COSMIC: SCALABLE CAUSAL DISCOVERY VIA NYSTRÖM MACRO-TIER PEELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recovering causal graphs from purely observational data is essential when experimental interventions are constrained by ethical or logistical barriers. However, ordering-based approaches face severe computational bottlenecks due to repeated kernel computations that scale cubically with sample size and quadratically with graph dimension. We introduce COSMIC, a highly scalable causal discovery framework for nonlinear additive noise models. Our approach integrates a low-rank Nyström Stein approximation with adaptive macro-tier peeling to eliminate batches of candidate sinks simultaneously, drastically reducing both the size and frequency of kernel computations. To combine ordering evidence across scales and inform intra-tier evidence allocation, COSMIC utilizes a Tri-Band Harmonic Stein Ensemble and evaluates leave-one-macro-tier-out parent scores with local coupling confidence. We support this framework by formally showing that terminal-block removals preserve the underlying Additive Noise Model structure. Extensive benchmarking demonstrates that COSMIC scales topological ordering and candidate parent screening to $d = 500$ variables in minutes, a regime where existing non-parametric methods stall. In full end-to-end discovery (including non-parametric regression pruning), COSMIC completes graphs through $d = 100$ before downstream GAM pruning reaches an empirical timeout at $d = 200$. Across small synthetic and biological benchmarks ($d = 10$–20), COSMIC achieves competitive recovery accuracy while evaluating candidate edges orders of magnitude faster than continuous-DAG and non-parametric baselines.

1 INTRODUCTION

Causal discovery aims to retrieve directed acyclic graphs (DAGs) of cause-and-effect relationships from data (Pearl, 2009). Observational data are especially valuable when interventions are impractical, but generally leave some directions unresolved (Spirtes et al., 1993). Additive noise models (ANMs) can identify a DAG under restrictions on the mechanisms and independent noise distributions (Peters et al., 2014).

Ordering-based methods construct a topological ordering of the features, placing parents before children, then screen and prune the edges permitted by that order. SCORE and CaPS construct the order by repeatedly estimating derivatives of the score/log-density function and removing a candidate sink, a variable with no remaining children (Rolland et al., 2022; Xu et al., 2024). Their kernel estimators face two costs: dense matrix solves over n observations and repeated estimation over d variables. Parent scoring through individual variable omissions adds further evaluations.

COSMIC (Continuous Ordering via Scalable Macro-tiers for Identifiable Causality) compresses both axes of this computation. Its contributions are:

- A Nyström Stein estimator using $m \ll n$ landmark observations, reducing computational complexity from $\mathcal{O}(n^3)$ to $\mathcal{O}(nm^2 + m^3)$ per score estimation.
- A Tri-Band Harmonic Stein Ensemble combining separately fitted curvature estimates at three bandwidths and retaining their score information for parent-score allocation.
- Macro-tier peeling and grouped omission: removing batches of candidate sinks reduces repeated fits, while Leave-One-Macro-Tier-Out scoring with local coupling allocates aggregate

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.

omission evidence to candidate edges while reducing the number of estimator evaluations from variable-wise to tier-wise omissions.

We formalize the conditions under which terminal-block removals preserve the underlying ANM and evaluate the practical accuracy and runtime trade-offs of low-rank estimation. Downstream pruning uses generalized additive models (GAMs) with generalized cross-validation (GCV) smoothing. Crucially, we establish a clear operational distinction: COSMIC scales topological ordering and candidate screening to $d = 500+$ in minutes, while complete end-to-end recovery is bounded at $d \leq 100$ by the scaling limits of serial downstream regression pruning.

2 Background

2.1 Scores and additive noise models

For observations $X \in \mathbb{R}^{n \times d}$ with density p, the score is $s(x) = \nabla_x \log p(x)$ and its Jacobian is the log-density Hessian $H(x) = \nabla_x^2 \log p(x)$. These derivatives are with respect to observed variables, not model parameters. Differentiation removes any constant normalizer. Kernel Stein estimators use integration-by-parts identities to estimate these quantities through linear algebra (Li & Turner, 2018). A regularized solve involving the Gram matrix $K_{ab} = k(x_a, x_b)$ costs $\mathcal{O}(n^3)$ time and $\mathcal{O}(n^2)$ kernel storage in the dense case. An exact Gram matrix is the full empirical kernel matrix, not exact population derivatives.

An ANM has equations $X_j = f_j(X_{\mathrm{pa}(j)}) + \epsilon_j$ with mutually independent noises $\epsilon_j \sim \mathcal{N}(0, \sigma_j^2)$. However, derivative-based ordering criteria rely on variance regularity to rank variables. Under SCORE's nonlinear Gaussian assumptions, sinks have constant Hessian diagonal across observations (Rolland et al., 2022). Removing a true sink preserves the structural equations of the remaining variables. CaPS instead ranks expected diagonal curvature $\mathbb{E}[H_{jj}] = -1/\sigma_j^2 - \sum_{k \in \mathrm{ch}(j)} \mathbb{E}[(\partial_j f_k)^2]/\sigma_k^2$ under separation conditions (Xu et al., 2024). This ranking criterion relies on equal error variances ($\sigma_j^2 \equiv \sigma^2$): a true sink s has no children, yielding the least negative expected curvature $-1/\sigma^2$. When noise variances are heterogeneous, an upstream root with large variance can exhibit less negative curvature than a downstream sink, inverting topological peeling. In practice, empirical feature standardization (X_j/σ_{X_j}) does not equalize error variances because downstream variables accumulate ancestral variance ($\mathrm{Var}(X_j) = \mathrm{Signal}_j + \sigma_j^2$), altering empirical score scales without restoring the homoscedastic ANM condition. The controlled noise-variance experiments in Appendix A.4.4 illustrate the resulting ordering failures. Equal error variance is thus a sufficient condition for curvature-based sink identification rather than a general requirement for ANM identifiability (Peters et al., 2014).

2.2 Parent scores

CaPS measures changes in average diagonal curvature when a variable is omitted (Xu et al., 2024). With row j denoting a candidate parent and column i a child, the signed leave-one-out (LOO) statistic is $\mathcal{P}_{j \to i}^{\mathrm{LOO}} = \bar{H}_j^{(-i)} - \bar{H}_j$ ($j \neq i$). Here LOO removes a variable, not an observation; bars denote empirical averages and indices remain aligned after omission. For a true sink in a Gaussian ANM, the population change measures noise-scaled squared mechanism derivatives. For general omissions, the change instead measures information lost by marginalizing variables; Appendix A.5 derives this identity and explains why it need not isolate direct parent–child effects. Scores screen edges before regression-based pruning and are neither causal-effect estimates nor calibrated edge probabilities. The final order permits only edges from earlier to later variables.

3 Related Work

SCORE uses score matching to identify sinks in nonlinear Gaussian ANMs and subsequently prunes an order-compatible graph (Rolland et al., 2022). DAS extends score-based discovery by using second-order information to identify candidate parents and reduce the cost of pruning (Montagna et al., 2023). CaPS develops expected-Hessian ordering and omission-based parent scores for linear and nonlinear relations under its stated assumptions (Xu et al., 2024). COSMIC follows this order–screen–prune strategy, adopting CaPS's mean diagonal curvature for ordering and omission-induced

curvature changes for parent evidence. It introduces landmark approximation and grouped removals to reduce the cost of estimating these statistics; DAS motivates screening candidates before expensive regression pruning.

DiffAN learns score information with a diffusion model and uses it for topological ordering (Sanchez et al., 2023). It offers a neural route to reducing sample-related costs, whereas COSMIC retains kernel linear algebra and introduces landmarks and grouped removals. LNMIX uses score-Jacobian information for mixed linear and nonlinear additive mechanisms (Liu et al., 2024). Its mechanism assumptions and recovery procedure differ from COSMIC's; computational approximation does not by itself extend identifiability to every mixed model.

CAM combines additive-model order search with regression pruning (Bühlmann et al., 2014); COSMIC instead obtains its order from Stein curvature estimates. DAGMA learns weighted DAGs through continuous optimization with a log-determinant acyclicity characterization (Bello et al., 2022), rather than explicit sink peeling.

Nyström approximation supplies the low-rank kernel foundation (Williams & Seeger, 2000), while score matching estimates unnormalized densities through log-density derivatives (Hyvärinen, 2005). Landmark-based score estimation predates COSMIC: Sutherland et al. (2018) fit kernel exponential-family densities with statistical guarantees. COSMIC instead approximates Stein scores and derivative sums, combining them with batched causal ordering and grouped parent scoring. Earlier kernel and density-estimation guarantees do not directly cover these derivatives or causal orders; we provide conditional removal guarantees and separate ordering, screening, and recovery diagnostics.

4 The COSMIC Architecture

COSMIC estimates a causal DAG $G = (\mathcal{V}, \mathcal{E})$ from observations $X \in \mathbb{R}^{n \times d}$ in three phases, as seen in Figure 1. A landmark is an observation; a macro-tier is a batch of features/variables.

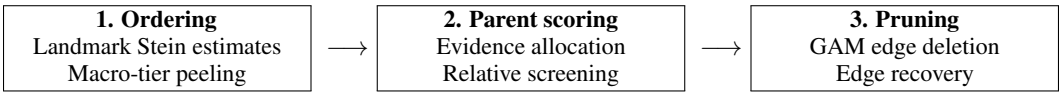

Figure 1: COSMIC phases yield an order, a screened candidate DAG, and the final DAG, respectively.

4.1 Landmark estimation and the three-band ensemble

Instead of constructing all n^2 observation similarities, COSMIC uses Nyström approximation (Williams & Seeger, 2000) to represent them through m landmarks and solves compact $m \times m$ systems. Farthest Point Sampling (FPS) (Gonzalez, 1985) selects observations that cover the sample geometry. The current target is $\min(300, \lfloor n/2 \rfloor, \max(25, 10q))$ for q active variables and $n \geq 2$. This limits the requested rank while adapting to dimension; it is an empirical budget, not an accuracy guarantee. Appendix B.1 further motivates the policy and its diagnostics.

A narrow, middle, and wide radial basis function (RBF) kernel use bandwidths $(s_0/2, s_0, 2s_0)$ around a median-distance estimate. COSMIC fits each Stein estimator separately and averages the estimated Hessian diagonals with weights $(1/4, 1/2, 1/4)$. The bandwidths have dyadic (geometric) spacing; the ensemble combines estimates by a weighted arithmetic average. Landmarks approximate both the kernel matrix and its derivative sums. The three bands provide multiple scales, without a general conditioning guarantee. Appendix A supplements this estimator specification with derivative formulas, regularization details, and numerical safeguards.

4.2 Macro-tier ordering and parent scoring

For a budget of k passes, pass t removes the $\lceil d_{\text{rem}}/(k-t+1) \rceil$ variables with largest estimated mean diagonal Hessian, then recomputes on the retained columns. Thus $k = d$ gives single-variable peeling. Reversing the removal sequence gives the causal topological order $\pi = (\pi_0, \ldots, \pi_{d-1})$ from root to sink, defining an initial tournament DAG $A^{\text{init}}_{ji} = 1 \iff j \prec_\pi i$ in canonical parent-by-child orientation ($A_{ji} = 1 \iff j \to i$). Mean-curvature ranking also orders members within a batch.

In adaptive mode, COSMIC identifies tiers by searching for gaps between sorted mean curvatures spanning at least 15% of the active dynamic range in the upper half of sorted curvatures, capped at $\lceil 0.4 d_{\text{rem}} \rceil$ with a $\lceil d_{\text{rem}}/4 \rceil$ fallback, as detailed alongside our threshold sensitivity analysis in Appendix C.1. Estimated batches can contain connected variables and need not be true sink layers.

To reduce the repeated fits required by variable-wise omissions, we introduce Leave-One-Macro-Tier-Out (LOMTO) scoring, which compares each tier omission with the same full-data reference. For candidate parent $j \notin \mathcal{S}$ and omitted tier $\mathcal{S}$, the aggregate evidence is $\Delta_{j \to \mathcal{S}} = \left| \bar{H}_j^{(-\mathcal{S})} - \bar{H}_j \right|$. For coordinates within the omitted tier ($j \in \mathcal{S}$), LOMTO perturbation is undefined; COSMIC therefore assigns a default score of 0.0, as co-batched members cannot be parents of one another.

Grouped omission supplies one score for an entire macro-tier, so parent scoring must resolve which member nodes receive that evidence. We introduce two allocation rules to address this within-macro-tier attribution problem for candidate parents outside the tier. LOMTO-Static (LOMTO-S) distributes this aggregate tier evidence across member nodes using normalized magnitudes of full-data score couplings. LOMTO with Local Coupling (LOMTO-LC) refines these weights by blending them with local score couplings captured during peeling. Refinement is weighted by the departure of local allocation from uniformity, the usable band weight, and the weighted pairwise agreement of signed couplings. Band disagreement continuously downweights refinement; it does not necessarily set the weight to zero. Uniform or degenerate local evidence, fewer than two valid bands, or zero average agreement retains the static allocation. This conserves aggregate evidence within each parent–tier block before applying the order mask, requiring no extra Stein fits. Transposing raw omission effects and masking by the tournament DAG yields canonical parent scores $S = S^{\text{raw}} \odot A^{\text{init}}$, where S_{ji} denotes evidence for $j \to i$ and intra-tier pairs satisfy $S_{ji} \equiv 0$. Appendix A.3 gives the complete multi-band allocation and fallback equations. First-order score couplings $\mathbb{E}[s_j s_p]$ can vanish on symmetric mechanism dependencies (e.g., $Y = X^2 + \epsilon$), misallocating tier evidence across co-batched members as demonstrated in the fidelity diagnostics of Appendix H.

4.3 Candidate screening, regression pruning, and recovery

To construct a sparse candidate graph prior to pruning, COSMIC follows CaPS's relative screening rule (Xu et al., 2024) with threshold divisor $\lambda_1 = 50.0$: candidate parent $j \in \text{pa}_{\text{init}}(i)$ is retained if $S_{ji} \geq \frac{1}{\lambda_1} \max_k S_{ki}$, defaulting to the single best candidate if all scores fall below threshold. When the largest candidate score is at most 10^{-8}, all order-compatible candidates are retained. Thus zero-scored intra-tier edges survive only in this degenerate-score case. The resulting screened DAG satisfies $A^{\text{screen}} \leq A^{\text{init}}$.

Next, nonparametric GAM pruning fits penalized cubic spline regression models with Generalized Cross-Validation (GCV) smoothing by Wood (2001) for each target i regressed on its screened candidates: $X_i = \beta_0 + \sum_{j \in \text{pa}_{\text{screen}}(i)} f_j(X_j) + \epsilon_i$. Edges with partial smooth significance $p \geq \alpha = 0.001$ are deleted, guaranteeing $A^{\text{pruned}} \leq A^{\text{screen}}$. Finally, post-pruning recovery evaluates pruned tournament non-edges ($A^{\text{init}}_{uv} = 1, A^{\text{pruned}}_{uv} = 0$) against an empirical threshold $\tau_{\text{rec}} = \lambda_2 \bar{S}_{\text{surv}}$ with $\lambda_2 = 50.0$, iteratively restoring candidate edges that preserve DAG acyclicity. This yields the final DAG A^{final}, establishing the hierarchical inclusion $\mathcal{E}^{\text{pruned}} \subseteq \mathcal{E}^{\text{final}} \subseteq \mathcal{E}^{\text{init}}$, as formalized alongside complete algorithmic steps and error attribution in Appendix A.4.

5 Theoretical Analysis

5.1 Computational complexity

COSMIC decouples ordering and parent scoring from downstream regression pruning. With landmark budget $m \ll n$ and $k \leq d$ realized macro-tiers, Phase 1 requires at most k Stein score evaluations. Phase 2 uses one full-data reference and k tier-omission evaluations, followed by coupling-based allocation and screening. With a fixed number of kernel bands, the estimator evaluations cost $\mathcal{O}(kdm^2n)$ under the conservative bound derived in Appendix D.2. Local cross-tier and full-data Fisher couplings cost $\mathcal{O}(nd^2)$; relative screening adds $\mathcal{O}(d^2)$. Together, Phases 1–2 require

$$\mathcal{O}\big(kdm^2n + nd^2\big) \quad \text{time}, \qquad \mathcal{O}(mn + m^2 + nd + d^2) \quad \text{memory}. \tag{1}$$

For fixed d, the cap $m \leq 300$ makes this bound linear in sample size n. Grouping reduces the number of estimator evaluations from $\mathcal{O}(d)$ to $\mathcal{O}(k)$, but the total bound remains quadratic in d because it includes dense pairwise coupling and score matrices. Near-linear dimension scaling over a measured range is therefore an empirical observation, not a sub-quadratic asymptotic guarantee. Dense single-variable methods such as SCORE and CaPS instead incur $\mathcal{O}(dn^3 + n^2d^2)$ time and quadratic kernel storage.

Phase 3 GAM pruning has a separate cost. With at most $B_{\max}$ spline coefficients per target and T fitting iterations, dense penalized regression has a conservative linear-algebra budget $\mathcal{O}\big(dT(nB_{\max}^2 + B_{\max}^3)\big)$. Here $B_{\max}$ grows with the number of screened parents and the basis size per smooth. Thus pruning does not inherit the landmark estimator's computational bound. Appendix D.2 gives the stage-by-stage derivation; Section 6.4 reports the empirical pruning cost and end-to-end limits. Appendix D.3 provides an extended pre-pruning asymptotic comparison with CaPS, SCORE, DAS, and DiffAN.

5.2 Macro-tier peeling under ANMs

Theorem 5.1 (Terminal-block marginalization invariance)**.** *Let X follow an acyclic additive noise model (ANM) with mutually independent noises ϵ_j. Let $\mathcal{B} \subset \mathcal{V}$ be a terminal block of variables, such that no directed edge points from $\mathcal{B}$ into the retained set $\mathcal{A} = \mathcal{V} \setminus \mathcal{B}$ ($\mathrm{de}_G(\mathcal{B}) \subseteq \mathcal{B}$). Then the marginal distribution $p(x_\mathcal{A})$ follows the exact ANM induced on $G[\mathcal{A}]$, with unchanged structural functions f_j, unchanged parent sets $\mathrm{pa}(j) \subseteq \mathcal{A}$, and mutually independent noise distributions for all $j \in \mathcal{A}$. Marginalizing $\mathcal{B}$ induces no unobserved confounding among the retained variables. Furthermore, if successive disjoint blocks partition $\mathcal{V}$ such that each block is terminal in the remaining graph and internally ordered correctly, reversing the block sequence yields a valid global topological order. For a block consisting exclusively of current sinks ($\mathrm{ch}(s) = \emptyset$ for all $s \in \mathcal{B}$), any internal order is valid.*

Sketch proof. Because no edge points from $\mathcal{B}$ into $\mathcal{A}$, each retained equation $X_j = f_j(X_{\mathrm{pa}(j)}) + \epsilon_j$ ($j \in \mathcal{A}$) depends only on retained variables and its independent noise ϵ_j. Integrating the joint density over $x_\mathcal{B}$ in reverse topological order of $\mathcal{B}$ evaluates each conditional density $\int p_b(x_b \mid x_{\mathrm{pa}(b)})\, dx_b = 1$, leaving the exact factorization $p(x_\mathcal{A}) = \prod_{j \in \mathcal{A}} p_j(x_j \mid x_{\mathrm{pa}(j)})$. Reversing successive terminal blocks respects all inter-block edges because an edge cannot point from an earlier removed block into a later one. □

Mathematical guarantee versus empirical heuristics. Theorem 5.1 guarantees preservation of the retained ANM when the removed block is terminal. A correct internal order is additionally required to obtain a global topological order from connected blocks. Crucially, it does *not* prove that empirical heuristics, such as adaptive gap thresholding on estimated mean curvatures, k-means clustering, or fixed schedules, will reliably select true terminal blocks in finite samples. If an estimated tier mistakenly groups an upstream non-sink while its child remains in $\mathcal{A}$, the terminal condition is violated, so preservation of the original retained equations is no longer guaranteed and latent confounding can arise. Moreover, when connected nodes are co-batched into the same tier, Theorem 5.1 requires an internal order; COSMIC's diagonal curvature heuristic is prone to finite-sample inversions, and co-batched edges receive an intra-tier score of 0.0, excluding them during screening whenever the child has a nondegenerate positive candidate score. Appendix D.1 provides full proofs and separation conditions. Appendix A.7 derives deterministic perturbation bounds relative to a matched dense empirical Stein estimator and branch-specific sufficient conditions for preserving adaptive peeling decisions. None of the 120 pilot runs satisfies these conditions at the production ridge. These bounds therefore do not certify dense-reference tracking in the tested settings; they neither establish nor rule out causal recovery, which we evaluate empirically.

6 Experiments

We evaluate three questions: how much recovery accuracy the scalable configuration retains, how its computational cost changes with problem size, and which components account for the observed trade-off. We distinguish recovery of a final graph from the intermediate ordering and parent-screening stages. This distinction is essential: completing an intermediate fit quickly does not by itself establish accurate or fast end-to-end causal discovery.

6.1 EXPERIMENTAL SETUP

Datasets and comparisons. We evaluate Synthetic ER, SynER1, and SynSF1 ($d = 10$, $n = 1000$), Sachs cellular signaling ($d = 11$, $n = 853$) against its published reference network (Sachs et al., 2005), and simulated fMRI ($d = 15$, $n = 200$) (Smith et al., 2011). Synthetic ER and scalability sweeps use nonlinear additive mechanisms with unit Gaussian noise; SynER1 and SynSF1 use the mixed linear/Gaussian-process generators of Xu et al. (2024). Linearity settings are held fixed, as these benchmarks evaluate scalability with graph size and sample budget rather than sensitivity to functional linearity. Appendix F.1 specifies mechanisms, preprocessing, and fMRI window selection.

Baseline methods and comparison scope. Figure 2 compares COSMIC with the closely related ordering methods CaPS, DAS, and DiffAN, as well as CAM and linear DAGMA, representing additive-model search and continuous DAG optimization. Exact-Gram COSMIC is included as an ablation of landmark approximation. The overview emphasizes CaPS as the closer mean-curvature and omission-scoring baseline. To keep it readable, SCORE, LNMIX, PC, GES, GOLEM, DAG-GNN, and DAGMA-MLP appear in the extended comparison in Appendix F.1.1; their omission from the main figure does not indicate missing evaluations.

Figure 3 compares COSMIC, CaPS, DAS, SCORE, DiffAN, and LNMIX to isolate the cost of ordering and candidate construction. SCORE provides the sequential dense-kernel peeling reference; LNMIX provides a score-based mixed-mechanism alternative. Figure 4 evaluates these methods plus CAM and linear DAGMA end to end. Exact-Gram COSMIC is omitted there because kernel approximation is isolated in the recovery and factorial ablations, while this experiment tests complete methods under a common timeout. Pruning follows each method's implementation; COSMIC uses GCV-smoothed GAMs, matching CaPS. Appendix F.1.1 details baseline configurations.

Metrics and execution scope. We use directed-edge F_1, the harmonic mean of precision and recall, and structural Hamming distance (SHD), which measures edits relative to the reference graph. Higher F_1 and lower SHD are preferable. In full-pipeline evaluations, both across the benchmark environments and in the end-to-end scaling experiment, these metrics describe the final recovered DAG after downstream pruning. In contrast, in the large-scale dimension and sample-size scaling sweeps, we stop COSMIC, CaPS, and DAS after candidate screening, and SCORE and DiffAN after ordering; LNMIX returns its complete graph because it has no separate pruning stage. Runtime measures execution up to that stopping point, and F_1 evaluates the resulting candidate or final graph. Their plotted F_1 values therefore describe different intermediate or final graph constructions. We retain both scopes to isolate the computational scaling of ordering and parent scoring, while separately evaluating complete end-to-end causal discovery. Appendix F provides extended empirical analyses, including Structural Intervention Distance (SID), runtime, additional baseline algorithms, and comparisons across other COSMIC variants.

Run accounting. The empirical comparison and component ablations use seed values 0–24. Dimension and sample-size scaling use the matched subset 0–9, except that the COSMIC runs at $n = 200{,}000$, $500{,}000$, and $1{,}000{,}000$ contain five seeds each. Means use successful runs with finite measurements; failures, timeouts, and skipped configurations are not assigned zero scores or completed runtimes. Error bars show sample standard deviations. These describe variation across runs, not independent replications of a biological dataset. In the scalability experiments, every algorithm is subject to a per-run timeout of 30 minutes (1800 seconds); the scaling driver also skips configurations deemed infeasible after an earlier timeout. Hardware and software descriptions are given in Appendix E; timings here refer to the recorded execution environments.

6.2 RECOVERY ACROSS BENCHMARK ENVIRONMENTS

Controlled synthetic benchmarks. On Synthetic ER, SynER1, and SynSF1, Figure 2 shows similar mean recovery for exact-Gram COSMIC and CaPS, supporting grouped omission in these settings. Scalable COSMIC has lower mean recovery on all three, identifying a cost associated with landmark approximation in these controlled nonlinear and mixed-mechanism benchmarks.

Biological data and realistic simulation. Sachs consists of real cellular-signaling measurements, whereas fMRI is a simulated hemodynamic benchmark. On Sachs, CaPS leads the displayed methods

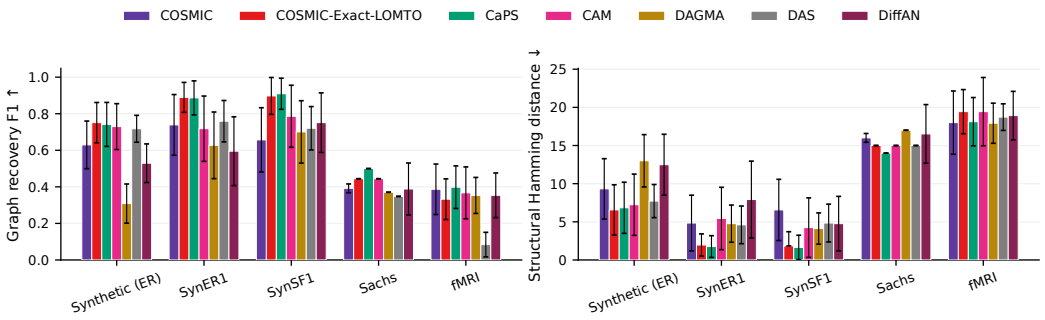

Figure 2: Final-graph recovery across five environments: directed-edge F_1 (left) and SHD (right), shown as mean $\pm$ sample standard deviation over 25 seeds. The two COSMIC configurations expose an accuracy trade-off that varies across datasets.

($F_1 = 0.500$), followed by exact-Gram COSMIC and CAM (0.444); scalable COSMIC reaches 0.392. On simulated fMRI, scalable COSMIC reaches $F_1 = 0.387$, close to CaPS (0.398), whereas exact-Gram COSMIC reaches 0.332. Linear DAGMA has the lowest mean SHD (17.92 versus COSMIC's 18.00), with overlapping variability. Thus the recovery gap is dataset-dependent: it persists on Sachs, but scalable COSMIC does not underperform its exact-Gram counterpart on simulated fMRI. Detailed numerical comparisons are discussed in Appendix F.

6.3 Ordering and Parent-Screening Scalability

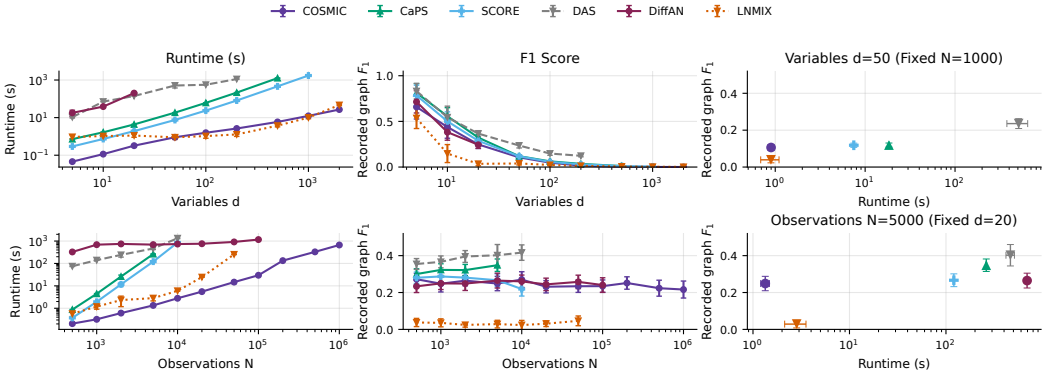

Figure 3: Scaling on nonlinear Synthetic ER data (unit Gaussian noise; more info in Appendix F.1) in dimension at $n = 1000$ (top) and sample size at $d = 20$ (bottom): runtime, intermediate graph F_1, and runtime–accuracy slices ($d = 50$ and $n = 5000$). Stages differ by method (Section 6.1); low F_1 for SCORE/DiffAN can reflect dense unpruned graphs rather than ordering errors. Points show mean $\pm$ sample SD over successful runs; truncated curves indicate a timeout.

Increasing the number of variables. At fixed $n = 1000$, COSMIC's screening runtime grows approximately proportionally to dimension over the larger tested settings, reaching $d = 2000$ with all ten selected seeds completing (Figure 3, top). Its curve remains well below the dense-kernel CaPS, SCORE, and DAS curves, which reach their time limits at smaller dimensions. LNMIX is the closest runtime comparator; at the $d = 50$ slice, both take about 0.9 seconds, although LNMIX has lower graph F_1. These measured trends show the benefit of capped landmark budgets and fewer peeling passes; they do not remove the quadratic coupling term in Section 5.1.

The quality curves show that faster candidate construction does not by itself preserve graph recovery as dimension grows. COSMIC's screening F_1 falls markedly, as do the other plotted graph scores. For SCORE and DiffAN, dense unpruned tournament graphs force precision to decrease as $\mathcal{O}(1/d)$

on these sparse benchmarks even with a correct order. The figure therefore establishes large-scale candidate construction, while final-graph recovery is evaluated separately below.

Increasing the number of observations. At fixed $d = 20$, COSMIC sustains screening through one million observations, while the dense-kernel baselines reach their time limits at much smaller sample sizes (Figure 3, bottom). The fixed landmark cap keeps estimator cost linear in n asymptotically, although measured runtimes also reflect memory and implementation overhead. Screening F_1 remains in a relatively narrow range across the sweep and does not improve monotonically with more data. This highlights the distinction between computational access to larger samples and improved statistical recovery.

The fixed-slice panels make the speed–accuracy trade-off explicit: COSMIC is substantially faster than CaPS and DAS, while DAS attains higher intermediate graph F_1.

6.4 End-to-end scalability

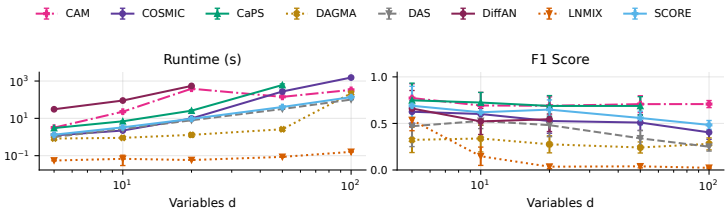

Figure 4: End-to-end runtime (left) and final DAG F_1 (right) on ER graphs with $n = \max(1000, 10d)$. Points show mean $\pm$ sample SD across 10 seeds; truncated curves indicate a timeout.

Figure 4 evaluates the full discovery pipeline, integrating ordering, parent screening, and edge pruning. In this full-pipeline regime, Phase 3 GAM pruning consumes 99.4% and 99.8% of total wall-clock time at $d = 50$ and $d = 100$, respectively. While CaPS records a timeout at $d = 100$, COSMIC, DAS, SCORE, CAM, and DAGMA complete, taking between 101 s and 336 s. At $d = 200$, the report records timeouts for all non-parametric and continuous methods previously mentioned. LNMIX is the sole method completing at $d = 500$ (31.60 ± 1.12 s), but its recovered graph is statistically uninformative ($F_1 = 0.0059 \pm 0.0005$, SHD $= 37{,}405$). These full-pipeline results demonstrate the operational boundary: COSMIC scales topological ordering and candidate screening to $d = 500+$ in minutes, but complete end-to-end recovery remains computationally bounded at $d \leq 100$ due to the scaling limits of downstream non-parametric GAM pruning.

6.5 Component ablations

Kernel approximation and grouped omission. Figure 5 (left and middle) shows contributions from both approximations, evaluated strictly with single-band estimation across all four cells to isolate Gram matrix compression and grouped omission without multi-band interaction. At $d = 100$, exact-Gram LOO takes 45.95 seconds, compared with 5.82 seconds for Nyström LOO and 8.87 seconds for exact-Gram LOMTO. Combining Nyström estimation and LOMTO reduces runtime to 2.13 seconds. At $d = 500$, Nyström LOO takes 399.41 seconds, while the combined configuration takes 38.64 seconds. Grouped omission therefore reduces a substantial cost that remains after kernel compression. This improvement is not free of an accuracy trade-off: at $d = 100$, screened-graph F_1 is 0.058 for the combined configuration, compared with 0.059 for exact LOO and 0.064 for exact LOMTO.

Single-band and tri-band estimation. The right panel of Figure 5 shows that three-band estimation improves mean final-graph recovery over a single band for both tested dimensions ($d = 10, 20$; $n = 500$). The gain is larger with Nyström estimation than with the exact Gram matrix, suggesting that combining scales is especially useful in the approximate estimator. Ordering/scoring and end-to-end runtimes increase moderately in these settings, giving a favorable empirical accuracy–runtime trade-off.

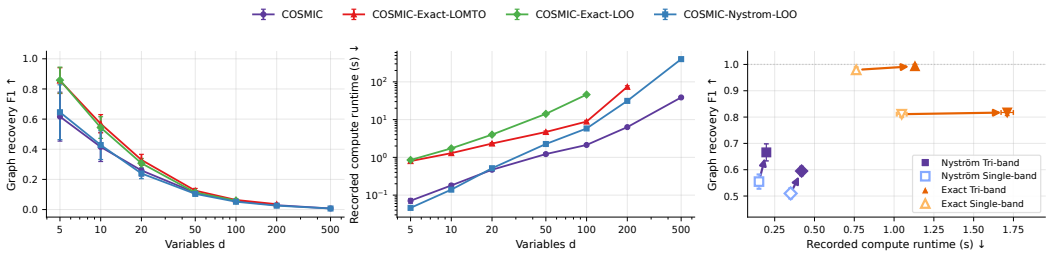

Figure 5: Algorithmic component ablations. Left and middle: Factorial comparison of exact versus Nyström Gram matrices and variable-wise LOO versus grouped LOMTO omission with $n = \max(1000, 10d)$, evaluated strictly under single-band estimation to isolate Gram matrix compression and grouped omission without multi-band confounding, showing screened-graph F_1 (left) and recorded ordering and parent-scoring compute runtime (middle); truncated curves indicate a timeout. Right: Single-band versus tri-band estimation ($n = 500$), showing final-graph recovery and ordering/scoring runtime. Arrows trace the shift from single-band (open) to tri-band (filled) estimation for $d = 10$ (left pair; squares/up-triangles) and $d = 20$ (right pair; diamonds/down-triangles)

7 Discussion

COSMIC makes derivative-based ordering and candidate screening accessible at substantially larger sample sizes and graph dimensions by combining two complementary reductions: landmarks compress kernel computations, and macro-tiers reduce repeated estimator evaluations. The factorial ablation shows benefits from both components, while the three-band ablation improves recovery at a moderate additional cost (Section 6.5). Grouped omission also retains recovery close to CaPS when paired with exact Gram matrices. These results support COSMIC as an efficient front end for ordering-based causal discovery.

Approximation quality remains dataset-dependent. The recovery gap on controlled synthetic benchmarks and Sachs is not observed relative to exact-Gram COSMIC on simulated fMRI (Section 6.2). While a full finite-sample analysis connecting landmark approximation to derivative error and ordering remains future work, the empirical ablations quantify where the present estimator offers a useful trade-off. The terminal-block theorem establishes when grouped removal preserves the ANM; it does not ensure that estimated tiers satisfy those conditions. Ordering errors can exclude true edges before pruning, and unequal noise variances can disrupt mean-curvature ranking (Section 2.1). Hidden confounding, dependent errors, and parent-dependent noise scales remain outside the stated guarantees.

The main opportunity for extending these gains to end-to-end discovery is downstream pruning: GAM fits consume 99.8% of runtime at $d = 100$ (Section 6.4). We've tested a B-spline/BIC replacement and two-worker R implementation but these did not meet the development criteria, leaving pruning acceleration open, described in more detail in Appendix A.4.4. Improving pruning, ordering reliability, and sparse candidate allocation are complementary directions; Appendix I develops these further.

8 Conclusion

COSMIC combines landmark Stein estimation, multi-band curvature estimates, and macro-tier peeling with grouped parent scoring to reduce the two central costs of derivative-based discovery: kernel size and repeated fitting. The terminal-block result establishes the structural conditions for valid grouped removal, and the experiments separate the contributions of kernel approximation, grouped omission, and bandwidth aggregation. Recorded screening is $38.9\times$ faster than CaPS at $d = 100$, and extends to $d = 2000$ and, in a separate sweep, one million observations. Recovery remains dataset-dependent, and full-pipeline runs complete through $d = 100$ before GAM pruning becomes limiting. Together, these contributions provide an efficient ordering and screening framework and identify the remaining work needed for scalable end-to-end causal discovery.

AI USE STATEMENT

The core conceptual framework of COSMIC, theoretical hypotheses, and experimental methodology were conceived and designed entirely by the authors. Generative AI tools were not used to generate research ideas, formulate hypotheses, generate datasets, or interpret scientific findings.

Generative AI was used strictly as an assistive tool for: (1) assisting with the mathematical typesetting and drafting of proofs in the Appendix based on author-derived results; (2) implementing and debugging software code in Python and PyTorch; (3) writing Matplotlib scripts to generate scientific figures from raw experimental logs; and (4) managing LaTeX formatting.

All software implementations were verified through automated test suites and benchmark execution; all proof steps, text, and theoretical claims were audited and verified line-by-line by the authors. The authors take full responsibility for the entirety of the work.

REPRODUCIBILITY STATEMENT

Section 4 specifies the estimator and parent-scoring procedure, and Section 5 states the assumptions and theoretical arguments. Section 6 describes the benchmarks and evaluation measures; Appendix E records the available computing environments; Workstation A hosts the empirical benchmark comparisons, while Workstation B hosts the remaining experiments, including end-to-end scalability. Appendices A and D give the complete formulas and proofs. Code is available through GitHub in the following repository: [Anonymized Link]

REFERENCES

Kevin Bello, Bryon Aragam, and Pradeep Ravikumar. DAGMA: Learning dags via m-matrices and a log-determinant acyclicity characterization. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in neural information processing systems*, volume 35, pp. 8226–8239. Curran Associates, Inc., 2022. doi: 10.52202/068431-0598. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/36e2967f87c3362e37cf988781a887ad-Paper-Conference.pdf.

Peter Bühlmann, Jonas Peters, and Jan Ernest. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6), December 2014. ISSN 0090-5364. doi: 10.1214/14-aos1260. URL http://dx.doi.org/10.1214/14-AOS1260.

Yutian Chen, Max Welling, and Alex Smola. Super-samples from kernel herding. In *Proceedings of the twenty-sixth conference on uncertainty in artificial intelligence*, UAI'10, pp. 109–116, Catalina Island, CA, 2010. AUAI Press. ISBN 978-0-9749039-6-5.

David Maxwell Chickering. Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, 3(null):507–554, March 2003. ISSN 1532-4435. doi: 10.1162/153244303321897717. URL https://doi.org/10.1162/153244303321897717.

Raaz Dwivedi and Lester Mackey. Kernel thinning. In Mikhail Belkin and Samory Kpotufe (eds.), *Proceedings of thirty fourth conference on learning theory*, volume 134 of *Proceedings of machine learning research*, pp. 1753–1753. PMLR, August 2021. URL https://proceedings.mlr.press/v134/dwivedi21a.html.

Ahmed El Alaoui and Michael Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in neural information processing systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/f3f27a324736617f20abbf2ffd806f6d-Paper.pdf.

Damien Garreau, Wittawat Jitkrittum, and Motonobu Kanagawa. Large sample analysis of the median heuristic, 2018. URL https://arxiv.org/abs/1707.07269. arXiv: 1707.07269 [math.ST].

Albert Gong, Kyuseong Choi, and Raaz Dwivedi. Supervised kernel thinning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in neural*

information processing systems, volume 37, pp. 6267–6322. Curran Associates, Inc., 2024. doi: 10.52202/079017-0203. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/0bd7c1a579459520e4d731a14b7bda7d-Paper-Conference.pdf.

Teofilo F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 38:293–306, 1985. ISSN 0304-3975. doi: https://doi.org/10.1016/0304-3975(85)90224-5. URL https://www.sciencedirect.com/science/article/pii/0304397585902245.

Maximilian Hahn, Alina Zajak, Dominik Heider, and Adèle Helena Ribeiro. Federated causal discovery across heterogeneous datasets under latent confounding, 2026. URL https://arxiv.org/abs/2603.05149. arXiv: 2603.05149 [cs.LG].

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005. URL http://jmlr.org/papers/v6/hyvarinen05a.html.

Yingzhen Li and Richard E. Turner. Gradient estimators for implicit models. In *International conference on learning representations*, 2018. URL https://openreview.net/forum?id=SJi9WOeRb.

Wenqin Liu, Biwei Huang, Erdun Gao, Qiuhong Ke, Howard Bondell, and Mingming Gong. Causal discovery with mixed linear and nonlinear additive noise models: a scalable approach. In Francesco Locatello and Vanessa Didelez (eds.), *Proceedings of the third conference on causal learning and reasoning*, volume 236 of *Proceedings of machine learning research*, pp. 1237–1263. PMLR, April 2024. URL https://proceedings.mlr.press/v236/liu24b.html.

Thomas A. Louis. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):226–233, January 1982. ISSN 0035-9246. doi: 10.1111/j.2517-6161.1982.tb01203.x. URL https://doi.org/10.1111/j.2517-6161.1982.tb01203.x. tex.eprint: https://academic.oup.com/jrsssb/article-pdf/44/2/226/49097595/jrsssb_44_2_226.pdf.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54, pp. 1273–1282. PMLR, 2017.

Francesco Montagna, Nicoletta Noceti, Lorenzo Rosasco, Kun Zhang, and Francesco Locatello. Scalable causal discovery with score matching. In Mihaela van der Schaar, Cheng Zhang, and Dominik Janzing (eds.), *Proceedings of the second conference on causal learning and reasoning*, volume 213 of *Proceedings of machine learning research*, pp. 752–771. PMLR, April 2023. URL https://proceedings.mlr.press/v213/montagna23b.html.

Francesco Montagna, Philipp Michael Faller, Patrick Blöbaum, Elke Kirschbaum, and Francesco Locatello. Score matching through the roof: linear, nonlinear, and latent variables causal discovery. In Biwei Huang and Mathias Drton (eds.), *Proceedings of the fourth conference on causal learning and reasoning*, volume 275 of *Proceedings of machine learning research*, pp. 552–605. PMLR, May 2025. URL https://proceedings.mlr.press/v275/montagna25a.html.

Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and DAG constraints for learning linear dags, 2021. URL https://arxiv.org/abs/2006.10201. arXiv: 2006.10201 [cs.LG].

Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979. doi: 10.1109/TSMC.1979.4310076.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, UK, 2nd edition, 2009.

Jonas Peters, Joris Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models, 2014. URL https://arxiv.org/abs/1309.6779. arXiv: 1309.6779 [stat.ML].

Patrik Reizinger, Siyuan Guo, Ferenc Huszár, Bernhard Schölkopf, and Wieland Brendel. Identifiable exchangeable mechanisms for causal structure and representation learning. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International conference on learning representations*, volume 2025, pp. 62196–62223, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/9b91ee0da3bcd61905fcd89e770168fc-Paper-Conference.pdf.

Paul Rolland, Volkan Cevher, Matthäus Kleindessner, Chris Russell, Dominik Janzing, Bernhard Schölkopf, and Francesco Locatello. Score matching enables causal discovery of nonlinear additive noise models. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th international conference on machine learning*, volume 162 of *Proceedings of machine learning research*, pp. 18741–18753. PMLR, July 2022. URL https://proceedings.mlr.press/v162/rolland22a.html.

Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in neural information processing systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/03e0704b5690a2dee1861dc3ad3316c9-Paper.pdf.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A. Lauffenburger, and Garry P. Nolan. Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data. *Science*, 308(5721): 523–529, 2005. doi: 10.1126/science.1105809.

Pedro Sanchez, Xiao Liu, Alison Q O'Neil, and Sotirios A. Tsaftaris. Diffusion models for causal discovery via topological ordering. In *The eleventh international conference on learning representations*, 2023. URL https://openreview.net/forum?id=Idusfje4-Wq.

Stephen M Smith, Karla L Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F Beckmann, Thomas E Nichols, Joseph D Ramsey, and Mark W Woolrich. Network modelling methods for FMRI. *NeuroImage*, 54(2):875–891, 2011. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2010.08.063.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*, volume 81. January 1993. ISBN 978-1-4612-7650-0. doi: 10.1007/978-1-4612-2748-9.

Danica J. Sutherland, Heiko Strathmann, Michael Arbel, and Arthur Gretton. Efficient and principled score estimation with Nyström kernel exponential families, 2018. URL https://arxiv.org/abs/1705.08360. arXiv: 1705.08360 [stat.ML].

Christopher Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. In T. Leen, T. Dietterich, and V. Tresp (eds.), *Advances in neural information processing systems*, volume 13. MIT Press, 2000. URL https://proceedings.neurips.cc/paper_files/paper/2000/file/19de10adbaa1b2ee13f77f679fa1483a-Paper.pdf.

S. Wood. mgcv: GAMs and generalized ridge regression for R. *R News*, 1, January 2001.

Zhuopeng Xu, Yujie Li, Cheng Liu, and Ning Gui. Ordering-based causal discovery for linear and nonlinear relations. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in neural information processing systems*, volume 37, pp. 4315–4340. Curran Associates, Inc., 2024. doi: 10.52202/079017-0141. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/081b08068e4733ae3e7ad019fe8d172f-Paper-Conference.pdf.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. DAG-GNN: DAG structure learning with graph neural networks, 2019. URL https://arxiv.org/abs/1904.10098. arXiv: 1904.10098 [cs.LG].

Kai Zhang, Ivor W. Tsang, and James T. Kwok. Improved Nyström low-rank approximation and error analysis. In *Proceedings of the 25th international conference on machine learning*, Icml '08, pp. 1232–1239, Helsinki, Finland, 2008. Association for Computing Machinery. ISBN 978-1-60558-205-4. doi: 10.1145/1390156.1390311. URL https://doi.org/10.1145/1390156.1390311.

A Estimator and Parent-Score Derivations

This section gives the complete estimator and allocation rules underlying Section 4.

A.1 Nyström Landmark Stein Estimator

The exact Stein estimator requires solving a linear system involving the full Gram matrix $K_{nn} \in \mathbb{R}^{n\times n}$, which incurs $\mathcal{O}(n^3)$ time and $\mathcal{O}(n^2)$ memory. To avoid this bottleneck, we employ the Nyström method (Williams & Seeger, 2000), which approximates the full kernel matrix through a set of $m \ll n$ landmark points $Z \in \mathbb{R}^{m\times d}$:

$$K_{nn} \approx \widetilde{K}_{nn} = K_{nm} K_{mm}^{\dagger} K_{nm}^{\top}, \tag{2}$$

where $K_{nm} \in \mathbb{R}^{n\times m}$ contains the kernel evaluations between all observations and the landmarks, and $K_{mm} \in \mathbb{R}^{m\times m}$ is the landmark Gram matrix; $\dagger$ denotes the Moore–Penrose pseudoinverse. This approximation is particularly effective when the kernel matrix has low effective rank, as occurs under rapid spectral decay or when the observations concentrate near a lower-dimensional manifold. In plain terms, low effective rank means that relatively few independent patterns explain most of the kernel matrix. The approximation retains those dominant patterns using fewer columns.

For regularization $\eta > 0$, when the landmark Gram matrix K_{mm} is non-singular ($K_{mm}^{\dagger} = K_{mm}^{-1}$), the Woodbury matrix identity equates the $n \times n$ regularized inverse to a solve involving the compact $m \times m$ matrix

$$M_\eta = \eta K_{mm} + K_{nm}^{\top} K_{nm} \in \mathbb{R}^{m\times m}, \tag{3}$$

via $(\widetilde{K}_{nn} + \eta I_n)^{-1} = \eta^{-1}[I_n - K_{nm} M_\eta^{-1} K_{nm}^{\top}]$. If K_{mm} is singular, algebraic equivalence fails because $(K_{mm}^{\dagger})^{-1}$ is undefined. While distinct FPS landmarks guarantee non-singularity in exact arithmetic, large bandwidths can induce high condition numbers in floating-point computation (Appendix A.6). Consequently, the Hessian estimate $\widehat{H}$ can be evaluated through the Nyström coefficients and the corresponding kernel derivatives without forming or inverting K_{nn}. The resulting computational cost is reduced from $\mathcal{O}(n^3)$ to

$$\mathcal{O}(nm^2 + m^3), \tag{4}$$

for the matrix factorization, with $\mathcal{O}(nm + m^2)$ kernel storage. Feature-dependent operations and storage are included in Section 5.1.

The adaptive landmark budget depends on the current dimension and is capped by the sample size, as specified in Section B.1. The current implementation uses Farthest Point Sampling (FPS) to choose landmark locations within this budget. Starting from an initial observation, FPS recursively selects the point farthest from the current landmark set:

$$z_\ell = \arg\max_{x_i \in X} \min_{z \in Z_{\ell-1}} \|x_i - z\|_2, \qquad \ell = 2, \ldots, m. \tag{5}$$

Unlike uniform sampling, which may repeatedly select observations from densely populated regions, FPS promotes geometrically diverse landmarks that cover the observed support. The budget limits how many points can be requested; geometric coverage alone does not control the numerical stability of derivative estimation.

For a bandwidth $s > 0$, the estimator uses the radial basis function (RBF) kernel, whose similarity decreases with squared Euclidean distance:

$$k_s(x, z) = s^{-1} \exp\left(-\frac{\|x - z\|_2^2}{2s^2}\right). \tag{6}$$

For a single bandwidth, we use a regularized Stein solve following the first-order estimator of Li & Turner (2018) and the second-order construction of Rolland et al. (2022). Write $C = K_{nm}$ and $W = K_{mm}$ for this bandwidth. For the default FPS landmarks, define derivative sums $D_1, D_2 \in \mathbb{R}^{n\times d}$ by

$$(D_1)_{ij} = \frac{n}{m} \sum_{a=1}^{m} \frac{z_{aj} - x_{ij}}{s^2} k_s(x_i, z_a), \tag{7}$$

$$(D_2)_{ij} = \frac{n}{m} \sum_{a=1}^{m} \left(\frac{(z_{aj} - x_{ij})^2}{s^4} - \frac{1}{s^2}\right) k_s(x_i, z_a). \tag{8}$$

These approximate the full empirical derivative sums using landmarks. They introduce an approximation in addition to the low-rank Gram matrix. In particular, multiplying by n/m does not make deterministic FPS an unbiased sample of the empirical distribution.

Here D_1 approximates first-derivative sums for score estimation, while D_2 approximates second derivatives along each variable for diagonal-curvature estimation. This landmark-based summation is also called quadrature. With positive regularizers η_G for the score and η_H for the second derivative, and $M_\eta = C^\top C + \eta W$, the estimates are

$$G = \eta_G^{-1}\left[D_1 - CM_{\eta_G}^{-1}C^\top D_1\right], \tag{9}$$

$$Q = \eta_H^{-1}\left[D_2 - CM_{\eta_H}^{-1}C^\top D_2\right], \tag{10}$$

$$\widehat{H} = Q - G \odot G. \tag{11}$$

Here G_{ij} estimates score component j at observation x_i, evaluating to $+\nabla \log p(x_i)$. Note that Stein's identity $\mathbb{E}[s(X)k(X,z)] = -\mathbb{E}[\nabla_X k(X,z)]$ integrates by parts with respect to the sample argument. For the RBF kernel, $-\nabla_z k(z,x) = \frac{z-x}{s^2}k_s(x,z)$, so differentiating with respect to landmark argument z_a in Equation 7 absorbs the negative sign, yielding consistent positive score estimates matching Li & Turner (2018) and Rolland et al. (2022). Q_{ij} estimates $\partial_j^2 p(x_i)/p(x_i)$. The identity $\partial_j^2 \log p = (\partial_j^2 p)/p - (\partial_j \log p)^2$ then gives the diagonal-curvature estimate $\widehat{H}_{ij} = Q_{ij} - G_{ij}^2$. Thus $\widehat{H}$ stores one diagonal entry per variable and observation, forming an $n \times d$ matrix. Storing the full Hessian at every observation would instead require an $n \times d \times d$ tensor. The symbol $\odot$ denotes entrywise multiplication; expressions involving matrix inverses are evaluated through linear solves.

A.2 Tri-Band Harmonic Stein Ensemble

A kernel bandwidth sets the distance over which observations contribute to an estimate. COSMIC uses three bandwidths to retain information at several scales, rather than committing to a single scale. The base bandwidth s_0 is the median pairwise Euclidean distance in a random subset of at most 300 observations. This follows the median heuristic studied by Garreau et al. (2018); their analysis concerns kernel two-sample testing and does not establish an optimal bandwidth for Hessian estimation.

For each active dataset, we use

$$(s_1, s_2, s_3) = (s_0/2, s_0, 2s_0), \tag{12}$$

$$(\omega_1, \omega_2, \omega_3) = (1/4, 1/2, 1/4). \tag{13}$$

The implementation replaces $s_0 < 10^{-3}$ by 1 and floors each resulting bandwidth at 10^{-2}. The narrow band gives greater relative weight to nearby points, while the wide band gives more weight to distant points. The middle band receives half of the total weight. These scales and weights are design choices in COSMIC. The bandwidths form a dyadic (geometric) progression. Harmonic is the ensemble name; the estimates use a weighted arithmetic average, not a harmonic mean.

We apply the estimator in Section A.1 independently at each of the three bandwidths and average the resulting Hessian diagonals:

$$\widehat{H}_{\mathrm{tri}} = \sum_{\ell=1}^{3} \omega_\ell \widehat{H}^{(\ell)}. \tag{14}$$

We combine the estimates after solving separately at each bandwidth. Averaging the kernels before solving would generally produce a different estimator, because the solution depends nonlinearly on the kernel matrix. We also retain the score estimates from each band for the macro-tier parent-scoring procedure introduced in Section A.3. That procedure uses the weighted average of the separate score-product matrices, $-\sum_{\ell=1}^{3} \omega_\ell (G^{(\ell)})^\top G^{(\ell)}/n$, rather than the product of averaged scores. The latter would introduce cross-band terms. Using several bandwidths provides estimates at different scales, but does not by itself guarantee better conditioning or lower estimation error.

A.3 Macro-Tier Peeling and Parent Scoring

In the algorithmic description, expectations denote sample averages of estimated Hessian entries on the current active dataset. Population quantities are used separately in Section 5.2.

Sequential leaf peeling recomputes the Stein Hessian after removing each individual variable, requiring up to d successive estimation passes. COSMIC reduces repeated estimation by removing variables in $k \leq d$ batches, called macro-tiers, and recomputing scores after each removal. Once the order is established, LOMTO parent scoring independently omits each discovered tier from the full dataset to measure its effect on the remaining variables.

Let $\mathcal{V}_t$ denote the active variable set at tier t and let $d_{\text{rem}} = |\mathcal{V}_t|$. The number of variables removed at that tier is

$$b_t = \max\left(1, \left\lceil \frac{d_{\text{rem}}}{k - t + 1} \right\rceil\right). \tag{15}$$

For a fixed budget $1 \leq k \leq d$, t indexes passes from 1 to k. COSMIC selects the b_t largest estimated mean diagonal Hessians (the least negative values when all are negative):

$$\mathcal{S}_t = \text{top-}\, b_t \left(\left\{ \mathbb{E}\left[H_{jj}^{(t)} \right] \right\}_{j \in \mathcal{V}_t} \right). \tag{16}$$

The active set is then updated as

$$\mathcal{V}_{t+1} = \mathcal{V}_t \setminus \mathcal{S}_t, \tag{17}$$

and the Hessian is recomputed only after the complete tier has been removed. Repeating this procedure produces the upstream-to-downstream tier ordering

$$\Pi = (\mathcal{S}_k, \mathcal{S}_{k-1}, \ldots, \mathcal{S}_1). \tag{18}$$

Macro-tier peeling therefore reduces the number of sequential Stein estimation passes from d to k. The parameter k directly controls the acceleration: $k = d$ recovers single-node peeling, whereas $k \ll d$ substantially reduces the number of Hessian recomputations.

These estimated tiers group variables for computation; they need not coincide with true causal generations, which are layers formed by repeatedly deleting all current sinks. The tier order alone does not identify individual parents of a downstream target. To estimate individual parent scores, COSMIC first computes static LOMTO scores (LOMTO-S), then optionally refines their allocation using Local Coupling (LOMTO-LC). "Static" means that every omitted tier is compared against the same full-data reference and uses full-data allocation weights. "Local" means that the allocation also uses the reduced set of variables present when the tier was peeled. Coupling measures how two score components vary together; it does not by itself establish a causal edge.

Let $\bar{H}_j$ be the sample mean of the estimated diagonal Hessian on the full variable set $\mathcal{V}$, and let $\bar{H}_j^{(-\mathcal{S}_t)}$ be its estimate after omitting tier $\mathcal{S}_t$ from that same full set. For $j \notin \mathcal{S}_t$, the static tier perturbation is:

$$\Delta_{j \to \mathcal{S}_t} = \left| \bar{H}_j^{(-\mathcal{S}_t)} - \bar{H}_j \right|. \tag{19}$$

Each omission is evaluated independently on $X_{\mathcal{V} \setminus \mathcal{S}_t}$; these are not consecutive marginalizations of the peeling active set.

An omission gives one aggregate change for candidate parent j, but does not identify how much to assign to each member of the removed tier. The following steps distribute that change, first using full-data couplings and then optionally refining the allocation using couplings from the peeling stage. For the full-data allocation, we define a negative score-product estimate of the expected Hessian:

$$F^{(0)} = -\sum_{\ell=1}^{3} \omega_\ell \frac{(G^{(\ell,0)})^\top G^{(\ell,0)}}{n}. \tag{20}$$

Each entry averages products of two score components across observations and combines the three bandwidths. Its magnitude measures coupling for allocation; the symmetric matrix does not supply an edge direction. Normalizing $|F_{jp}^{(0)}|$ across $p \in \mathcal{S}_t$ gives static weights a_{jp}, with a uniform fallback for a numerically zero sum. LOMTO-S assigns $\mathcal{P}_{j \to p}^{S} = \Delta_{j \to \mathcal{S}_t} a_{jp}$: it divides the total change associated with a tier among that tier's variables. For example, a change of 2 with weights $(0.75, 0.25)$ gives scores $(1.5, 0.5)$. These weights allocate evidence; they do not prove which tier members are children of j.

The alternative local allocation uses couplings obtained during peeling, after earlier tiers have been removed. This can change the relative strengths assigned to tier members, but does not guarantee a more accurate allocation. Let $G^{(\ell,t)}$ denote the score matrix obtained at kernel band ℓ over the active set $\mathcal{V}_t$. For retained candidate j and tier member $p \in \mathcal{S}_t$, their local coupling is:

$$c^{(t)}_{j\to p} = \left| \sum_{\ell=1}^{3} \omega_\ell \frac{\left(G^{(\ell,t)}_{:j}\right)^\top G^{(\ell,t)}_{:p}}{n} \right|, \tag{21}$$

where ω_ℓ is the weight assigned to band ℓ. Normalizing $c^{(t)}_{j\to p}$ across $p \in \mathcal{S}_t$ provides the normalized local weights u_{jp}.

COSMIC blends these allocations to define the final parent score:

$$\mathcal{P}_{j\to p} = \Delta_{j\to\mathcal{S}_t}\left[(1-\rho_{jt})a_{jp} + \rho_{jt}u_{jp}\right], \qquad j \in \mathcal{V}_t \setminus \mathcal{S}_t. \tag{22}$$

The refinement weight $\rho_{jt} \in [0,1]$ controls the blend: zero retains the static weights a_{jp}, while one uses only the local weights u_{jp}. For example, with $\Delta = 2$, $a = (0.75, 0.25)$, $u = (0.9, 0.1)$, and $\rho = 0.4$, the final scores are $(1.62, 0.38)$ instead of the static $(1.5, 0.5)$. We choose ρ_{jt} using three factors: local preference, usable band weight, and agreement between bands.

First, local preference measures departure from equal allocation. For $b = |\mathcal{S}_t| > 1$, define

$$s_{jt} = \frac{\frac{1}{2}\sum_{p\in\mathcal{S}_t} |u_{jp} - 1/b|}{1 - 1/b}. \tag{23}$$

This is zero for equal shares and one when a single tier member receives all the weight. For two members, $u = (0.9, 0.1)$ gives $s_{jt} = 0.8$. Concentration expresses a preference, not evidence that the preference is correct.

Second, we check which bandwidths provide usable couplings for this particular parent–tier pair. Let $h_{\ell p} = -(G^{(\ell,t)}_{:j})^\top G^{(\ell,t)}_{:p}/n$ be the signed coupling for band ℓ and tier member p. A band is valid if these entries are finite, their absolute values have a nonzero sum, and its weight ω_ℓ is positive. Let L contain the valid bands. Their combined weight, divided by the total band weight, discounts missing or degenerate estimates.

Third, we compare the coupling patterns across valid bands. Normalize each signed vector by its sum of absolute entries: $v_{\ell p} = h_{\ell p}/\sum_{q\in\mathcal{S}_t} |h_{\ell q}|$. For example, $(2, -1, 1)$ becomes $(0.5, -0.25, 0.25)$. This removes overall magnitude while retaining signs and relative strengths. Pairwise agreement is $A_{\ell r} = 1 - \frac{1}{2}\|v_\ell - v_r\|_1$, where the ℓ_1 norm sums absolute entrywise differences. The distance lies between zero and two, so agreement lies between zero and one. Identical normalized vectors give one; opposite vectors give zero. Thus sign reversals reduce agreement even when coupling magnitudes match. These signs concern score products, not causal directions.

Combining the three factors gives

$$\begin{aligned} \rho_{jt} &= s_{jt}\, w_{jt}\, A_{jt}, \\ w_{jt} &= \frac{\sum_{\ell\in L}\omega_\ell}{\sum_{\ell=1}^{3}\omega_\ell}, \\ A_{jt} &= \frac{\sum_{\ell<r,\,\ell,r\in L}\omega_\ell\omega_r A_{\ell r}}{\sum_{\ell<r,\,\ell,r\in L}\omega_\ell\omega_r}. \end{aligned} \tag{24}$$

Here w_{jt} is the usable band weight and A_{jt} is the weighted average agreement. The restriction $\ell < r$ counts each pair once; the denominator normalizes the pair weights. For example, local preference 0.8, usable weight 1, and average agreement 0.75 give $\rho_{jt} = 0.6$: the blend uses 60% local and 40% static allocation. Band agreement is a consistency check, not a calibrated probability of correct allocation.

We set $\rho_{jt} = 0$ for singleton tiers ($|\mathcal{S}_t| \le 1$), fewer than two valid bands ($|L| < 2$), or numerically zero total local coupling ($\sum_p c^{(t)}_{j\to p} \le 10^{-12}$), avoiding undefined normalizations. Uniform local allocation or zero average agreement also retains the static scores. Both allocation vectors sum to

one, so their blend redistributes rather than increases the total evidence. Before the final order mask, the total score, called tier mass, is strictly preserved for each retained parent and removed tier:

$$\sum_{p \in \mathcal{S}_t} \mathcal{P}_{j \to p} = \Delta_{j \to \mathcal{S}_t}. \tag{25}$$

A.3.1 Matrix Lifecycle and Complete Piecewise Score Definition

Operationally, parent scores transition through three distinct matrix representations:

1. **Raw Omission Effects ($M \in \mathbb{R}^{d \times d}$):** Evaluated in `compute_parent_scores`, where rows index omitted children $i \in \mathcal{S}_t$ and columns index retained parents $j \in \mathcal{V} \setminus \mathcal{S}_t$: $M_{ij} = a_{ji}\Delta_{j \to \mathcal{S}_t}$. Intra-tier entries ($i, j \in \mathcal{S}_t$), unobserved pairs, self-loops ($i = j$), and the full-graph omission boundary ($|\mathcal{S}_t| = d$) are initialized and held at 0.0.
2. **Canonical Transposition ($S^{\text{raw}} = M^\top \in \mathbb{R}^{d \times d}$):** Transposing M aligns the matrix with the canonical parent-by-child orientation (S^{raw}_{ji} has row j = parent, column i = child). Local-coupling refinement is applied across active retained parent rows ($j \in \mathcal{V}_t \setminus \mathcal{S}_t$) to blend static and local attributions: $S^{\text{raw}}_{j,\mathcal{S}_t} = \Delta_{j \to \mathcal{S}_t}[(1 - \rho_{jt})a_{j,\mathcal{S}_t} + \rho_{jt}u_{j,\mathcal{S}_t}]$.
3. **Topological Order Masking ($S = S^{\text{raw}} \odot A^{\text{init}}$):** The initial tournament DAG $A^{\text{init}}_{ji} = 1 \iff j \prec_\pi i$ masks the matrix entrywise. Order-reversed pairs have $A^{\text{init}}_{ji} = 0$ and are removed by the mask. Forward intra-tier pairs have $A^{\text{init}}_{ji} = 1$, but their masked scores remain zero because their raw omission scores are zero.

Combining these rules yields the complete closed-form piecewise parent score $S_{ji} = \mathcal{P}_{j \to i}$ for every pair $(j, i) \in \mathcal{V}^2$:

$$S_{ji} = \begin{cases} \Delta_{j \to \mathcal{T}_{\tau(i)}}\left[(1 - \rho_{j,\tau(i)})a_{ji} + \rho_{j,\tau(i)}u_{ji}\right] & \text{if } \tau(j) < \tau(i) \text{ (forward inter-tier)}, \\ 0.0 & \text{if } \tau(j) = \tau(i),\ j \neq i \text{ (intra-tier)}, \\ 0.0 & \text{if } \tau(j) > \tau(i) \text{ (reverse inter-tier)}, \\ 0.0 & \text{if } j = i \text{ (self-loop)}, \\ 0.0 & \text{if } k = 1 \text{ (all-nodes tier, } |\mathcal{S}_t| = d). \end{cases} \tag{26}$$

where $\tau(v) \in \{1, \ldots, K\}$ denotes the unique root-to-sink macro-tier index containing variable v ($\mathcal{T}_1$ contains roots, $\mathcal{T}_K$ contains sinks). Because Fisher coupling is symmetric, orientation is supplied strictly by the estimated topological order π. Crucially, this local refinement requires no additional Stein evaluations beyond the initial ordering and LOMTO passes.

A.4 Candidate Screening, Nonparametric GAM Pruning, and Post-Recovery

Following parent scoring, COSMIC constructs the final causal DAG through three sequential edge-assembly operators: relative pre-screening (Phase 2), nonparametric regression pruning (Phase 3), and post-pruning recovery (Phase 3).

A.4.1 Phase 2: Score-to-Edge Pre-Screening

Given the tournament DAG A^{init} and masked score matrix S, pre-screening prunes weak ancestor candidates for each child $i \in \{0, \ldots, d-1\}$:

1. **Candidate Set:** $\text{pa}_{\text{init}}(i) = \{j \in \mathcal{V} : A^{\text{init}}_{ji} = 1\}$. If $\text{pa}_{\text{init}}(i) = \emptyset$, child i is a topological root; $A^{\text{screen}}_{:i} = \mathbf{0}$.
2. **Relative Thresholding:** Compute the maximum candidate score $M_i = \max_{k \in \text{pa}_{\text{init}}(i)} S_{ki}$.
   - If $M_i > 10^{-8}$: Candidate parent j is retained if and only if

$$S_{ji} \geq \frac{M_i}{\lambda_1} = \frac{\max_{k \in \text{pa}_{\text{init}}(i)} S_{ki}}{50.0}. \tag{27}$$

   If strict thresholding eliminates all candidates, the procedure executes an **empty-survivor fallback**, retaining the single best candidate $j^* = \text{argmax}_{k \in \text{pa}_{\text{init}}(i)} S_{ki}$.

- If $M_i \leq 10^{-8}$ (all candidate scores near zero): The procedure executes a **degenerate zero-score fallback**, retaining all candidate parents ($A_{ji}^{\text{screen}} = 1$ for all $j \in \text{pa}_{\text{init}}(i)$).

3. **Parent Cap Clamping:** If a maximum parent budget $p_{\max}$ is specified and $|\,\text{survivors}\,| > p_{\max}$, only the top $p_{\max}$ candidates sorted descending by S_{ji} are retained.

Because $A_{ji}^{\text{screen}} = 1 \implies A_{ji}^{\text{init}} = 1$, screening satisfies the strict subgraph inclusion $A^{\text{screen}} \leq A^{\text{init}}$.

A.4.2 Phase 3: Nonparametric GAM Pruning

Candidate edges surviving pre-screening are subjected to statistical regression pruning. For each target child i with screened parents $\mathcal{P}_i = \{j : A_{ji}^{\text{screen}} = 1\}$:

1. If $|\mathcal{P}_i| = 0$, column i requires no regression.
2. If $|\mathcal{P}_i| \geq 1$, COSMIC fits a Generalized Additive Model via R's `mgcv::gam(selGam.R)`:
$$X_i = \beta_0 + \sum_{j \in \mathcal{P}_i} f_j(X_j) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma_i^2), \tag{28}$$
where each f_j is a penalized cubic regression spline ($k = 10$ basis functions) with Generalized Cross-Validation smoothing parameter selection (`GCV.Cp`).
3. **Hypothesis Testing & Deletion:** The partial smooth effect of each candidate parent $j \in \mathcal{P}_i$ is tested via an ANOVA F-test / Wald test:
$$A_{ji}^{\text{pruned}} = \begin{cases} 1 & \text{if } A_{ji}^{\text{screen}} = 1 \text{ and } p\text{-value}(f_j) < \alpha \quad (\alpha = 0.001), \\ 0 & \text{if } p\text{-value}(f_j) \geq \alpha \quad (\text{edge is pruned}). \end{cases} \tag{29}$$

Phase 3 operates as a **pure edge deletion operator**, strictly enforcing $A^{\text{pruned}} \leq A^{\text{screen}}$.

A.4.3 Phase 3 (continued): Post-Pruning Recovery

To mitigate false negatives induced by conservative GCV smoothing, post-recovery re-evaluates pruned tournament non-edges:

1. **Candidate Recovery Non-Edges:**
$$\mathcal{E}_{\text{rec}} = \{(u, v) \in \mathcal{V} \times \mathcal{V} : A_{uv}^{\text{init}} = 1 \text{ and } A_{uv}^{\text{pruned}} = 0\}. \tag{30}$$
2. **Adaptive Recovery Threshold:** Compute the average parent score among surviving pruned edges:
$$\bar{S}_{\text{surv}} = \frac{1}{\sum_{a,b} A_{ab}^{\text{pruned}}} \sum_{a,b : A_{ab}^{\text{pruned}} = 1} S_{ab}. \tag{31}$$
If no edges survived pruning ($\sum_{a,b} A_{ab}^{\text{pruned}} = 0$), recovery terminates immediately ($A^{\text{final}} = A^{\text{pruned}}$). Otherwise, set threshold $\tau_{\text{rec}} = \lambda_2 \cdot \bar{S}_{\text{surv}}$ with $\lambda_2 = 50.0$.
3. **Iterative Addition with Acyclicity Guard:** Sort all candidate pairs $(u, v) \in \mathcal{E}_{\text{rec}}$ descending by parent score S_{uv}. For each candidate pair:
   - If $S_{uv} \geq \tau_{\text{rec}}$: Tentatively set $A_{uv} \leftarrow 1$. Test DAG acyclicity via depth-first cycle detection. If a cycle is formed, revert $A_{uv} \leftarrow 0$; otherwise retain $A_{uv} = 1$.

Because candidate pairs are drawn exclusively from A^{init}, recovery respects the causal order π ($u \prec_\pi v$). Combining Phases 1–3 establishes the global hierarchical DAG inclusion:

$$\boxed{\mathcal{E}^{\text{pruned}} \subseteq \mathcal{E}^{\text{final}} \subseteq \mathcal{E}^{\text{init}} \quad \text{and} \quad \mathcal{E}^{\text{pruned}} \subseteq \mathcal{E}^{\text{screen}} \subseteq \mathcal{E}^{\text{init}}.} \tag{32}$$

A.4.4 Stage-by-Stage Structural Error Attribution

To decouple the interacting failure modes across the five pipeline phases, controlled stage-by-stage edge ledger instrumentation partitions all causal discovery errors into mutually exclusive categories: topological order inversions (FN_{order}), screening exclusions ($\text{FN}_{\text{screen}}$), nonparametric pruning deletions (FN_{prune}), and surviving false positives (FP). This diagnostic framework reveals that error generation is strictly topology-dependent:

Topology-Dependent Bottleneck Partitioning. Rather than a single universal failure point, causal discovery errors bifurcate into distinct structural regimes across benchmark graph topologies:

1. **Chains and Low-Degree Topologies (Phase 1 Bottleneck):** On unbranched chains and trees (e.g., F3 sine-chains), topological ordering inversions ($\mathrm{FN}_{\mathrm{order}}$) account for the vast majority of missed edges. The estimated diagonal Stein Hessians sometimes rank upstream variables as candidate sinks. This is an empirical ordering failure; under equal noise variances, the population identity in Equation 49 makes larger derivative energy more negative and does not itself explain premature sink selection. These exclusions are predominantly cross-tier false sink selections (accounting for 75%–100% of ordering errors across confirmatory runs), with occasional within-tier ordering exclusions occurring when co-batched non-sinks are peeled simultaneously (e.g., F3 seed 4). Crucially, raw LOMTO parent scores already contain sufficient signal to resolve the true graph: when an oracle topological order mask $A^{\mathrm{oracle_init}}$ is applied to unmodified raw scores S^{raw} (holding macro-tiers and scoring fixed), the pipeline achieves $F_1 = 1.000$ and $\mathrm{SHD} = 0$, with zero screening or pruning losses.
2. **Multi-Parent Nonlinear SEMs (Partitioned Ordering and Pruning):** In multi-parent nonlinear SEMs (e.g., SynER1 Gaussian Process SEMs), causal discovery errors are partitioned across upstream ordering and downstream pruning. In confirmatory runs (seeds 3 and 4), Phase 1 continuous ordering inversions account for 60%–80% of missed edges, while Phase 3 GAM pruning conservatism selectively deletes remaining available true parents (accounting for 20% of false negatives). When multiple parents jointly drive a child, explained target variance is partitioned, elevating individual partial smooth p-values into the range $p \in [0.011, 0.264]$, where the rigid significance threshold $\alpha = 0.001$ eliminates genuine parents. While relaxing the cutoff to $\alpha = 0.025$ rescues marginal parents ($p \in [0.01, 0.02]$), its effect on graph precision varies across seeds: preserving 100% precision on seed 3 while inflating false positives from 1 to 4 on seed 4 ($F_1 : 0.625 \to 0.600$, precision $0.833 \to 0.600$).
3. **Branched Collider Networks:** On branched structures with colliders, LOMTO-LC reliably preserves true collider parents when sink layers are correctly separated (achieving $F_1 = 1.000$ on Branch/Collider seed 3); however, ordering stochasticity can still induce upstream sink inversions on individual seeds (e.g., seed 4, $F_1 = 0.667, \mathrm{SHD} = 4.0$).

Screening Sensitivity and Local Coupling Allocation. Phase 2 candidate screening with relative divisor $\lambda_1 = 50.0$ retains the vast majority of order-compatible true edges, achieving 100% retention on chains and colliders and 85.7%–100% on multi-parent SEMs (e.g., SynER1 seed 4 lost 1 edge during screening), with a median score safety margin of $+50.5\%$ above threshold. Sweeping $\lambda_1 \in [5, 200]$ confirms that $\lambda_1 = 50.0$ balances sensitivity and candidate density: stricter cutoffs ($\lambda_1 \leq 10$) prematurely discard weak true collider parents, whereas looser cutoffs ($\lambda_1 \geq 100$) inflate candidate density by $+6.4\%$ with zero recall gain. Meanwhile, LOMTO-LC refines within-tier attribution via multi-band cross-blocks, rescuing true collider edges that static attribution misallocates (elevating Branch/Collider accuracy from $F_1 = 0.800$ to 1.000), whereas uniform allocation inflates thresholds by $15.6\times$.

Phase 3 Computational Profile and Evaluated Interventions. Phase 3 GAM pruning consumes 96.3%–98.9% of total pipeline wall-clock time on small benchmark fixtures ($d \leq 10, n = 256$). Profiling reveals that this monolithic measurement reflects the combined overhead of serial R process invocation, disk CSV I/O, spline basis construction, and GCV smoothing parameter optimization. To address this computational share and the fixed-α pruning conservatism in multi-parent SEMs, two prospective interventions were formally evaluated:

1. **Python B-spline Step-Down Model Selection:** A pure Python B-spline backward elimination candidate optimizing the Bayesian Information Criterion (BIC, $\gamma = 0$) was tested on paired development cases. While eliminating external R process overhead, step-down BIC reduced mean conditional true-parent retention on SynER1 development seeds from 0.817 (production GCV) to 0.717 (-10.0 percentage points) and introduced two additional surviving false positives on Branch/Collider seed 5. Because model selection penalized weak nonlinear partial effects under finite sample sizes, it was rejected before confirmation.

2. **Parallel Multi-Worker Child Execution:** Executing unchanged R/`mgcv` child regressions across two worker processes achieved bit-for-bit identical graphs but yielded modest warm Phase 3 speedups of only $1.04\times$ ($d = 3$) and $1.31\times$ ($d = 20$), held back by the fixed ≈ 0.28 s per-worker overhead of loading the R runtime and `mgcv` library.

Consequently, production COSMIC retains the sequential R/`mgcv` GCV pruner with partial-smooth cutoff $\alpha = 0.001$, and reserved confirmation holdout seeds (seeds 7–11) were preserved unconsumed.

Noise Heterogeneity and Feature Standardization. To evaluate sensitivity to violation of the homoscedastic error assumption, a controlled 12-cell noise experiment was conducted on SynER1 Gaussian Process SEMs ($d = 10, n = 256$, seeds 5 and 6) across three noise regimes: homogeneous additive Gaussian noise, moderate heterogeneous noise ($\sigma_i \in [0.5, 1.5]$), and severe heterogeneous noise ($\sigma_i \in [0.2, 2.5]$), evaluated both with raw features and with unit-variance empirical standardization. Under raw scaling, mean recovery recall degraded from 0.694 (homogeneous) to 0.194 (severe heterogeneous). Applying empirical feature standardization did not reverse this loss, yielding recall of 0.111 on the severe heterogeneous cells. Because observational scaling alters the relative magnitudes of empirical score functions without equalizing conditional innovation variances, standardization does not restore the homoscedastic ANM score condition.

Post-Pruning Recovery Limitations. In production COSMIC, the recovery threshold $\tau_{\text{rec}} = \lambda_2 \bar{S}_{\text{surv}}$ evaluates average score exclusively over surviving pruned edges ($\lambda_2 = 50.0$), yielding a threshold $7.2\times$ –$50.0\times$ higher than CaPS (which averaged over all matrix entries). Counterfactual simulations demonstrate that lowering τ_{rec} to CaPS levels is unviable: it recovers at most 0–2 true edges while admitting 8–10 false positives, severely degrading graph precision and SHD. Consequently, post-pruning threshold adjustments cannot rescue multi-parent GAM deletions without substantial false-positive inflation.

A.5 Interpreting LOMTO Parent Scores

CaPS motivates parent scoring through changes in expected curvature after a variable is omitted (Xu et al., 2024). For a set $\mathcal{S}$ of true sinks in a Gaussian ANM, Equation 49 and Theorem D.1 give the population identity

$$\mathbb{E}[H_{jj}^{(-\mathcal{S})}] - \mathbb{E}[H_{jj}] = \sum_{p \in \mathcal{S} \cap \text{ch}(j)} \frac{\mathbb{E}[(\partial_j f_p)^2]}{\sigma_p^2}, \quad j \notin \mathcal{S}. \tag{33}$$

The terms are nonnegative squared local derivatives, weighted by child noise precision. They are not generally the square of an average derivative: averaging before squaring can cancel positive and negative effects. Equation 33 is the tier extension of the single-child calculation; it identifies aggregate evidence, not its allocation among children.

A LOMTO omission starts from the full variable set, even when its tier was identified later in peeling. Such a tier need not consist of sinks in the full graph, so Equation 33 does not automatically apply to every static omission. More generally, under differentiation-under-the-integral and integrability conditions, the marginal score is the conditional mean of the joint score. Differentiating again gives

$$H_{jj}^{(-\mathcal{S})}(X_{-\mathcal{S}}) = \mathbb{E}[H_{jj}(X) \mid X_{-\mathcal{S}}] + \text{Var}(s_j(X) \mid X_{-\mathcal{S}}). \tag{34}$$

Consequently the population expected-curvature change is $\mathbb{E}[\text{Var}(s_j(X) \mid X_{-\mathcal{S}})] \geq 0$. This is the same differentiation identity underlying the missing-information decomposition of Louis (1982), here applied to retained data coordinates. It describes information lost through marginalization, which can extend beyond direct parent-child effects. COSMIC takes the absolute estimated change $\Delta_{j\to\mathcal{S}} = |\bar{H}_j^{(-\mathcal{S})} - \bar{H}_j|$; this retains a nonnegative magnitude but introduces an empirical noise floor. For non-parents where the population shift is zero, finite-sample fluctuation $D = \bar{H}_j^{(-\mathcal{S})} - \bar{H}_j \sim \mathcal{N}(0, \sigma_D^2)$ folds under absolute value into a strictly positive expectation $\mathbb{E}[|D|] = \sigma_D\sqrt{2/\pi} \approx 0.798\sigma_D > 0$. Because tier mass is conserved, this positive bias distributes across tier members. One-sided positive truncation $\max(0, D)$ does not eliminate this bias (retaining $\mathbb{E}[\max(0, D)] = \sigma_D/\sqrt{2\pi} \approx 0.399\sigma_D$); only variance-aware thresholding or shrinkage can suppress zero-mean sample fluctuations to zero.

For the spatial score $s = \nabla \log p$, integration by parts gives $-\mathbb{E}[H] = \mathbb{E}[ss^\top]$ when the boundary terms vanish (Li & Turner, 2018). This motivates the score-product coupling. The negative matrix used in the method estimates the expected Hessian; its negative is the corresponding Fisher information matrix. Symmetry provides no direction, and the cancellation example in Section C.1 shows why coupling magnitude alone is not an edge detector. The estimated total order supplies orientation.

Our local allocation combines the static weights $a_{j:}$ with local weights $u_{j:}$ through a confidence ρ_{jt}. This is a COSMIC construction, rather than a published statistical guarantee. Both allocations are nonnegative and sum to one, so their convex combination preserves the tier mass:

$$\sum_{p \in \mathcal{S}_t} \mathcal{P}_{j \to p} = \Delta_{j \to \mathcal{S}_t}. \tag{35}$$

It also bounds the change from static allocation:

$$\|\mathcal{P}_{j \to \mathcal{S}_t} - \Delta_{j \to \mathcal{S}_t} a_{j:}\|_1 \leq 2\rho_{jt} \Delta_{j \to \mathcal{S}_t}. \tag{36}$$

These follow directly from normalization and the triangle inequality. They hold on retained-parent-to-removed-tier blocks, all of which respect the tier order. Scores elsewhere keep their static allocation before the final total-order mask. Zero confidence retains the static allocation, but mass preservation cannot establish that the redistributed evidence reaches the correct children. The confidence is an agreement weight, not a calibrated probability that an edge exists.

A.6 Numerical stability versus score fidelity

The single/tri-band recovery ablation in Section 6.5 does not measure kernel spectra. Separate controlled development diagnostics used eight-variable sine chains, $n = 256$, seeds 0–2, shared float64 geometry, and the same three bandwidths. At $m = 100, \eta = 0.001$ in one example, the narrow-band landmark system had estimated condition number 4.21×10^4 but a dense-Nyström score norm of 5435, versus 87.9 for the full-kernel score. Wider bands had estimated conditions 3.44×10^8 and 7.46×10^{10} and were more sensitive to float32 arithmetic. Thus an accurately solved low-rank system can still estimate a poor score, while other bands suffer numerical error.

In those controls, full-landmark float64 Woodbury and direct float64 solves returned identical orders in 27 matched comparisons; this agreement did not imply that the orders were correct. To formalize this effect, decompose the landmark quadrature derivative vector $D_1 = D_{1,\|} + D_{1,\perp}$, where $D_{1,\|} \in \mathrm{col}(C)$ and $D_{1,\perp} \in \mathrm{null}(C^\top)$. Under the Woodbury regularized inverse, $(\widetilde{K} + \eta_G I_n) D_{1,\perp} = \eta_G D_{1,\perp}$, so $G = (\widetilde{K} + \eta_G I_n)^{-1} D_{1,\|} + \eta_G^{-1} D_{1,\perp}$. Because $(D_1)_{ij}$ in Equation 7 contains variable-coordinate weights $(z_{aj} - x_{ij})$, its columns generically have non-trivial projections onto $\mathrm{null}(C^\top)$. These off-subspace derivative components are unregularized by the kernel, scaling by $\eta_G^{-1} = 10^3$ in the score estimate G, while squaring G in $\widehat{H} = Q - G \odot G$ scales curvature errors by $\eta_G^{-2} = 10^6$, which can dominate the curvature statistic.

Four-cell error decomposition and off-subspace amplification. To isolate the exact mechanisms governing derivative error, controlled experiments cross dense and low-rank kernel representations against full and landmark derivative quadrature (2×2 factorial design). Across standard Gaussian, additive noise model, and sine-chain fixtures ($n = 256, m = 64, \eta = 10^{-3}$), the compact linear solves achieve machine-precision backward accuracy (relative residuals $< 10^{-13}$). However, because the landmark derivative right-hand sides D_1, D_2 are evaluated via landmark quadrature, their columns do not lie entirely within the m-dimensional column span $\mathrm{col}(K_{nm})$. Under the regularized Nyström inverse, any component $D_\perp \in \mathrm{null}(K_{nm}^\top)$ is annihilated from the projection term and scaled purely by the ridge reciprocal:

$$(\widetilde{K} + \eta I_n)^{-1} D = (\widetilde{K} + \eta I_n)^{-1} D_\| + \eta^{-1} D_\perp. \tag{37}$$

With $\eta = 10^{-3}$, $D_\perp$ is amplified by a factor of 10^3 in the score G, and its contribution is squared in $\widehat{H} = Q - G \odot G$, yielding error amplification up to $\eta^{-2} = 10^6$. Consequently, the low-rank estimator's normalized Hessian error relative to analytic population truth reaches $\sim 10^3$–10^4, even though the linear systems are solved accurately.

Bandwidth prefactors and effective regularization. Factoring out the scalar RBF kernel prefactor $a_s = 1/s$ as $K_s = a_s \bar{K}_s$ reveals that the effective ridge regularizer in the solve is $\lambda_s = \eta/a_s = \eta s$. Thus, under fixed ridge $\eta = 10^{-3}$, the narrow band ($s = s_0/2$) operates with half the effective regularization of the base band, magnifying its off-subspace derivative amplification. In the tri-band ensemble, the narrow band contributes disproportionately to raw Hessian variance, while wider bands exhibit higher landmark condition numbers (up to 10^{10}) and increased sensitivity to floating-point precision.

Subspace projection intervention and ordering robustness. To test whether eliminating off-subspace amplification improves causal discovery, an isolated research prototype evaluated thin-SVD projection of derivative right-hand sides onto $\mathrm{col}(K_{nm})$ prior to the compact solve. In held-out evaluations, this projection reduced normalized derivative errors by four to six orders of magnitude (from $\sim 10^4$ down to ~ 0.1, a $40{,}000\times$ to $800{,}000\times$ error reduction). Crucially, however, eliminating this derivative error did not improve downstream topological ordering (the number of edge-order violations on the chain benchmark remained identical). Macro-tier peeling depends on relative curvature rankings, so reducing absolute derivative error need not improve ordering. Equal edge-order violation counts do not establish that the complete rankings were unchanged. In accordance with pre-declared acceptance criteria requiring downstream ordering improvements, the production baseline estimator is retained.

Fixed-rank scaling guarantees versus statistical consistency. Capping the landmark budget at $m \leq 300$ guarantees that estimation-stage time and memory scale as $\mathcal{O}(n)$ for fixed dimension. This is strictly a computational upper bound. Existing Nyström consistency theorems in the literature require growing landmark counts ($m \to \infty$) and decaying regularization ($\eta \to 0$) to guarantee convergence to true derivatives. Capping m does not guarantee statistical consistency as $n \to \infty$, nor does it prove a universal non-zero error floor across all data distributions. The non-monotone screening performance observed in large-sample empirical sweeps reflects the combined interaction of fixed regularization, derivative quadrature, curvature ranking margins, macro-tier batching, and relative thresholding.

A.7 Deterministic Approximation and Peeling Stability

This section asks whether approximation preserves the decisions of a matched dense empirical Stein estimator. This is distinct from approximating population curvature and from identifying true causal sinks. The latter claims need additional statistical and model assumptions. All perturbation results below hold in exact arithmetic; the pilot evaluates them numerically with residual checks, not validated floating-point certificates.

A.7.1 From Stein Solve Error to Mean Curvature Error

Fix active observations $X \in \mathbb{R}^{n\times d}$, a bandwidth $s > 0$, and ridge $\eta > 0$. Let K be the dense RBF Gram matrix and $C = K_{nm}$, $W = K_{mm}$ its landmark blocks. Assume W is positive definite, as for distinct landmark locations in exact arithmetic, and set $\widetilde{K} = CW^{-1}C^\top$. For full and landmark derivative sums $D_j, \widetilde{D}_j \in \mathbb{R}^n$, write $A = K + \eta I$, $\widetilde{A} = \widetilde{K} + \eta I$, $u_j = A^{-1}D_j$, and $\tilde{u}_j = \widetilde{A}^{-1}\widetilde{D}_j$.

Lemma A.1 (Stein solve discrepancy)**.** *The exact identity and its Euclidean/spectral-norm bound are*

$$\tilde{u}_j - u_j = \widetilde{A}^{-1}(\widetilde{D}_j - D_j) + \widetilde{A}^{-1}(K - \widetilde{K})u_j, \tag{38}$$

$$\|\tilde{u}_j - u_j\|_2 \leq \eta^{-1}\|\widetilde{D}_j - D_j\|_2 + \eta^{-2}\|K - \widetilde{K}\|_2\|D_j\|_2. \tag{39}$$

Proof. Subtract the two solves and use $\widetilde{A}^{-1} - A^{-1} = \widetilde{A}^{-1}(K - \widetilde{K})A^{-1}$. Both Gram matrices are positive semidefinite, so both inverse norms are at most η^{-1}. The triangle inequality and submultiplicativity give the bound. □

As explained in Appendix A.6, the inverse multiplies derivative components in $\mathrm{null}(C^\top)$ by η^{-1}. Accurate linear solves therefore need not imply accurate scores. Applying Equation 39 separately to G and Q, and using $\|v\|_\infty \leq \|v\|_2$, supplies entrywise discrepancy bounds for the next lemma.

Lemma A.2 (Propagation to curvature). *If $\|\widetilde{G}_j - G_j\|_\infty \le B_{G,j}$ and $\|\widetilde{Q}_j - Q_j\|_\infty \le B_{Q,j}$, then*

$$\|\widetilde{H}_j - H_j\|_\infty \le B_{Q,j} + (2\|\widetilde{G}_j\|_\infty + B_{G,j})B_{G,j} \equiv B_{H,j}. \tag{40}$$

For separately fitted bands with weights $w = (1/4, 1/2, 1/4)$, let $\bar{h}, \tilde{h} \in \mathbb{R}^d$ be the sample-mean ensemble curvatures. Then

$$\|\tilde{h} - \bar{h}\|_\infty \le \max_j \sum_{b=1}^{3} w_b B_{H,b,j} \equiv B_h. \tag{41}$$

Proof. With $\delta_j = \widetilde{G}_j - G_j$, factor $\widetilde{G}_j^2 - G_j^2 = \delta_j(2\widetilde{G}_j - \delta_j)$ entrywise. Apply the triangle inequality to $H = Q - G \odot G$, then average over observations and bands. □

A.7.2 Branch-Specific Adaptive Peeling Stability

Let $s_0(h) \le \cdots \le s_{d-1}(h)$ be sorted curvatures, $g_j(h) = s_{j+1}(h) - s_j(h)$, and $R(h) = s_{d-1}(h) - s_0(h)$. Under $\|\tilde{h} - \bar{h}\|_\infty \le e$, each order statistic moves by at most e, so each gap and the range move by at most $2e$. For $d \ge 3$, define $J = \{\lfloor d/2 \rfloor, \ldots, d-2\}$, let j^* maximize $g_j(\bar{h})$ over J, and set $g^* = g_{j^*}(\bar{h})$ and $\Psi = g^* - 0.15R(\bar{h})$. The reference batch size follows the complete production rule:

$$b^* = \begin{cases} \lceil d/4 \rceil, & R(\bar{h}) \le 10^{-6}, \\ \max(1, \min(d-1-j^*, \lceil 0.4d \rceil, d-1)), & R(\bar{h}) > 10^{-6},\ \Psi \ge 0, \\ \lceil d/4 \rceil, & R(\bar{h}) > 10^{-6},\ \Psi < 0. \end{cases} \tag{42}$$

The rule removes the b^* largest coordinates in descending curvature order. For $d = 2$ it removes the argmax; the last remaining node is appended directly.

Proposition A.3 (Adaptive peeling tier stability). *Suppose $\|\tilde{h} - \bar{h}\|_\infty \le e$. For $d = 2$, $|\bar{h}_1 - \bar{h}_0| > 2e$ guarantees the same removed node. For $d \ge 3$, the removed sets agree if:*

1. *$|R(\bar{h}) - 10^{-6}| > 2e$;*
2. *only when $R(\bar{h}) > 10^{-6}$, $|\Psi| > 2.30e$ and, if $|J| \ge 2$, $g^* - \max_{j \in J \setminus \{j^*\}} g_j(\bar{h}) > 4e$;*
3. *at the actual boundary from Equation 42, $\Delta_{\text{boundary}} = s_{d-b^*}(\bar{h}) - s_{d-b^*-1}(\bar{h}) > 2e$.*

For a multi-node batch, its ordered sequence also agrees if every adjacent gap within the selected batch exceeds $2e$.

Proof. For $d = 2$, the signed difference changes by at most $2e$, preserving either sign and hence the argmax. For $d \ge 3$, the first margin preserves the range-guard branch. In the low-range branch, both batch sizes are $\lceil d/4 \rceil$ without any gap-search condition. Otherwise, each upper gap changes by at most $2e$, so the $4e$ competitor margin preserves j^*. The statistic Ψ changes by at most $2e + 0.15(2e) = 2.30e$, preserving the threshold decision and therefore the capped or fallback batch size. At this common size, any selected coordinate remains above any retained coordinate because their difference is at least $\Delta_{\text{boundary}} - 2e > 0$. The selected sets thus coincide. Applying the same argument to adjacent selected coordinates proves sequence agreement under the internal-gap condition. □

The cutoff must be checked after the range fallback and any batch cap: the winning search gap can differ from the actual selection boundary. For example, $\bar{h} = 10^{-8}(0, 1, 2, 3, 4, 4.001, 4.002, 9)$ has a large final gap relative to $e = 10^{-10}$, but its range triggers a two-node fallback. Increasing coordinate 5 by e and decreasing coordinate 6 by e changes the removed set from $\{7, 6\}$ to $\{7, 5\}$ (zero-based indices). The actual boundary gap is $10^{-11} < 2e$, so the proposition correctly declines to certify this case.

A complete peeling trajectory is preserved only if these conditions hold at every shared active set, with matched reference geometry and estimator settings at each step. Induction then preserves the active sets; internal separation additionally preserves the total removal sequence. Agreement with this empirical reference does not establish a valid causal order. We do not derive a default LOMTO-LC screening or final-DAG guarantee from these curvature bounds.

A.7.3 Pilot Coverage and Interpretation

The pilot contains 120 runs: 24 valid fixture/sample/budget configurations, each with seeds $\{0, 1, 3, 4, 5\}$ (seed 2 held out). It uses Gaussian (F1), nonlinear ANM chain (F2), and sine-chain (F3) fixtures, $n \in \{128, 256, 512\}$, and $m \in \{32, 64, 128\}$ with $m \leq n/2$. Both regularizers are 10^{-3} and the initial dense and approximate estimators use matched bandwidths.

Table 1: Selected configurations from the 120-run pilot (24 configurations, five seeds each). Discrepancies $e_h^{\text{meas}} = \|\tilde{h} - \bar{h}\|_\infty$ and theoretical bounds B_h are reported as medians across seeds. Step-0 Set counts initial tier-set matches with the dense reference. Certification coverage denotes the fraction of runs satisfying Proposition A.3.

Fixture	n	m	**Median** e_h^{meas}	**Median** B_h	**Step-0 Set**	**Cert.**
F1 (Gaussian $d = 6$)	128	64	5.17×10^3	1.11×10^{14}	0/5	0.0%
F1 (Gaussian $d = 6$)	256	128	1.96×10^3	3.03×10^{14}	0/5	0.0%
F1 (Gaussian $d = 6$)	512	128	1.66×10^4	4.58×10^{16}	0/5	0.0%
F2 (ANM chain $d = 5$)	128	64	1.93×10^3	3.36×10^{13}	1/5	0.0%
F2 (ANM chain $d = 5$)	256	64	1.45×10^4	2.69×10^{15}	3/5	0.0%
F2 (ANM chain $d = 5$)	256	128	4.07×10^2	5.72×10^{13}	2/5	0.0%
F2 (ANM chain $d = 5$)	512	128	4.38×10^3	6.04×10^{15}	2/5	0.0%
F3 (Sine-chain $d = 8$)	128	64	5.31×10^3	1.32×10^{13}	1/5	0.0%
F3 (Sine-chain $d = 8$)	256	128	1.74×10^3	6.14×10^{13}	1/5	0.0%
F3 (Sine-chain $d = 8$)	512	128	1.06×10^4	7.10×10^{15}	1/5	0.0%

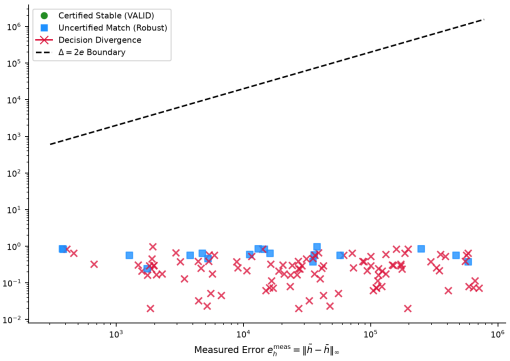

Figure 6: Reference boundary cutoff gap Δ_{boundary} versus measured error e_h^{meas}. All runs lie below the $\Delta = 2e$ certification boundary.

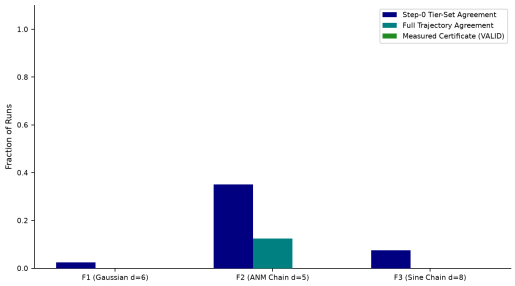

Figure 7: Decision agreement versus certification coverage across fixtures F1, F2, and F3.

Across the pilot, measured discrepancies range from 3.77×10^2 to 7.08×10^5, whereas the evaluated bounds range from 1.04×10^{13} to 2.32×10^{19} and exceed measured error in all runs. The initial boundary gaps are below 1, so no run satisfies the boundary condition, whether e is the measured error or the theoretical bound. The corrected proposition therefore retains zero certification coverage for this pilot.

Nevertheless, the initial tier sets agree in $18/120$ runs, their ordered sequences in $17/120$, and complete ordered-tier trajectories in $5/120$. Failure of these sufficient conditions does not imply that decisions must differ. The result establishes that these bounds do not certify dense-reference tracking in the tested settings; it neither rules out sharper analyses nor establishes or rules out causal recovery. Population approximation and causal validity require separate evidence, and the main experiments evaluate the integrated pipeline empirically.

B BUDGETING AND SAMPLE LIMITS

The dimension rule is intended to avoid paying for landmarks that offer little additional recovery, while retaining enough representation for the active variables. The sweeps compare historical policies and motivate the current bounded rule; they do not prove a unique optimal knee or zero accuracy loss beyond $10d$.

In Figure 9, at $n = 5000, d = 5$, the recorded $10d$ policy reaches screened $F_1 = 0.708$ in 0.184 seconds, compared with 0.453 in 2.688 seconds for $N/10$. In Figure 8, at $n = 1000, d = 20$, however, $10d$ gives 0.239 in 0.435 seconds, whereas $N/10$ gives 0.246 in 0.216 seconds. Thus a larger budget need not improve accuracy, and $10d$ is not uniformly the best policy. The sample-size sweep in Figure 10 likewise shows the historical adaptive policy outperforming unrestricted FPS at $n = 5000$ ($F_1 = 0.453$ versus 0.225, 0.433 versus 167.70 seconds). The legend's historical adaptive policy must not be silently equated with the current capped rule. Figure 11 complements these policy comparisons with the realized landmark and tier counts.

B.1 CHOOSING LANDMARKS AND THEIR BUDGET

The landmark count controls the size of each Stein solve. We choose this budget using the number of active variables, while using sample size as an upper-bound safeguard. This separates the desired count from the geometric procedure used to place landmarks. Nyström approximation quality also depends on the kernel spectrum and regularization; the budget is an empirical design choice rather than a universal rank or accuracy bound.

B.1.1 DIMENSION-BASED BUDGET.

For a fit with $n \geq 2$ observations and q active variables, define

$$\begin{aligned} u_n &= \min(300, \lfloor n/2 \rfloor)\,, \\ m_{\text{target}} &= \min(u_n, \max(25, 10q))\,. \end{aligned} \tag{43}$$

Initially $q = d$; during peeling and tier omissions, q is the number of columns in the current fit. The factor $10q$ increases the budget with dimension, the ceiling of 300 limits solve size, and the half-sample ceiling takes priority over the nominal minimum of 25. Thus the intended count never exceeds $\lfloor n/2 \rfloor$, including when $n < 50$. For example, $(n, q) = (1000, 5)$ gives 50 landmarks, $(1000, 50)$ gives 300, and $(100, 50)$ gives 50. The last case prevents a high-dimensional, small-sample fit from allocating nearly every observation as a landmark. A fit with fewer than two observations is outside this rule's scope.

We also evaluate historical and alternative sample-size scaling policies in Appendix J. The historical `old-adaptive` policy, inspired by Rudi et al. (2015), sets $m = \text{clip}(\lfloor 2.5\sqrt{n}\log_{10}(\max(10, n)) \rfloor, \min(25, n), \min(300, n))$. The `sqrt(N)ln(N)` policy evaluates $m = \max(25, \lfloor \sqrt{n}\ln(n) \rfloor)$, and fixed-fraction rules evaluate $m = \max(25, \lfloor n/10 \rfloor)$ and $m = \max(25, \lfloor n/5 \rfloor)$. Internal sweeps motivated the current dimension-dependent budget ($10q$) for low-dimensional graphs with a fixed ceiling for larger ones. More landmarks increased runtime and sometimes reduced recovery accuracy in those studies. The constants 10, 25, 300, and the half-sample cap are practical choices to be assessed experimentally, not guarantees of optimality.

B.1.2 AVOIDING EXCESSIVE LANDMARK COUNTS.

Coverage-only FPS can continue selecting points until most or all observations are landmarks if its distance threshold remains unmet. This removes the low-rank computational advantage and can make the Woodbury evaluation numerically sensitive. If all observations are used in the same order, $C = W = K$ and the compact system becomes $M_\eta = K^2 + \eta K$. Its eigenvalues are $\lambda_i(K)(\lambda_i(K) + \eta)$, so small kernel eigenvalues can make this system poorly conditioned. The subsequent subtraction in the score formula can further amplify rounding error, especially in single precision. This explains a possible mechanism for the degradation recorded in the FPS study; it does not imply that every large landmark set must fail.

Equation 43 bounds the target away from the full sample and prevents it from growing without limit as n or q increases. This is a guarantee on the requested count, not on the condition number or the

accuracy of the resulting Hessian. Numerical stability still depends on landmark locations, bandwidth, regularization, and precision. The adaptive formula determines the landmark count, while FPS selects the landmark locations.

B.1.3 CLUSTERING AND LEVERAGE SAMPLING.

Using k-means centroids emphasizes average squared reconstruction error. Zhang et al. (2008) relate landmark encoding error to Nyström approximation and motivate clustering-based landmarks. COSMIC's option uses centroids as kernel landmarks and cluster populations as derivative-sum weights. This makes the quadrature reflect cluster size, although replacing observations by centroids still approximates their nonlinear kernel derivatives.

The kernel spectrum is the set of its eigenvalues and indicates how strongly different patterns are represented. Ridge leverage scores instead measure a point's contribution to a regularized kernel representation. For a Gram matrix K and ridge $\lambda > 0$, define

$$\ell_i(\lambda) = [K(K + \lambda I)^{-1}]_{ii}, \qquad d_{\text{eff}}(\lambda) = \sum_i \ell_i(\lambda). \tag{44}$$

The effective dimension d_{eff} counts represented kernel directions after regularization, allowing fractional contributions. It is different from the number d of observed variables. Sampling with these scores can give budgets governed by effective dimension under the conditions of El Alaoui & Mahoney (2015). COSMIC approximates them using an FPS pilot at the base bandwidth, then samples landmarks without replacement. For the leverage-sampling alternative, it uses a separate uniform sample for derivative quadrature, so the importance-sampled basis is not also treated as a uniform empirical measure. This pilot approximation and the shared choice across bands do not automatically satisfy every hypothesis of the published sampling guarantees.

The three selectors consequently address different objectives: FPS controls sample coverage, k-means summarizes average geometry, and leverage sampling targets the regularized kernel representation. None alone establishes accurate causal ordering. We treat the landmark budget and the number of macro-tiers as separate design parameters, with no assumed relationship between them.

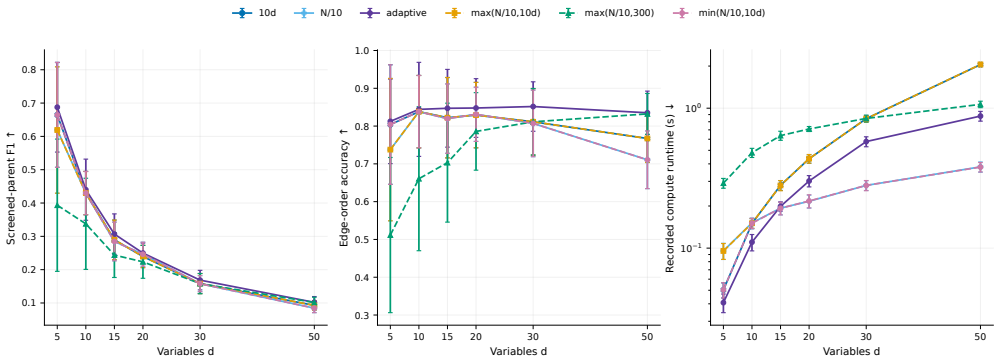

Figure 8: Landmark-policy comparison across dimensions at fixed $n = 1000$. Screening quality, ordering accuracy, and compute time use successful-run means with sample standard deviations. The two sample-size regimes are kept separate.

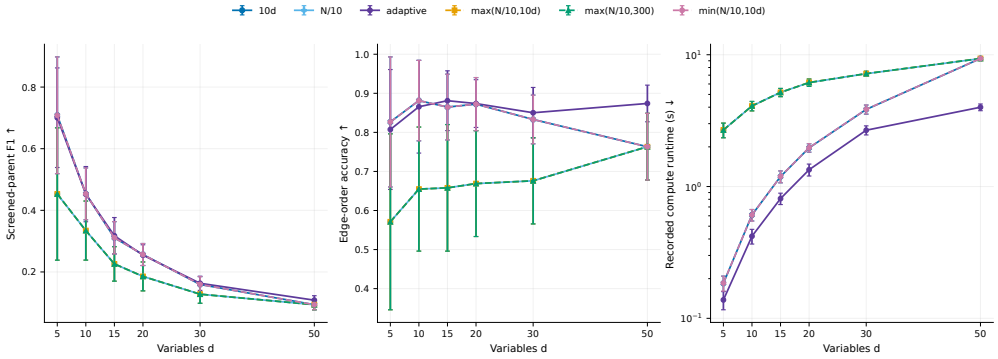

Figure 9: Landmark-policy comparison across dimensions at fixed $n = 5000$. Screening quality, ordering accuracy, and compute time use successful-run means with sample standard deviations. The two sample-size regimes are kept separate.

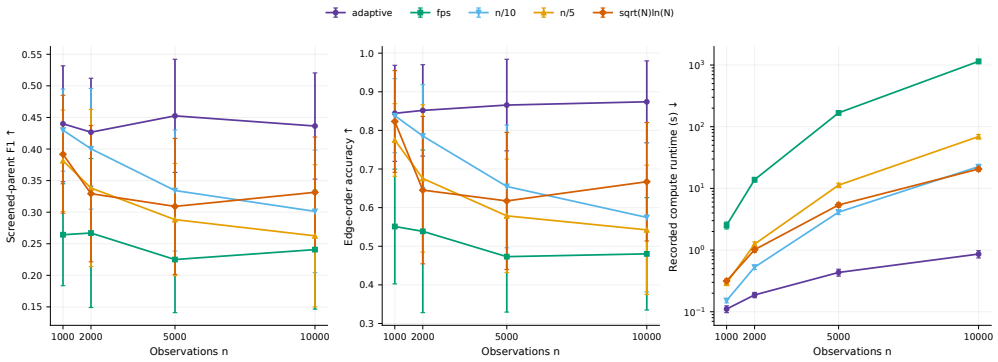

Figure 10: Sample-size sweep of the recorded landmark policies. Quality and runtime need not improve together as the requested rank grows; labels describe the historical configurations.

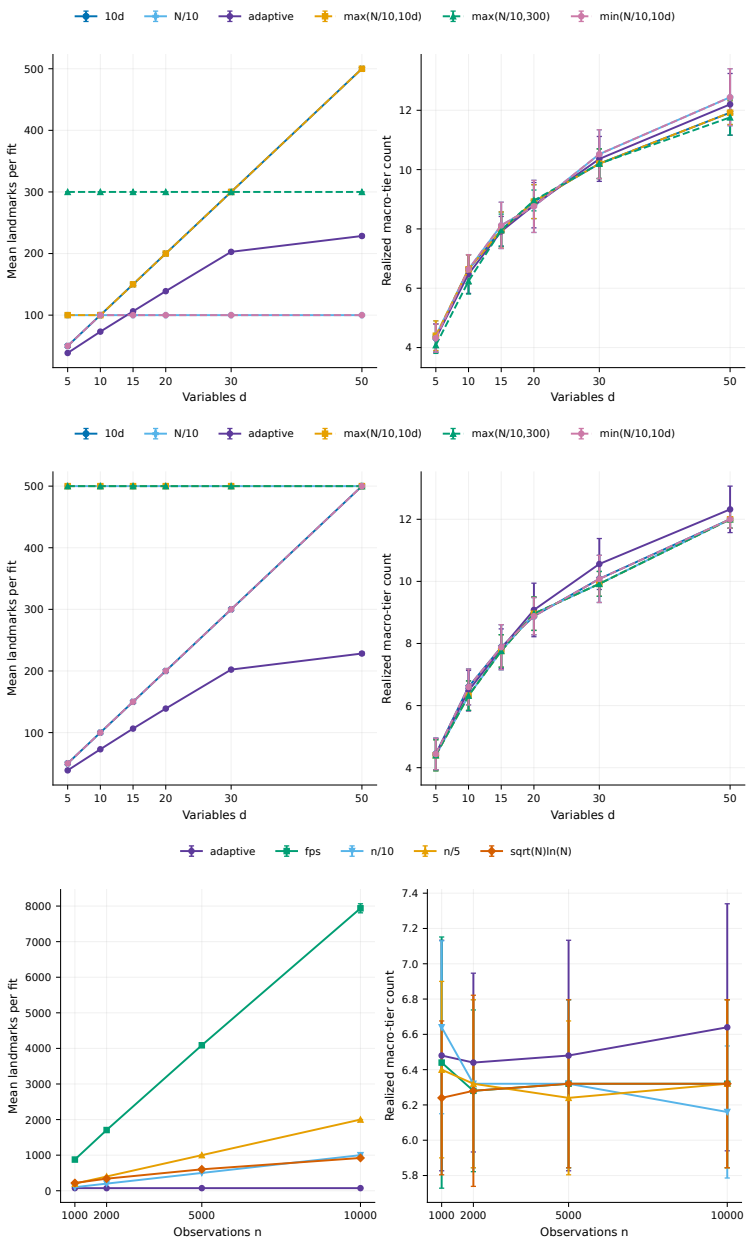

Figure 11: Realized landmark and macro-tier counts: the top row uses $n = 1000$, the middle row $n = 5000$, and the bottom row varies n. Requested policy labels and realized per-fit counts describe different quantities.

C ADAPTIVE HEURISTICS

C.1 ESTIMATING MACRO-TIERS IN PRACTICE

The ideal selector in Theorem D.3 requires a reliable distinction between sinks and nonsinks. Practical selectors must infer that distinction from estimated statistics. Fixed batching avoids estimating the number of sinks, but can place a parent and its child in the same tier. Adaptive batching tries to find a boundary in the statistics which does not necessarily coincide with a causal boundary.

C.1.1 FIXED BATCHES AND HESSIAN GAPS.

The fixed schedule divides the remaining variables across a prescribed number of passes: for step $\in \{k, k-1, \ldots, 2\}$, the scheduled batch size is $b_t = \min(\max(1, \lceil d_{\text{rem}}/\text{step} \rceil), d_{\text{rem}} - 1)$. If $d_{\text{rem}} > 1$ after the loop finishes, all remaining variables are bundled into a single final macro-tier; if $d_{\text{rem}} = 1$, the remaining node forms a singleton tier. When $k \geq d$, this collapses to single-variable sequential peeling ($K = d$); when $k = 1$, it yields a single dense macro-tier ($K = 1$).

The adaptive rule sorts active variables ascending by sample-mean diagonal Hessian, $h_{(0)} \leq h_{(1)} \leq \ldots \leq h_{(d_{\text{rem}}-1)}$ where $h_j = \frac{1}{n}\sum_{i=1}^{n} \widehat{H}_{ij}$, and searches for the largest adjacent gap $\delta_\alpha = h_{(\alpha+1)} - h_{(\alpha)}$ strictly in the upper half of sorted values ($\alpha \in \{\lfloor d_{\text{rem}}/2 \rfloor, \ldots, d_{\text{rem}} - 2\}$). Gaps below the median are excluded. An early version accepted a gap exceeding 1.3 times the median adjacent gap; our development notes record that a near-zero median made this rule accept small fluctuations and produce many small tiers.

The current rule evaluates gaps relative to the full score range $\mathcal{R} = h_{(d_{\text{rem}}-1)} - h_{(0)}$. A gap qualifies as a macro-tier boundary if it spans at least 15% of this dynamic range:

$$\frac{\delta_{j^*}}{\mathcal{R}} = \frac{h_{(j^*+1)} - h_{(j^*)}}{h_{(d_{\text{rem}}-1)} - h_{(0)}} \geq 0.15 \quad \text{with } \mathcal{R} > 10^{-6}. \tag{45}$$

If a qualifying gap j^* is found, the candidate batch size is $b_{\text{cand}} = d_{\text{rem}} - 1 - j^*$, subject to a 40% batch cap:

$$b_t = \max\left(1, \min\left(b_{\text{cand}}, \lceil 0.4 d_{\text{rem}} \rceil, d_{\text{rem}} - 1\right)\right). \tag{46}$$

Post-hoc cap truncation and cluster slicing. Importantly, the 40% batch cap does not constrain the candidate gap search index range a priori: argmax searches unconstrained across all upper gaps $\alpha \geq \lfloor d_{\text{rem}}/2 \rfloor$. When a dominant gap occurs near the median such that $b_{\text{cand}} > \lceil 0.4 d_{\text{rem}} \rceil$, the algorithm does not reject the gap or search for an alternate cut; instead, it clamps $b_t = \lceil 0.4 d_{\text{rem}} \rceil$ post-hoc. Consequently, the actual tier cut is forced at index $d_{\text{rem}} - \lceil 0.4 d_{\text{rem}} \rceil$, which severs the cluster above the gap, slicing right through variables with identical or near-identical curvatures ($\delta \approx 0$). If no qualifying gap exists, or if $\mathcal{R} \leq 10^{-6}$, the procedure falls back to peeling the top $b_t = \max(1, \lceil d_{\text{rem}}/4 \rceil)$ candidates. When $d_{\text{rem}} = 2$, COSMIC bypasses gap search and directly peels $\text{argmax}(\bar{H})$ as a single-variable tier; when $d_{\text{rem}} = 1$, the remaining node is appended as the terminal singleton.

Reversal and intra-tier ordering asymmetry. Within each peeled tier, candidate sinks enter the removal accumulator in descending curvature order (largest first). Reversing the accumulator sequence at termination yields the causal topological order $\pi = (\pi_0, \ldots, \pi_{d-1})$, which places variables in *ascending* curvature order within each tier (smallest curvature = earliest = parent). In contrast, reversing the tier partition list $\mathcal{T} = \text{reversed}(\mathcal{S})$ reverses only the outer tier order (root tier first, sink tier last), while preserving descending curvature ordering within each tier block. In the initial tournament DAG $A^{\text{init}}_{ji} = 1 \iff j \prec_\pi i$, intra-tier edges are directed strictly from smaller to larger curvature.

C.1.2 ALTERNATIVE SELECTORS.

Our exploratory work considered clustering mean or variance statistics, relative and ratio thresholds, bootstrap variability estimates, and compatibility checks between candidates. We also considered Otsu thresholding, which maximizes between-group variance (Otsu, 1979). An optimum for this one-dimensional objective does not certify two causal groups. Likewise, clustering can force a split

when sink and nonsink statistics overlap, while a conservative ratio threshold can leave almost every tier as a singleton.

The recorded bootstrap pilots and sequential-support checks did not establish a generally better selector. Bootstrap variation does not remove a persistent estimator bias. A projected score-matching alternative was also tested to improve calibration of the derivative statistics. In the recorded held-out comparisons, better agreement with the reference partition in some settings came with worse cross-tier edge ordering; the candidate did not meet the promotion criterion. These are observations from our development studies, not guarantees supplied by the cited literature.

C.1.3 WHY COUPLING CHECKS ARE INSUFFICIENT.

Mean Fisher coupling can cancel even along a direct nonlinear edge. For the illustrative model $X \sim \mathcal{N}(0,1)$, $Y = X^2 + \epsilon$, with independent $\epsilon \sim \mathcal{N}(0,1)$, direct differentiation gives $H_{XY} = 2X$. Thus $\mathbb{E}[H_{XY}] = 0$ despite the edge. A root-mean-square (RMS) cross-derivative, which squares values before averaging and taking the square root, retains this signal, but its finite-sample threshold must distinguish estimation noise from weak dependence. The recorded RMS pilot became conservative and produced many small tiers. Moreover, pairwise compatibility alone does not establish that candidates have no children among the retained variables.

Equation 49 supplies another limitation: unequal noise variances can obscure the sink boundary, and linear Gaussian mechanisms eliminate the variance signal. These population effects cannot be resolved solely by changing a gap threshold. We therefore retain batching as a computational approximation. Within a selected tier, the implementation orders candidates by their estimated mean Hessian before reversing the removal sequence. Exact ties follow the numerical sorting convention. This produces a total order, but cannot certify the direction of an edge inside an incorrectly merged tier. Neither a fixed tier count nor any tested adaptive rule is claimed to recover the ideal layers universally.

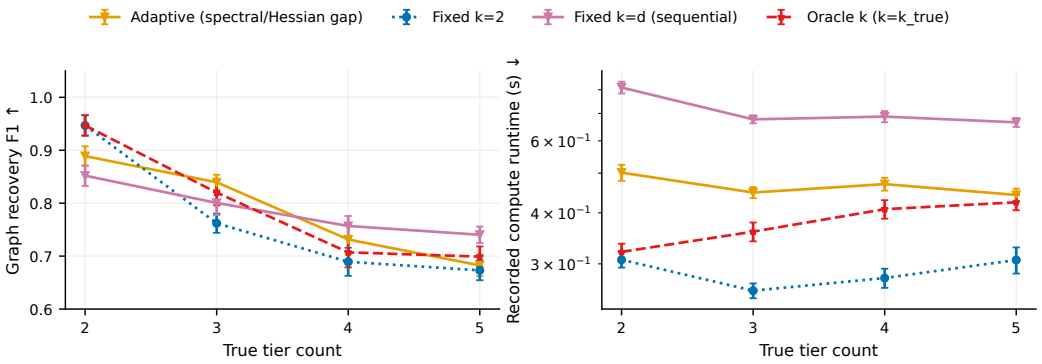

Figure 12: Adaptive versus fixed and count-informed tier policies on synthetic graphs with $d = 12$, $n = 1000$, and two to five generating tiers. Curves show final-graph F_1 and total runtime over 25 seeds; error bars are standard errors of the mean. "Oracle k" supplies the generating tier count, not the correct partition or internal order.

Adaptive macro-tier selection. Figure 12 evaluates this trade-off: adaptive selection takes 0.44–0.50 seconds across the four settings, compared with 0.67–0.81 seconds for sequential peeling. Its recovery relative to sequential peeling depends on the graph: mean F_1 improves from 0.852 to 0.889 with two generating tiers and from 0.801 to 0.839 with three, but decreases from 0.757 to 0.731 with four and from 0.740 to 0.683 with five. The count-informed policy reaches 0.947 with two tiers, while adaptive selection reaches 0.889; at three tiers, adaptive selection instead has the higher mean (0.839 versus 0.820). Supplying a tier count is therefore neither a guarantee of optimal recovery nor an accuracy upper bound. The evidence supports adaptive batching as a runtime-saving policy whose recovery cost depends on structure, rather than as an established exact tier selector.

D PROOFS AND COMPUTATIONAL COMPLEXITY

The next results give the full argument behind Theorem 5.1 and distinguish ideal sink identification from practical fixed and adaptive batches.

D.1 MACRO-TIER PEELING UNDER ANMS

The theoretical question is whether removing several correctly identified sinks changes the causal model of the retained variables. We extend the leaf-removal argument underlying SCORE to sets of sinks (Rolland et al., 2022). Throughout this subsection, H denotes a population Hessian, rather than a finite-sample estimate.

Theorem D.1 (Removal of a terminal block)**.** *Let X follow an acyclic ANM with mutually independent noises and observational density $p(x) = \prod_{j\in\mathcal{V}} p_j(x_j \mid x_{\mathrm{pa}(j)})$. Let $\mathcal{B} \subseteq \mathcal{V}$ be a nonempty block of variables and write $\mathcal{A} = \mathcal{V} \setminus \mathcal{B}$. Suppose no directed edge points from $\mathcal{B}$ into $\mathcal{A}$; equivalently, $\mathcal{B}$ contains every descendant of each of its members in the current DAG ($\mathrm{de}_G(\mathcal{B}) \subseteq \mathcal{B}$). Then $X_\mathcal{A}$ follows the ANM induced by deleting $\mathcal{B}$, with unchanged structural functions f_j, unchanged parent sets $\mathrm{pa}(j)$, and mutually independent noise distributions for all retained nodes $j \in \mathcal{A}$. Edges within $\mathcal{B}$ are permitted. In particular, the result applies to any subset of current sinks.*

Proof. Every parent of a node in $\mathcal{A}$ also belongs to $\mathcal{A}$. Thus each retained structural equation involves only retained variables and its original noise, and these noises remain mutually independent. To verify the corresponding marginal factorization, write

$$p(x_\mathcal{A}) = \prod_{j\in\mathcal{A}} p_j(x_j \mid x_{\mathrm{pa}(j)}) \int \prod_{b\in\mathcal{B}} p_b(x_b \mid x_{\mathrm{pa}(b)})\, dx_\mathcal{B} = \prod_{j\in\mathcal{A}} p_j(x_j \mid x_{\mathrm{pa}(j)}). \tag{47}$$

The integral equals one by integrating the variables in $\mathcal{B}$ in reverse topological order. At each integration step, the variable has no remaining child in $\mathcal{B}$ and no child in $\mathcal{A}$, so it occurs only in its own conditional density, which integrates to one. Marginalizing $\mathcal{B}$ leaves the structural noise independence intact, inducing no unobserved confounding or selection bias among $\mathcal{A}$. A set of current sinks has no outgoing edges at all and therefore trivially satisfies this condition. □

Corollary D.2 (Ordering by terminal blocks)**.** *Suppose successive nonempty blocks $\mathcal{B}_1, \ldots, \mathcal{B}_K$ partition the variables, each block has no edge to the retained variables in the active DAG at its removal step, and a valid internal topological order $\pi_{\mathcal{B}_t}$ is recovered within each block. Reversing the sequence of removed blocks, while preserving each block's internal topological order, yields a valid topological order $\pi = (\pi_{\mathcal{B}_K}, \ldots, \pi_{\mathcal{B}_1})$ of the original DAG G. If each block consists exclusively of current sinks, any internal order is valid.*

Proof. Theorem D.1 preserves the induced ANM after every removal. For any edge joining distinct blocks, its source cannot be removed while its target remains, so reversing the block sequence places the source before the target. Edges within a block are respected by its assumed internal topological order. A block of current sinks has no internal edges, so every internal order is valid in that case. □

Scope and limitations of the guarantee. Theorem D.1 establishes the closure of ANMs under terminal marginalization: integrating out sinks or descendant-closed blocks leaves upstream causal mechanisms untouched. However, this theoretical result must not be conflated with a correctness guarantee for COSMIC's empirical pipeline:

1. **Heuristic tier selection:** The theorem assumes that an estimated batch $\mathcal{S}_t$ is truly terminal. Practical estimators infer tier boundaries from sample-mean Stein curvatures $\bar{H}$. If estimation error or weak mechanism gradients cause an upstream node to be co-peeled while its child remains, the terminal condition fails. The original retained-equation guarantee no longer applies, and marginalization can induce confounding or alter the model class; these outcomes are not inevitable for every nonterminal removal.
2. **Intra-tier ordering errors:** For connected terminal blocks (e.g., removing $\{X_3, X_4\}$ in a chain $X_1 \to X_2 \to X_3 \to X_4$), correctness strictly requires ordering $X_3 \prec X_4$ internally. Empirical curvature sorting within a batch is heuristic and vulnerable to noise inversions.

3. **Screening blindness on co-batched edges:** LOMTO counterfactual omission evaluates tier-wide shifts $\Delta_{j\to\mathcal{S}_t}$, leaving intra-tier perturbation undefined. COSMIC sets intra-tier candidate scores identically to 0.0. Consequently, if an edge exists within an estimated tier, it is eliminated during Phase 2 screening unless the maximum candidate score is at most 10^{-8}.

For identifying the sinks, assume the nonlinear Gaussian ANM conditions of SCORE: independent Gaussian noises with positive variances and twice continuously differentiable mechanisms nonlinear in every parent coordinate. Assume also finite Hessian variances. SCORE's Lemma 1 then characterizes sinks by $\mathrm{Var}(H_{jj}(X)) = 0$ (Rolland et al., 2022).

Theorem D.3 (Ideal macro-tier recovery)**.** *Under these conditions, at each step select all active variables satisfying* $\mathrm{Var}(H_{jj}(X)) = 0$*, and recompute the population Hessian after their removal. This procedure recovers the successive sink layers of the DAG. Reversing the layer sequence and choosing any order within each layer yields a valid topological order.*

Proof. The SCORE criterion identifies exactly the sinks at the first step. Theorem D.1 leaves a model with the same assumptions on the retained equations. The criterion therefore applies again. Induction gives the successive sink layers until the graph is empty. For every edge $i \to j$, node i cannot be removed while j remains. Thus j is removed in an earlier layer, and reversing the layers puts i before j. No edge lies within a sink layer. □

This is an ideal selector, with access to population quantities. A finite-sample version needs a separation condition. For example, write $\mathbb{V}_j = \mathrm{Var}(H_{jj})$ and suppose that, at every correct removal step, $|\widehat{\mathbb{V}}_j - \mathbb{V}_j| \le e$, while every nonsink has $\mathbb{V}_j \ge \Delta > 4e$. Selecting

$$\widehat{\mathbb{V}}_j \le \min_i \widehat{\mathbb{V}}_i + 2e \tag{48}$$

then selects exactly the sinks: a sink lies within $2e$ of the estimated minimum, whereas a nonsink lies at least $\Delta - 2e > 2e$ above it. This sufficient condition is a consequence of the error bounds, not a bound proved for COSMIC's estimator.

COSMIC instead ranks expected Hessian diagonals, following CaPS (Xu et al., 2024). Under Gaussian ANMs and integrable derivatives,

$$\mathbb{E}[H_{jj}] = -\sigma_j^{-2} - \sum_{c\in\mathrm{ch}(j)} \sigma_c^{-2}\mathbb{E}[(\partial_j f_c)^2]. \tag{49}$$

Here $\mathrm{ch}(j)$ is the child set of j and σ_j^{-2} is the reciprocal noise variance (noise precision). Each child contributes a nonpositive term, making nonsinks more negative than sinks when noise variances are equal ($\sigma_j^2 \equiv \sigma^2$). In that homoscedastic setting, a positive derivative contribution for every nonsink ensures that sinks strictly maximize expected diagonal curvature.

We emphasize that equal error variance is a sufficient condition for score-based curvature ranking, not a necessary condition for causal identifiability: general nonlinear ANMs with mutually independent errors are generically identifiable from observational distributions without equal variances (Peters et al., 2014). Under unequal noise variances, however, the baseline precision term $-\sigma_j^{-2}$ varies across coordinates, which can cause an upstream root with large noise variance to exhibit less negative curvature than a downstream sink. Feature standardization ($X_j/\hat{\sigma}_{X_j}$) equalizes total marginal variance ($\mathrm{Var}(X_j) = 1$) rather than structural error variance σ_j^2; as shown in our benchmark sensitivity analysis (Appendix A.4.4), standardization did not repair ordering failures in the tested heterogeneous-noise cases. Scaling a variable rescales its signal and noise variances equally, leaving its signal-to-noise ratio unchanged; it changes relative noise variances and curvature scales across coordinates. Neither this expectation criterion nor the SCORE variance criterion justifies taking an arbitrary fixed number of high-ranking variables as simultaneous sinks. Our fixed and adaptive batching rules approximate the ideal partition and do not inherit Theorem D.3 without correct tier selection.

D.2 Computational Complexity

We give stage-by-stage arithmetic and working-memory bounds, with a conservative accounting for regression pruning. The three kernel bands multiply Stein evaluation work by a constant factor of 3,

suppressed below. Let m be the landmark budget, k the realized number of macro-tiers, n the sample size, and d the variable count.

Phase 1: Macro-Tier Peeling (Ordering). For a fit on $q \leq d$ active variables, Farthest Point Sampling (FPS) costs $\mathcal{O}(nmq)$. Constructing the $n \times m$ cross-kernel and evaluating landmark quadrature derivative sums adds $\mathcal{O}(nmq)$. Factoring the regularized $m \times m$ landmark Gram matrix costs $\mathcal{O}(m^3)$, and applying the factored systems to q coordinate columns costs $\mathcal{O}(nm^2 + m^2 q)$. The total cost per Stein solve is:

$$C_{\text{Stein}}(n, m, q) = \mathcal{O}(nm^2 + m^3 + nmq). \tag{50}$$

Summing over at most k peeling iterations on active subsets $d_t \leq d$ yields total Phase 1 ordering time $\mathcal{O}(kdm^2 n)$. Sequential execution stores one $n \times m$ kernel matrix and one $m \times m$ landmark factorization, requiring $\mathcal{O}(nm + m^2)$ working memory (plus $\mathcal{O}(nd)$ for data).

Phase 2: LOMTO-LC Parent Evidence Scoring. LOMTO requires one full-data reference solve and k independent tier-omission fits on subsets of size $d - |\mathcal{S}_t|$. These $k + 1$ Stein fits require $\mathcal{O}(kdm^2 n)$ time. Multi-scale local coupling (LOMTO-LC) computes cross-blocks $(G_{:j}^{(\ell,t)})^\top G_{:p}^{(\ell,t)}/n$ between retained candidates and tier members. Because each cross-block is evaluated once from cached Stein score matrices, local coupling requires $\mathcal{O}(nr_t b_t)$ work per tier. Summing over all tiers gives $\sum_t r_t b_t \leq d(d-1)/2$, contributing $\mathcal{O}(nd^2)$ operations, alongside $\mathcal{O}(nd^2)$ for the full-data static Fisher matrix.

Phase 2 (continued): Candidate Edge Pre-Screening. Pre-screening applies the relative threshold $S_{ji} \geq \frac{1}{\lambda_1} \max_k S_{ki}$ ($\lambda_1 = 50.0$) across the masked score matrix $S \in \mathbb{R}^{d \times d}$. This requires $\mathcal{O}(d^2)$ comparisons.

Pre-Pruning Complexity Summary. Combining Phases 1–2 establishes the pre-pruning budget:

$$\mathcal{O}\big(kdm^2 n + nd^2\big) \quad \text{time}, \qquad \mathcal{O}(nm + m^2 + nd + d^2) \quad \text{memory}. \tag{51}$$

For fixed d, the production budget $m \leq \min(300, \max(25, 10d))$ is bounded independently of n, ensuring strictly linear scaling $\mathcal{O}(n)$ in sample size.

Phase 3: Nonparametric GAM Pruning Exclusion. Candidate edges entering Phase 3 are pruned using penalized cubic spline GAMs (`mgcv::gam`). Let B_i be the total spline design dimension, including the intercept, for target i, and $B_{\max} = \max_i B_i$. With a fixed basis size per parent, $B_i = \mathcal{O}(1 + |\mathcal{P}_i|)$. Dense design-matrix factorization costs $\mathcal{O}(nB_i^2)$ and coefficient-space matrix operations cost $\mathcal{O}(B_i^3)$ per fitting step. If T bounds the number of such steps, a conservative linear-algebra budget is

$$C_{\text{GAM}} = \mathcal{O}\left(T \sum_{i=1}^{d} (nB_i^2 + B_i^3)\right) \subseteq \mathcal{O}\big(dT(nB_{\max}^2 + B_{\max}^3)\big)\,. \tag{52}$$

Here T denotes fitting iterations, including smoothing-parameter updates; it is distinct from the macro-tier notation $\mathcal{T}$. Factorization reuse can reduce this budget, while basis construction, R startup, and data transfer also contribute to measured runtime. In particular, fixed basis size per parent does not justify a bound linear in the total number of parents, because the smooths are fitted jointly. This regression-pruning stage does *not* share the low-rank $\mathcal{O}(kdm^2 n)$ Stein bound. In our empirical benchmarks, serial execution of R's `mgcv` consumes 96.3%–98.9% of pipeline wall-clock time and times out at $d = 200$ (1800 s). Large-dimension screening results concern Phases 1–2 only. Their total asymptotic bound remains quadratic in d because of dense pairwise coupling; near-linear measured runtime over a finite range does not establish a sub-quadratic bound.

D.3 Pre-pruning Comparison with CaPS, SCORE, DAS, and DiffAN

Here the pre-pruning comparison covers ordering for SCORE and DiffAN, and ordering plus Hessian-based parent screening for CaPS, DAS, and COSMIC. It includes COSMIC's Phases 1–2 (ordering and parent scoring), but excludes downstream regression-based pruning for every method. Thus the bounds describe different available outputs, not interchangeable final-DAG estimators.

SCORE and CaPS. SCORE removes one variable at a time, requiring $\mathcal{O}(d)$ dense Stein fits and $\mathcal{O}(dn^3)$ matrix work (Rolland et al., 2022). Including kernel derivatives and application of the solved systems to q active coordinates gives $\mathcal{O}(n^3 + n^2 q)$ per diagonal-Hessian fit, hence $\mathcal{O}(dn^3 + n^2 d^2)$ over the ordering. CaPS uses mean diagonal curvature for ordering and adds a full-data reference and variable-wise omission estimates for parent scores (Xu et al., 2024). Counting these separately gives at most $2d + 1$ fits; reusing a reference can reduce this constant. Its ordering-and-scoring cost remains $\mathcal{O}(dn^3 + n^2 d^2)$, with quadratic sample memory for each sequential dense solve. COSMIC replaces these dense systems with landmark systems and variable-wise fits with tier-wise fits; the benefit depends on the realized m and k.

DAS. DAS retains SCORE's sequential ordering and uses off-diagonal Hessian entries to screen candidate parents, rather than fitting one omitted-variable model per parent score (Montagna et al., 2023). Its published leading accounting is $\mathcal{O}(dn^3 + d^2)$ for dense ordering solves and candidate selection with a fixed parent cap. The d^2 term is a search cost, not the complete sample-dependent cost of constructing Hessian entries. As with SCORE's dn^3 expression, derivative evaluation and feature-dependent matrix operations must also be paid; this convention should not be confused with COSMIC's dimension-explicit bound above. No CAM or GAM regression-pruning cost is included here.

DiffAN. DiffAN trains a neural score model and uses automatic differentiation and deciduous score updates to recover an order (Sanchez et al., 2023). Its reported $\mathcal{O}(n + d^3)$ accounting assumes a fixed training schedule and an ordering-evaluation batch independent of n, suppressing network-operation costs. More explicitly, the training term is $\mathcal{O}(Enc_{\text{train}}(d))$ for E epochs and per-observation network training cost $c_{\text{train}}(d)$; ordering also depends on evaluation-batch size, architecture, and differentiation costs. The masking-only variant reports $\mathcal{O}(n + d^2)$ but omits deciduous corrections. Training is included in this comparison, and neither variant supplies the parent scores compared above.

Parallel parent scores and accelerator implementations. Once ordering fixes the omitted tiers and the full-data reference is available, the additional Hessian evaluations for LOMTO parent scores are independent across omissions and kernel bands. They can therefore be scheduled concurrently, as can CaPS's variable-wise omission fits. This reduces potential wall-clock time, not total arithmetic work; sequential peeling remains dependent on earlier removals, and concurrent fits increase memory use. A prospective JAX implementation could batch compatible tensor operations and compile them for GPU or TPU execution. Differing active dimensions may require grouping or padding, and numerical precision, data transfer, and compilation overhead must be assessed. Such software improvements are opportunities rather than speedups established by the present experiments.

E HARDWARE AND COMPUTE ENVIRONMENT SPECIFICATIONS

The empirical benchmark results, including the extended baselines and COSMIC variants in Appendix F, were obtained on Workstation A (GTX 1660). All other COSMIC experiments and scalability tests, including screening and end-to-end scalability, were run on Workstation B (RTX 4090). This attribution is supplied by the authors. A workstation's GPU inventory does not imply that every algorithm used its GPU. The following specifications describe the available workstations and their recorded computing environments.

E.1 WORKSTATION A: AMD RYZEN 5 3600 (GTX 1660)

- **Central Processing Unit (CPU):** AMD Ryzen 5 3600 6-Core Processor featuring 12 logical processing threads via Simultaneous Multithreading (SMT). Cache topology includes:
  - L1 Cache: 384 KiB total (6 physical cores $\times$ 32 KiB instruction + 32 KiB data);
  - L2 Cache: 3.0 MiB total (6 physical cores $\times$ 512 KiB dedicated per-core cache);
  - L3 Cache: 32 MiB unified high-speed cache distributed across two Core Complex (CCX) clusters.
- **System Memory (RAM):** 16 GB DDR4 dual-channel system memory (15.5 GiB physical address space recognized by the Linux kernel, with approximately 12.0 GiB available for user processes under standard background load). Swap memory consists of a 4.0 GB swap partition configured on high-throughput solid-state storage. Peak Resident Set Size (RSS) memory consumption was tracked throughout large-scale sweeps via the Linux `procfs` interface (`/proc/[pid]/status`).
- **Graphics Processing Unit (GPU):** NVIDIA GeForce GTX 1660 based on the 12 nm TU116 Turing GPU architecture (1408 CUDA execution cores, 88 texture units, 48 ROPs). The GPU is equipped with 6144 MiB (6 GB) of high-speed GDDR5 video memory operating on a 192-bit memory bus with a peak memory bandwidth of 192.1 GB/s. The system operates on NVIDIA Linux display driver version 610.57.04 supporting the CUDA 13.3 computing runtime.
- **Operating System & Linux Kernel:** Ubuntu Linux 26.04.1 LTS (Resolute Raccoon), 64-bit architecture (`x86_64`), running Linux kernel version 7.0.0-29-generic with preemptive dynamic symmetric multiprocessing (`#29-Ubuntu SMP PREEMPT_DYNAMIC`).

E.2 WORKSTATION B: INTEL CORE I9-12900F (RTX 4090)

The second workstation has the following specifications, as reported by the operating system and NVIDIA driver utilities:

- **Central Processing Unit (CPU):** A 12th-generation Intel Core i9-12900F in a single socket, with 16 physical cores and 24 online logical processors on the `x86_64` architecture. The operating system reports a maximum CPU frequency of 5.10 GHz; this is not a measurement of sustained benchmark frequency. Aggregate cache capacities are 640 KiB for L1 data, 768 KiB for L1 instructions, 14 MiB for L2, and 30 MiB for L3. The system reports one NUMA node.
- **System Memory (RAM):** Approximately 125 GiB of physical memory is recognized by the operating system, with 122 GiB of configured swap. These are the rounded totals reported by `free -h`; memory available to an individual experiment depends on other running processes.
- **Graphics Processing Unit (GPU):** An NVIDIA GeForce RTX 4090 with 24,564 MiB of reported device memory (approximately 24 GiB), using NVIDIA driver version 550.163.01.
- **Operating System and Kernel:** GNU/Linux on `x86_64`, running kernel `6.12.96+deb13-amd64`.

E.3 SOFTWARE SPECIFICATIONS

- **Python Execution Environment & Dependencies:** Experiments were executed using Python 3.11.15 managed deterministically by the fast `uv` workspace virtual environment. The core package versions reported for this workstation include:
  - **PyTorch:** Version 2.x with compiled CUDA 13.3 support (utilized for tensor linear algebra, score-matching Jacobians, and neural baselines);
  - **Causal-Learn:** Version 0.1.4.5 (constraint-based and score-based benchmarks, including PC and GES);
  - **Causal Discovery Toolbox (CDT):** Version 0.6.0 (SHD evaluation and causal evaluation utilities);
  - **NetworkX:** Version 3.4.2 (directed acyclic graph reachability, topological sorting, and canonical d-separation validation);
  - **SciPy / NumPy / Scikit-Learn:** NumPy 2.x, SciPy 1.15, and Scikit-Learn 1.6.1 (Nyström landmarks, distance calculations, and GAM cubic spline regressions).
- **External R Interoperability for Baseline Pruning:** Generalized Cross-Validation (GCV) baseline pruning utilizes an external `R` runtime (`mgcv::gam`) invoked directly from Python via `Rscript` subprocesses. The historical inventory associated this pathway with Workstation A; the author-supplied attribution places the scalability runs on Workstation B. Per-run backend records should distinguish execution through R from any Python replacement.
- **Reported Execution Configuration:** The recorded configuration for this workstation uses single-threaded CPU execution (`device='cpu'`) with seed values $seed \in \{0, \ldots, 24\}$. Seed initialization resets Python's `random`, NumPy's `random.seed`, and PyTorch's `torch.manual_seed` across CPU and CUDA devices.

F Extended Empirical Results

This section expands Section 6.2 with the full baseline comparison and the recorded COSMIC variants. All empirical results here were obtained on Workstation A. Figures aggregate original seeds 0–24, averaging repeated source-batch measurements within each method/environment/seed before computing the mean and sample standard deviation. Unsuccessful or nonfinite values do not become zeros. The 25 seeds are repeated algorithm/data-generation settings, not 25 independent replications of Sachs or the fMRI study.

F.1 Benchmark Data Generation and Preprocessing

Synthetic ER and Scalability Sweeps. For the standard Synthetic ER benchmark and scaling experiments, directed acyclic graphs are generated as Erdős–Rényi graphs with edge probability $p_{\text{ER}} = \min(1.0, 3.0/(d-1))$, with $d > 1$, topologically sorted so that all edges satisfy $u < w$. For each node i, data are generated via additive structural equations:

$$X_i = \sin(v) + \exp(-|v|) + av + Z_i, \quad v = \sum_{j \in \text{Pa}(i)} X_j, \tag{53}$$

where $Z_i \sim \mathcal{N}(0,1)$ is independent standard Gaussian noise. All reported synthetic and scalability experiments use $a = 0$ (recorded as `linear_factor = 0.0`), yielding purely nonlinear additive mechanisms.

SynER1 and SynSF1 Mixed SEMs. SynER1 and SynSF1 replicate the benchmark suite of Xu et al. (2024), generating Erdős–Rényi ($d = 10$, average degree 1) and Barabási–Albert scale-free graphs ($d = 10$, 1 edge per attachment), respectively. In these benchmarks, the parameter $p = 0.5$ (configured as `linear_rate = 0.5`) represents the independent Bernoulli probability that each non-root node's structural equation is linear Gaussian ($X_i = \sum_{j \in \text{Pa}(i)} W_{ji} X_j + Z_i$). The remaining non-root mechanisms are sampled from Gaussian process regressions with unit Gaussian noise.

We keep the linearity settings fixed because these experiments target scalability with sample size and dimension, rather than a sweep of linear versus nonlinear mechanisms. COSMIC adopts CaPS's mean-Hessian and parent-score rationale: squared first derivatives can retain evidence from linear effects, unlike a criterion relying only on variation in curvature.

Simulated fMRI (`sim3.mat`). The simulated blood-oxygen-level-dependent (BOLD) fMRI benchmark uses the realistic synthetic hemodynamic dataset `sim3.mat` from Smith et al. (2011) ($d = 15$ brain regions, 50 simulated subjects). For each benchmark seed, an individual subject is randomly sampled, and a contiguous temporal window of $n = 200$ timepoints is extracted without pooling across subjects. The recovered graph is evaluated against that subject's ground-truth directed connectivity matrix.

F.1.1 Extended baseline comparison

The expanded comparison adds PC, GES, GOLEM, SCORE, LNMIX, DAG-GNN, and DAGMA-MLP to the primary baseline set. In Figures 13 and 14, the baselines and their operational configurations are:

- **COSMIC (Default):** Scalable landmark Stein estimation with three-bandwidth ensemble, adaptive macro-tier peeling, LOMTO-LC parent scoring, relative screening ($\lambda_1 = 50.0$), and sequential R/`mgcv` GAM pruning ($\alpha = 0.001$).
- **COSMIC-Exact-LOMTO:** Grouped macro-tier omission using the full $n \times n$ empirical Gram matrix, isolating the effect of grouped omission without Nyström approximation error.
- **CaPS (Xu et al., 2024):** Exact-Gram score matching with variable-wise leave-one-out omission scoring and downstream GAM pruning ($\alpha = 0.001$).
- **DAS (Montagna et al., 2023):** Dense SCORE ordering combined with off-diagonal Hessian thresholding for candidate screening, followed by GAM regression pruning.
- **SCORE (Rolland et al., 2022):** Sequential single-leaf removal using exact second-order Stein scores with downstream GAM pruning.

- **DiffAN (Sanchez et al., 2023):** Diffusion-based score-matching neural network with autograd Jacobians and deciduous score updating, followed by regression pruning.
- **CAM (Bühlmann et al., 2014):** Nonparametric causal additive model implemented via `SEMGAM` with feature significance threshold $p = 0.001$.
- **DAGMA (Linear) (Bello et al., 2022):** Continuous optimization DAG learner using the log-determinant acyclicity characterization (`DagmaLinear` with ℓ_1 penalty $\lambda_1 = 0.03$ and edge threshold $w_{\text{thresh}} = 0.3$).
- **DAGMA-MLP (Bello et al., 2022):** Nonlinear neural network variant parameterized by multi-layer perceptrons with hidden dimension $d_{\text{hidden}} = 10$, warm-start iterations $50{,}000$, maximum iterations $80{,}000$, and edge threshold $w_{\text{thresh}} = 0.3$.
- **PC & GES (Spirtes et al., 1993; Chickering, 2003):** Constraint-based (PC) and score-based (GES) algorithms from `causal-learn` (v0.1.4.5). Because PC and GES output Completed Partially Directed Acyclic Graphs (CPDAGs), directed metrics describe the benchmark wrapper's heuristic cycle-avoidance DAG completion of unoriented edges (`endpoint_to_amat`). In all extended plots, PC is styled in royal blue (`#4169E1`) to distinguish it visually from LNMIX's orange (`#D55E00`).
- **GOLEM (Ng et al., 2021):** Soft-constrained continuous optimization with ℓ_1 penalty 0.02.
- **DAG-GNN (Yu et al., 2019):** A graph-neural-network variational autoencoder with an acyclicity constraint, evaluated through `gCastle`'s `DAG_GNN`. The benchmark wrapper uses `epochs=10`, `device_type='cpu'`, and a fixed model seed of 42; the 25 benchmark seeds vary the sampled data rather than this model seed. The plotted results describe this short training configuration, not a convergence-certified fit.
- **LNMIX (Liu et al., 2024):** Score matching for mixed linear-nonlinear models without an explicit pruning stage.

Structural Intervention Distance (SID) and Runtime. Structural intervention distance (SID) counts ordered variable pairs for which the estimated graph provides an incorrect adjustment set relative to the ground-truth DAG; lower values are preferable. Reported runtimes reflect total wall-clock execution time on Workstation A across 25 original seeds.

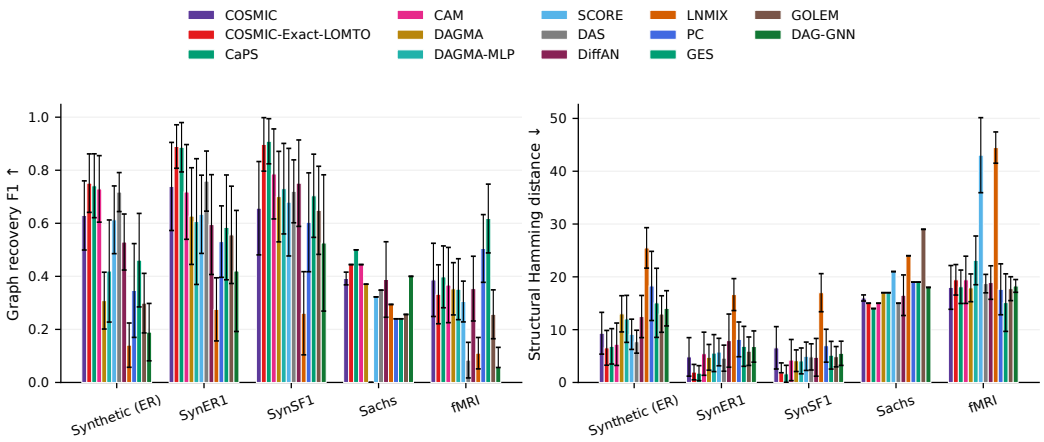

Figure 13: Extended empirical baseline comparison: final-graph F_1 and SHD across five environments on Workstation A. Bars and error bars show means and sample standard deviations over 25 original seeds. PC is plotted in royal blue (`#4169E1`); DAGMA-MLP represents the nonlinear neural variant.

Performance Analysis across Environments. On Synthetic ER ($d = 10, n = 1000$), exact-Gram COSMIC and CaPS achieve mean $F_1 = 0.751$ and 0.742, compared with 0.718 for DAS, 0.613 for SCORE, 0.529 for DiffAN, 0.461 for GES, and 0.347 for PC. Their mean SID values are 13.64 and 14.72, versus 22.84 for scalable COSMIC. Linear DAGMA obtains $F_1 = 0.308$, while DAGMA-MLP achieves $F_1 = 0.420 \pm 0.189$ with runtime 104.1 ± 33.0 s.

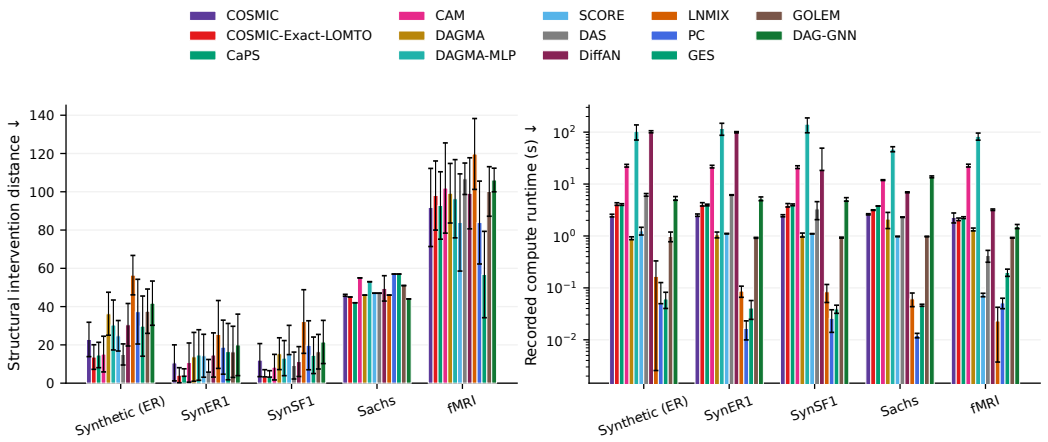

Figure 14: SID and recorded full-algorithm runtime for the extended baseline set on Workstation A. Lower values are preferable. Runtime reflects complete execution, including downstream pruning where applicable.

On the mixed SEMs SynER1 and SynSF1, CaPS reaches $F_1 = 0.887$ and 0.909 (SHD $= 1.76$ and 1.64), while exact-Gram COSMIC reaches $F_1 = 0.889$ and 0.897. Exact-Gram COSMIC has slightly higher mean F_1 on SynER1, whereas CaPS has higher mean F_1 on SynSF1 and lower mean SHD on both. DAGMA-MLP reaches $F_1 = 0.606 \pm 0.232$ on SynER1 and $F_1 = 0.731 \pm 0.167$ on SynSF1, outperforming linear DAGMA ($F_1 = 0.701$) on scale-free graphs, but requiring approximately two minutes per run (117.5 s and 142.6 s).

On Sachs ($d = 11, n = 853$), CaPS reaches $F_1 = 0.500$ and SID 42.0, followed by exact-Gram COSMIC and CAM (both $F_1 = 0.444$, with SID 45.0 and 55.0, respectively). Default COSMIC reaches $F_1 = 0.392$ and SID 45.92, competitive with DiffAN ($F_1 = 0.388$) and Linear DAGMA ($F_1 = 0.370$). In contrast, DAGMA-MLP collapses entirely on Sachs: across all 25 seeds, it outputs an empty graph with 0 directed edges, yielding $F_1 = 0.000 \pm 0.000$, precision 0.000, recall 0.000, and SHD $= 17.0$. The recorded outputs establish failure to recover edges under this configuration, but do not identify its cause. Without optimization or gradient diagnostics, we cannot attribute the empty graphs to a particular penalty, gradient behavior, or noise distribution. In end-to-end scalability sweeps (Figure 4), DAGMA-MLP completes $d = 50$ (788.4 s, $F_1 = 0.394$) and times out at $d = 100$ (> 1800 s on seed 0).

On simulated fMRI ($d = 15, n = 200$), GES attains the highest mean F_1 (0.618) and PC reaches 0.505, both outperforming scalable COSMIC (0.387) and CaPS (0.398). GES also achieves the lowest mean SID (56.76 versus COSMIC's 91.80) with runtime 0.198 s versus 2.277 s. DAGMA-MLP achieves $F_1 = 0.352 \pm 0.113$ and SHD $= 23.12$ in 83.0 s. These results confirm that no single method dominates across all biological, hemodynamic, and synthetic regimes.

F.2 Fixed-Slice Runtime Comparisons

The fixed-slice scatter plots (Figure 3, right column) highlight this operational trade-off. At $d = 50$, COSMIC completes screening in sub-second time (0.899 s) with $F_1 = 0.106$, compared with CaPS at $F_1 = 0.119$ at a fraction of its runtime (18.40 s); DAS achieves higher accuracy ($F_1 = 0.236$) but requires 510.03 s ($> 500\times$ longer), while LNMIX is comparably fast (0.896 s) with near-zero recovery ($F_1 = 0.039$). Similarly at $n = 5000$, COSMIC returns candidate graphs in 1.34 s, compared to over four minutes for CaPS (261.61 s) and nearly eight minutes for DAS (465.48 s). COSMIC thus occupies an aggressive fast-screening regime, while the higher-accuracy alternatives illustrate why runtime alone cannot rank the methods.

F.3 COSMIC VARIANTS

The figure names refer to the following estimator and allocation choices. Except where specified, the named alternatives use the Nyström Gram representation and adaptive macro-tier peeling.

COSMIC: Default landmark estimator and LOMTO-LC allocation, combining static omission evidence with local couplings.

COSMIC-FPS: Coverage-only farthest-point selection, without the adaptive landmark-budget stopping rule.

COSMIC-LOMTO-S: Independent full-data tier omissions with static coupling-based allocation.

COSMIC-LOMTO-U: The same grouped omissions, distributing each retained parent's tier evidence uniformly.

COSMIC-SPA: Sequential Peeling Attribution; computes parent scores dynamically from consecutive active-set Hessian shifts during the peeling passes ($|\bar{H}_j(\mathcal{V}_t \setminus \mathcal{S}_t) - \bar{H}_j(\mathcal{V}_t)|$), eliminating the additional full-data omission passes required by standard LOMTO.

COSMIC-Hessian: Direct score-product/Hessian coupling for screening, avoiding omission-based parent scoring.

COSMIC-Nystrom-LOO: Variable-wise omissions with the low-rank estimator.

COSMIC-Exact-LOO: Variable-wise omissions using the full empirical Gram matrix.

COSMIC-Exact-LOMTO: Full-Gram estimation with LOMTO-LC allocation.

COSMIC-Exact-LOMTO-S: Full-Gram estimation with static LOMTO allocation.

These are recorded configurations, not necessarily one-factor ablations. All COSMIC variants use GCV pruning and tri-band by default. Most algorithms apply post-recovery using calculated parent scores; the Hessian variant uses the spline-check recovery rule. The merged report omits these effective per-run options, so the names establish the intended score variant but do not verify that pruning, bandwidths, recovery, and landmark policies were identical historically. In particular, empirical runtime differences should not be attributed solely to the kernel representation. The factorial and bandwidth experiments provide the more targeted component comparisons.

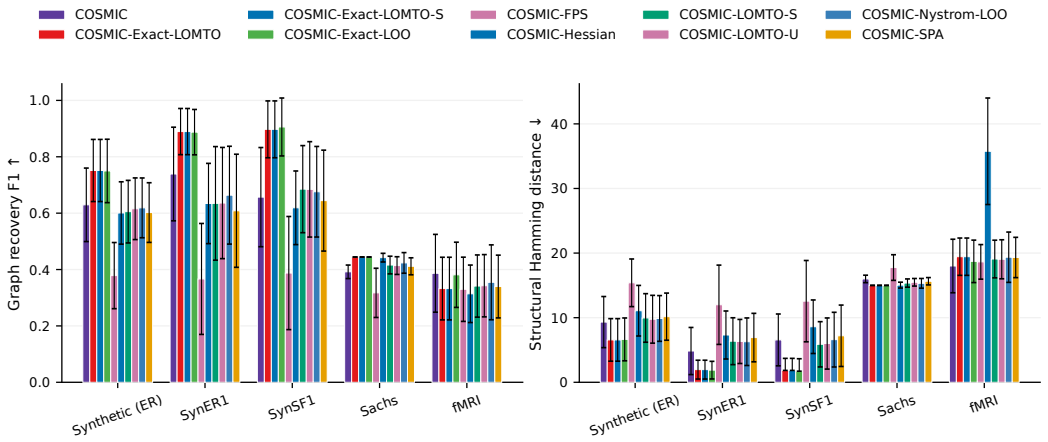

Figure 15: COSMIC variants: final-graph F_1 and SHD. Means and sample standard deviations use 25 original seeds on Workstation A. Variant names describe estimator/allocation choices, while historical downstream settings are not fully recorded.

On synthetic ER, exact LOO reaches mean $F_1 = 0.750$ and SID 13.64, close to exact LOMTO's 0.751 and 13.64. Their runtimes are 5.005 and 4.125 seconds. Coverage-only FPS is slower than default COSMIC (4.968 versus 2.475 seconds) and less accurate ($F_1 = 0.378$ versus 0.630); increasing landmark coverage is therefore not automatically beneficial. Static, uniform, and cascaded low-rank allocations reach $F_1 = 0.605$, 0.616, and 0.602, respectively. Their lower runtimes include

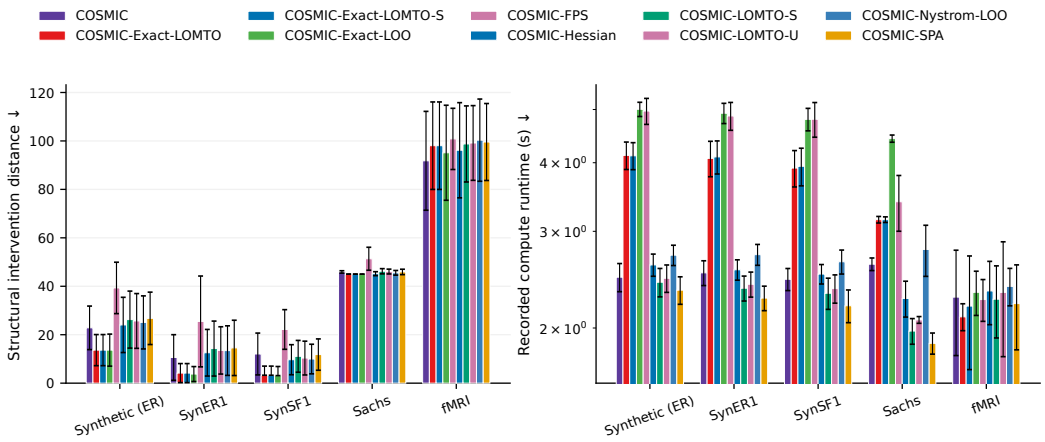

Figure 16: COSMIC variants: SID and recorded full-algorithm runtime. These curves include the downstream procedures of the recorded variants; they are not isolated estimator timings.

differences in downstream configuration and should not be interpreted as a controlled advantage over the default.

On Sachs, the three displayed exact variants all have mean $F_1 = 0.444$ and SID 45.0, while default COSMIC reaches 0.392 and 45.92. On fMRI, Nyström LOO has mean $F_1 = 0.355$, below default COSMIC's 0.387 and above exact LOMTO's 0.332. Default COSMIC also has lower mean SID (91.80) than either the Hessian variant (96.16) or Nyström LOO (100.32). Exact LOMTO is faster than default COSMIC on fMRI (2.096 versus 2.277 seconds), despite its lower mean F_1. These comparisons motivate evaluating recovery, intervention distance, and runtime together rather than treating a better parent-score approximation as sufficient for better final recovery.

G COMPONENT ISOLATION

These diagnostics evaluate ordering, parent screening, and tier recovery separately. They expose component-specific errors rather than assuming that a final recovery score identifies which phase succeeded. Means and sample standard deviations use 25 seeds.

G.1 ORDERING

G.2 FIXED-SLICE RUNTIME COMPARISONS

The fixed-slice scatter plots (Figure 3, right column) highlight this operational trade-off. At $d = 50$, COSMIC completes screening in sub-second time ($0.899\,\text{s}$) with $F_1 = 0.106$, compared with CaPS at $F_1 = 0.119$ at a fraction of its runtime ($18.40\,\text{s}$); DAS achieves higher accuracy ($F_1 = 0.236$) but requires $510.03\,\text{s}$ ($> 500\times$ longer), while LNMIX is comparably fast ($0.896\,\text{s}$) with near-zero recovery ($F_1 = 0.039$). Similarly at $n = 5000$, COSMIC returns candidate graphs in $1.34\,\text{s}$, compared to over four minutes for CaPS ($261.61\,\text{s}$) and nearly eight minutes for DAS ($465.48\,\text{s}$). COSMIC thus occupies an aggressive fast-screening regime, while the higher-accuracy alternatives illustrate why runtime alone cannot rank the methods.

COSMIC has mean edge-order accuracy 0.896 and a fully valid order in 60% of runs in this diagnostic. Exact-Gram COSMIC and CaPS reach 1.0 on both measures; SCORE reaches 0.992 edge-order accuracy and 0.96 valid-order fraction. Edge-level ordering accuracy can therefore hide at least one wrong direction in many graphs.

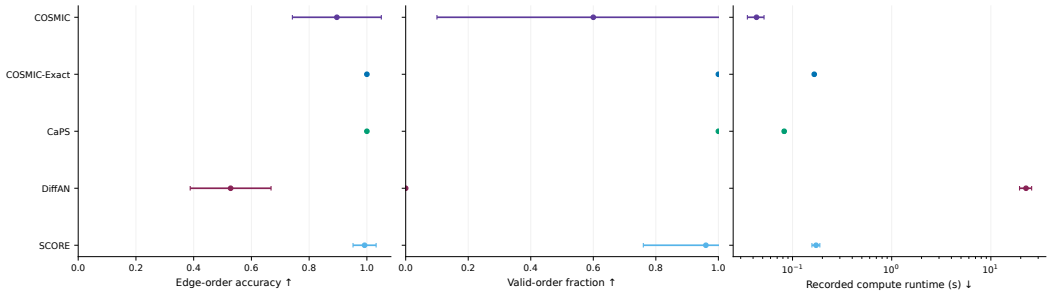

Figure 17: Isolated ordering quality and runtime on a 6-node permuted branching/collider DAG ($d = 6, n = 1000, 5$ directed edges) with $1.5\sin(\cdot)$ mechanisms and unit Gaussian noise across 25 seeds. Edge-order accuracy measures the fraction of true edges consistent with the returned order; valid-order fraction requires every true edge to be consistent. No downstream pruning can change these ordering measurements.

G.3 PARENT SCREENING

The COSMIC and CaPS score comparisons in this diagnostic retain an edge only if $S_{ji} > 0$ and $S_{ji} \geq 0.02 \max_k S_{ki}$; DAS retains its native screened graph. This positive-evidence rule differs from production COSMIC's near-zero fallback in Appendix A.4, which retains all order-compatible candidates when the maximum score is at most 10^{-8}. The diagnostic isolates positive score evidence and should not be interpreted as an exact reproduction of production screening in that regime.

LOMTO-LC yields screened $F_1 = 0.469$, compared with 0.465 for static LOMTO. Supplying the oracle order raises LOMTO-LC to 0.513, while its exact-Gram variant reaches 0.550. The oracle-order diagnostic isolates the scoring stage from estimated-order mistakes, but leaves score estimation and screening imperfect. These results support evaluating both stages independently, not claiming either stage has solved the other's errors.

G.4 MACRO-TIER RECOVERY

Matched membership accuracy, same-tier pair F_1, and tier-count error answer different questions. For synthetic graphs, COSMIC LOMTO-LC has membership accuracy 0.556 and same-tier pair

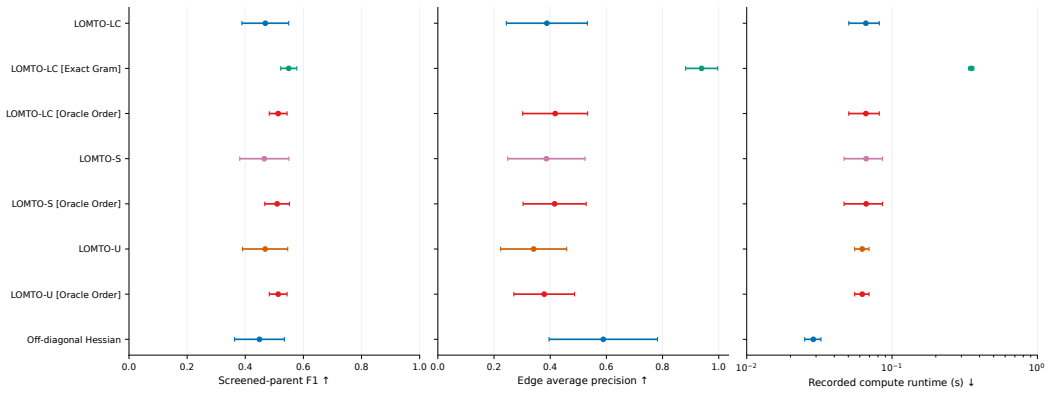

Figure 18: Parent-screening F_1, average precision, and compute runtime before pruning on the same 6-node benchmark ($d = 6, n = 1000$). Oracle-order and exact-Gram variants are identified explicitly and alter different components.

$F_1 = 0.219$, compared with 0.684 and 0.469 for exact-Gram COSMIC. On Sachs, membership accuracy is 0.545 for COSMIC LOMTO-LC and 0.455 for exact-Gram COSMIC. These imperfect matches reinforce the distinction between computational batches and causal sink layers. Allocation variants can share the same tiers because allocation follows ordering.

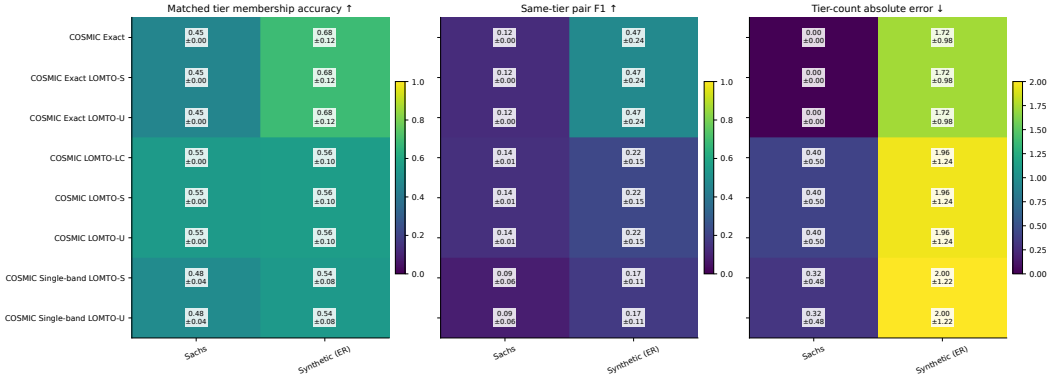

Figure 19: Macro-tier recovery against the reference layers. Matching resolves arbitrary tier labels; pairwise agreement and count error retain information that membership accuracy alone omits. These measurements concern computational tiers, unlike the recovered-DAG layer counts in the adaptive-policy benchmark.

H SCORE FIDELITY

This diagnostic asks whether fast parent scores reproduce the ranking of an exact-Gram, variable-wise LOO reference. It is distinct from testing whether either score recovers true edges. Figure 20 therefore reports both reference similarity and edge discrimination over 25 seeds per environment. The reference is an omission-based screening statistic; this experiment does not establish equivalence to every score used by DAS or SCORE.

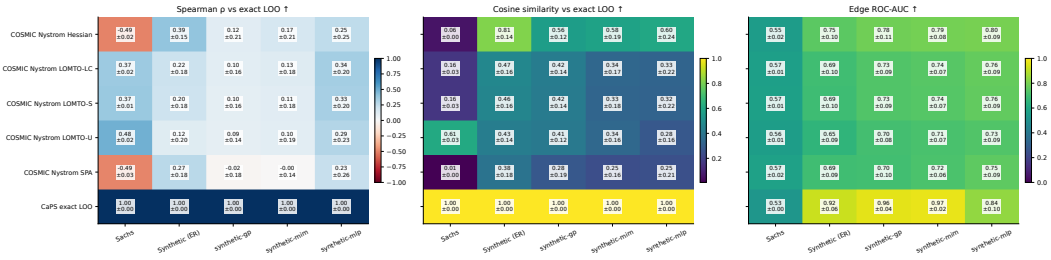

Figure 20: Parent-score fidelity and edge discrimination across five environments. Spearman correlation and cosine similarity compare each recorded estimator with exact LOO; edge metrics use the reference graph. Means are aggregated over 25 seeds. Negative correlations are retained. Recorded estimator names distinguish Hessian, static LOMTO, and local-coupling variants.

For Nyström LOMTO-LC (default COSMIC), mean Spearman correlations are 0.216 on synthetic ER, 0.102 on synthetic GP, 0.128 on synthetic MIM, 0.341 on synthetic MLP, and 0.365 on Sachs. The synthetic GP, MLP, and MIM benchmarks evaluate Gaussian Process, Multi-Layer Perceptron, and Multiple Interaction Model mechanisms on Erdős–Rényi ER1 topologies with $d = 10$ variables, $n = 2000$ observations, and unit Gaussian noise across 25 seeds (Xu et al., 2024). Static LOMTO gives similar values (0.205, 0.098, 0.113, 0.331, and 0.373). These results show positive but limited rank agreement, rather than uniformly strong fidelity. On symmetric or even mechanism components (e.g., MIM), first-order score coupling $\mathbb{E}[s_j s_p]$ can vanish despite direct causal links, suppressing static and local allocation weights and reducing correlation. The Nyström Hessian comparator has negative correlation on Sachs (-0.494), illustrating why clipping negative values would obscure a failure case. Local refinement is not uniformly better than static allocation.

The algebraic guarantees are narrower than an approximation-error bound: the blended allocation is nonnegative, preserves omission mass within its blocks, and has a bounded deviation from static allocation. Appendix A.5 derives these properties and the population omission identity. Neither mass conservation nor correlation with an estimated reference proves correct parent recovery.

I Extended Discussion and Future Research Directions

The development of COSMIC demonstrates that score-based causal discovery can scale to hundreds of variables and large sample sizes by replacing dense $n \times n$ kernel operations with low-rank landmark approximations and multi-node macro-tier peeling. However, empirical and theoretical investigations reveal several foundational bottlenecks that outline a concrete roadmap for future causal discovery research.

I.1 Bounding Landmark Quadrature Error and Regularized Score Solves

Appendix A.7 gives deterministic discrepancy bounds relative to a dense empirical reference, but the pilot finds them too conservative to certify peeling decisions. Useful distribution-dependent bounds on derivative quadrature and population curvature remain open. Appendix A.6 identifies one difficulty: derivative components outside the landmark span are multiplied by η_G^{-1} in the score, and the squared-score contribution to $\widehat{H} = Q - G \odot G$ can scale as η_G^{-2}. Here Q estimates the normalized second density derivative, not the log-density curvature by itself.

Kernel thinning (Dwivedi & Mackey, 2021; Gong et al., 2024) and kernel herding (Chen et al., 2010) offer possible alternatives to geometric FPS quadrature. For a suitable RKHS and compactly supported distribution, the original kernel-thinning result gives an $\mathcal{O}_d(n^{-1/2}\sqrt{\log n})$ integration discrepancy with $\sqrt{n}$ retained points and $\mathcal{O}(n^2)$ construction time; other distribution classes have different rates. Applying such results to COSMIC requires control of the derivative function class, its norms, and the regularized inverse. A $\sqrt{n}$ coreset also exceeds the current 300-landmark cap when $n > 90{,}000$. Whether a derivative-aware coreset improves ordering at a comparable total budget must therefore be tested rather than inferred from a generic integration guarantee.

I.2 Resolving the Upstream Ordering Bottleneck

The stage-error analysis in Appendix A.4.4 identifies both cross-tier false sink selections and within-tier inversions. Thus the current macro-tiers cannot be assumed to recover correct coarse causal strata. Because downstream pruning respects the estimated order, ordering reliability remains a separate target from better pruning.

One possible research direction is conditional or residualized refinement within estimated tiers. However, subtracting ancestral conditional means does not generally produce a new ANM on the residuals. For example, let $Z, \epsilon_X, \epsilon_Y$ be independent standard Gaussian variables, $X = Z + \epsilon_X$, and $Y = X^2 + \epsilon_Y$. If Z is in an earlier tier, population residualization gives

$$X - \mathbb{E}[X \mid Z] = \epsilon_X, \qquad Y - \mathbb{E}[Y \mid Z] = 2Z\epsilon_X + \epsilon_X^2 - 1 + \epsilon_Y. \tag{54}$$

The residual for Y still depends on Z, and its conditional variance given ϵ_X is $4\epsilon_X^2 + 1$. It therefore cannot be represented as a function of the residual X plus independent additive noise. A valid refinement scheme would need additional mechanism restrictions, conditional score estimation, or a model that explicitly retains this dependence. Its evaluation should measure both cross-tier and within-tier errors rather than presuming that only the latter remain.

I.3 Eliminating the Downstream GAM Pruning Bottleneck

At $d = 100$, the matched end-to-end runs spend 2.85 s on ordering, scoring, and screening out of 1537.69 s total; GAM pruning accounts for 99.8% of runtime. This stage includes spline construction, fitting and smoothing-parameter selection, as well as R process and data-transfer overhead. Its measured share does not isolate which of these costs dominates.

Two possible improvements are:

- **Spline basis caching.** With a fixed basis dimension, all univariate basis matrices Φ_j require $\mathcal{O}(nd)$ storage and can potentially be reused across candidate fits. A joint design Gram matrix includes both $\Phi_j^\top \Phi_j$ and generally nonzero cross-parent blocks $\Phi_j^\top \Phi_k$. It is not block diagonal merely because the smooth penalties are. Selective caching must balance these cross-block costs, memory, and compatibility with each fit's smoothing and centering constraints.

- **Joint group-penalized selection.** A solver penalizing spline coefficient blocks could evaluate groups of candidate edges using batched linear algebra. It would change the selection procedure relative to partial-smooth significance testing, so both runtime and recovery must be compared. The unsuccessful BIC and two-worker interventions in Appendix A.4.4 show why reduced implementation overhead alone is insufficient evidence of a better pipeline.

I.4 Sparse Candidate Screening at Larger Dimensions

Landmark estimation reduces the sample-dependent cost of a Stein fit, but dense score-product couplings still require $\mathcal{O}(nd^2)$ work. A prospective hierarchical allocation could first score candidate parent tiers against omitted child tiers. Consistent with LOMTO's omission direction, an aggregate statistic for $a < b$ is

$$S_{\mathcal{T}_a \to \mathcal{T}_b} = \left\| \bar{H}_{\mathcal{T}_a}^{(-\mathcal{T}_b)} - \bar{H}_{\mathcal{T}_a} \right\|_2 . \tag{55}$$

Here $\bar{H}_{\mathcal{T}_a}$ is the vector of mean diagonal curvatures on the retained parent tier. Like the individual omission statistic, this quantity need not isolate direct causal effects. Computing it, selecting tier pairs, and refining the surviving blocks all have costs. Subquadratic scaling would require explicit sparsity or screening assumptions controlling these operations; it does not follow merely from grouping variables into tiers. Any computational gain should be evaluated together with the true-parent recall lost by early block rejection.

I.5 Broader Methodological Extensions

Several extensions remain possible research directions:

- **Federated discovery.** Decentralized causal learning motivates avoiding raw-data centralization (McMahan et al., 2017; Hahn et al., 2026). A distributed Stein estimator would need to combine kernel and derivative-sum information under compatible geometry and account for heterogeneous client distributions. Observation-level score matrices and landmarks can disclose information; transmitting them is not by itself a privacy guarantee. The estimator, communication costs, and any formal privacy protection would need a separate design and evaluation.
- **Latent confounders.** Connecting low-rank estimation with score-based approaches that address departures from causal sufficiency (Montagna et al., 2025) is a possible direction. COSMIC's present terminal-block and ordering arguments do not supply guarantees for latent-variable models; the target graph and assumptions would need to be specified anew.
- **Causal representations.** Fast Stein estimates could be explored within representation-learning systems (Reizinger et al., 2025). Computational efficiency alone would not make the learned representation identifiable; that would require appropriate data and model assumptions and a separate argument.
- **Noise heterogeneity.** Unequal noise variances across nodes, parent-dependent noise scales, and non-Gaussian errors raise distinct challenges. Normalizing empirical scores may change ordering cues without preserving causal asymmetry. Any proposed noise-adaptive criterion should therefore establish its population ordering conditions before evaluating finite-sample robustness.

J Hyperparameter Robustness

We vary the landmark count m and the macro-tier count k separately and jointly. Here k counts removal passes, not nearest neighbors. The joint grid uses fixed values $m \in \{50, 200\}$ and $k \in \{2, 5\}$ alongside each adaptive policy. Adaptive labels denote data-dependent choices and are not numerical coordinates between the fixed settings. The report evaluates ordering, parent scores, and relative screening without pruning or recovery. These grids use the positive-evidence diagnostic rule $S_{ji} > 0$ and $S_{ji} \geq 0.02 \max_k S_{ki}$, rather than production COSMIC's near-zero all-retention fallback (Appendix A.4).

Each heatmap compares screened-graph F_1, runtime, the paired change from the adaptive reference, and the paired interaction. For each matched seed, the interaction subtracts the two single-parameter

changes from the joint change. This tests whether the effect of changing one budget depends on the other; it does not presume that m and k should obey a fixed relationship.

Across the nine grid cells, mean F_1 ranges from 0.172 to 0.301 in the Synthetic ER environment, 0.171 to 0.269 on SynER1, 0.184 to 0.261 on SynSF1, 0.150 to 0.262 on Sachs, and 0.110 to 0.157 on fMRI. Neighboring choices can give similar results, but these ranges do not establish a universal broad plateau. The single-budget sweeps below show where extra computation changes recovery and where sensitivity remains.

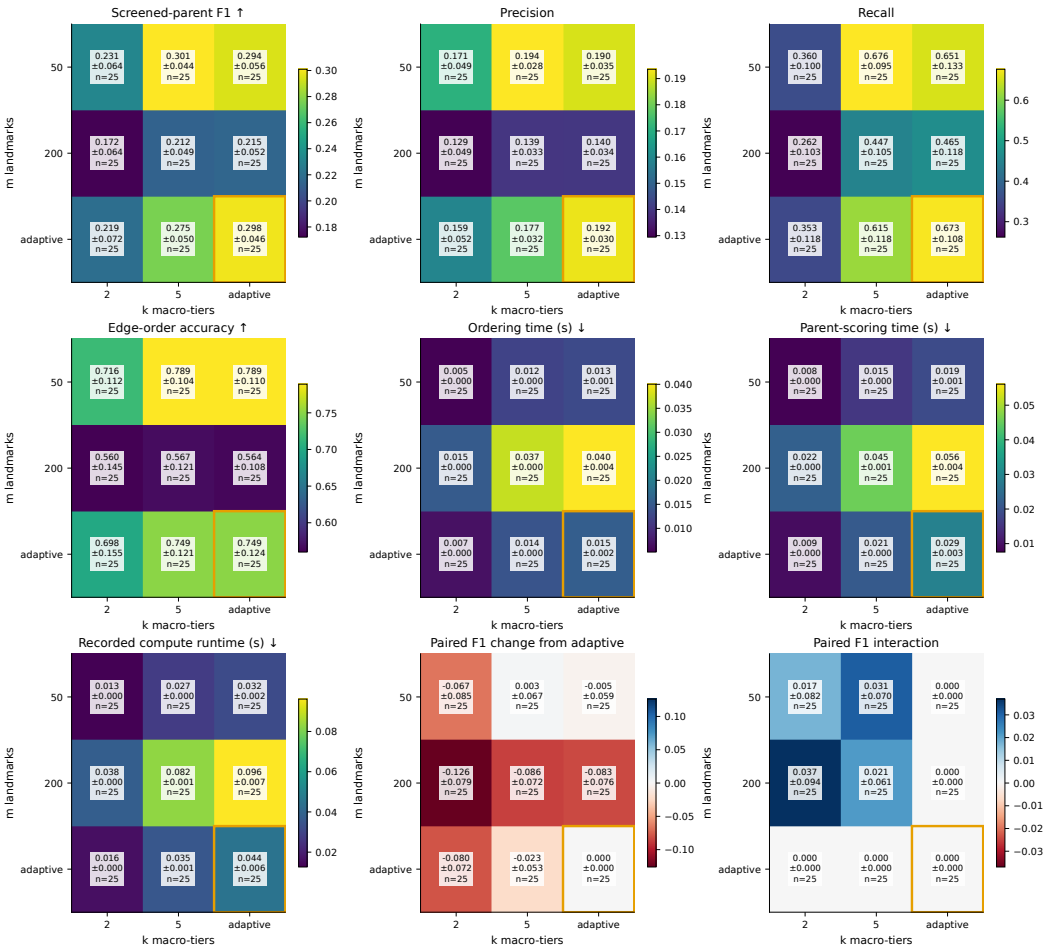

Figure 21: Synthetic ER: joint landmark and tier-budget sensitivity. Cells use 25 matched seeds; differences and interactions are formed per seed before aggregation. These are screening diagnostics, not final-graph results.

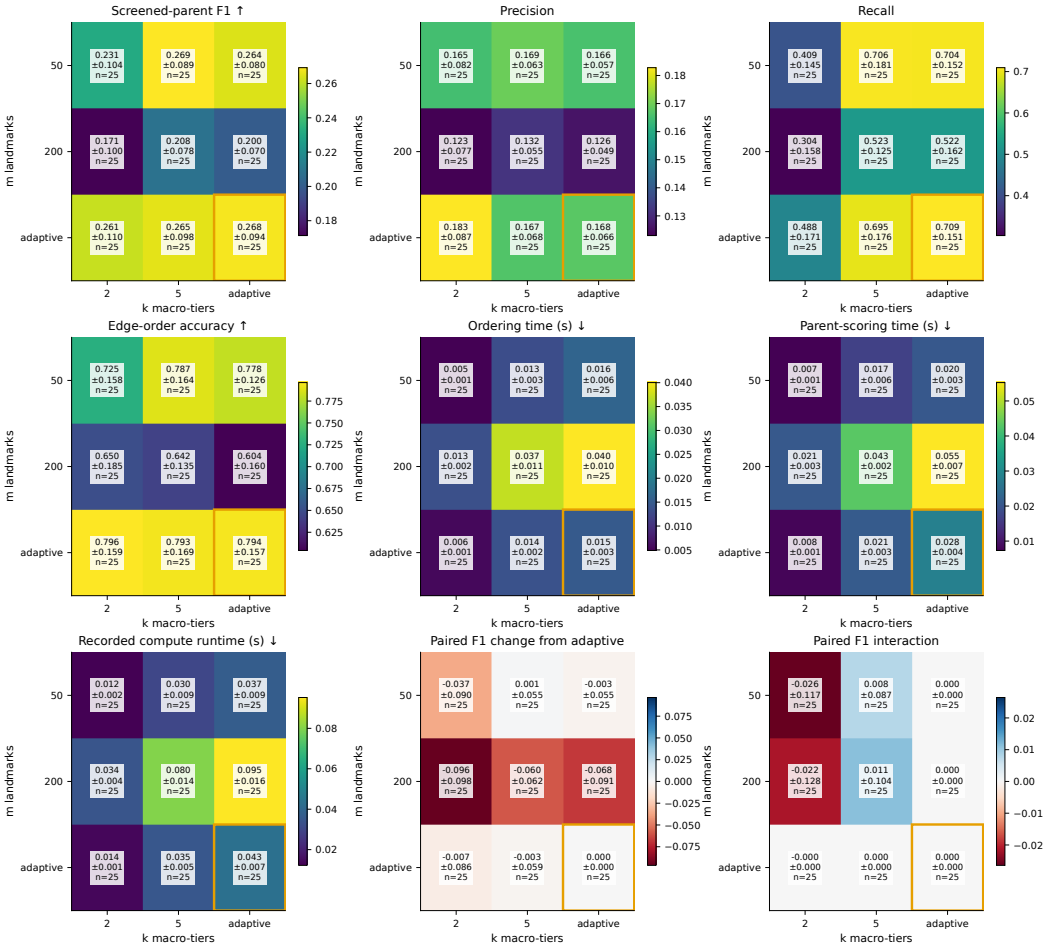

Figure 22: SynER1: joint landmark and tier-budget sensitivity. Cells use 25 matched seeds; differences and interactions are formed per seed before aggregation. These are screening diagnostics, not final-graph results.

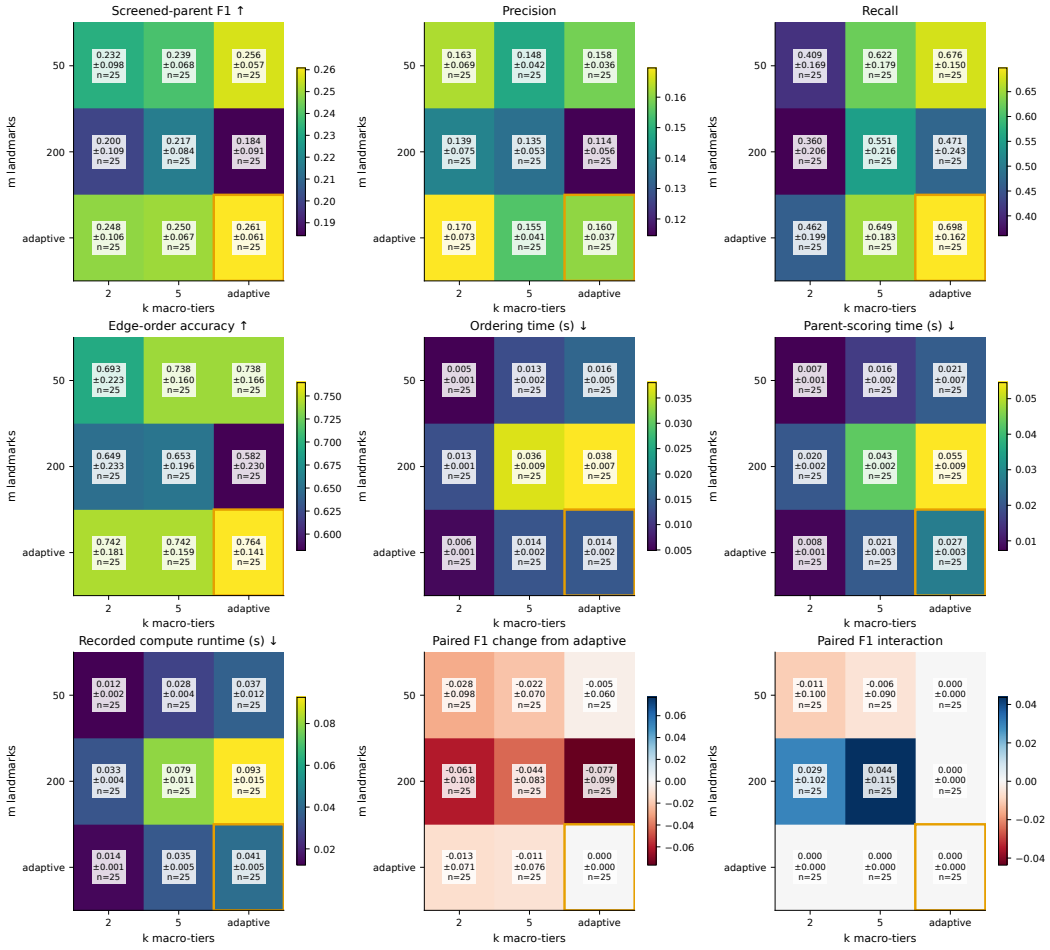

Figure 23: SynSF1: joint landmark and tier-budget sensitivity. Cells use 25 matched seeds; differences and interactions are formed per seed before aggregation. These are screening diagnostics, not final-graph results.

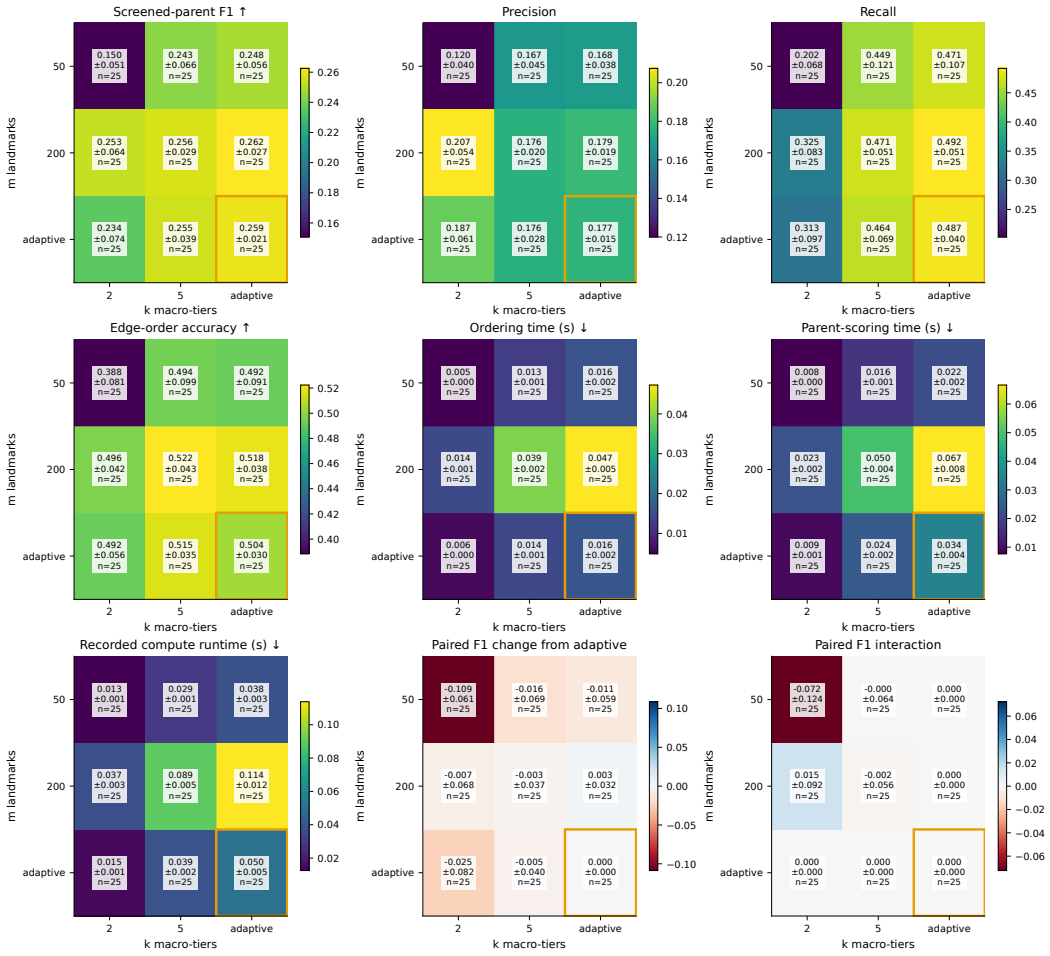

Figure 24: Sachs: joint landmark and tier-budget sensitivity. Cells use 25 matched seeds; differences and interactions are formed per seed before aggregation. These are screening diagnostics, not final-graph results.

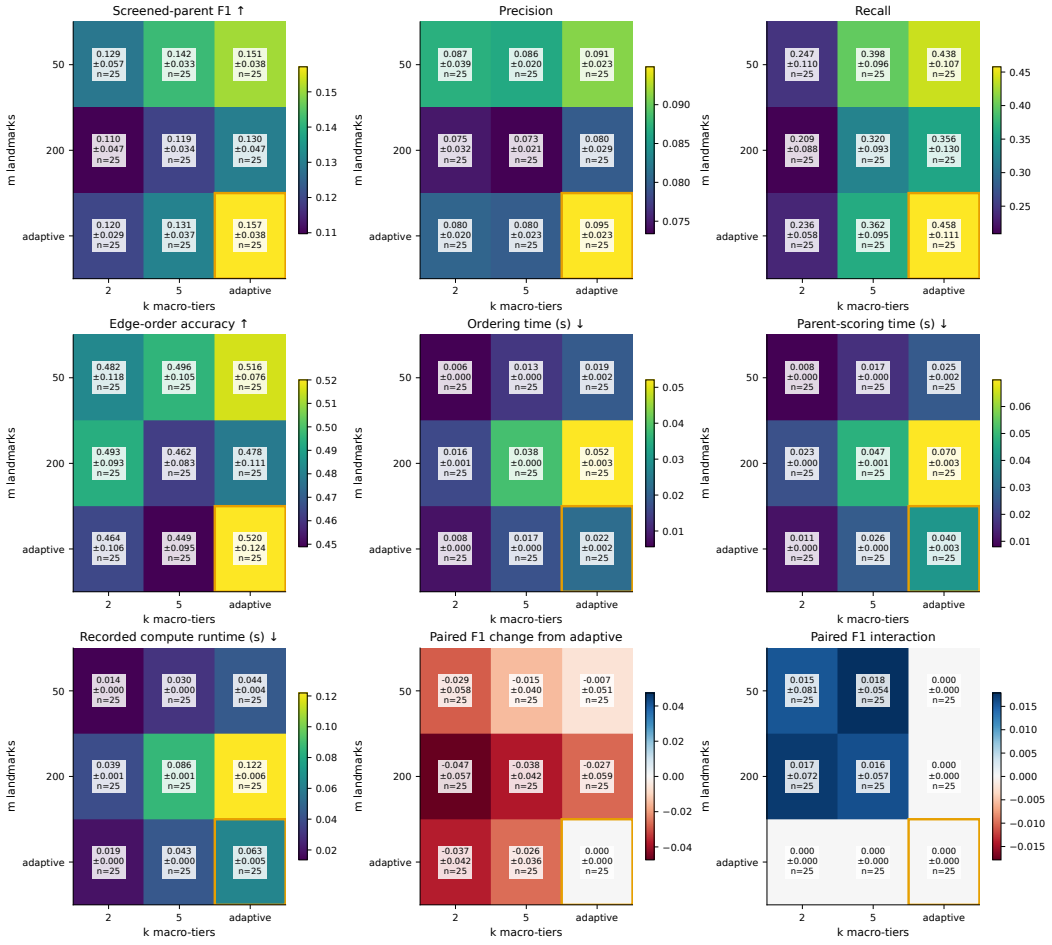

Figure 25: simulated fMRI: joint landmark and tier-budget sensitivity. Cells use 25 matched seeds; differences and interactions are formed per seed before aggregation. On fMRI ($n = 200$), the requested budget $m = 200$ is clamped by the upper-bound safeguard $u_n = \lfloor n/2 \rfloor = 100$, so the nominal $m = 200$ column evaluates $m = 100$ landmarks. These are screening diagnostics, not final-graph results.

J.1 SINGLE-BUDGET SWEEPS

Fixed-count curves are shown against the named policy references in the reports. Error bars are sample standard deviations over successful seeds. In the tier-count sweeps, $k = 1$ assigns all d variables to a single macro-tier, leaving no retained variables for omission scoring and hence an all-zero parent-score matrix. The diagnostic rule rejects these zero scores and yields an empty graph ($F_1 = 0.00$), despite high edge-order accuracy (> 0.90). This is a property of the diagnostic screen: production COSMIC would instead retain the order-compatible tournament under its near-zero fallback, before pruning and any configured parent cap. The m sweeps and k sweeps isolate budget sensitivity within their recorded configurations; they should not be read as controlled comparisons with every historical main-paper run.

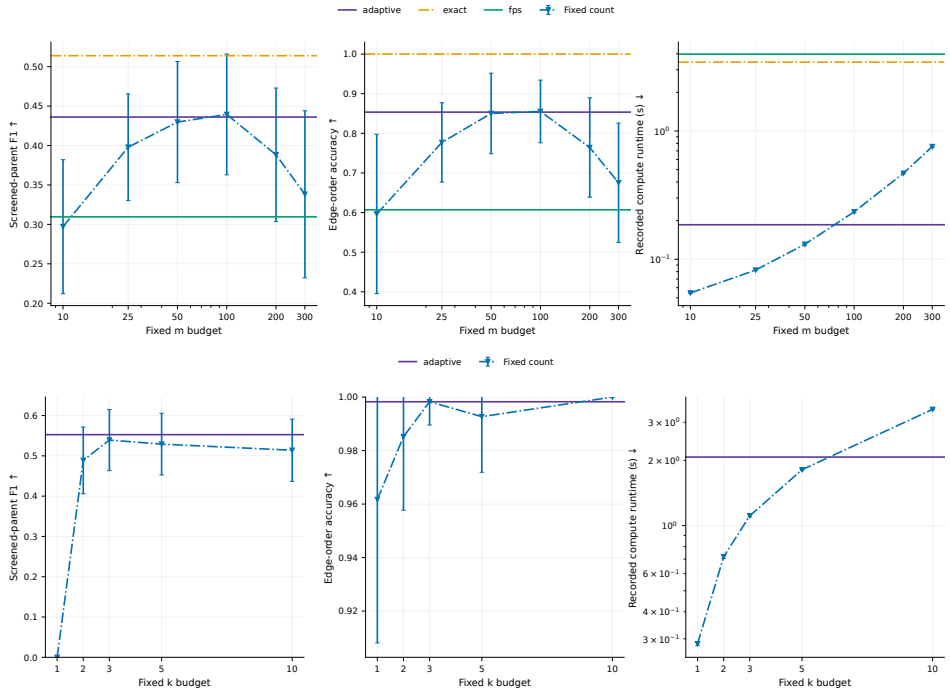

Figure 26: Synthetic ER: landmark-count sweep (top) and tier-count sweep (bottom). Panels show screened-graph F_1, edge-order accuracy, and recorded compute runtime. Named policies remain separate references; they are not fixed counts.

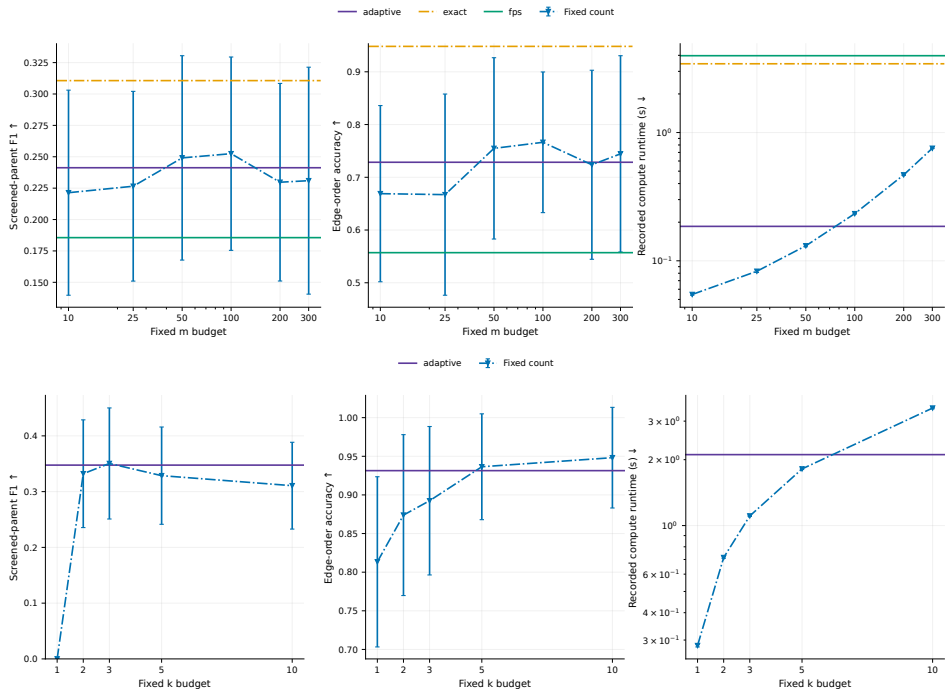

Figure 27: SynER1: landmark-count sweep (top) and tier-count sweep (bottom). Panels show screened-graph F_1, edge-order accuracy, and recorded compute runtime. Named policies remain separate references; they are not fixed counts.

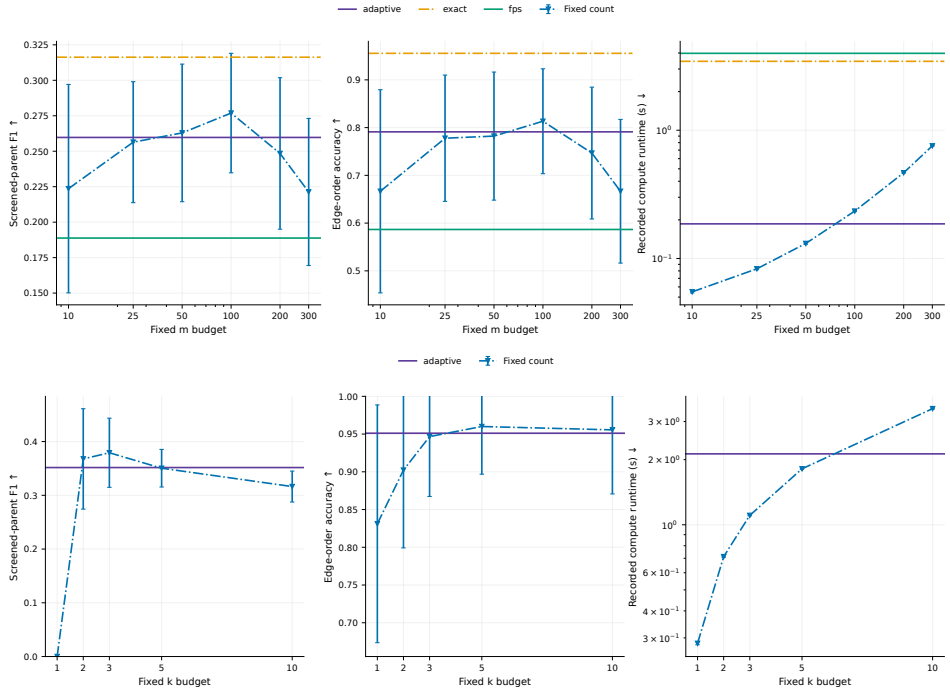

Figure 28: SynSF1: landmark-count sweep (top) and tier-count sweep (bottom). Panels show screened-graph F_1, edge-order accuracy, and recorded compute runtime. Named policies remain separate references; they are not fixed counts.

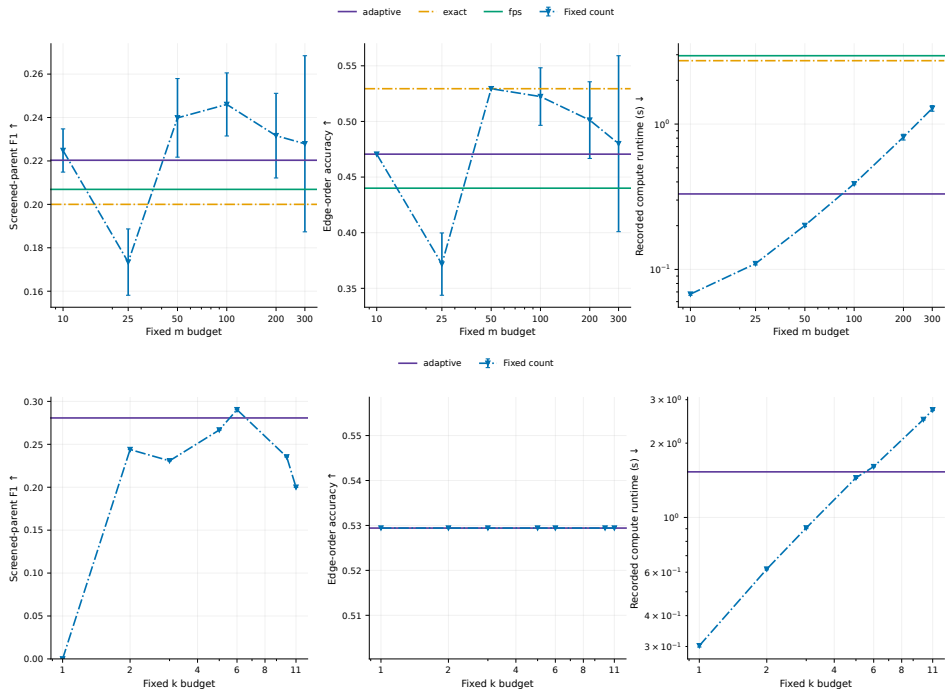

Figure 29: Sachs: landmark-count sweep (top) and tier-count sweep (bottom). Panels show screened-graph F_1, edge-order accuracy, and recorded compute runtime. Named policies remain separate references; they are not fixed counts.

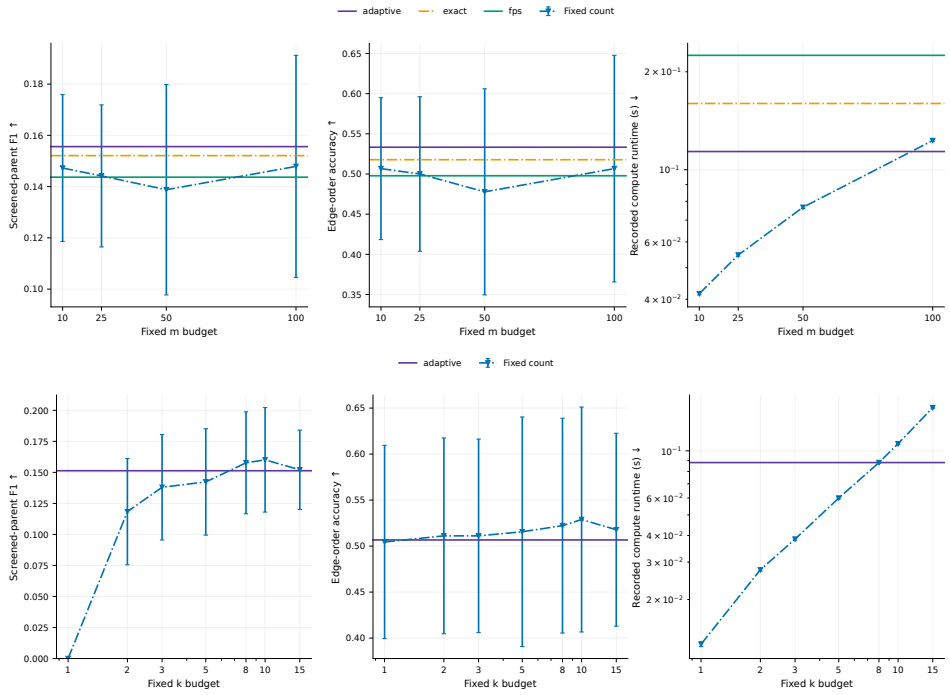

Figure 30: simulated fMRI: landmark-count sweep (top) and tier-count sweep (bottom). Panels show screened-graph F_1, edge-order accuracy, and recorded compute runtime. Named policies remain separate references; they are not fixed counts.