 to 62 and includes chance. BDeu is higher in every band and has the same shape, falling from 87% to 63%; the edge penalty changes no band by more than one point.

On these pairs the score difference between two candidates, the search's only runtime signal, carries little information. Agreement with the truth is 57% in the bottom decile of that difference and 64% in the top, and normalising the difference by the sample size or by the BIC penalty unit does not widen the spread. Cut by F1 distance instead, which the search cannot observe, agreement runs from 55% in the bottom decile to 98% in the top.

Larger networks put more comparisons in the near band. On hepar2 and win95pts agreement rises in every band, to 66%, 80% and 99%, and the near band grows from 22% of comparisons on the smaller networks to 47%.

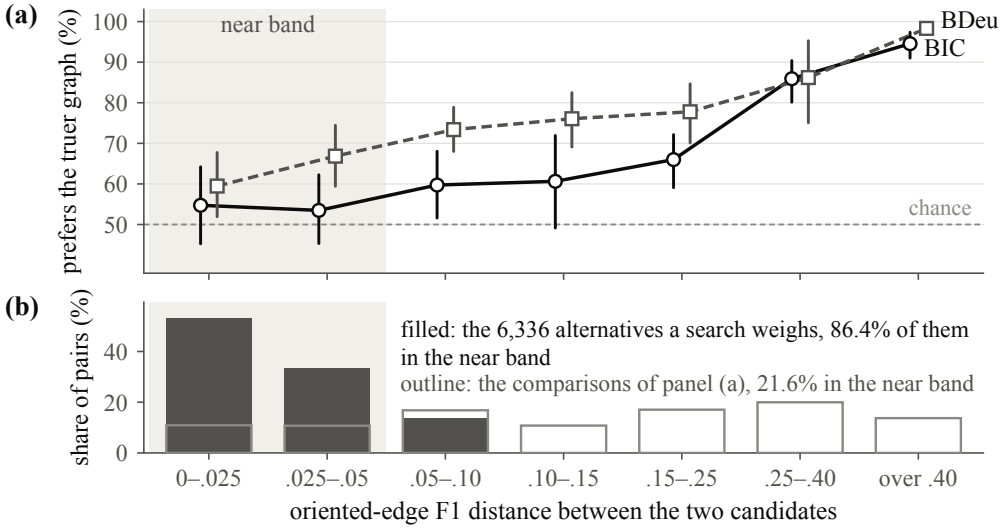

Figure 4: **Fidelity on pairs of finished graphs, and where a search's alternatives fall. (a)** How often BIC and BDeu prefer the truer of two graphs that PC, GES and NOTEARS produce on the same rows, judged by oriented-edge F1 and grouped by F1 distance, over the 1,742 comparisons of the smaller networks; bars are 95% intervals from a bootstrap that resamples cases, and chance is 50%. **(b)** The 6,336 alternatives a greedy search weighs against the edit it takes (filled), and the comparisons of panel (a) (outline), on the same axis. Measured on those alternatives, the taken edit is truer 71.2% of the time within the near band (Table 7); on the pairs of panel (a), the near-band level depends on the truth metric (Table 9).

Table 8: **How often the score prefers the candidate closer to the true network, by how far apart the two candidates are.** Comparisons between graphs from PC, GES and NOTEARS at twelve settings on the same rows, grouped by their difference in oriented-edge F1; percentages with 95% intervals clustered on the case. A score with no information about the truth sits at 50%. The near band, under 0.05, is where a single-edit search makes 86% of its choices. The shaded column is the near band of Figure 4.

score	under 0.05	0.05 to 0.15	over 0.15
child, insurance and alarm: 1,742 comparisons, n = 377, 481, 884 by band			
BIC	54 [47, 62]	60 [52, 68]	82 [77, 85]
BIC, $\lambda = 2$ per edge	54 [47, 62]	60 [52, 68]	81 [77, 85]
BDeu	63 [58, 69]	74 [70, 79]	87 [81, 92]
hepar2 and win95pts: 378 comparisons, n = 177, 116, 85 by band			
BIC	66 [57, 75]	80 [68, 90]	99 [96, 100]
BIC, $\lambda = 2$ per edge	66 [55, 75]	78 [63, 90]	99 [96, 100]
BDeu	64 [57, 71]	85 [75, 95]	98 [93, 100]

Other truth metrics. Table 9 judges the same pairs by CPDAG SHD and skeleton F1: by CPDAG SHD the near-band agreement is 72.5% for BIC and 50.5% for BDeu. The sparsity preference of Section 5 also depends on the metric: by CPDAG SHD the sparser graph is the truer one in 55.5% of the 1,586 pairs that differ in edge count, $[50.7, 60.3]$, close to the 57.8% in which BIC takes the sparser graph, and taking the sparser, worse graph accounts for 56.1% of its errors, $[47.2, 64.6]$; it makes 296 errors here, against 519 under oriented-edge F1.

Table 9: **The finished-graph pairs of Section 3 under three truth metrics.** Agreement of each score with the truer graph, by oriented-edge F1 distance band, when the truer graph is the one with the higher oriented-edge F1, the smaller CPDAG SHD, or the higher skeleton F1; the oriented rows are those of Table 8. CPDAG SHD drops the pairs involving the 3 of the 324 pool graphs that contain a cycle or a two-way edge. Percentages with 95% case-bootstrap intervals.

score	truth judged by	under 0.05	0.05 to 0.15	over 0.15
BIC	oriented F1	54.1 [46.7, 62.3]	60.1 [52.2, 68.3]	81.6 [77.2, 85.4]
	CPDAG SHD	72.5 [63.0, 81.1]	75.8 [70.2, 81.3]	87.1 [82.2, 91.4]
	skeleton F1	47.9 [40.8, 55.9]	53.5 [45.0, 63.1]	78.8 [74.2, 82.9]
BIC, $\lambda = 2$ per edge	oriented F1	54.1 [46.7, 62.3]	60.1 [52.2, 68.3]	81.4 [77.0, 85.3]
	CPDAG SHD	72.5 [63.0, 81.1]	75.8 [70.2, 81.3]	87.0 [82.0, 91.3]
	skeleton F1	47.9 [40.8, 55.9]	53.5 [45.0, 63.1]	78.7 [74.0, 82.9]
BDeu	oriented F1	63.1 [58.0, 68.7]	74.4 [69.8, 78.8]	86.7 [81.0, 91.7]
	CPDAG SHD	50.5 [42.0, 60.1]	57.6 [51.1, 64.3]	79.9 [74.5, 84.8]
	skeleton F1	62.5 [55.7, 69.7]	75.2 [70.3, 80.3]	87.1 [81.8, 91.9]

THE DECOMPOSITION, BY NETWORK AND BY CASE

The decomposition of Section 4 in full: the values plotted in Figure 3, the shares by network, and every case.

Table 10: **The values plotted in Figure 3.** Ratios of sums with 95% case-bootstrap intervals; contrasts are computed within each replicate.

	share	95% interval
(a) 39 cases, against the single-edit ceiling		
greedy: the score's own order	10.1	[3.9, 16.7]
ties settled by the truth	21.4	[14.9, 28.9]
all accepted edits reordered by the truth	32.6	[26.1, 39.5]
(b) gpt-5.6-luna, 9 child cases, against the reranking ceiling		
real names	64.8	[49.3, 82.1]
plain descriptions	65.6	[46.8, 85.9]
pseudonyms	−22.6	[−36.1, −11.5]
pseudonyms, second map	15.7	[−14.1, 38.2]
greedy, the score's own order	48.3	[26.7, 64.3]
plain descriptions − real names	0.8	[−12.6, 15.7]
plain descriptions − pseudonyms	88.3	[64.6, 115.4]

Table 11: **The decomposition and the tie-break, by network.** Shares of the distance from uniform to the single-edit ceiling for greedy, greedy with its ties settled by the truth, and reranked; the gap is reranked minus greedy. The last column replaces the enumeration tie-break with a coin flip averaged over ten draws and is measured against the reranking ceiling. Intervals are 95% case-bootstrap intervals. The grey row pools the three networks.

network	cases	greedy	ties resolved	reranked	gap	reranked wins	enumeration − coin
child	15	24.7	45.8	56.7	32.0 [24.3, 40.7]	15/15	+32.2 [+14.4, +46.1]
insurance	15	5.9	7.8	17.2	11.2 [8.2, 14.1]	14/15	+35.9 [+25.2, +44.2]
alarm	9	−5.0	10.1	25.6	30.5 [25.1, 36.2]	9/9	−21.8 [−51.2, −9.0]
smaller, pooled	39	10.1	21.4	32.6	22.5 [18.4, 27.0]	38/39	+20.1 [+5.3, +32.1]

Table 12: **Every decomposition case.** Final oriented-edge F1 of each policy from the same start, for the 39 cases of the smaller networks and the 12 of hepar2 and win95pts (Section 4). uniform and the coin tie-break are means over ten draws; a dash is a policy not run on that case. Bold marks the 13 cases in which greedy ends below uniform; Figure 5 draws them.

network	rows	seed	uniform	greedy	ties resolved	coin tie-break	prior	reranked	oracle
child	250	42	0.516	0.605	0.698	0.569	0.727	0.773	1.000
child	250	43	0.663	**0.651**	0.744	0.628	0.682	0.773	1.000
child	250	44	0.543	0.727	0.727	0.543	0.711	0.844	1.000
child	500	42	0.667	0.756	0.756	–	0.756	0.844	1.000
child	500	43	0.703	**0.591**	0.773	–	0.773	0.773	1.000
child	500	44	0.683	0.711	0.773	–	0.756	0.844	1.000
child	1,000	42	0.686	0.750	0.851	0.750	0.766	0.851	1.000
child	1,000	43	0.779	0.800	0.800	0.783	0.800	0.844	1.000
child	1,000	44	0.654	0.917	0.958	0.763	0.776	0.958	1.000
child	2,000	42	0.789	0.816	0.939	–	0.816	0.939	1.000
child	2,000	43	0.459	0.720	0.760	–	0.694	0.760	1.000
child	2,000	44	0.800	0.857	0.939	–	0.939	0.939	1.000
child	5,000	42	0.553	0.731	0.769	0.638	0.680	0.863	1.000
child	5,000	43	0.903	**0.880**	0.960	0.904	0.880	0.960	1.000
child	5,000	44	0.814	0.880	0.960	0.810	0.920	0.960	1.000
insurance	250	42	0.455	0.506	0.506	0.430	0.625	0.617	1.000
insurance	250	43	0.438	0.453	0.474	0.422	0.571	0.571	1.000
insurance	250	44	0.484	0.579	0.579	0.505	0.649	0.649	1.000
insurance	500	42	0.491	0.548	0.548	–	0.548	0.571	1.000
insurance	500	43	0.546	0.561	0.561	–	0.619	0.619	1.000
insurance	500	44	0.522	0.548	0.548	–	0.595	0.595	1.000
insurance	1,000	42	0.607	0.621	0.621	0.602	0.659	0.667	1.000
insurance	1,000	43	0.612	0.621	0.621	0.611	0.659	0.659	1.000
insurance	1,000	44	0.594	0.621	0.621	0.600	0.659	0.659	1.000
insurance	2,000	42	0.701	0.733	0.733	–	0.711	0.733	1.000
insurance	2,000	43	0.616	0.621	0.621	–	0.659	0.659	1.000
insurance	2,000	44	0.699	0.703	0.733	–	0.667	0.733	1.000
insurance	5,000	42	0.632	0.667	0.688	0.660	0.688	0.688	1.000
insurance	5,000	43	0.755	**0.745**	0.766	0.736	0.766	0.766	1.000
insurance	5,000	44	0.740	**0.731**	0.753	0.720	0.753	0.753	1.000
alarm	250	42	0.380	**0.350**	0.400	0.367	0.380	0.506	0.989
alarm	250	43	0.401	**0.390**	0.416	0.384	0.462	0.610	0.989
alarm	250	44	0.399	0.405	0.456	0.425	0.400	0.568	0.989
alarm	1,000	42	0.757	**0.750**	0.795	0.773	0.764	0.787	0.989
alarm	1,000	43	0.655	**0.652**	0.727	0.672	0.644	0.764	0.989
alarm	1,000	44	0.738	**0.705**	0.795	0.739	0.750	0.818	0.989
alarm	5,000	42	0.881	**0.857**	0.901	0.888	0.879	0.901	0.989
alarm	5,000	43	0.822	**0.800**	0.822	0.802	0.791	0.822	0.989
alarm	5,000	44	0.881	**0.857**	0.901	0.881	0.857	0.901	0.989
hepar2	250	42	0.239	0.368	–	–	0.368	0.380	0.944
hepar2	250	43	0.260	0.344	–	–	0.344	0.380	0.944
hepar2	250	44	0.226	0.358	–	–	0.358	0.368	0.944
hepar2	1,000	42	0.451	0.462	–	–	0.462	0.509	0.944
hepar2	1,000	43	0.418	0.465	–	–	0.453	0.488	0.944
hepar2	1,000	44	0.434	0.491	–	–	0.491	0.539	0.944
win95pts	250	42	0.373	0.429	–	–	0.429	0.404	0.591
win95pts	250	43	0.484	0.543	–	–	0.543	0.550	0.659
win95pts	250	44	0.444	0.461	–	–	0.461	0.486	0.600
win95pts	1,000	42	0.722	0.725	–	–	0.725	0.729	0.783
win95pts	1,000	43	0.622	0.646	–	–	0.646	0.628	0.762
win95pts	1,000	44	0.703	0.705	–	–	0.705	0.730	0.815

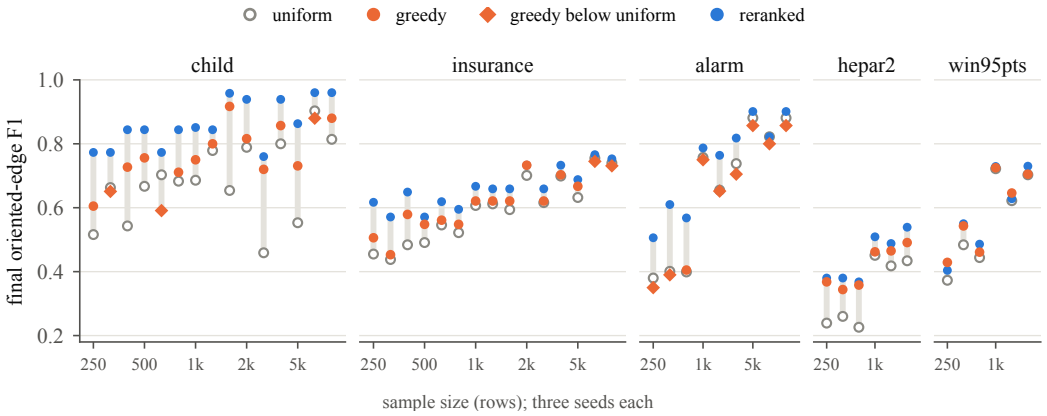

Figure 5: **Every decomposition case, drawn.** The values of Table 12: one column per case, three seeds at each sample size, with the grey bar spanning the three policies. reranked ends above greedy in 38 of the 39 cases of the smaller networks. greedy ends below uniform in 3 cases of child, 2 of insurance and 8 of the 9 of alarm, the network whose enumeration tie-break points the wrong way. On hepar2 and win95pts the three policies sit close together, and reranked, which chooses one step at a time, ends below greedy in two cases of win95pts.

LANGUAGE MODELS

The ordering role of Section 7 for every model, in both roles and under both presentations, and split by network.

Table 13: **Every model in both roles, under both presentations.** Share of the distance from a uniform draw among accepted edits to the reranking ceiling, pooled over the 18 cases of child and insurance; greedy reaches 43.5. "Every accepted edit" is the ordering role of Section 7; "exact ties only" leaves every other step to greedy. Contrasts carry 95% case-bootstrap intervals and are bold when the interval excludes zero. Grey type marks an arm that fails the audit gate, whose share is shown and not read: gpt-oss-20b parses 69% of its answers, llama3.2-1b answers position 3 on 65% of choices, and qwen3.8-27b under pseudonyms answers position 1 on 52%, which on a shuffled presentation is a uniform draw. A dash is an arm that was not run.

	every accepted edit			exact ties only		
model	names	pseud.	names − greedy	names	pseud.	names − greedy
deepseek-v2-lite	−18.2	–	**−61.6** [−78.5, −36.6]	–	–	–
glm-4.7-flash	0.4	–	**−43.1** [−63.0, −25.2]	–	–	–
gpt-5.6-luna	72.5	−8.0	**+29.1** [+9.9, +54.6]	58.8	−8.4	**+15.4** [+0.9, +34.9]
gpt-oss-20b	39.1†	–	–	–	–	–
llama3.1-8b	−12.6	−11.9	**−56.1** [−79.1, −27.7]	37.6	5.5	−5.8 [−19.5, +10.1]
llama3.2-1b	−2.1†	–	–	–	–	–
olmo2-7b	−2.5	–	**−46.0** [−69.2, −17.0]	17.5	17.5	−26.0 [−45.3, +0.6]
qwen2.5-7b	26.4	−12.2	−17.0 [−44.7, +15.0]	25.8	−14.5	−17.7 [−35.9, +1.0]
qwen3-1.7b	19.4	−5.7	−24.1 [−45.6, +1.4]	–	–	–
qwen3-14b	13.4	−17.1	**−30.1** [−56.5, −2.2]	29.3	−0.9	−14.2 [−30.7, +4.8]
qwen3-30b-a3b	12.5	−13.3	**−30.9** [−49.9, −6.6]	–	–	–
qwen3-32b-fp8	40.6	−19.3	−2.9 [−25.6, +25.2]	22.4	17.5	**−21.0** [−37.3, −1.2]
qwen3-4b	13.4	−10.0	**−30.1** [−48.2, −5.4]	–	–	–
qwen3-4b-2507	31.2	5.3	−12.3 [−33.4, +15.6]	36.8	0.4	−6.6 [−21.3, +10.4]
qwen3-8b	10.3	13.0	**−33.1** [−57.6, −6.5]	12.8	25.9	**−30.7** [−49.1, −11.6]
qwen3-coder-30b	26.3	–	−17.2 [−44.1, +16.4]	33.2	–	−10.2 [−24.8, +10.0]
qwen3.8-27b	46.5	−5.9†	+3.0 [−15.0, +23.5]	45.1	5.8	+1.7 [−10.7, +18.7]

Table 14: **Every model that passes the audit gate, by network.** The ordering role with real names, as a share of the distance from a uniform draw among accepted edits to the reranking ceiling; "both" is the 18-case pool of Table 2. Models differ more between the two networks than their pooled shares suggest. The grey row is greedy; bold marks a share above it in the same column. Figure 6 draws them.

model	child	insurance	both
greedy	48.3	31.7	43.5
gpt-5.6-luna, first run	**64.8**	**91.0**	**72.5**
qwen3.8-27b	42.7	**55.7**	**46.5**
qwen3-32b-fp8	35.1	**53.8**	40.6
qwen3-4b-2507	28.1	**38.6**	31.2
qwen2.5-7b	15.9	**51.8**	26.4
qwen3-coder-30b	23.8	**32.3**	26.3
qwen3-1.7b	22.9	10.8	19.4
qwen3-4b	13.7	12.7	13.4
qwen3-14b	6.1	30.9	13.4
qwen3-30b-a3b	−0.7	**44.4**	12.5
qwen3-8b	12.0	6.3	10.3
glm-4.7-flash	−2.7	7.8	0.4
olmo2-7b	3.7	−17.5	−2.5
llama3.1-8b	−23.4	13.2	−12.6
deepseek-v2-lite	−24.3	−3.3	−18.2

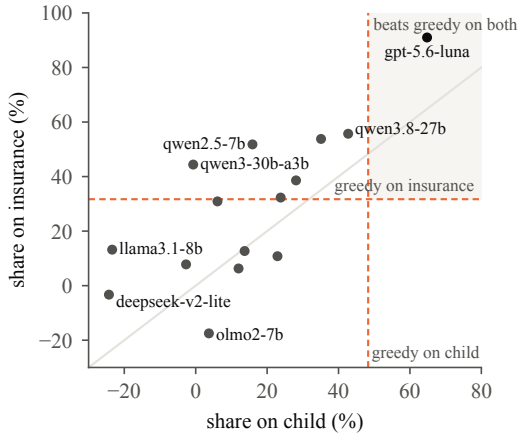

Figure 6: **Every gated model on the two networks, drawn.** The values of Table 14; the dashed lines are greedy on each network and the grey line marks equal shares. Only gpt-5.6-luna lies above greedy on child, where seven models, six of them open-weight, lie above it on insurance. Eleven of the fifteen models do better on insurance than on child, so a model's pooled share depends on which network carries it.

THE COMPLEXITY TERM IN THE SEARCH

The search of Section 6 with its complexity term scaled, cell by cell.

Table 15: **The complexity term in the search** (Section 6). Final oriented-edge F1 of the shipped search with its complexity term multiplied by γ, as the mean over three seeds, with the final edge count and the number of steps in parentheses. The shaded column is the shipped setting; bold is the best setting of each row.

network, rows	$\gamma = 0.25$	0.5	0.75	1 (shipped)
alarm 250	**0.718** (32, 22)	0.558 (34, 22)	0.480 (35, 22)	0.410 (34, 22)
alarm 1,000	**0.743** (40, 17)	0.722 (41, 17)	0.724 (43, 17)	0.724 (43, 17)
alarm 5,000	**0.904** (45, 11)	**0.904** (45, 11)	**0.904** (45, 11)	0.890 (45, 10)
child 250	0.564 (16, 17)	**0.726** (20, 17)	0.692 (19, 17)	0.692 (19, 17)
child 1,000	0.804 (21, 13)	**0.836** (22, 14)	**0.836** (22, 14)	**0.836** (22, 14)
child 5,000	**0.867** (25, 14)	**0.867** (25, 14)	0.809 (25, 16)	0.809 (25, 16)
insurance 250	**0.478** (25, 22)	0.451 (26, 22)	0.451 (26, 22)	0.451 (26, 22)
insurance 1,000	**0.640** (33, 13)	**0.640** (33, 13)	0.618 (35, 13)	0.618 (35, 13)
insurance 5,000	0.719 (41, 11)	0.719 (41, 11)	**0.728** (41, 11)	**0.728** (41, 11)

E LIMITATIONS

Scale. The ordering loss is 22.5 points on the smaller networks and 4.3 on hepar2 and win95pts. Section 4 attributes the difference to how much oriented-edge F1 one arrow is worth, and the tie measurements support that, but only two networks have more than 56 variables. We state the headroom for the regime the standard benchmarks occupy and do not extend it to larger graphs.

Move set. The decomposition concerns searches that move one edge at a time. GRaSP and BOSS move through permutations instead (Section 8), and nothing here bounds what they lose.

The complexity term. The best factor differs across networks and sample sizes, and choosing it requires the graph the search is trying to recover. Section 6 establishes that the term produces the sparsity preference, and offers no rule for setting it.

Truth metric. The decomposition and every share are measured in oriented-edge F1, which scores orientations the data cannot identify. Fidelity at the point of choice (Section 3) is reported under both metrics; the finished-graph near band and the sparsity preference change with the metric (Appendix D). Repeating the decomposition with an equivalence-class ceiling is the next measurement.

Data. Five discrete networks with 20 to 76 variables, observational samples only, and no real-world data. The samples on disk are frozen: the current environment, with pgmpy 0.1.24, reproduces none of the rows of the tables generated with the original one, so the release is the data files themselves.

Models. Seventeen models are measured in the ordering role, sixteen open-weight models served locally and one frontier model reached through an API. The frontier model is run three times with real names in both roles, where its shares were 72.5%, 67.3% and 88.7% in the full role and 58.8%, 56.6% and 56.4% at the ties; the pseudonym, description and alarm conditions are run once, so their intervals carry case-level uncertainty only. Of child and insurance, only child has descriptions, and the description condition was not run on alarm. Pseudonym results use one map from names to labels, except on child, where a second map moved the frontier model's share by 38.3 points, $[12.2, 62.6]$; the level a model reaches under pseudonyms depends on the draw, and the comparison with real names, which is what we read, holds under both maps. Descriptions separate a variable's meaning from its benchmark spelling, and cannot separate reasoning from recall of a network's structure through what its variables mean; a network whose variables and structure no model has seen would be needed for that.