CERTIFIED END-TO-END GRAPH-INDUCED SURROGATE COMPRESSION FOR BLACK-BOX LOCAL INTERVENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

In expensive graph intervention, the search domain is often not designed by the optimizer. A cheap graph heuristic or learned front end proposes a shortlist, and a black-box optimizer then spends simulator calls inside that domain. This fixed-domain view hides the central risk: the shortlist may be simple enough to compress, or it may be missing the edits that matter. Returning to the full library exposes this risk but destroys the simulator-call budget. We formulate the missing layer as auditable post-screening: the system returns not only a deployment, but also queried interactions, surrogate fit, omitted-candidate risk, and a deploy / expand / enrich / fallback recommendation. Local footprints induce a shortlist interaction graph; clique queries and Möbius inversion recover a graph-induced surrogate whose query complexity is controlled by local overlap and bag load rather than ambient library size. Outside, joint, and sparse-clique certificates turn residual shortlist risk into one-sided measurable quantities. The practical system is pairwise, with general-order theory as the structural query template; on moderate real tasks and localized large graphs, it reaches high-call greedy references in the main giant-connected-component (GCC) rows using up to about 70% fewer simulator calls, while the same run records shortlist credibility, certificate evidence, and next-call repair decisions under the charged budget.

1 INTRODUCTION

The best intervention among a few candidate changes can still be a poor choice for the full network. As high-dimensional data grow and new network architectures emerge, larger graphs make this choice harder to check. Under cost constraints, applications such as epidemic containment and infrastructure protection seek to improve network outcomes by changing a limited number of nodes or edges (Tong et al., 2010). Choosing which changes to make requires costly simulations to evaluate how they jointly improve the network outcome. Local graph structure offers a cheap way to identify promising changes. Structural rules or learned models screen candidates using this information and retain a small set, called a shortlist. Established black-box optimization methods then use simulation results to choose which changes to make together within this set (Jones et al., 1998; Frazier, 2018). But screening may exclude changes that work well together. No optimizer restricted to the shortlist can recover them, while searching over all possible changes can exhaust the budget. How can we assess what screening has missed without paying to search every possibility?

We answer with auditable post-screening, positioned between shortlist construction and black-box optimization. The system builds a predictive model, called a surrogate, by simulating selected groups of changes. The model guides the selection of changes to execute, called a deployment. The system then checks the model's prediction errors and assesses whether changes excluded from the shortlist could yield a better solution. These checks guide whether to execute the deployment, expand the shortlist, evaluate additional interactions, or switch to direct search within the shortlist. All additional simulations count toward the same evaluation budget.

The inputs are a graph G_0, a library of possible changes $\mathcal{A}$, and an initial shortlist $S_0 \subseteq \mathcal{A}$. Each change is assigned a footprint computed from the graph using a specified rule, such as

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

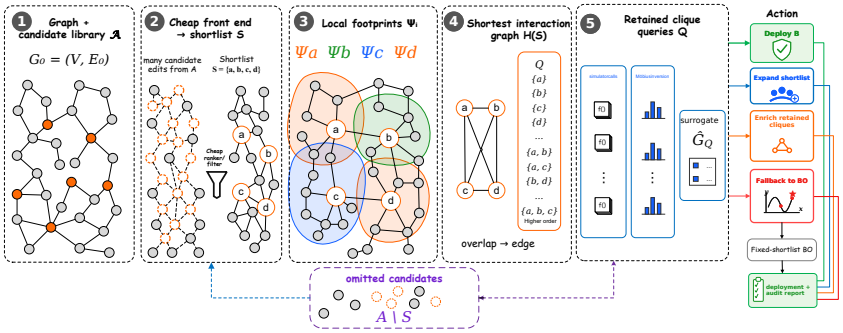


Figure 1: Auditable graph-induced post-screening. A cheap front end proposes a shortlist; local footprints induce a shortlist interaction graph; clique queries recover a compressed surrogate; outside and sparse-clique certificates bound shortlist risk when their stated conditions hold, trigger charged expansion, enrich the retained query set, or hand off to fixed-shortlist search. Output is a deployment plus a record of the queries and diagnostics.

the two-hop neighborhood of the modified node. The deployment budget k limits executed changes, while the simulator-call budget limits all simulator evaluations.

Our framework addresses three questions left open by existing methods (Section 2). How can the shortlist be evaluated without querying all candidate groups? How can excluded changes be judged without exhaustive search? What should happen when either test fails? **First**, a tree decomposition (Cygan et al., 2015) maps each footprint to the bags it intersects (its bag shadow). Overlapping shadows induce an interaction graph whose cliques support the model's nonzero interactions. Query cost is therefore controlled by shadow overlap and bag load, the number of footprints meeting each bag, allowing efficient recovery when this structure is sparse. **Next**, we bound the potential gains from excluded changes. Together with a bound on surrogate error, this gives a one-sided upper bound on screening loss, making omission risk measurable. **Finally**, these checks guide shortlist expansion, further interaction queries, or fallback to direct search. Our contributions are:

1. **Auditable post-screening.** We make the shortlist itself an object of evaluation. Our framework selects an intervention and assesses the risk of excluding other changes. This assessment guides shortlist expansion, additional interaction queries, or fallback within a shared simulation budget.
2. **Graph-induced surrogate recovery.** For any interaction order, we prove that clique queries are sufficient and worst-case necessary for exact coefficient recovery. For connected footprints, bag-shadow overlap and bag load bound the queries needed for exact recovery.
3. **Bounds on screening and approximation loss.** We derive certificates that bound the potential gains from excluded changes and the error from omitted interaction terms. Value intervals bound reconstruction error even when omitted interactions have high order. Together, these bounds control deployment loss when the stated model and certificate assumptions hold.
4. **Evaluation efficiency across scales and interaction orders.** On six moderate graph-intervention tasks, the 35-call system matches the exact shortlist optimum and a 90-call greedy reference. On localized intervention tasks in ogbn-arxiv (1.16 million edges), ogbn-products (61.9 million), and Orkut (117.2 million), the system matches a 91-call greedy reference with 36 calls. On real five-drug combination data, increasing the model order from two to five progressively reduces the average gap between the selected intervention and the measured optimum.

2 Related Work

Black-box, Bayesian, and combinatorial optimization. Bayesian optimization allocates expensive evaluations within a specified domain (Jones et al., 1998; Shahriari et al., 2016; Frazier, 2018; Garnett, 2023). High-dimensional Bayesian optimization (BO) improves efficiency through additive structure, trust regions, or sparsity (Kandasamy et al., 2015; Rolland et al., 2018; Eriksson et al., 2019; Eriksson and Jankowiak, 2021). Active-subspace evolution strategies also reduce black-box query cost (Choromanski et al., 2019). Heteroscedastic Bayesian optimization handles varying

noise and non-stationarity in hyperparameter optimization (Cowen-Rivers et al., 2022). Graph and combinatorial BO extend this machinery to discrete domains: COMBO uses diffusion kernels on graph products (Oh et al., 2019), BayesOptG optimizes over graph nodes (Wan et al., 2023), node-subset BO searches subset graphs (Liang et al., 2024), Bayesian optimization of combinatorial structures (BOCS) uses semidefinite relaxations (Baptista and Poloczek, 2018), and Bounce adapts trust regions (Papenmeier et al., 2023). These methods guide search within the optimization domain; our certificates bound the loss from excluding actions from a shortlist under stated assumptions.

Graph intervention and vulnerability. Graph intervention begins one step earlier: which edits deserve consideration? Centrality, vulnerability, and task-specific heuristics rank candidate edits cheaply (Tong et al., 2010). Seed-selection methods also optimize diffusion objectives (Wang et al., 2017). This speed makes large libraries manageable, but offers no guarantee that discarded candidates are harmless. We keep these front ends and check their output before accepting the resulting search domain. Post-screening audits omissions and charges every repair call to the shared evaluation budget.

Sparse recovery and graph decompositions. Auditing a shortlist requires learning joint effects from a limited set of queries. Möbius inversion, exact learning, group testing, and sparse Boolean recovery show how targeted queries recover structured functions (Hwang, 1972; Du and Hwang, 1999; Aldridge et al., 2026; Kushilevitz and Mansour, 1993; Bshouty, 2013; O'Donnell, 2014); tree decompositions and chordal subtree intersection graphs show how geometry controls complexity (Gavril, 1974; Koller and Friedman, 2009; Cygan et al., 2015). We bring them to the shortlist interface: footprints determine which groups to query, and certificates guide whether to deploy or repair the shortlist.

3 Graph-Induced Certified Post-Screening

3.1 Post-screening contract

To model joint effects with fewer queries, we use the footprint structure introduced above. Our comparator permits interactions only between edits whose bag shadows overlap, forming a graph on the shortlisted candidates. A nonzero joint interaction in this model must lie on a clique. Empty-set and clique queries therefore recover the model; its other values decompose into these terms. Möbius inversion turns these values into a surrogate that can score every feasible deployment within the shortlist.

Let $G_0 = (V, E_0)$ be a fixed graph and $\mathcal{A} = \{a_1, \dots, a_M\}$ a candidate edit library of ambient size M. A front end returns a shortlist S of size r, and the deployment family is

$$\mathcal{X}_k(S) = \{B \subseteq S : |B| \le k\}.$$

Each candidate has an expanded footprint $\Psi_i \subseteq V$. Given a tree decomposition $\mathcal{T} = (T, \{X_t\})$, the bag shadow of a connected footprint consists of all bags that intersect it:

$$K_i^{\mathcal{T}} = \{t \in V(T) : X_t \cap \Psi_i \neq \varnothing\}.$$

Two candidates interact when their shadows intersect. This gives the shortlist interaction graph

$$H_{\mathcal{T}}(S) = (S, E_{\mathcal{T}}), \qquad \{i, j\} \in E_{\mathcal{T}} \iff K_i^{\mathcal{T}} \cap K_j^{\mathcal{T}} \neq \varnothing.$$

For an interaction order d, let $\mathcal{C}_d(H)$ be all nonempty cliques of H of size at most d, and define

$$\mathcal{G}_{H,d} = \left\{ G_\theta(B) = \theta_\varnothing + \sum_{U \in \mathcal{C}_d(H)} \theta_U \mathbf{1}\{U \subseteq B\} \right\}.$$

This is an overlapping additive model with graph-induced clique groups: each factor depends on $B \cap U$, with a converse for clique-supported multilinear models (Proposition 16). Thus the graph supplies groups assumed known in the theoretical analysis of additive BO (Kandasamy et al., 2015).

These definitions specify the simulator queries. The interaction graph determines the candidate groups, d limits their interaction order, and k limits which terms can affect deployment. We do not solve candidate generation or full-library black-box optimization. Algorithm 1 routes diagnostics into deploy, expand, enrich, or fallback using thresholds (τ_o, τ_s, τ_v), call cap $B_{\max}$, validation budget B_{val}, and retained-clique budget L when the full clique family is too large to query.

Algorithm 1 Auditable graph-induced post-screening

Input: $G_0, \mathcal{A}, S_0, k, d, \Psi, \tau, B_{\max}, B_{\mathrm{val}}, L$, MODE
Output: $\widehat{B}, \mathcal{Q}, \widehat{G}_{\mathcal{Q}}, C_{\mathrm{out}}, \mathsf{Tail}_{L,k}(G), \widehat{\varepsilon}_{\mathrm{val}}$ and call report

$S \leftarrow S_0$; initialize query cache and call count
reserve one call for final deployment at every stage
  update local $\mathcal{T}(S)$; compute $K_i^{\mathcal{T}}$; build $H_{\mathcal{T}}(S)$
  score cliques by singleton/front-end and footprint-overlap bounds
  $\mathcal{Q}_0 \leftarrow \textsc{RetainedCliques}(H_{\mathcal{T}}(S), d, L, \text{MODE})$
  $\mathcal{Q} \leftarrow \textsc{DownwardClosure}(\mathcal{Q}_0)$
  reuse cached y_U; query all missing $U \in \{\varnothing\} \cup \mathcal{Q}$ if budget permits; otherwise return FALLBACKREPORT
  $\widehat{\theta}_U \leftarrow \sum_{W \subseteq U} (-1)^{|U|-|W|} y_W$ (Möbius inversion)
  $\widehat{G}_{\mathcal{Q}}(B) \leftarrow \widehat{\theta}_{\varnothing} + \sum_{U \in \mathcal{Q}: U \subseteq B} \widehat{\theta}_U$
  compute $C_{\mathrm{out}}(S)$, $\mathsf{Tail}_{L,k}(G)$, and $\widehat{\varepsilon}_{\mathrm{val}}$; charge diagnostics within B_{val} and the remaining budget
  if $C_{\mathrm{out}}(S) > \tau_o$ and calls remain **then** $S \leftarrow S \cup \textsc{TopOmitted}(C_{\mathrm{out}})$
  if S changed **then** return to step 3
if $\widehat{\varepsilon}_{\mathrm{val}} > \tau_v$ **then**
  if calls remain, enrich $\mathcal{Q}$ toward the next order and return to step 3; otherwise return FALLBACKREPORT with $\widehat{B}$ marked uncertified
if $\mathsf{Tail}_{L,k}(G) > \tau_s$ and $\widehat{\varepsilon}_{\mathrm{val}} \le \tau_v$ and calls remain **then**
  if a larger downward-closed family fits, increase L and return to step 3; otherwise return FALLBACKREPORT
if calls are exhausted or both probes stay below thresholds without a fitted family **then return** FALLBACKREPORT with $\widehat{B}$ marked uncertified
choose deployment $\widehat{B}$ maximizing $\widehat{G}_{\mathcal{Q}}$ over $\mathcal{X}_k(S)$
charge one evaluation of $F(\widehat{B})$; **return** the outputs listed above, with $\widehat{B}$ marked certified only when no fallback branch fired

Modes and actions. The retained-clique rule selects $\mathcal{Q}$: exact mode queries every admissible clique; sparse mode keeps singletons and high-score cliques; validation-driven mode enlarges $\mathcal{Q}$ only when residuals stay large. Expansion recomputes the local decomposition before new shadows are audited. A large validation residual enriches toward the next order when calls remain, and otherwise forces a fallback with the deployment marked uncertified. The sparse tail first increases L when fit is acceptable. Fallback handles large tail plus large residual, exhausted budgets, or missing validation, so the same contract covers exact recovery, practical thinning, charged expansion, and residual-triggered refinement.

Diagnostics allocate the next simulator call: large outside mass targets omitted candidates, large sparse tail enriches the retained clique family, and large validation residual recommends fallback rather than forced compression.

For $d > 2$, retained cliques include downward closure so Möbius coefficients are identifiable; pairwise mode uses the empty set, singletons, and retained pairs. Sparse mode scores cliques by footprint overlap and singleton mass, while the outside proxy $C_{\mathrm{out}}^{\mathrm{proxy}}(S)$ sums Jaccard-weighted singleton envelopes over the top-k omitted candidates (definitions in Appendix E.1). This proxy ranks expansion candidates; it acts as the certificate $C_{\mathrm{out}}(S)$ of Corollary 5 only when its envelopes upper-bound the comparator marginals, a condition the exhaustive offline audit checks separately (Table 23), and the online routing study (Appendix E.4.3) calibrates residual thresholds as an empirical proxy for this decision rule rather than implementing it. If $B_{\mathrm{val}} = 0$, $\widehat{\varepsilon}_{\mathrm{val}}$ is ex-post audit only and cannot trigger online fallback.

3.2 Clique recovery and exact-query optimality

The interaction graph tells us which groups to query. We now ask how accurately these queries determine deployment values. Write H for the shortlist interaction graph. For objective F, define $\varepsilon^{\star}_{H,d}(F)$ as the best uniform approximation error by $\mathcal{G}_{H,d}$ on queried cliques and feasible deployments, and let $\Gamma_{d,k}(H) = \max_{B \in \mathcal{X}_k(S)} |\mathcal{C}_d(H[B])|$.

Theorem 1 (Agnostic clique recovery)**.** *Assume $d \le k$, so every queried clique is a feasible deployment. Query all cliques in $\mathcal{C}_d(H)$ and reconstruct $\widehat{G}$ by Möbius inversion. Then*

$$\sup_{B \in \mathcal{X}_k(S)} |F(B) - \widehat{G}(B)| \le \big(2 + \Gamma_{d,k}(H)\, 2^d\big)\varepsilon^{\star}_{H,d}(F),$$

and the deployment regret is at most twice this quantity. The number of simulator queries is exactly $1 + |\mathcal{C}_d(H)|$, including one empty-set query and one query per admissible clique.

In words, if the objective is close to a graph-induced factor structure, the recovered surrogate carries a uniform error certificate, and the deployment chosen from it loses at most twice that error. The reconstruction is Möbius inversion on the clique-query poset: querying $F(\varnothing)$ and every admissible

clique recovers the coefficients θ_U by inclusion-exclusion, exact on every feasible deployment when F lies in the factor class. The factor 2^d comes from inversion over subsets of a queried clique. The theorem thus separates two issues: the graph-induced query count, which is observable, and the approximation error, which experiments measure by exact pool accounting and validation residuals.

Proposition 2 (Exact-query optimality). *For $\mathcal{G}_{H,d}$ with $d \le k$, deterministic coefficient identification requires $D = 1 + |\mathcal{C}_d(H)|$ queries. If a randomized algorithm uses $q < D$ noiseless queries, then*

$$\sup_{\theta} \mathbb{E}\|\widehat{\theta} - \theta\|_2^2 = \infty,$$

so no finite uniform mean-squared coefficient-recovery guarantee is possible.

Thus clique queries are sufficient and worst-case necessary for exact or uniformly controlled approximate coefficient recovery. This bound concerns coefficient recovery; matched-budget BO comparisons separately test deployment quality. Approximation leaves a validation residual directing enrichment or handoff to a generic optimizer as the pipeline's fallback route.

3.3 Topology-controlled query budgets

The recovery result makes the number of cliques our query cost. We next bound this count using footprint overlap. Let $\ell_t^{\mathcal{T}} = |\{i \in S : t \in K_i^{\mathcal{T}}\}|$ and $\eta^{\mathcal{T}} = \max_t \ell_t^{\mathcal{T}}$.

Theorem 3 (Bag-load compression). *For connected footprints, $H_{\mathcal{T}}(S)$ is chordal, $\mathrm{tw}(H_{\mathcal{T}}(S)) \le \eta^{\mathcal{T}} - 1$, and the following query-count bound holds in terms of the observed bag loads:*

$$|\mathcal{C}_d(H_{\mathcal{T}}(S))| \le \sum_{t \in V(T)} \sum_{s=1}^{d} \binom{\ell_t^{\mathcal{T}}}{s}.$$

If every shadow spans at most h bags and every bag load is at most η, then $|\mathcal{C}_d(H_{\mathcal{T}}(S))| \le hr \sum_{q=0}^{d-1} \binom{\eta-1}{q}$ for connected footprints under the stated shadow-span and bag-load conditions.

Topology replaces library size: query cost follows shortlisted-footprint overlap density, not the number of library edits. The ambient graph need not have small treewidth; the shortlist interaction graph is relevant. Long footprints increase the shadow-span bound h, concentrated candidates increase the bag-load bound η, and larger d expands the clique family; these measured quantities determine whether the system uses exact cliques, sparse retained cliques, or fallback.

For disconnected footprints, a measured extension forms component shadows and reports component-load or quotient-width diagnostics. The connected-shadow bound requires component overlaps satisfying the same bag-load conditions; a high-width quotient graph prescribes sparse thinning or fallback under the stated diagnostic rule when component bounds fail.

3.4 Outside-shortlist certificates

A small query budget within the shortlist does not show that it contains a good intervention. To bound the loss from excluded candidates, take a full-library comparator G and define the positive marginal envelope of $u \notin S$:

$$\mu_G(u) = \max_{A \subseteq [M] \setminus \{u\}, |A| \le k-1} [G(A \cup \{u\}) - G(A)]_+.$$

Any upper bound $c(u) \ge \mu_G(u)$ is a valid outside certificate for candidate u.

Theorem 4 (Conditional shortlist completion). *If G uniformly approximates F within ε on all deployments of size at most k, and $c_{(1)} \ge c_{(2)} \ge \cdots$ are the omitted-candidate certificates, then*

$$F(B^\star_{\mathrm{full}}) - F(B^\star_S) \le 2\varepsilon + \sum_{m=1}^{k} c_{(m)}.$$

For pairwise G, the minimal one-step certificate is

$$\mu_G(u) = \Big[\theta_u + \sum_{m=1}^{k-1} \theta^+_{u,(m)}\Big]_+.$$

The outside term is conditional but operational: it assumes a bounded-error comparator surrogate calibrated on an enlarged audit pool. The pairwise expression is the smallest possible one-step certificate: no smaller per-candidate score upper-bounds every marginal contribution of u. For a monotone submodular comparator, diminishing returns reduces this envelope to a singleton gain (Corollary 11; Nemhauser et al., 1978). When comparator error and certificate mass are small, the bound rules out any outside deployment improving on the shortlist optimum beyond the stated slack. Top-k mass relaxes joint feasibility, and the pairwise surrogate trades tightness for call economy; certificate mass is a repair trigger, not an omitted-gain estimate.

A joint feasibility-aware certificate directly solves the omitted-subset problem over the omitted pool. It is no larger than the top-k one-step mass and explains its conservatism: the decoupled certificate can count several omitted candidates as simultaneously useful even when the budget or pair structure prevents that combination from being jointly deployable.

Corollary 5 (Pairwise deployed post-screening guarantee)**.** *Consider the practical pairwise system on an audit pool $P \supseteq S$. Suppose a pairwise comparator approximates F on P within ε, the outside certificate on $P \setminus S$ has top-k mass $C_{\mathrm{out}}(S)$, the exact pairwise comparator on S has uniform error at most δ_S over $\mathcal{X}_k(S)$, and the retained-pair surrogate has omitted-pair tail at most $\mathsf{Tail}_{L,k}(G)$. If $\widehat{B}$ maximizes the retained surrogate over $\mathcal{X}_k(S)$, then*

$$F(B_P^\star) - F(\widehat{B}) \le 2\varepsilon + C_{\mathrm{out}}(S) + 2\delta_S + 2\mathsf{Tail}_{L,k}(G).$$

Corollary 5 is the deployed theorem: the report separates pairwise risk into the measured outside mass, sparse retained-pair tail, and validation residual. The statement is pool-conditional when P is an audit pool and full-library only when comparator and certificates are available there. Experiments estimate δ_S on audit/validation subsets rather than assert it uniformly, reporting any persistent residual as an observable diagnostic alongside the certificate terms.

Charged expansion removes the largest outside certificates and reuses prior queries. Exact fitting on S' costs $|\mathcal{C}_d(H(S')) \setminus \mathcal{Q}_{\mathrm{seen}}|$ new calls. The one-step and multi-round bounds, including separate audit and deployment charges, are stated in Corollaries 14 and 15.

3.5 Sparse fallback and diagnostic actions

Outside certificates guide expansion, but querying every clique in the expanded shortlist may exceed the remaining budget. The algorithm then keeps all singletons and a scored subset $\mathcal{Q} \subseteq \mathcal{C}_d(H)$, using $\widehat{G}_{\mathcal{Q}}$ to predict the values of feasible deployments from retained terms.

Proposition 6 (Certified sparse clique compression)**.** *For an exact $G \in \mathcal{G}_{H,d}$,*

$$\sup_{B \in \mathcal{X}_k(S)} |G(B) - \widehat{G}_{\mathcal{Q}}(B)| \le \max_{B \in \mathcal{X}_k(S)} \sum_{U \in \mathcal{C}_d(H[B]) \setminus \mathcal{Q}} |\theta_U|.$$

The deployment selected by the sparse surrogate loses at most twice this omitted-tail bound.

Retaining selected pairs gives the $d = 2$ case of this bound. Even querying every pair can leave higher-order effects unmodeled. We next ask when those effects are small enough to preserve the deployment value or the selected optimum.

Proposition 7 (Pairwise sufficiency certificate)**.** *Let G be exact and let $T_{>2}(B) = \sum_{U \subseteq B, |U| \ge 3} |\theta_U|$ be the higher-order mass inside a deployment. If $\max_{B \in \mathcal{X}_k(S)} T_{>2}(B) \le \tau$, then the deployment chosen by the pairwise (d=2) surrogate has loss at most 2τ relative to the optimum of G. If its best-to-second-best deployment gap exceeds 2τ, the pairwise surrogate also selects an optimum of G.*

This is Proposition 6 with $\mathcal{Q}$ equal to all cliques of size at most two: truncating the order drops exactly the terms in $T_{>2}$. Validation residuals check pairwise fit; a uniform bound on the higher-order tail needs separate evidence. Table 3 instantiates the boundary: on the spectral tasks third-order querying does not change the deployment, while the query count grows by 6–8×.

Increasing d adds interactions to the model but also requires more queries. When calls remain, persistent validation error can trigger these extra queries before fallback.

A failed condition does not invalidate the pipeline; it determines the next action. Table 1 summarizes this operational interpretation throughout the empirical evaluation.

Table 1: Diagnostics guide deployment, shortlist expansion, clique enrichment, and fallback.

Quantity	Observable diagnostic	Action when large
$\lvert\mathcal{C}_d(H)\rvert$, width	query count, load proxy	thin cliques or lower d
$\varepsilon^{\star}_{H,d}(F)$ proxy	validation residual	richer cliques or BO fallback
Outside mass	top-k omitted certificate	expand shortlist
Sparse tail	omitted-clique tail	query more retained cliques
Both residual and tail	persistent model-fit failure	fixed-shortlist BO fallback

These results separate query cost (Theorem 3), model error (Theorem 1), shortlist loss (Theorem 4), and sparse approximation (Proposition 6). We next test when these distinctions matter for deployment under a shared call budget on graph-intervention tasks.

4 Experiments

4.1 Experiment Setting

Benchmarks and objectives. We evaluate across two regimes. The moderate suite uses CA-HepPh, Email-Enron, and com-DBLP from the Stanford Network Analysis Project (SNAP) (Leskovec and Krevl, 2014), with node-deletion objectives for giant-connected-component (GCC) attack and spectral suppression; breadth tests add edge deletion and radius-2 maximum coverage. The large suite uses ogbn-arxiv and ogbn-products from the Open Graph Benchmark (Hu et al., 2020), and Orkut from SNAP, testing capped local regions and charged block post-screening on the corresponding full graphs.

Candidate construction and footprints. Candidates are node deletions for main tasks; breadth uses edge deletions or coverage centers. Main defaults are $r = 24$, $k = 4$, $L = 9$, and radius-2 footprints. The 35 calls comprise one empty-set query, 24 singletons, nine pairs, and one deployment. Appendix sweeps also use $L = 10$, costing 36 calls. The implementation uses direct footprint overlap; bag-shadow overlap defines the theoretical comparator class. The front end combines degree, PageRank, k-core, and bridge scores; sensitivity tests isolate individual rankings or sample from the top pool.

Baselines and call accounting. We report objective calls separately from optimizer requests. On the shared shortlist, Bounce (Papenmeier et al., 2023) uses 10 seeds and 35 calls per seed; greedy is reported both at call 35 and as a 90-call reference. Expansion singletons are charged, whereas enlarged-pool diagnostics are ex-post. COMBO, GraphComBO, and BayesOptG use five seeds and 35 optimizer requests on the same domain; their actual objective-call counts are reported separately in Table 22. COMBO uses truncated posterior sampling; its settings and the implementations of all comparison methods are given in Appendix E.3.

Evaluation scope and metrics. Unless marked adaptive, main rows use the fixed-shortlist interface. Localized large-graph experiments operate on capped regions; full-graph block experiments partition the 24 candidates into four graph-local blocks and use 36 local-objective calls; full-graph deployment values are checked offline. Appendix tables specify protocols, seeds, means, and standard deviations, while deterministic wrappers are reported as point estimates. We report local and full-graph results separately because their objective calls evaluate different regions.

4.2 Results

We first compare deployment quality at fixed call budgets. We then test when pairwise models need repair and whether local evaluation remains useful on large graphs.

Call-efficient post-screening. With the shared shortlist and call accounting, the 35-call pairwise system reaches both the greedy-90 value and the exact shortlist optimum on all six rows of Table 2, using 61% fewer calls (Table 19). A 26-call singleton control also reaches these values; the water-network and interaction tests below isolate when pair queries improve deployment (Table 13). The 35-call system also exceeds Bounce's 35-call mean throughout, although Bounce's best seed ties on Email-Enron/spectral and com-DBLP/GCC. Table 2 reports deployment values and total calls together.

Table 2: Fixed-shortlist benchmark with shared call accounting. Ours uses 35 calls; singletons tie all six values in 26. Bounce: mean ± std, 10 seeds. Optimal checks all $C(24, 4) = 10{,}626$ subsets.

Task Calls	Greedy 35	Greedy 90	Bounce 35	Ours 35	Optimal —
CA-HepPh / GCC	18.000	**35.000**	30.600 ± 4.789	**35.000**	**35.000**
CA-HepPh / spectral	2.408	**4.812**	4.621 ± 0.104	**4.812**	**4.812**
Email-Enron / GCC	1421.0	**1995.0**	1681.5 ± 308.4	**1995.0**	**1995.0**
Email-Enron / spectral	4.378	**7.946**	7.551 ± 0.543	**7.946**	**7.946**
com-DBLP / GCC	57.0	**107.0**	93.1 ± 7.8	**107.0**	**107.0**
com-DBLP / spectral	2.210	**4.590**	4.271 ± 0.310	**4.590**	**4.590**

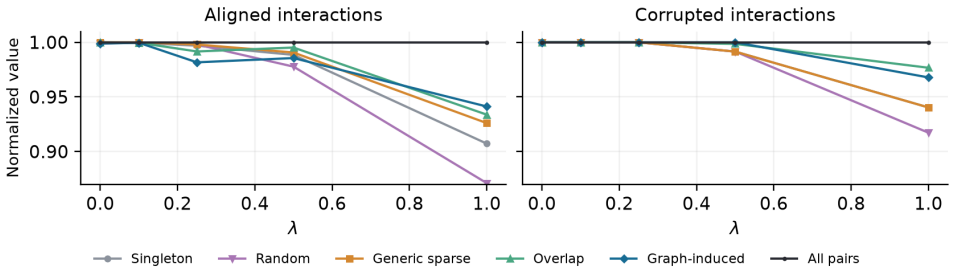


Figure 2: Pair-query mechanism check. Values use each task's exact shortlist optimum for normalization and are averaged over four tasks. For full curves under both interaction modes, see Figure 16.

Charged adaptive audit. Within the 35-call cap, proxy-ranked expansion tests whether the audit repairs rather than merely diagnoses the domain. It raises Email-Enron/GCC from 1995 to 2050, com-DBLP/GCC from 107 to 143, and com-DBLP/spectral from 4.590 to 4.650; it ties the other spectral rows and moves CA-HepPh/GCC from 35 to 27. Every added singleton is charged, and three deployments leave the shortlist. Table 12 reports all six outcomes and the charged calls.

Pair queries in water-network sensor placement. On Network 1 of the Battle of the Water Sensor Networks (BWSN) benchmark (Ostfeld et al., 2008), pair queries change deployment under all three front ends. With 504 training and 504 held-out contamination scenarios, the demand-ranked shortlist improves held-out volume reduction from 463.003 to 485.457 m^3, closing 64.0% of the gap from the singleton baseline to the held-out shortlist optimum. Greedy reaches 487.949 m^3 at 35 calls; pairs improve on singletons under all three front ends (Appendix E.4.1). Calls evaluate sensor subsets on a shared bank of EPANET water-network simulations (Rossman et al., 2020).

Interaction-aligned pair querying. The water-network result shows that pair queries can help. To isolate the role of graph structure, we add a footprint-aligned pair term of strength λ. At $\lambda = 1$, graph-induced pairs reach mean oracle-normalized value 0.941, compared with 0.907 for singletons and 0.926 for generic sparse pairs (Figure 2).

The eight-selector, four-task sweep tests both aligned and corrupted interactions (Table 17), separating the value of overlap information from the choice of scoring rule.

Pairwise deployment and higher-order diagnostics. When is higher order useful? Table 3 tests the two corrected-coordinate spectral tasks. Moving from $d = 2$ to $d = 3$ reduces maximum error but leaves deployed value unchanged at roughly 7× more queries; at $k = 4$ both orders are exact-optimal. On 760 constructed pure-triadic cases, order routing reaches zero regret in all cases at 387.03 mean calls, versus 1063.39 for full clique-$d = 3$ recovery (Appendix E.4.2).

Measured higher-order data provide a further check. Across 13,608 five-drug dose settings (Tekin et al., 2018), mean deployment regret falls from 11.475 to 7.504, 3.955, and 0 as d rises from two to five. Calls rise from 17 to 27, 32, and 33, including deployment. Fifth-order fitting queries every subset, so zero regret reflects full recovery on measured data (Appendix E.4.2).

Table 3: Order diagnostics on the corrected main shortlist; fitting calls exclude deployment. Support and width statistics for the separate appendix-union shortlist appear in Table 24.

Task	Order	Value	Max error	Queries
CA-HepPh	2	4.8124	0.00048	291
CA-HepPh	3	4.8124	0.0000032	2116
Email-Enron	2	7.9465	0.0243	296
Email-Enron	3	7.9465	0.0016	2220

End-to-end mechanism checks. Holding the shortlist fixed isolates pair selection from expansion. Degree priority and 100 random selectors at $L = 10$ match the graph rule's deployment values; their surrogate errors differ (Table 13). Across 36 further dataset/objective/front-end configurations, pairs improve one deployment; the water task above shows larger gains (Appendix E.4.1).

Calibrated audit-pool certificates. The 612-case audit-pool study measures coverage and threshold behavior separately: all proxies cover their reference loss, but false-positive and false-negative rates vary across tasks and pools (Table 18). An exhaustive audit over 32 candidates then checks both certificate bounds using uniform model error (Table 23).

Localized large-graph deployment. On capped regions of ogbn-arxiv, ogbn-products, and Orkut, the 36-call system matches the 91-call greedy reference (Table 21). These regions contain 40,193, 89,220, and 120,000 nodes, respectively. Each result uses one seed and a local objective; the full-graph test below checks whether these gains carry over to the full-graph objective.

Breadth, stress, and scale. Additional tasks test whether the diagnostics identify a need for more interaction queries. At 36 calls, ours reaches the shortlist oracle on $3/6$ tasks: all three edge-deletion/GCC tasks. The three coverage tasks reveal higher-order redundancy beyond $L = 10$ pairs (Table 10). Full-graph block post-screening and greedy-91 reach $210/210$ on ogbn-arxiv, $215/271$ on Orkut, and $44/60$ on ogbn-products. Nonzero residuals and tails still signal error in the fitted model after the outside mass has been exhausted (Table 15).

A 50-call adaptive test uses 29 calls to choose an action and 21 for repair and deployment. Frozen thresholds agree with the held-out reference action in $8/12$ cases. On the five cases labeled fallback by the offline reference, mean regret is 6.97, versus 23.35 for forced pairwise fitting and 13.75 for direct search from the start (Appendix E.4.3). This measures both the value and cost of auditing.

5 Conclusion

Auditable post-screening makes the search domain part of the decision. Local graph structure determines which interactions to query, and the resulting model guides deployment. Under the stated assumptions, certificates connect excluded candidates and unqueried interactions to deployment loss. These checks also guide expansion, further queries, or direct search within a shared budget. Experiments show fewer calls for equal deployment values, including on localized large-graph tasks, and better decisions when pair or higher-order interactions matter. The central contribution is a deployment backed by evidence about the shortlist that produced it.

ETHICS STATEMENT

This work studies graph interventions through simulations and evaluations on recorded data. Objectives such as component size or spectral change do not by themselves establish safety, fairness, or benefit to affected people. Network vulnerability analysis can also be misused to identify disruptive changes. Any operational use therefore requires domain review, authorization, and checks of consequences beyond the modeled objective. The certificates apply only under their stated assumptions and provide no guarantee about these broader consequences.

REPRODUCIBILITY STATEMENT

The formal assumptions, additional results, and proofs are provided in Appendices C and D. Appendix E describes the algorithms, data preparation, baseline settings, and simulator-call accounting. Appendices F and G report additional results and diagnostics. The accompanying artifact contains Python code, pinned dependencies, a Makefile, and saved CSV/JSON results. The README lists the experiment entry points. The evaluation protocols distinguish queries available to each method from offline checks used to assess its results.

AI USE STATEMENT

We used AI assistants to help revise manuscript text, review theoretical arguments and proofs, check references, and prepare LaTeX layouts. These tools also assisted with code refactoring, experiment execution, and result checks. The reported numerical results come from saved experimental outputs. AI assistance supports manuscript preparation and review; it does not replace author verification of mathematical claims, references, or experimental evidence. The authors take responsibility for the final content of the paper and all reported results.

REFERENCES

Matthew Aldridge, Oliver Johnson, and Jonathan Scarlett. Group testing: An information theory perspective. *Foundations and Trends in Communications and Information Theory*, 23(1–2):1–221, 2026. Second edition. https://doi.org/10.1108/FTCIT-11-2025-0150.

Ricardo Baptista and Matthias Poloczek. Bayesian optimization of combinatorial structures. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of Proceedings of Machine Learning Research, pages 462–471, 2018. https://proceedings.mlr.press/v80/baptista18a.html.

Nader H. Bshouty. Exact learning from membership queries: Some techniques, results and new directions. In *Algorithmic Learning Theory*, volume 8139 of Lecture Notes in Computer Science, pages 33–52. Springer, 2013. https://doi.org/10.1007/978-3-642-40935-6_4.

Krzysztof M. Choromanski, Aldo Pacchiano, Jack Parker-Holder, Yunhao Tang, and Vikas Sindhwani. From Complexity to Simplicity: Adaptive ES-Active Subspaces for Blackbox Optimization. In *Advances in Neural Information Processing Systems*, volume 32, 2019. https://proceedings.neurips.cc/paper_files/paper/2019/file/88bade49e98db8790df275fcebb37a13-Paper.pdf.

Alexander I. Cowen-Rivers, Wenlong Lyu, Rasul Tutunov, Zhi Wang, Antoine Grosnit, Ryan Rhys Griffiths, Alexandre Max Maravel, Hao Jianye, Jun Wang, Jan Peters, and Haitham Bou-Ammar. HEBO: Pushing The Limits of Sample-Efficient Hyperparameter Optimisation. *Journal of Artificial Intelligence Research*, 74:1269–1349, 2022. https://www.jair.org/index.php/jair/article/view/13643.

Marek Cygan, Fedor V. Fomin, Łukasz Kowalik, Daniel Lokshtanov, Dániel Marx, Marcin Pilipczuk, Michał Pilipczuk, and Saket Saurabh. *Parameterized Algorithms*. Springer, 2015. https://doi.org/10.1007/978-3-319-21275-3.

Ding-Zhu Du and Frank K. Hwang. *Combinatorial Group Testing and Its Applications*, 2nd edition. World Scientific, 1999. https://doi.org/10.1142/4252.

David Eriksson and Martin Jankowiak. High-dimensional Bayesian optimization with sparse axis-aligned subspaces. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of Proceedings of Machine Learning Research, pages 493–503, 2021. https://proceedings.mlr.press/v161/eriksson21a.html.

David Eriksson, Michael Pearce, Jacob Gardner, Ryan D. Turner, and Matthias Poloczek. Scalable Global Optimization via Local Bayesian Optimization. In *Advances in Neural Information Processing Systems*, volume 32, 2019. https://proceedings.neurips.cc/paper_files/paper/2019/file/6c990b7aca7bc7058f5e98ea909e924b-Paper.pdf.

Julie Fournet and Alain Barrat. Contact patterns among high school students. *PLOS ONE*, 9(9):e107878, 2014. https://doi.org/10.1371/journal.pone.0107878.

Peter I. Frazier. Bayesian Optimization. In *Recent Advances in Optimization and Modeling of Contemporary Problems*, INFORMS TutORials in Operations Research, pages 255–278. INFORMS, 2018. https://doi.org/10.1287/educ.2018.0188.

Roman Garnett. *Bayesian Optimization*. Cambridge University Press, 2023. https://doi.org/10.1017/9781108348973.

Fanica Gavril. The intersection graphs of subtrees in trees are exactly the chordal graphs. *Journal of Combinatorial Theory, Series B*, 16(1):47–56, 1974. https://doi.org/10.1016/0095-8956(74)90094-X.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open Graph Benchmark: Datasets for Machine Learning on Graphs. In *Advances in Neural Information Processing Systems*, volume 33, pages 22118–22133, 2020. https://proceedings.neurips.cc/paper_files/paper/2020/file/fb60d411a5c5b72b2e7d3527cfc84fd0-Paper.pdf.

F. K. Hwang. A method for detecting all defective members in a population by group testing. *Journal of the American Statistical Association*, 67(339):605–608, 1972. https://doi.org/10.1080/01621459.1972.10481257.

Donald R. Jones, Matthias Schonlau, and William J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, 1998. https://doi.org/10.1023/A:1008306431147.

Kirthevasan Kandasamy, Jeff Schneider, and Barnabas Poczos. High dimensional Bayesian optimisation and bandits via additive models. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of Proceedings of Machine Learning Research, pages 295–304, 2015. https://proceedings.mlr.press/v37/kandasamy15.html.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009. https://mitpress.mit.edu/9780262013192/probabilistic-graphical-models/.

Eyal Kushilevitz and Yishay Mansour. Learning decision trees using the Fourier spectrum. *SIAM Journal on Computing*, 22(6):1331–1348, 1993. https://doi.org/10.1137/0222080.

Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford Large Network Dataset Collection. http://snap.stanford.edu/data, 2014.

Huidong Liang, Xingchen Wan, and Xiaowen Dong. Bayesian Optimization of Functions over Node Subsets in Graphs. In *Advances in Neural Information Processing Systems*, volume 37, pages 38199–38234, 2024. https://proceedings.neurips.cc/paper_files/paper/2024/file/437cd2749391ad40f67e4dd1d87c4596-Paper-Conference.pdf.

George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 14:265–294, 1978. https://doi.org/10.1007/BF01588971.

Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014. https://doi.org/10.1017/CBO9781139814782.

Changyong Oh, Jakub Tomczak, Efstratios Gavves, and Max Welling. Combinatorial Bayesian Optimization using the Graph Cartesian Product. In *Advances in Neural Information Processing Systems*, volume 32, 2019. https://proceedings.neurips.cc/paper_files/paper/2019/file/2cb6b10338a7fc4117a80da24b582060-Paper.pdf.

Avi Ostfeld et al. The Battle of the Water Sensor Networks (BWSN): A Design Challenge for Engineers and Algorithms. *Journal of Water Resources Planning and Management*, 134(6):556–568, 2008. https://doi.org/10.1061/(ASCE)0733-9496(2008)134:6(556).

Leonard Papenmeier, Luigi Nardi, and Matthias Poloczek. Bounce: Reliable High-Dimensional Bayesian Optimization for Combinatorial and Mixed Spaces. In *Advances in Neural Information Processing Systems*, volume 36, pages 1764–1793, 2023. https://proceedings.neurips.cc/paper_files/paper/2023/file/05d2175de7ee637588d1b5ced8b15b32-Paper-Conference.pdf.

Paul Rolland, Jonathan Scarlett, Ilija Bogunovic, and Volkan Cevher. High-dimensional Bayesian optimization via additive models with overlapping groups. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of Proceedings of Machine Learning Research, pages 298–307, 2018. https://proceedings.mlr.press/v84/rolland18a.html.

Lewis A. Rossman, Hyoungmin Woo, Michael Tryby, Feng Shang, Robert Janke, and Terranna Haxton. *EPANET 2.2 User Manual*. U.S. Environmental Protection Agency, EPA/600/R-20/133, 2020. https://usepa.github.io/EPANET2.2/.

Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P. Adams, and Nando de Freitas. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2016. https://doi.org/10.1109/JPROC.2015.2494218.

Elif Tekin, Cynthia White, Tina Manzhu Kang, Nina Singh, Mauricio Cruz-Loya, Robert Damoiseaux, Van M. Savage, and Pamela J. Yeh. Prevalence and patterns of higher-order drug interactions in Escherichia coli. *npj Systems Biology and Applications*, 4:31, 2018. https://doi.org/10.1038/s41540-018-0069-9.

Hanghang Tong, B. Aditya Prakash, Charalampos Tsourakakis, Tina Eliassi-Rad, Christos Faloutsos, and Duen Horng Chau. On the vulnerability of large graphs. In *2010 IEEE International Conference on Data Mining*, pages 1091–1096, 2010. https://doi.org/10.1109/ICDM.2010.54.

Xingchen Wan, Pierre Osselin, Henry Kenlay, Binxin Ru, Michael A. Osborne, and Xiaowen Dong. Bayesian Optimisation of Functions on Graphs. In *Advances in Neural Information Processing Systems*, volume 36, pages 43012–43040, 2023. https://proceedings.neurips.cc/paper_files/paper/2023/file/86419aba4e5eafd2b1009a2e3c540bb0-Paper-Conference.pdf.

Zhefeng Wang, Yu Yang, Jian Pei, Lingyang Chu, and Enhong Chen. Activity maximization by effective information diffusion in social networks. *IEEE Transactions on Knowledge and Data Engineering*, 29(11):2374–2387, 2017. https://doi.org/10.1109/TKDE.2017.2740284.

Appendix

Contents

A Notation . . . 15

B Graph-Mechanism Figure Atlas . . . 16

B.1 Formal Mechanism Schematics . . . 16

B.2 Real-Graph Footprints and Interactions . . . 20

C Additional Formal Results . . . 20

D Proofs . . . 22

D.1 Connected Shadows and Agnostic Reconstruction . . . 22

D.2 Bag-Load Compression and Exact Lower Bound . . . 23

D.3 Shortlist Completion and Certificate Minimality . . . 25

D.4 Joint, Noisy, Sparse, and Thin Certificates . . . 27

D.5 Scaling and Validation Proxy Proofs . . . 28

E Experimental Details . . . 29

E.1 Algorithmic and Implementation Details . . . 29

E.3 Evaluation Protocol and Baseline Implementations . . . 31

F Supplementary Tables . . . 35

F.1.1 Main-Text Deferred Tables . . . 38

F.1.2 Sensitivity, Robustness, and Stress Tests . . . 39

G Supplementary Figures . . . 45

A NOTATION

Table 4: Notation for the graph, candidate edits, surrogate, certificate terms, and call budgets.

Symbol	Meaning
Problem setting	
$G_0 = (V, E_0)$	fixed base graph
$\mathcal{A}$, $M = \lvert\mathcal{A}\rvert$	candidate edit library and its ambient size
S, r	front-end shortlist and its size
k, $\mathcal{X}_k(S)$	deployment budget; feasible deployments $\{B \subseteq S : \lvert B\rvert \le k\}$
F	black-box simulator objective
$\widehat{B}$	returned deployment
Footprints and interaction graph	
Ψ_i	local footprint of candidate i
$\mathcal{T} = (T, \{X_t\})$	tree decomposition
$K_i^{\mathcal{T}}$	bag shadow of a footprint
$H_{\mathcal{T}}(S)$	shortlist interaction graph (edge = shadows intersect)
d, $\mathcal{C}_d(H)$	interaction order; cliques of H of size at most d
$\ell_t^{\mathcal{T}}$, $\eta^{\mathcal{T}}$, h	bag load, maximum bag load, shadow span
Surrogate	
G_θ, θ_U	graph-induced surrogate and its Möbius coefficients
$\varepsilon^\star_{H,d}(F)$	best uniform approximation error by $\mathcal{G}_{H,d}$
$\Gamma_{d,k}(H)$	maximum clique count inside a size-k deployment
$\mathcal{Q}$, L	retained clique family and its budget
Certificates and diagnostics	
$\mu_G(u)$, $c(u)$	marginal envelope of omitted u; its computable upper bound
$C_{\mathrm{out}}(S)$	top-k outside certificate mass
$\mathsf{Tail}_{L,k}(G)$	omitted-pair (sparse retained) tail
δ_S, $\widehat{\varepsilon}_{\mathrm{val}}$	audit/validation residual and its estimate
P	audit pool ($P \supseteq S$)
Budgets and thresholds	
$B_{\max}$, B_{val}	simulator-call cap; validation-call budget
τ_o, τ_s, τ_v	action thresholds on outside mass, sparse tail, residual

B GRAPH-MECHANISM FIGURE ATLAS

B.1 FORMAL MECHANISM SCHEMATICS

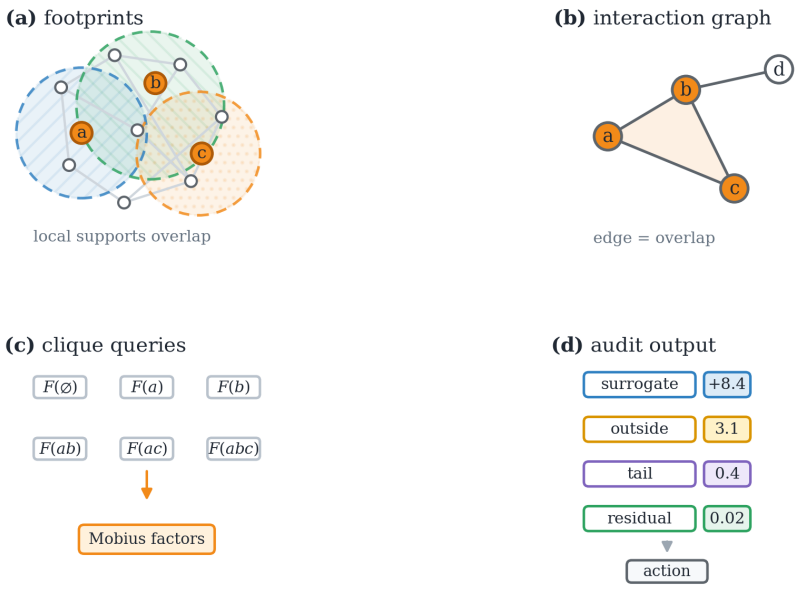


Figure 3: Footprint overlap is the visual origin of admissible clique queries. Local candidate footprints induce a shortlist interaction graph; retained cliques in that graph are the only candidate groups queried before Möbius reconstruction and certificate reporting.

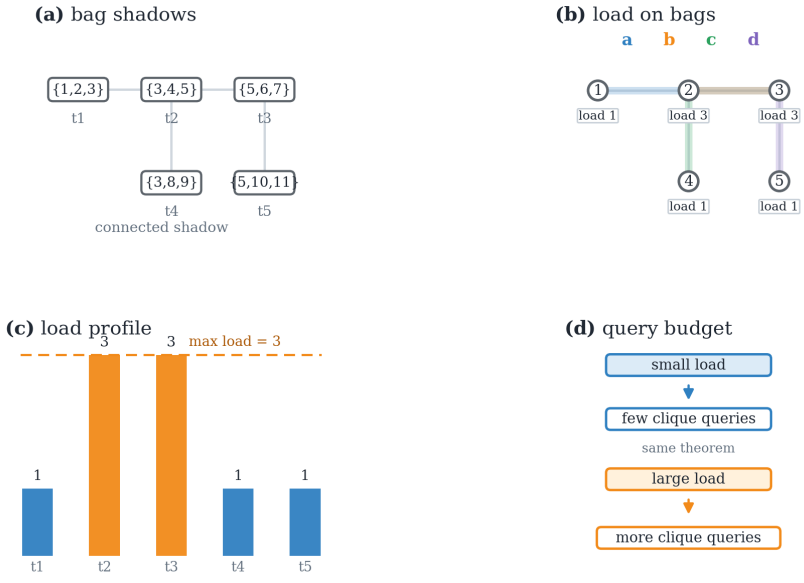


Figure 4: Bag shadows turn local graph structure into a measurable query budget. Candidate shadows live on tree-decomposition bags, bag load counts how many shortlisted candidates meet each bag, and high load is exactly where admissible clique counts grow.

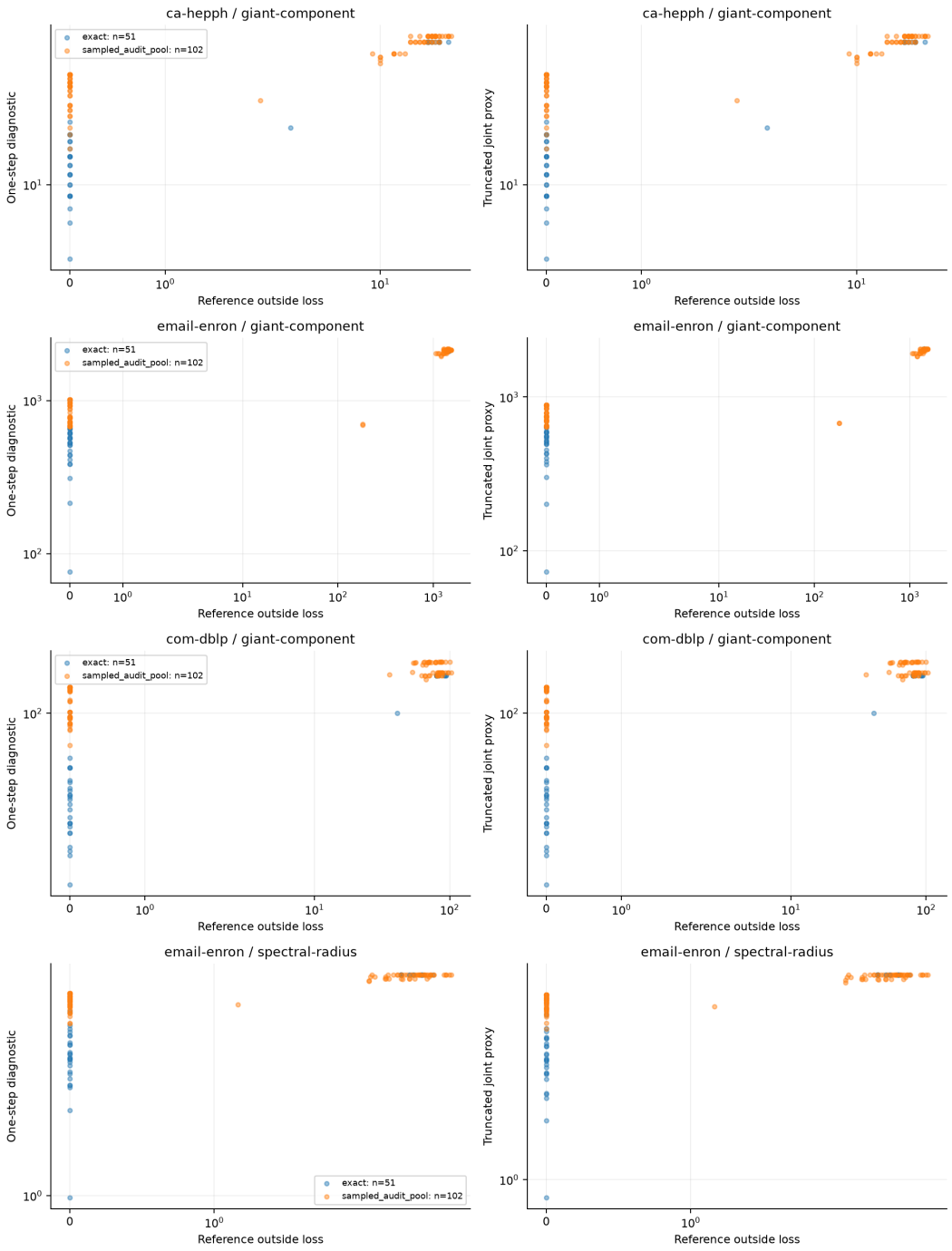


Figure 5: Audit-pool proxy values versus reference outside loss. Panels separate tasks and exact or sampled references; threshold behavior is evaluated in Table 18.

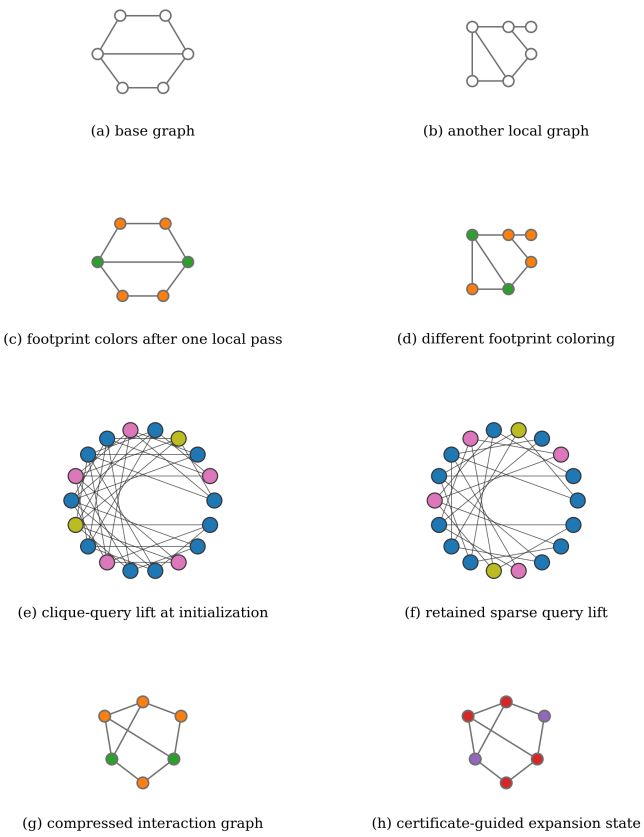


Figure 6: Graph-induced surrogate compression as a formal mechanism. The panels show local footprints, bag shadows, the induced shortlist interaction graph, admissible clique factors, sparse retained-clique fallback, and certificate-driven expansion.

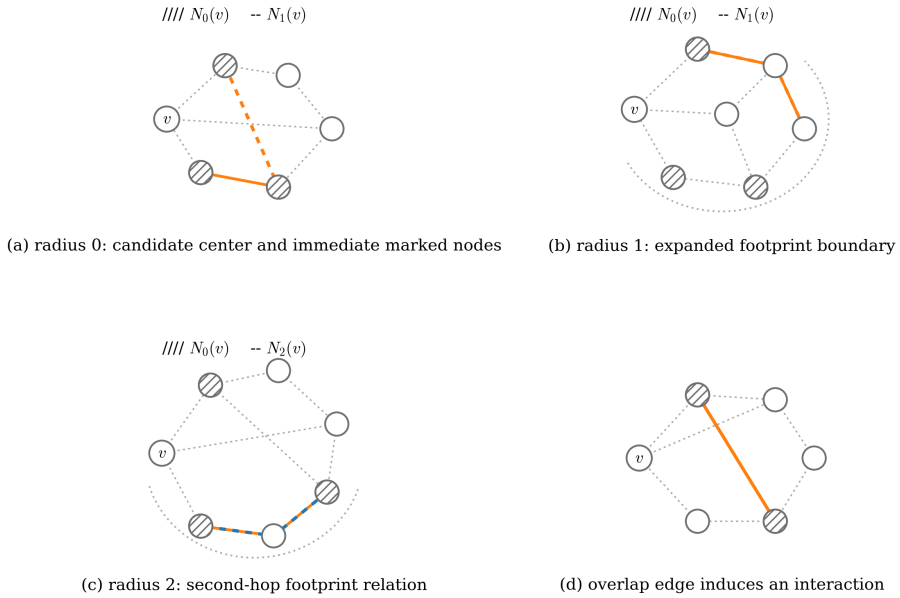
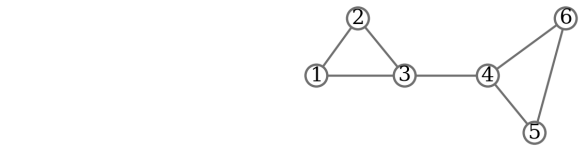
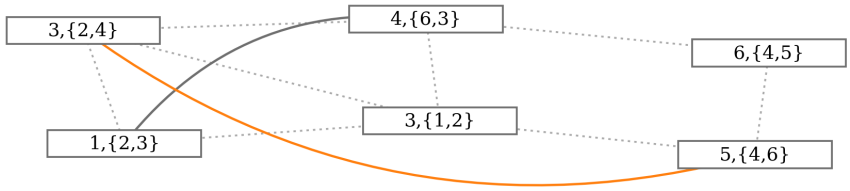


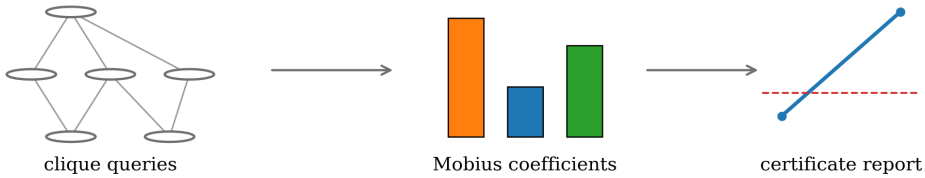
Figure 7: Formal graph examples underlying the assumptions. The panels separate footprint overlap, non-overlap, tree-decomposition witness bags, and certificate-driven shortlist expansion.



(a) input graph and local candidates



(b) bag-shadow witness graph and retained repair edge



(c) query reconstruction and outside-certificate decision

Figure 8: From clique queries to deployment and certificates. The queries define a poset, Möbius inversion produces coefficients, and omitted candidates receive certificate mass to guide expansion of the shortlist or selection of the final feasible deployment.

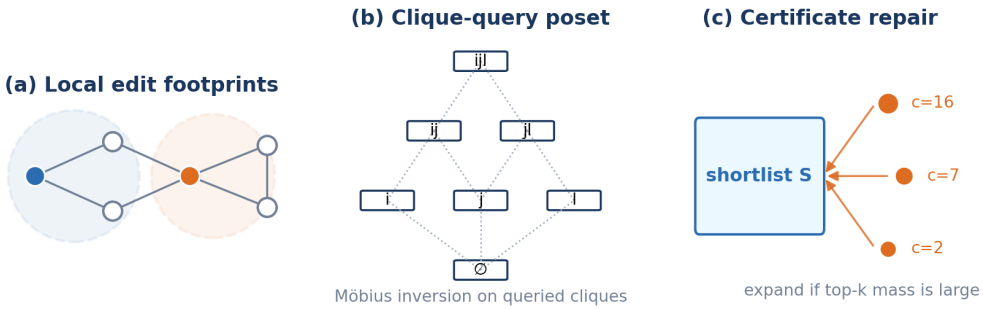


Figure 9: Clique queries and certificate repair. Footprint overlap determines which candidate groups are queried; the resulting coefficients support both deployment and expansion decisions.

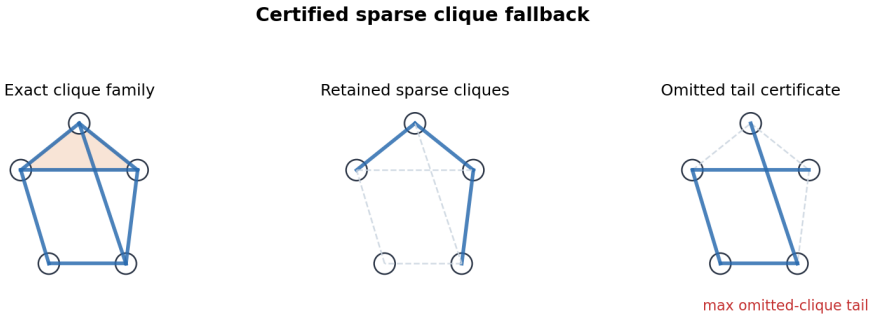


Figure 10: Sparse retained-clique fallback. At small interaction width, query all cliques. At larger width, retain fewer cliques and bound the omitted tail.

B.2 Real-Graph Footprints and Interactions

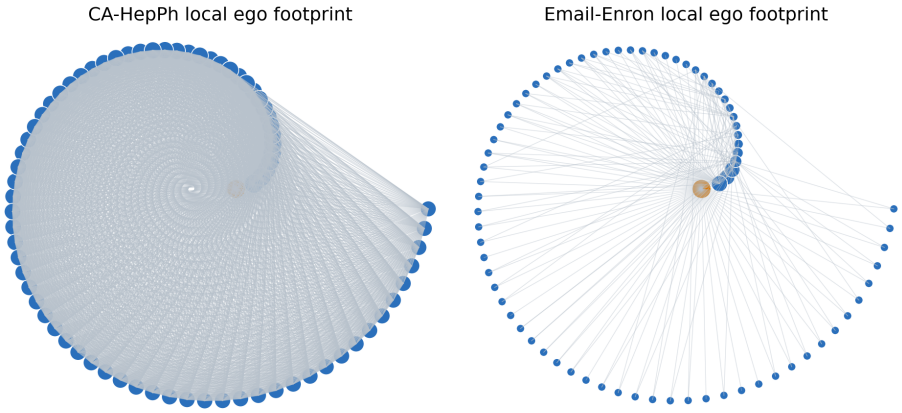


Figure 11: Local graph views behind the footprint assumption. Orange nodes are candidate centers and blue nodes are nearby vertices in real graph neighborhoods.

C Additional Formal Results

We state the additional results used by the proofs, including the charged expansion bounds.

Proposition 8 (Disconnected-footprint diagnostic extension)**.** *If footprints are disconnected, replace each candidate by its component copies and form the component-intersection graph over bag shadows. The component graph is chordal, but identifying component copies of the same candidate is a quotient operation and need not preserve chordality or treewidth. The same bag-load-style bound is therefore valid only after measuring the induced quotient width or imposing explicit inter-component overlap bounds; otherwise disconnectedness is treated as a diagnostic that can trigger thinning or fallback.*

Corollary 9 (Linear shortlist scaling)**.** *If each shortlisted shadow spans at most h bags and every bag has load at most η, the number of admissible cliques satisfies the following bound*

$$|\mathcal{C}_d(H_{\mathcal{T}}(S))| \le hr \sum_{q=0}^{d-1} \binom{\eta-1}{q}.$$

For disconnected footprints, the same expression applies only to the measured component-shadow graph before quotienting, or to the candidate graph when the quotient width is separately bounded.

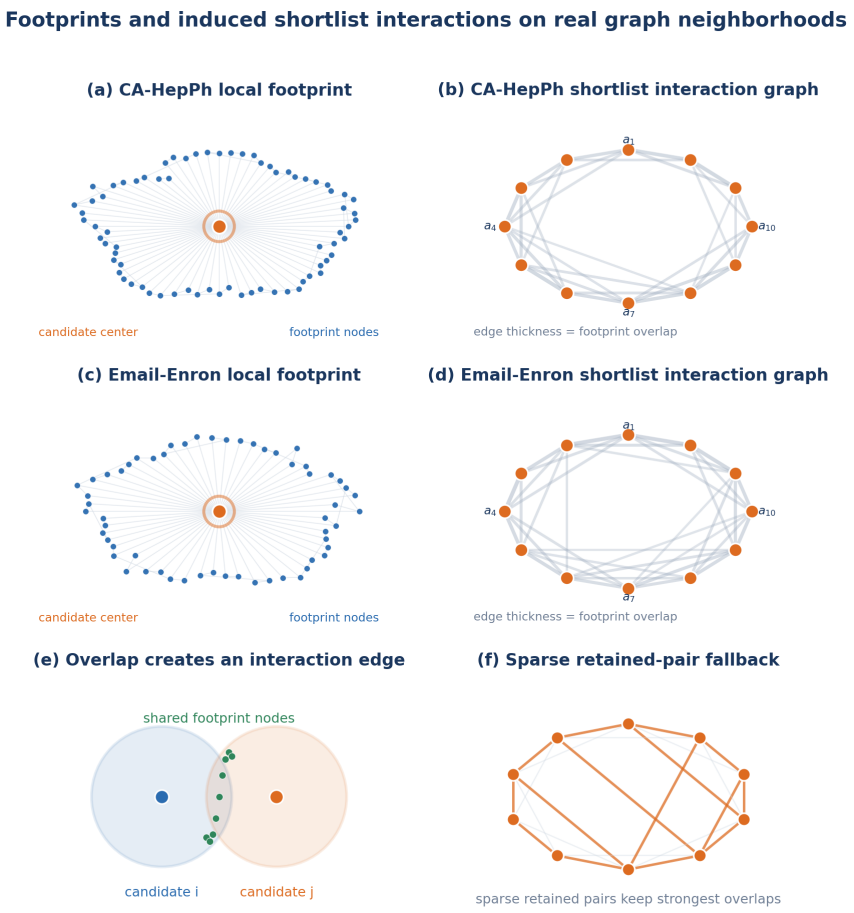


Figure 12: Schematic footprint and interaction examples. Panels (a,c) depict local footprints; panels (b,d) depict shortlist interaction graphs, with edge thickness representing overlap. Panels (e,f) show the interactions and the pairs retained for sparse recovery.

Corollary 10 (Closed-form pairwise outside envelope)**.** *For pairwise G, let θ_u be the singleton coefficient of u and $\theta^+_{u,(m)}$ its sorted positive incident pair coefficients. Then*

$$\mu_G(u) = \left[\theta_u + \sum_{m=1}^{k-1} \theta^+_{u,(m)}\right]_+.$$

Corollary 11 (Submodular singleton certificate)**.** *If the comparator G is monotone submodular, then*

$$c_G(u) = G(\{u\}) - G(\varnothing)$$

is a valid one-step outside certificate. Thus Theorem 4 applies using the top-k omitted singleton gains without any need to estimate the pair coefficients incident to omitted candidates.

Proposition 12 (Minimal one-step outside certificate)**.** *For an exact pairwise comparator, the positive marginal envelope $\mu_G(u)$ is the pointwise smallest valid one-step outside certificate for candidate u. Therefore summing the top-k such envelopes gives the smallest universal upper bound obtainable using a separate one-step marginal bound for each omitted candidate.*

Corollary 13 (Pairwise end-to-end guarantee)**.** *In the pairwise setting, the full-library regret decomposes into an outside-shortlist term, controlled by $2\varepsilon + \sum_{m=1}^{k} c_{(m)}$, plus a post-screening surrogate term controlled by the agnostic reconstruction error on the shortlist interaction graph.*

Corollary 14 (Charged one-step expansion)**.** *Let $c_{(1)} \geq c_{(2)} \geq \cdots$ be the ordered outside certificates of Theorem 4 for a shortlist S, and suppose the system expands S to S' by adding the ℓ omitted*

candidates with the largest certificates. Applying Theorem 4 to S' yields

$$F(B^\star_{\text{full}}) - F(B^\star_{S'}) \le 2\varepsilon' + \sum_{m=\ell+1}^{\ell+k} c_{(m)},$$

where ε' is the comparator error on the enlarged pool. Let $\mathcal{Q}_{\text{seen}}$ be the nonempty subsets already queried. Exact recovery on S' needs $|\mathcal{C}_d(H(S')) \setminus \mathcal{Q}_{\text{seen}}|$ new fitting calls. This equals the clique-count difference when the families are nested and only the old family has been queried. The new family size is bounded by Theorem 3; audit and deployment calls are charged separately.

One charged expansion therefore replaces the top-ℓ certificate mass by the next-k tail, and the expansion decision compares this gain against the clique-growth call cost, giving the adaptive loop both a call budget and a bound on the remaining outside loss.

Corollary 15 (Charged multi-round convergence)**.** *Fix ordered valid certificates $c_{(1)} \ge c_{(2)} \ge \cdots$. From S_0, each round adds the next ℓ omitted candidates and reuses prior queries. For the resulting S_t, the following deployment-gap bound holds:*

$$F(B^\star_{\text{full}}) - F(B^\star_{S_t}) \le 2\varepsilon_t + \sum_{m=t\ell+1}^{t\ell+k} c_{(m)}.$$

For $\gamma > 2\sup_t \varepsilon_t$, let q_γ be the least q with $\sum_{m=q+1}^{q+k} c_{(m)} \le \gamma - 2\sup_t \varepsilon_t$. The gap is at most γ after $\lceil q_\gamma/\ell \rceil$ rounds. Exact clique fitting uses $1 + |\bigcup_{j=0}^{t} \mathcal{C}_d(H(S_j))|$ distinct calls. For nested clique families this reduces to $1 + |\mathcal{C}_d(H(S_t))|$, bounded by Theorem 3. The total budget also charges every audit query and final deployment evaluation, including calls outside the clique-fitting cache.

Hence vanishing certificate tails and bounded comparator error give finite stopping without assuming that expansion improves fit at any intermediate round.

D Proofs

D.1 Connected Shadows and Agnostic Reconstruction

D.1.1 Proof of the Connected-Shadow Observation

We first prove the connected-shadow statement used throughout the compression analysis. For every vertex $v \in V(G_0)$, the running-intersection property of the tree decomposition implies that

$$T_v := \{t \in V(T) : v \in X_t\}$$

induces a connected subtree of T. Now let Ψ_i be connected in G_0 and write

$$K_i^{\mathcal{T}} = \bigcup_{v \in \Psi_i} T_v.$$

To prove that this union is connected, it suffices to show that the subtrees $\{T_v : v \in \Psi_i\}$ form a connected family in the intersection graph. If $(u, v) \in E(G_0[\Psi_i])$, then some bag of the tree decomposition contains both endpoints of that edge, so $T_u \cap T_v \neq \varnothing$. Because Ψ_i is connected, any two vertices of Ψ_i can be joined by a path in $G_0[\Psi_i]$, and consecutive subtrees along that path intersect. Hence the union $\bigcup_{v\in\Psi_i} T_v = K_i^{\mathcal{T}}$ is connected. Therefore $K_i^{\mathcal{T}}$ induces a connected subtree of T.

D.1.2 Proof of the Agnostic Reconstruction Theorem

Let $G^\star \in \mathcal{G}_{H,d}$ be any comparator achieving error at most $\varepsilon^\star_{H,d}(F) + \delta$ on every queried clique and feasible deployment, where $\delta > 0$ is arbitrary. Write

$$G^\star(B) = \sum_{U \in \mathcal{C}_d(H)\cup\{\varnothing\}} \theta^\star_U \mathbf{1}\{U \subseteq B\}, \qquad \widehat{G}(B) = \sum_{U \in \mathcal{C}_d(H)\cup\{\varnothing\}} \widehat{\theta}_U \mathbf{1}\{U \subseteq B\}.$$

For each admissible clique U, let $f_U := F(U)$ and $g_U := G^\star(U)$. By construction,

$$|f_U - g_U| \le \varepsilon^\star_{H,d}(F) + \delta.$$

The Möbius inversion formula on the finite poset $\mathcal{C}_d(H) \cup \{\varnothing\}$ gives

$$\widehat{\theta}_U = \sum_{W \subseteq U} \mu(W, U) f_W, \qquad \theta^\star_U = \sum_{W \subseteq U} \mu(W, U) g_W,$$

where $\mu(W, U) = (-1)^{|U \setminus W|}$ is the Boolean-lattice Möbius function. Hence

$$|\widehat{\theta}_U - \theta^\star_U| \le \sum_{W \subseteq U} |f_W - g_W| \le 2^{|U|}(\varepsilon^\star_{H,d}(F) + \delta) \le 2^d(\varepsilon^\star_{H,d}(F) + \delta).$$

For the empty-set coefficient, the same comparison gives

$$|\widehat{\theta}_\varnothing - \theta^\star_\varnothing| \le \varepsilon^\star_{H,d}(F) + \delta.$$

Now fix any feasible deployment $B \in \mathcal{X}_k(S)$. Then

$$\widehat{G}(B) - G^\star(B) = (\widehat{\theta}_\varnothing - \theta^\star_\varnothing) + \sum_{U \in \mathcal{C}_d(H[B])} (\widehat{\theta}_U - \theta^\star_U).$$

Using the definition of $\Gamma_{d,k}(H)$,

$$|\widehat{G}(B) - G^\star(B)| \le \big(1 + \Gamma_{d,k}(H)\, 2^d\big)(\varepsilon^\star_{H,d}(F) + \delta).$$

Adding and subtracting $G^\star(B)$ yields

$$|F(B) - \widehat{G}(B)| \le |F(B) - G^\star(B)| + |G^\star(B) - \widehat{G}(B)| \le \big(2 + \Gamma_{d,k}(H)\, 2^d\big)(\varepsilon^\star_{H,d}(F) + \delta).$$

Letting $\delta \downarrow 0$ proves the first claim. For the regret bound, if $\widehat{B} \in \operatorname{argmax} \widehat{G}$ and $B^\star \in \operatorname{argmax} F$, then

$$\begin{aligned} F(B^\star) - F(\widehat{B}) &\le \big(F(B^\star) - \widehat{G}(B^\star)\big) + \big(\widehat{G}(\widehat{B}) - F(\widehat{B})\big) \\ &\le 2 \sup_{B \in \mathcal{X}_k(S)} |F(B) - \widehat{G}(B)| \\ &\le 2\big(2 + \Gamma_{d,k}(H)\, 2^d\big)\varepsilon^\star_{H,d}(F). \end{aligned}$$

Finally, exact reconstruction of all coefficients requires one empty-set query and one query for each admissible clique, hence exactly $1 + |\mathcal{C}_d(H)|$ simulator evaluations.

D.2 Bag-Load Compression and Exact Lower Bound

D.2.1 Proof of the Bag-Load Compression Theorem

For each candidate $i \in S$, the connectedness of Ψ_i implies that the set of bags intersecting Ψ_i forms a connected subtree of T; this is the standard running-intersection property of tree decompositions. Thus $K_i^{\mathcal{T}}$ is a subtree of T for every i. By definition, $H_{\mathcal{T}}(S)$ joins i and j whenever these two subtrees intersect. Gavril's theorem states that intersection graphs of subtrees of a tree are exactly the chordal graphs, so $H_{\mathcal{T}}(S)$ is chordal.

For a chordal graph represented as an intersection graph of subtrees, every clique has a common witness bag. Concretely, if $Q \subseteq S$ is a clique in $H_{\mathcal{T}}(S)$, then the subtrees $\{K_i^{\mathcal{T}} : i \in Q\}$ pairwise intersect; Helly's property for subtrees of a tree implies

$$\bigcap_{i \in Q} K_i^{\mathcal{T}} \neq \varnothing.$$

Hence there exists a bag $t \in V(T)$ whose shadow contains every $i \in Q$, and therefore $|Q| \le \ell_t^{\mathcal{T}}$. This shows both that

$$\operatorname{tw}(H_{\mathcal{T}}(S)) \le \eta^{\mathcal{T}} - 1$$

and that every admissible clique of size at most d is counted inside at least one bag. Summing over bags gives

$$|\mathcal{C}_d(H_{\mathcal{T}}(S))| \le \sum_{t \in V(T)} \sum_{s=1}^{d} \binom{\ell_t^{\mathcal{T}}}{s} = \Lambda_d(\ell^{\mathcal{T}}).$$

Minimizing over all decompositions proves the claimed query upper bound. The dynamic-programming complexity statement is standard for bounded-treewidth subset optimization: once a width-w nice decomposition of the interaction graph is available, each bag stores a table indexed by local states and remaining budget, which yields complexity $O(|T_H| 2^{w+1} k^2)$.

D.2.2 DISCONNECTED-FOOTPRINT DIAGNOSTICS

Let the footprint of candidate i have connected components $\Psi_i^{(1)}, \dots, \Psi_i^{(q_i)}$. For each component we define its bag shadow as the set of bags intersecting that component. Because each component is connected, each component shadow is a connected subtree of T, so the component-intersection graph is chordal and its clique count is bounded by component load.

This component graph is not, by itself, a proof that the candidate-level graph has the same width. Identifying non-adjacent component-copy vertices belonging to the same candidate is a quotient operation, not an edge contraction in the component graph, and it need not preserve chordality or treewidth. We therefore use disconnected footprints only with an explicit measured quotient-width or inter-component-overlap bound. If that measured candidate-level width is large, the algorithm treats disconnectedness as a diagnostic and uses sparse thinning or fallback rather than invoking the connected-footprint theorem.

D.2.3 PROOF OF THE EXACT LOWER BOUND

Let

$$D := 1 + |\mathcal{C}_d(H)|$$

denote the number of coefficients in the exact class $\mathcal{G}_{H,d}$. Consider the deterministic algorithm run on the zero surrogate $G \equiv 0$. Because the algorithm is deterministic, this fixes a query sequence

$$Q_1, \dots, Q_q \in \mathcal{X}_k(S),$$

where q is the number of simulator calls. For each queried set Q_ℓ, define the row vector

$$r_\ell := \big(\mathbf{1}\{U \subseteq Q_\ell\}\big)_{U \in \mathcal{C}_d(H) \cup \{\varnothing\}} \in \mathbb{R}^D.$$

If $q < D$, then the $q \times D$ matrix R with rows r_ℓ has a nontrivial nullspace. Choose $\alpha \neq 0$ with $R\alpha = 0$, and let

$$G_\alpha(B) := \sum_{U \in \mathcal{C}_d(H) \cup \{\varnothing\}} \alpha_U \mathbf{1}\{U \subseteq B\}.$$

Then $G_\alpha \in \mathcal{G}_{H,d}$ and, for every query Q_ℓ on the zero-function path,

$$G_\alpha(Q_\ell) = r_\ell^\top \alpha = 0.$$

Hence the deterministic algorithm receives exactly the same answers on $G \equiv 0$ and on G_α, so it must output the same coefficient vector on both inputs.

That is impossible because the coefficient representation in $\mathcal{G}_{H,d}$ is unique. Indeed, if the coefficient vector were not unique, then some nonzero coefficient vector would induce the zero function on all admissible clique queries. But the evaluation matrix

$$M_{W,U} := \mathbf{1}\{U \subseteq W\}, \qquad W, U \in \mathcal{C}_d(H) \cup \{\varnothing\},$$

is triangular with ones on the diagonal under any ordering by increasing set size, hence invertible by Möbius inversion. Therefore $G_\alpha \not\equiv 0$ whenever $\alpha \neq 0$. The algorithm cannot exactly identify the coefficients of both 0 and G_α, which is a contradiction. Thus every deterministic exact-identification algorithm must use at least $D = 1 + |\mathcal{C}_d(H)|$ queries in the worst case.

For the randomized approximate statement, place the isotropic Gaussian prior $\theta \sim \mathcal{N}(0, \sigma^2 I_D)$ on the coefficient vector and condition on the algorithm's internal randomness. Every noiseless function query is one linear measurement $r^\top \theta$. Adaptivity adds no observation beyond the returned scalar because the next row is a function of the internal seed and earlier query–answer pairs. After q measurements, their realized row span has dimension at most q, so at least $D - q$ orthogonal Gaussian directions remain unobserved. Consequently every estimator has Bayes squared error at least $(D - q)\sigma^2$. If an algorithm using $q < D$ had finite uniform risk $R = \sup_\theta \mathbb{E}\|\widehat{\theta} - \theta\|_2^2$, its Bayes risk under every such prior would be at most R, contradicting $(D - q)\sigma^2 > R$ as $\sigma \to \infty$. Hence the uniform mean-squared recovery risk is infinite.

D.2.4 GRAPH-INDUCED ADDITIVE-MODEL EMBEDDING

Proposition 16 (Clique-supported additive equivalence)**.** *Every $G \in \mathcal{G}_{H,d}$ is an overlapping additive model*

$$G(B) = \theta_{\varnothing} + \sum_{U \in \mathcal{C}_d(H)} g_U(B \cap U), \qquad g_U(A) = \theta_U \mathbf{1}\{A = U\}.$$

Conversely, suppose an overlapping additive set function has factors on groups $U \in \mathcal{C}_d(H)$ and expand each factor in its Boolean Möbius basis. If every nonzero Möbius term is supported on a clique of size at most d, collecting equal terms gives the defining representation of $\mathcal{G}_{H,d}$. Thus the graph-induced class is exactly the multilinear additive class with clique-supported Möbius coefficients; the interaction graph supplies the overlapping groups rather than assuming them known.

D.3 SHORTLIST COMPLETION AND CERTIFICATE MINIMALITY

D.3.1 PROOF OF SHORTLIST COMPLETION AND THE PAIRWISE ENVELOPE

Let $B^\star_{\text{full}}$ be an optimal deployment over the full library $[M]$, and let $I = B^\star_{\text{full}} \cap S$, $O = B^\star_{\text{full}} \setminus S$. Enumerate the omitted candidates as $O = \{u_1, \dots, u_m\}$ with $m \le k$, and define the decreasing chain

$$A_0 := I \cup O, \qquad A_t := I \cup \{u_{t+1}, \dots, u_m\} \quad (t = 1, \dots, m).$$

Then $A_{t-1} = A_t \cup \{u_t\}$ and $|A_t| \le k-1$ for every t. By definition of the positive marginal envelope,

$$G(A_{t-1}) - G(A_t) = G(A_t \cup \{u_t\}) - G(A_t) \le [G(A_t \cup \{u_t\}) - G(A_t)]_+ \le \mu_G(u_t) \le c(u_t).$$

Summing over t telescopes to

$$G(B^\star_{\text{full}}) - G(I) \le \sum_{u \in O} c(u) \le \sum_{m=1}^{k} c_{(m)}.$$

Because $I \subseteq S$, the restricted optimum $B^\star_S$ satisfies $F(B^\star_S) \ge F(I)$. The uniform comparator error transfers the comparison through G with additive penalty 2ε:

$$F(B^\star_{\text{full}}) - F(B^\star_S) \le F(B^\star_{\text{full}}) - F(I) \le G(B^\star_{\text{full}}) - G(I) + 2\varepsilon \le 2\varepsilon + \sum_{m=1}^{k} c_{(m)}.$$

Now let $\widehat{B}$ be the output of post-screening surrogate optimization on the shortlist. Applying Theorem 1 on S yields

$$F(B^\star_S) - F(\widehat{B}) \le 2\big(2 + \Gamma_{d,k}(H[S])\, 2^d\big)\varepsilon^\star_{H[S],d}(F).$$

Adding the two displays proves the end-to-end bound. The adaptive-expansion statement is immediate once the residual outside mass is forced below τ.

For the multi-round statement, the fixed certificate ordering remains valid after removing candidates already added to the shortlist. After t rounds, the first $t\ell$ candidates are no longer omitted, so applying Theorem 4 to S_t gives the following bound

$$F(B^\star_{\text{full}}) - F(B^\star_{S_t}) \le 2\varepsilon_t + \sum_{m=t\ell+1}^{t\ell+k} c_{(m)}.$$

Padding the finite certificate sequence by zeros makes the display valid after the omitted pool is exhausted. By the definition of q_γ, any $t \ge \lceil q_\gamma/\ell \rceil$ makes the right-hand side at most γ. Exact clique fitting reuses cached values, so its distinct cumulative calls are $1 + |\bigcup_{j=0}^{t} \mathcal{C}_d(H(S_j))|$. Nested shortlists alone do not imply nested clique families when the decomposition changes. If the families are nested, the union equals the final family and Theorem 3 bounds its size. Separate audit and deployment evaluations remain charged.

For the pairwise closed form, write

$$G(A \cup \{u\}) - G(A) = \theta_u + \sum_{v \in A} \theta_{uv}.$$

Among sets A of size at most $k-1$, the maximum raw marginal is achieved by selecting the $k-1$ largest positive incident pair coefficients, which gives the equality

$$\max_{A\subseteq[M]\setminus\{u\},\,|A|\le k-1}\Big(\theta_u+\sum_{v\in A}\theta_{uv}\Big)=\theta_u+\sum_{m=1}^{k-1}\theta^+_{u,(m)}.$$

Taking the positive part proves

$$\mu_G(u)=\Big[\theta_u+\sum_{m=1}^{k-1}\theta^+_{u,(m)}\Big]_+.$$

D.3.2 Proof of Minimality and the Pairwise Master Bound

If $c(u)$ is any valid one-step outside certificate, then by definition

$$[G(A\cup\{u\})-G(A)]_+\le c(u)\qquad\text{for all } A\subseteq[M]\setminus\{u\},\ |A|\le k-1.$$

Taking the maximum over all such A yields

$$\mu_G(u)=\max_{A\subseteq[M]\setminus\{u\},\,|A|\le k-1}[G(A\cup\{u\})-G(A)]_+\le c(u).$$

Hence $\mu_G(u)$ is the pointwise smallest valid one-step outside certificate.

For the top-k mass claim, let $d(u)$ be any valid one-step certificate family. Sorting both sequences in decreasing order, the pointwise inequality $\mu_G(u)\le d(u)$ for every u implies

$$\sum_{m=1}^{k}\mu_{G,(m)}\le\sum_{m=1}^{k}d_{(m)}.$$

Therefore no universal upper bound obtained by summing k valid one-step certificates can be smaller than the sum of the k largest omitted-candidate marginal envelopes μ_G.

Combining Theorem 4 and corollary 10 at $d=2$ proves Corollary 13.

D.3.3 Proof of the Submodular Singleton Certificate

For any $A\subseteq[M]\setminus\{u\}$, diminishing returns gives

$$G(A\cup\{u\})-G(A)\le G(\{u\})-G(\varnothing).$$

Monotonicity makes the right-hand side nonnegative. Taking the positive part and maximizing over all $|A|\le k-1$ therefore yields $\mu_G(u)\le c_G(u)$, so $c_G(u)$ is a valid outside certificate. Substitution into Theorem 4 proves the stated top-k bound on outside-shortlist loss.

D.3.4 Proof of the Pairwise Deployed Post-Screening Guarantee

Let $B^\star_P$ be the best deployment in the audit pool P and $B^\star_S$ the best deployment in the shortlist S. The outside certificate argument above, applied on the audit pool, gives

$$F(B^\star_P)-F(B^\star_S)\le 2\varepsilon+C_{\mathrm{out}}(S).$$

It remains to compare $B^\star_S$ with the retained-pair output $\widehat{B}$. Let G be the pairwise comparator on S and let $\widehat{G}_{\mathcal{Q}}$ be the retained-pair surrogate. The validation term gives

$$\sup_{B\in\mathcal{X}_k(S)}|F(B)-G(B)|\le\delta_S,$$

on the reported validation/audit domain, while the sparse-tail certificate gives

$$\sup_{B\in\mathcal{X}_k(S)}|G(B)-\widehat{G}_{\mathcal{Q}}(B)|\le\mathsf{Tail}_{L,k}(G).$$

Therefore

$$\sup_{B\in\mathcal{X}_k(S)}|F(B)-\widehat{G}_{\mathcal{Q}}(B)|\le\delta_S+\mathsf{Tail}_{L,k}(G).$$

Since $\widehat{B}$ maximizes $\widehat{G}_{\mathcal{Q}}$ over $\mathcal{X}_k(S)$, the standard surrogate-regret transfer yields

$$F(B^\star_S)-F(\widehat{B})\le 2\delta_S+2\mathsf{Tail}_{L,k}(G).$$

Adding the outside-pool bound to this shortlist regret bound proves Corollary 5.

D.4 Joint, Noisy, Sparse, and Thin Certificates

D.4.1 Proof of the Joint Certificate Refinement

Let $B_P^\star$ be a maximizer over $P = S \cup U$ and write

$$I := B_P^\star \cap S, \qquad O := B_P^\star \cap U.$$

Then $I \subseteq S$, $O \subseteq U$, and $|I| + |O| \le k$. By the definition of the joint certificate,

$$G(B_P^\star) - G(I) = G(I \cup O) - G(I) \le C_{\mathrm{joint}}(U; S).$$

Since I is feasible for the shortlist problem, $F(B_S^\star) \ge F(I)$. The uniform comparator error transfers the remaining comparison to G with the standard additive 2ε penalty:

$$F(B_P^\star) - F(B_S^\star) \le F(B_P^\star) - F(I) \le 2\varepsilon + G(B_P^\star) - G(I) \le 2\varepsilon + C_{\mathrm{joint}}(U; S).$$

It remains to compare the joint certificate with the decoupled top-k envelope. Fix any feasible $O = \{u_1, \dots, u_m\} \subseteq U$ and any $A \subseteq S$ with $|A| + m \le k$. Define

$$A_0 := A, \qquad A_t := A \cup \{u_1, \dots, u_t\}, \quad t = 1, \dots, m.$$

Then

$$G(A \cup O) - G(A) = \sum_{t=1}^{m} \big(G(A_t) - G(A_{t-1})\big) \le \sum_{t=1}^{m} \mu_G(u_t) \le \sum_{j=1}^{k} \mu_{G,(j)}.$$

Taking positive parts and then the maximum over all feasible (A, O) proves

$$C_{\mathrm{joint}}(U; S) \le \sum_{j=1}^{k} \mu_{G,(j)}.$$

D.4.2 Proof of the Noisy-Query Result

Let $Q := 1 + |\mathcal{C}_d(H)|$ denote the number of queried clique values. Each query is repeated n times and averaged. Standard sub-Gaussian concentration implies that for any fixed queried clique,

$$\Pr\Big(\big|\bar{Y}_U - F(U)\big| > t\Big) \le 2\exp\Big(-\frac{nt^2}{2\sigma^2}\Big).$$

Applying a union bound over all Q queried cliques and setting

$$t = \sigma\sqrt{\frac{2\log(2Q/\delta)}{n}}$$

shows that with probability at least $1 - \delta$ all averaged query values satisfy $|\bar{Y}_U - F(U)| \le t$. Repeating the proof of Theorem 1 with this additional perturbation term gives

$$|\widehat{\theta}_U - \theta_U^\star| \le 2^d(\varepsilon_{H,d}^\star(F) + t)$$

for every admissible clique U, and therefore

$$\sup_{B \in \mathcal{X}_k(S)} |F(B) - \widehat{G}_n(B)| \le \big(2 + \Gamma_{d,k}(H)\, 2^d\big)\varepsilon_{H,d}^\star(F) + \big(1 + \Gamma_{d,k}(H)\, 2^d\big)t.$$

Substituting the displayed value of t proves the proposition, and the regret bound again follows by the standard factor-of-two transfer from surrogate error to deployment suboptimality.

D.4.3 Proof of Certified Sparse Clique Compression

Fix an exact comparator

$$G(B) = \theta_\varnothing + \sum_{U \in \mathcal{C}_d(H)} \theta_U \mathbf{1}\{U \subseteq B\}.$$

More generally, let $\mathcal{Q} \subseteq \mathcal{C}_d(H)$ contain all singleton cliques and define

$$\widehat{G}_{\mathcal{Q}}(B) = \theta_\varnothing + \sum_{U \in \mathcal{Q}} \theta_U \mathbf{1}\{U \subseteq B\}.$$

Then

$$G(B) - \widehat{G}_{\mathcal{Q}}(B) = \sum_{U \in \mathcal{C}_d(H[B]) \setminus \mathcal{Q}} \theta_U.$$

Taking absolute values and maximizing over feasible deployments gives

$$\sup_{B \in \mathcal{X}_k(S)} |G(B) - \widehat{G}_{\mathcal{Q}}(B)| \le \max_{B \in \mathcal{X}_k(S)} \sum_{U \in \mathcal{C}_d(H[B]) \setminus \mathcal{Q}} |\theta_U| = \mathsf{Tail}^{(d)}_{\mathcal{Q},k}(G).$$

If $B^\star$ maximizes G and $\widehat{B}_{\mathcal{Q}}$ maximizes $\widehat{G}_{\mathcal{Q}}$, then

$$\begin{aligned} G(B^\star) - G(\widehat{B}_{\mathcal{Q}}) &\le \big(G(B^\star) - \widehat{G}_{\mathcal{Q}}(B^\star)\big) + \big(\widehat{G}_{\mathcal{Q}}(\widehat{B}_{\mathcal{Q}}) - G(\widehat{B}_{\mathcal{Q}})\big) \\ &\le 2\,\mathsf{Tail}^{(d)}_{\mathcal{Q},k}(G). \end{aligned}$$

For the certificate bound, if $u_U \ge |\theta_U|$ for every omitted higher-order clique candidate, then for every feasible deployment,

$$\sum_{U \in \mathcal{C}_d(H[B]) \setminus \mathcal{Q}} |\theta_U| \le \sum_{U \in \mathcal{C}_d(H[B]) \setminus \mathcal{Q}} u_U \le \sum_{m > L} u_{(m)},$$

where the last step is the crude universal omitted-tail bound obtained by retaining the top-L non-singleton clique certificates. This proves the stated bound for the retained clique family.

D.4.4 Proof of the Thin Pairwise Corollary

Apply Proposition 6 with $d = 2$ and

$$\mathcal{Q} = \big\{\{i\} : i \in S\big\} \cup E_L.$$

The retained family contains all singleton cliques, and the omitted non-singleton cliques are exactly the pairs in $E(H) \setminus E_L$. Therefore

$$\mathsf{Tail}^{(2)}_{\mathcal{Q},k}(G) = \max_{B \in \mathcal{X}_k(S)} \sum_{\{i,j\} \in E(H[B]) \setminus E_L} |\theta_{ij}| = \mathsf{Tail}_{L,k}(G),$$

which yields the displayed error and regret bounds. The certificate statement is the same specialization with pair certificates $u_e \ge |\theta_e|$. In the agnostic setting, decompose $F - G$ and add the Theorem 1 term.

D.4.5 Proof of Certificate-Optimal Pair Truncation

Because every certificate weight is nonnegative,

$$\sum_{e \notin E_L} u_e = \sum_{e \in E(H)} u_e - \sum_{e \in E_L} u_e.$$

Therefore minimizing the omitted-weight bound $\sum_{e \notin E_L} u_e$ over all L-edge subsets is equivalent to maximizing the retained weight $\sum_{e \in E_L} u_e$. Any top-L subset under u_e achieves that maximum, so it minimizes the universal omitted-tail upper bound.

For every E_L,

$$\mathfrak{U}(E_L) = \max_{B \in \mathcal{X}_k(S)} \sum_{e \in E(H[B]) \setminus E_L} u_e \le \sum_{e \notin E_L} u_e.$$

Hence any choice of E_L that minimizes $\sum_{e \notin E_L} u_e$ also minimizes this crude worst-case upper bound. This proves the stated bound for the retained clique family.

D.5 Scaling and Validation Proxy Proofs

D.5.1 Proof of Linear Shortlist Scaling

From Theorem 3,

$$|\mathcal{C}_d(H_{\mathcal{T}}(S))| \le \sum_t \sum_{s=1}^{d} \binom{\ell_t^{\mathcal{T}}}{s}.$$

For every $s \geq 1$ and every load value $\ell \leq \eta$,

$$\binom{\ell}{s} \leq \ell \binom{\eta-1}{s-1},$$

because choosing an s-subset through any distinguished element yields at most $\binom{\eta-1}{s-1}$ possibilities. Therefore

$$\sum_t \sum_{s=1}^{d} \binom{\ell_t^{\mathcal{T}}}{s} \leq \sum_t \ell_t^{\mathcal{T}} \sum_{s=1}^{d} \binom{\eta-1}{s-1}.$$

Since every shortlisted candidate shadow covers at most h bags,

$$\sum_t \ell_t^{\mathcal{T}} = \sum_{i \in S} |K_i^{\mathcal{T}}| \leq hr.$$

Combining the two displays gives

$$|\mathcal{C}_d(H_{\mathcal{T}}(S))| \leq hr \sum_{q=0}^{d-1} \binom{\eta-1}{q}.$$

For disconnected footprints, this corollary applies to the component-shadow graph before quotienting. It applies to the candidate graph only when the quotient width is measured or otherwise bounded.

D.5.2 Validation Proxy Diagnostic

Assume F is L_F-Lipschitz on all of $\mathcal{X}_k(S)$ in Hamming distance. Restrict the comparison class to models that are L_G-Lipschitz on this same domain, and define $\widehat{\varepsilon}_{H,d}^{\mathrm{val},L_G}$ as its validation optimum. Let $G^\dagger$ attain this restricted optimum up to $\delta > 0$:

$$\max_{B \in \mathcal{V}} |F(B) - G^\dagger(B)| \leq \widehat{\varepsilon}_{H,d}^{\mathrm{val},L_G}(F; \mathcal{V}) + \delta.$$

These are full-domain assumptions; an empirical check on validation points alone does not establish them. Let $\mathcal{V} \subseteq \mathcal{X}_k(S)$ be a ρ-net. For any $B \in \mathcal{X}_k(S)$, it contains V_B with $d_H(B, V_B) \leq \rho$. Thus

$$\begin{aligned} |F(B) - G^\dagger(B)| &\leq |F(B) - F(V_B)| + |F(V_B) - G^\dagger(V_B)| \\ &\quad + |G^\dagger(V_B) - G^\dagger(B)| \\ &\leq (L_F + L_G)\rho + \widehat{\varepsilon}_{H,d}^{\mathrm{val},L_G}(F; \mathcal{V}) + \delta. \end{aligned}$$

Taking the supremum over B and then the infimum over all admissible G proves

$$\varepsilon_{H,d}^{\star}(F) \leq \widehat{\varepsilon}_{H,d}^{\mathrm{val},L_G}(F; \mathcal{V}) + (L_F + L_G)\rho.$$

Substituting this into Theorem 1 yields the stated regret bound after letting $\delta \downarrow 0$.

E Experimental Details

E.1 Algorithmic Details

E.1.1 Dynamic Program for Bounded-Treewidth Deployment

Given a nice tree decomposition of the interaction graph H, each bag stores a table indexed by a local subset state $\sigma \subseteq X_t$ and by the remaining budget $b \in \{0, \ldots, k\}$. The table entry records the best surrogate value achievable inside the processed part of the decomposition subject to agreeing with σ on the current bag. Introduce, forget, and join transitions are the standard ones for subset optimization on bounded-treewidth graphs. Since every bag contains at most $w+1$ vertices, the number of local states is 2^{w+1} and the convolution over budgets contributes the additional factor k^2 in the running time.

E.1.2 CHOOSING THE THIN BUDGET

In experiments we sweep a short geometric grid of values for L and choose the smallest one whose certificate tail satisfies the target threshold. This mirrors Theorem 4: the user specifies a tolerable omitted-tail mass and the procedure stops as soon as the empirical certificate falls below it. For pairwise thinning, retained-pair scores are

$$s_{ij} = J(\Psi_i, \Psi_j)(|\widehat{\theta}_i| + |\widehat{\theta}_j|),$$

where J is footprint Jaccard overlap and $\widehat{\theta}_i$ is the singleton coefficient recovered from the singleton query. The sparse-tail proxy is the sum of omitted scores after the top L retained pairs. No extra simulator calls are required because these ranking certificates use singleton coefficients and footprint overlaps only. For omitted-candidate expansion, let $\tilde{\theta}_u$ denote the queried singleton coefficient for omitted candidate u if it is available, and otherwise the calibrated singleton score supplied by the front end. We use the following graph-local envelope to rank candidates omitted from the shortlist:

$$q_{uv} = J(\Psi_u, \Psi_v)(|\tilde{\theta}_u| + |\widehat{\theta}_v|), \qquad \tilde{c}(u) = [\tilde{\theta}_u]_+ + \sum_{m=1}^{k-1} q_{u,(m)},$$

where $q_{u,(m)}$ are the largest $k-1$ scores over $v \in S$. The online outside proxy is

$$C_{\text{out}}^{\text{proxy}}(S) = \sum_{u \in \text{top-}k(P \setminus S)} \tilde{c}(u).$$

Without a singleton-call value for $\tilde{\theta}_u$, only a heuristic audit-pool score is available.

E.1.3 CONNECTED-FOOTPRINT ASSUMPTION IN PRACTICE

The connected-footprint assumption is natural for local node or edge interventions whose effect propagates through a radius-ρ neighborhood. Examples include node removal, local vaccination, local edge reinforcement, and small routing changes. For disconnected footprints, Proposition 8 is used as a diagnostic extension: component shadows are measured explicitly, and the candidate-level quotient width is checked before applying clique-count claims. The framework is therefore sensitive to the geometric spread of an intervention and falls back when disconnected components create high-width interactions.

E.1.4 BLOCK POST-SCREENING FOR GLOBAL STRESS SETTINGS

We test the operational block extension on ogbn-arxiv, Orkut, and ogbn-products. The same 24-candidate full-graph library is partitioned into four deterministic graph-local blocks of six candidates. One local empty-set call and four singleton calls per block cost 20 calls; eight charged singleton expansions exhaust the libraries, and the remaining seven calls enrich the block with the largest sparse tail. We merge local reports additively and charge one deployment call on that summed objective, for 36 local calls. A separate offline check evaluates the chosen deployment on the full graph. This procedure instantiates the allocation logic suggested by Theorem 3, but it is not a certificate for arbitrary global objectives: cross-block interactions and the difference between summed local GCC reductions and one full-graph GCC remain unmodeled. Table 15 reports both local and full-graph values so this boundary is observable.

E.2 DATASET, NOISE, AND PREPROCESSING DETAILS

E.2.1 DECOMPOSITION STATISTICS

For each dataset we report the shortlist size, the measured interaction-graph width under the chosen decomposition, and the two bag-load bounds used in the decomposition-sensitivity table below. The main moderate tasks have min-fill interaction-width upper bounds from 20 to 22; exact or thin pairwise post-screening remains tractable in this low-to-mid-width regime.

E.2.2 NOISE MODEL

The synthetic epidemic-suppression diagnostic is a noisy simulator stress test on the public-network intervention suite. The objective is the reduction in a Monte Carlo independent-cascade (IC) outbreak

proxy after deleting the selected nodes; each simulator call uses 200 rollouts. For the synthetic noisy validation table below, we also use additive Gaussian perturbations with variance chosen so that the single-query signal-to-noise ratio is approximately one at the median singleton coefficient magnitude. The repeated-query theorem bounds the noise term by a quantity of order $n^{-1/2}$. We report the measured errors and regrets for each repeat count.

E.2.3 Preprocessing Overhead

The graph-induced compression layer is not free: it pays for footprint construction, shortlist-domain interaction-graph formation, and tree decomposition. In our implementation this cost is small compared with the simulator on the public-network tasks, but we still record it. Decomposition costs depend on the support graph. The width-stress table reports its measured running time separately from the number of objective calls.

E.2.4 Breadth-Scenario Protocol

The breadth suite adds edge deletion for GCC attack and radius-2 maximum coverage on each moderate graph. Every task uses $r = 24$, $k = 4$, seed 0, and $35 = 1 + 24 + 10$ model-fitting calls plus one deployment call. Edge candidates are the 24 bridges with largest exact singleton split; a weighted bridge tree evaluates the full-graph GCC reduction. Coverage candidates use the default union shortlist and maximize the union of radius-2 closed neighborhoods. Greedy uses $24 + 23 + 22 + 21 = 90$ search calls plus one deployment call. The Oracle enumerates all $C(24, 4) = 10{,}626$ deployments, yielding $63{,}756$ stored objective values across the six tasks.

E.3 Evaluation Protocol and Baseline Implementations

E.3.1 Official BO Baselines

We use these open-source implementations with task adapters and the stated settings:

1. COMBO (`QUVA-Lab/COMBO`): diffusion-kernel BO over the shortlist subset space.
2. GraphComBO (`LeonResearch/GraphComBO`): shortlist graph search.
3. BayesOptG (`xingchenwan/bo-on-graph`): graph BO on the same shortlist subset graph.
4. Bounce (`lpapenme/bounce`): trust-region BO adapted to fixed-cardinality subsets.

The corrected-coordinate comparison reruns Bounce at commit `738b9bd` with the same shortlist, initial design size (12), full 24-dimensional trust regions, seeds 0–9, and exactly 35 charged invocations per seed. Its binary proposal is projected to a legal size-4 subset by retaining the first four active coordinates and filling missing positions in coordinate order; projected duplicates remain charged. COMBO, GraphComBO, and BayesOptG are also rerun on the corrected shortlist, with five seeds and 35 optimizer requests. Invalid-cardinality COMBO proposals receive a penalty without an objective evaluation, so COMBO uses 12.0–12.6 actual objective calls on average; the other two use 35. Table 22 reports these counts. COMBO caps posterior burn-in at 3 steps and retained samples at 1 per proposal, following the archived run configuration. Its row describes this truncated-sampling, request-budget protocol; it does not establish a comparison with standard COMBO at an equal objective-call budget. We did not identify an official public repository for BoGrape, so it is omitted.

E.3.2 Expanding-Pool BO Control

We also include an appendix-only generic expanding-pool control. It starts from the same front-end prefix of size 12, expands along the same ranking until size 24, fits a ridge-regularized quadratic surrogate from observed subsets, and selects the next subset by a simple upper-confidence acquisition over sampled feasible deployments. This baseline is deliberately not labeled as an official adaptive-domain BO method: it tests whether generic expansion plus a quadratic model explains the results, without importing a method whose assumptions do not match shortlist-audit graph intervention.

Table 5: Expanding-pool quadratic control with 35 calls. Ratios use the same-run 36-call fixed-shortlist reference on this protocol; calls are not matched. Each row has three seeds.

Task	Seeds	Mean value / ours	Min	Max
CA-HepPh / GCC	3	0.876	0.714	1.000
com-DBLP / GCC	3	1.000	1.000	1.000
Email-Enron / GCC	3	0.860	0.793	0.899
Email-Enron / spectral	3	1.000	1.000	1.000

E.3.3 Greedy and Beam Baselines

Greedy expands the best current partial deployment by all admissible one-step additions, re-evaluating only unseen subsets. Beam search uses beam width 20 and the same shortlist domain. The query counts reported in the main text include all simulator calls made during expansion and tie-breaking; this is why the total can exceed the shortlist size even when k is small.

E.3.4 Wall-Clock Accounting

Every method is charged for every simulator call. For the proposed approach, the accounting includes shortlist construction, pair querying, and surrogate optimization; for BO baselines, it includes model fitting and acquisition optimization. On the real large-graph suite, the full five-method runs complete in 339.1 seconds on ogbn-arxiv, 894.8 seconds on ogbn-products, and 1230.5 seconds on Orkut for one seed each. The synthetic BA-5M stress test completes in 885.4 seconds. These times are reported only as secondary context; all main tables compare methods under matched simulator-call budgets.

E.4 Additional Decision and Routing Experiments

E.4.1 Water-Network Sensor Placement

We use BWSN Network 1, a hypothetical engineering benchmark with 126 eligible junctions[1]. Demand, topology betweenness, and an equal-weight combination of demand, betweenness, and hydraulic reach supply three shortlists of 24 sensors. The front ends do not observe contamination outcomes. We select four passive sensors and measure the reduction in contaminated water consumed before detection, at a threshold of $0.01\,\mathrm{mg/L}$.

Each of 126 source junctions contributes four training and four held-out scenarios. Training injections start at 0, 6, 12, and 18 hours, last two hours, and use demand scale 1.0. Held-out settings use (start hour, duration in hours, demand scale) $(3, 1, 0.9)$, $(9, 3, 1.1)$, $(15, 1, 1.1)$, and $(21, 3, 0.9)$. Thus each bank contains 504 scenarios, with shared source locations and different injection and demand settings. One baseline hydraulic run and 1008 contamination runs produce the shared EPANET 2.2 bank (Rossman et al., 2020). Each logical objective call evaluates one sensor subset on this bank; it does not rerun the hydraulics.

The singleton and pairwise methods use 26 and 35 logical calls, including deployment. The 35-call greedy control evaluates all 24 singletons, then three singleton-ranked proposals at each of the three remaining steps. We enumerate all 10,626 size-four subsets on the held-out bank only for evaluation. Gap closed is $(F_{\mathrm{pair}} - F_{\mathrm{singleton}})/(F^*_{\mathrm{heldout}} - F_{\mathrm{singleton}})$.

Table 6: Held-out BWSN volume reduction (m^3). Pairs change all three deployments.

Front end	Singleton 26	Pairwise 35	Greedy 35	Held-out optimum	Gap closed
Demand	463.003	485.457	487.949	498.089	64.0%
Bridging	461.209	464.130	464.127	495.622	8.5%
Combined	460.716	461.674	462.285	495.456	2.8%

[1]BWSN Network 1: https://uknowledge.uky.edu/wdst_models/3.

The BWSN audit finds large omitted-pair terms and held-out residuals (maximum residual 1695.45–1922.67 m^3 across front ends). These errors still support enrichment or fallback.

On the natural graph objectives, a separate 3-dataset × 2-objective × 6-front-end grid changes one of 36 deployments. This is com-DBLP/spectral with degree ranking: value increases from 3.519440 to 3.519487, and deployment residual falls from 0.014799 to 0.000467. With absolute tolerance 10^{-10}, residual decreases in 17 configurations, increases in two, and is unchanged in 17.

E.4.2 Order Selection under Triadic Interactions

A controlled CA-HepPh spectral test adds a known footprint-supported third-order term to a pairwise comparator. Below the pairwise-margin bound, both orders select the optimum. At multiplier 1.25, retained pairs incur regret 0.019982; sparse third-order recovery removes it in 89 total calls, compared with 2117 for full clique recovery. The known added coefficients give the exact higher-order tail for this controlled test.

A separate stress suite contains 760 constructed pure-triadic cases with $r = 24$ and $k = 3$. The seven static networks are CAIDA-AS, Gnutella08, and Epinions from SNAP (Leskovec and Krevl, 2014), Power Grid[2], SocioPatterns HS2012 (Fournet and Barrat, 2014)[3], Sioux Falls[4], and IEEE118[5], each with five shortlist seeds. BWSN supplies three front ends. Each of these 38 panels uses ten susceptibility draws from each of Uniform$[0.75, 1]$ and Beta$(0.05, 1)$, giving $38 \times 2 \times 10 = 760$ cases. For each triple U, its coefficient is $\theta_U = \sum_{v \in \cap_{i \in U} \Psi_i} p_v$, with $0 \le p_v \le 1$; all lower-order coefficients are zero. The known envelope $u_U = |\cap_{i \in U} \Psi_i|$ bounds each coefficient. A value-free preflight sends 80 cases directly to enumeration at call 0. In the other 680 cases, a fixed positive-support triple and its pairwise closure use eight calls and expose a nonzero third-order residual. The construction guarantees this witness; it is not a general detector of unknown interaction order. Queried closures are reused. Recovery stops once the best observed coefficient is at least every unqueried envelope.

Table 7: Pure-triadic stress: 760 cases. Routing includes the cost of its direct-search handoffs. Zero regret is measured against the exact size-three optimum.

Method	Mean total calls	Zero-regret cases
Forced $d = 2$	182.00	3/760
Sparse $d = 3$ (24 triples)	61.26	669/760
Matched random search	387.03	169/760
Order routing	387.03	760/760
Full clique $d = 3$	1063.39	760/760

Measured five-drug interactions. We use the measured *E. coli* growth data of Tekin et al. (2018). The 56 five-drug sets and 3^5 fixed dose assignments per set give 13,608 tasks. Each task has five candidate drugs and budget $k = 5$; the objective is 100 minus the median relative growth percentage. For each order d, we query all subsets of size at most d, including the empty set, apply Möbius inversion, and maximize the surrogate over all subsets. The final deployment query is charged even if it repeats a fitting query. For $d = 2, 3, 4, 5$, the total calls are $17, 27, 32, 33$ and the mean regrets are $11.475, 7.504, 3.955, 0$ growth-percentage points. These means decrease with order; individual tasks need not improve at each step. Order five uses all 32 subset values and is exact for this measured objective. The entry point is `src/main/real_higher_order.py`, with saved results in `result/real_higher_order/raw.csv`.

E.4.3 A Fully Charged Adaptive Gate

We compare forced pairwise fitting, fixed sparse third-order fitting, audit-triggered random-search handoff, and random search from scratch under 50 logical calls. The CA-HepPh simulator uses a 12-candidate shortlist, $k = 4$, two known simulation horizons, and three activation

[2]Power Grid archive: https://networkdata.ics.uci.edu/data/power/power.gml.

[3]We acknowledge the SocioPatterns collaboration: HS2012 school-contact data.

[4]Transportation Networks for Research: Sioux Falls data, revision d1639b4.

[5]MATPOWER case118: case118.m, revision 5f1b706.

thresholds. Calibration and evaluation use disjoint Monte Carlo seeds; routing thresholds are fixed before evaluation, separately for each simulation horizon.

The gate uses one empty-set query, 12 singletons, 12 pairs, and four deployment probes: 29 calls. Each absolute probe residual is divided by its own $\max(|F(B)|, |\widehat{G}(B)|, 1)$, and the largest ratio is used. The range of signed probe residuals is divided by the largest of these scales across the four probes. This spread is an empirical routing proxy, not the certified sparse tail. This experiment tests a separately calibrated order-selection policy; its thresholds do not implement the certificate-based routing in Algorithm 1. If this spread is at most its frozen threshold, the method continues at $d = 2$. Otherwise, it enriches to $d = 3$ when normalized absolute residual reaches its threshold, and uses random-search handoff in the remaining cases. Handoff evaluates 20 uniformly sampled size-four deployments without replacement and spends one final call on the best feasible deployment observed across both the 29-call prefix and the 20 new queries.

The reference action is computed offline from exact order-two, order-three, and order-four models, with regret tolerance $\max\{0.5, 0.02|F_4^*|\}$ relative to the order-four optimum. These models and labels are excluded from online decisions. Across 12 held-out cases, the gate agrees with this action label in eight; its false-negative rate is $1/7$ and false-positive rate is $3/5$. In two difficult cases, regret falls from 42.18 to 11.50 and from 41.19 to 17.41 compared with forced pairs.

Table 8: Regret on five held-out reference-fallback cases, with 50 logical calls per method.

Method	Mean regret	Adaptive wins / cases
Adaptive gate	6.97	–
Forced pairwise	23.35	4/5
Random search from scratch	13.75	4/5

E.4.4 Bounds, Proxies, and Thresholds

A formal outside bound needs a valid marginal envelope over the stated candidate pool and a uniform comparator error bound. A sparse-tail bound needs bounds on omitted coefficients. The 32-candidate audit in Table 23 checks these quantities offline. A finite validation residual, overlap score, or unqueried front-end score provides a routing proxy. To set an operational threshold, specify a tolerable loss and call cost, normalize within the task, choose the threshold on separate calibration cases, then freeze it before held-out evaluation. Report coverage, false-positive and false-negative rates, and the calls left for repair together.

Table 9: Post-screening signals, their required evidence, and the decisions they support.

Signal	Required evidence and use
Outside / joint bound	Valid marginal bounds on the stated pool; add 2ε for uniform comparator error.
Sparse-clique tail	Bounds on absolute omitted coefficients; controls truncation loss.
Finite validation residual	Observed model error on queried subsets; guides order or method changes.
Overlap / front-end score	Ranks unqueried candidates or pairs for the next charged evaluation.

The distinction from fixed-domain BO is the decision being made. BO chooses the next subset within a supplied domain; post-screening also decides whether to change that domain or its query family. Two objectives can agree on every queried subset of S and differ on an unqueried candidate outside S. Their in-shortlist transcripts are identical. Assessing that omission therefore requires outside information or an explicit bound. Graph-induced clique queries supply the recovery structure; the charged audit connects it to this domain decision.

F SUPPLEMENTARY TABLES

Table 10: Breadth tasks: ours uses 36 calls and greedy uses 91. Optimal enumerates all $C(24, 4) = 10{,}626$ size-four deployments within each shortlist.

Dataset	Scenario	Ours	Greedy 91	Optimal	Ours − Optimal
CA-HepPh	Edge deletion / GCC	45	45	45	0
Email-Enron	Edge deletion / GCC	162	162	162	0
com-DBLP	Edge deletion / GCC	111	111	111	0
CA-HepPh	Maximum coverage / r2	4212	4826	4834	−622
Email-Enron	Maximum coverage / r2	23244	23967	24169	−925
com-DBLP	Maximum coverage / r2	14147	14973	15046	−899

Table 11: Corrected-coordinate Bounce under the same 35-call fixed-shortlist budget. Entries are mean ± sample standard deviation and range over 10 seeds; projected duplicates remain charged.

Task	Ours	Bounce 35	Bounce range
CA-HepPh / GCC	35.000	30.600 ± 4.789	$[22, 35]$
CA-HepPh / spectral	4.8124	4.6214 ± 0.1044	$[4.3732, 4.7386]$
Email-Enron / GCC	1995.0	1681.5 ± 308.4	$[840, 1951]$
Email-Enron / spectral	7.9465	7.5511 ± 0.5431	$[6.0222, 7.9465]$
com-DBLP / GCC	107.0	93.1 ± 7.80	$[76, 107]$
com-DBLP / spectral	4.5895	4.2708 ± 0.3102	$[3.4209, 4.4535]$

Table 12: Corrected-coordinate fixed versus charged adaptive post-screening under the same 35-call cap. "Outside" counts selected candidates absent from the fixed shortlist.

Task	Fixed	Adaptive	Difference	Outside
CA-HepPh / GCC	35.000	27.000	−8.000	0
CA-HepPh / spectral	4.8124	4.8124	0	0
Email-Enron / GCC	1995.0	2050.0	+55.0	1
Email-Enron / spectral	7.9465	7.9465	0	0
com-DBLP / GCC	107.0	143.0	+36.0	1
com-DBLP / spectral	4.5895	4.6496	+0.0600	1

Table 13: Corrected-coordinate retained-pair selection at $L = 10$. Paper, degree-priority, and all 100 random selectors deploy equal values; errors are maxima over all $C(24, 4) = 10{,}626$ deployments.

Task	Paper error	Degree error	Random median	Equal deployments
CA-HepPh / GCC	0	0	0	100/100
CA-HepPh / spectral	0.0143	0.0190	0.0205	100/100
Email-Enron / GCC	15.0	46.0	46.0	100/100
Email-Enron / spectral	0.2234	0.2239	0.2772	100/100
com-DBLP / GCC	0	0	0	100/100
com-DBLP / spectral	0.0120	0.0088	0.0088	100/100

Table 14: Pair-selector ablation on the appendix-union shortlists. Each thin selector uses ten pairs; calls include deployment. All-pair querying includes pairs outside the overlap graph.

Pair variant	CA-GCC	Email-GCC	Email-spec	DBLP-GCC	Calls
Singleton	35.000	1995.000	7.428	107.000	26
Random	35.000	1995.000	7.461	107.000	36
Degree product	35.000	1995.000	7.428	107.000	36
PageRank product	35.000	1995.000	7.428	107.000	36
Overlap	35.000	1995.000	7.428	107.000	36
Generic sparse	35.000	1995.000	7.461	107.000	36
Graph-induced	35.000	1995.000	7.428	107.000	36
All pairs	35.000	1995.000	7.461	107.000	302

Table 15: Charged four-block post-screening on full large graphs. Block values optimize the summed local objective; Full columns are offline checks of the selected deployment and a separate 91-call greedy reference on the true full-graph GCC objective. The 36 calls are charged to the summed local objective. Final outside mass is zero after exhausting each six-candidate block. Residual is the deployed local prediction error; tail sums the four blocks' remaining pair-score proxies.

		Block greedy		Full greedy		Final		
Graph	Calls	Block ours	91	Full ours	91	outside	Residual	Sparse tail
ogbn-arxiv	36	761	761	210	210	0	48	257.866
Orkut	36	3429	3429	215	271	0	13	919.234
ogbn-products	36	354	354	44	60	0	0	79.368

Table 16: Benchmark summary. Moderate rows use the corrected main shortlist; large rows use their localized domains. Width is a min-fill upper bound. Greedy reference costs 90 calls for moderate rows and 91 for large rows. Certificate experiments use separate pools.

Dataset	Objective	Width	Ours	Calls	Domain
CA-HepPh	gcc	20	35.000	35	Fixed shortlist
CA-HepPh	spectral	20	4.812	35	Fixed shortlist
Email-Enron	gcc	22	1995.000	35	Fixed shortlist
Email-Enron	spectral	22	7.946	35	Fixed shortlist
com-DBLP	gcc	20	107.000	35	Fixed shortlist
com-DBLP	spectral	20	4.590	35	Fixed shortlist
ogbn-arxiv	Local GCC	–	545	36	Localized
ogbn-products	Local GCC	–	298	36	Localized
Orkut	Local GCC	–	1392	36	Localized

Dataset	Objective	Greedy 35	Greedy ref.	Best BO mean	Cert.
CA-HepPh	gcc	18.000	35.000	30.600	–
CA-HepPh	spectral	2.408	4.812	4.664	–
Email-Enron	gcc	1421.000	1995.000	1742.400	–
Email-Enron	spectral	4.378	7.946	7.551	–
com-DBLP	gcc	57.000	107.000	93.100	–
com-DBLP	spectral	2.210	4.590	4.295	–
ogbn-arxiv	Local GCC	–	545	442	–
ogbn-products	Local GCC	–	298	237	–
Orkut	Local GCC	–	1392	1288	–

Table 17: Synergy sweep: exact-shortlist-optimum-normalized values averaged over four tasks. Aligned and corrupted modes are separate; all eight selectors are shown by task in Figure 16.

Mode / λ	Singleton	Random	Generic quad.	Overlap	Graph-induced	Oracle
Aligned / 0.00	0.999	0.999	1.000	1.000	0.999	1.000
Aligned / 0.10	1.000	1.000	1.000	1.000	1.000	1.000
Aligned / 0.25	0.997	0.998	0.998	0.992	0.982	1.000
Aligned / 0.50	0.988	0.978	0.991	0.995	0.986	1.000
Aligned / 1.00	0.907	0.871	0.926	0.934	0.941	1.000
Corrupted / 0.00	1.000	1.000	1.000	1.000	1.000	1.000
Corrupted / 0.10	1.000	1.000	1.000	1.000	1.000	1.000
Corrupted / 0.25	1.000	1.000	1.000	1.000	1.000	1.000
Corrupted / 0.50	0.991	0.991	0.991	0.999	1.000	1.000
Corrupted / 1.00	0.940	0.917	0.940	0.977	0.968	1.000

Table 18: Audit-pool calibration of the truncated joint proxy: 612 cases (four tasks, three pool sizes, 51 cases each). Pool 32 uses all 41,449 subsets of size at most four; pools 64 and 128 use 20,001 sampled subsets each. The false-positive rate (FPR) and false-negative rate (FNR) use the recorded task threshold; AUROC is the area under the receiver operating characteristic curve. Held-out routing under a 50-call budget is evaluated separately in Appendix E.4.3.

Task	Pool	Cases	Coverage	AUROC	FPR	FNR
CA-HepPh / GCC	32	51	1.00	0.998	0.64	0.00
CA-HepPh / GCC	64	51	1.00	0.977	1.00	0.00
CA-HepPh / GCC	128	51	1.00	1.000	1.00	0.00
Email-Enron / GCC	32	51	1.00	1.000	1.00	0.00
Email-Enron / GCC	64	51	1.00	0.975	1.00	0.00
Email-Enron / GCC	128	51	1.00	0.962	1.00	0.00
com-DBLP / GCC	32	51	1.00	1.000	1.00	0.00
com-DBLP / GCC	64	51	1.00	1.000	1.00	0.00
com-DBLP / GCC	128	51	1.00	1.000	1.00	0.00
Email-Enron / spectral	32	51	1.00	1.000	0.00	1.00
Email-Enron / spectral	64	51	1.00	0.982	0.00	1.00
Email-Enron / spectral	128	51	1.00	1.000	0.00	1.00

Table 19: Corrected-coordinate budget-matched greedy and query efficiency. Ours uses 35 calls; the singleton control reaches the same value in 26, including deployment. "Greedy 35" is the best value observed by call 35, not merely the last completed greedy round.

Task	Ours @35	Greedy 35	Greedy @90	Singleton total calls
CA-HepPh / GCC	35.000	18.000	35.000	26
CA-HepPh / spectral	4.8124	2.4076	4.8124	26
Email-Enron / GCC	1995.000	1421.000	1995.000	26
Email-Enron / spectral	7.9465	4.3778	7.9465	26
com-DBLP / GCC	107.000	57.000	107.000	26
com-DBLP / spectral	4.5895	2.2101	4.5895	26

Table 20: Separate large-graph audit pools ($|P| = 32$, $|S| = 24$, $k = 4$): means over three seeds. The localized regions differ from the deployment benchmark. These are pairwise-comparator proxies; this audit includes no uniform bound on real-objective error.

Dataset	Mean outside loss	One-step proxy	Joint proxy	Ratio of means
ogbn-arxiv	0.00	44.00	35.00	0.795
ogbn-products	96.33	206.33	199.33	0.966
Orkut	64.00	217.67	200.67	0.922

Table 21: Localized large-graph results from one seed, shown as objective value / actual calls. BO receives 36 optimizer requests; invalid-cardinality COMBO proposals receive penalties without objective calls. The greedy reference uses 91 objective calls including deployment.

Dataset / Scale	Ours (value / calls)	Greedy	COMBO	GraphComBO	BayesOptG
ogbn-arxiv (region 40,193)	545/36	545/91	374/12	442/36	424/36
ogbn-products (region 89,220)	298/36	298/91	112/12	207/36	237/36
Orkut (region 120,000)	1392/36	1392/91	861/12	1288/36	895/36

F.1 Experimental Diagnostics

F.1.1 Main-Text Deferred Tables

Table 22: Corrected-coordinate official BO means over five seeds. All BO methods receive 35 optimizer requests; GraphComBO and BayesOptG each use 35 actual objective calls. Invalid-cardinality COMBO proposals receive penalties without objective evaluation; its actual mean calls are shown.

Task	Ours	COMBO	GraphComBO	BayesOptG	COMBO calls
CA-HepPh / GCC	35.000	22.000	24.200	27.000	12.4
CA-HepPh / spectral	4.812	4.318	4.600	4.664	12.0
Email-Enron / GCC	1995.000	1640.400	1739.200	1742.400	12.6
Email-Enron / spectral	7.946	4.886	6.886	6.924	12.4
com-DBLP / GCC	107.000	66.400	80.000	87.600	12.4
com-DBLP / spectral	4.590	3.041	4.295	3.991	12.2

Table 23: Exhaustive offline audit on six 32-candidate pools, $|S| = 24$, $k = 4$. The real-objective outside-loss bound is $2\varepsilon + C$, where ε is the maximum comparator error over all 41,449 feasible subsets. Joint mass decreases on four tasks and ties on two. These offline queries are separate from the main protocol and its 35-call online budget.

Task	Outside loss	One-step C	Joint C	ε
CA-HepPh / GCC	3.0000	17.0000	17.0000	0.000000
CA-HepPh / spectral	0.0000	4.1278	4.1018	0.000476
Email-Enron / GCC	0.0000	76.0000	73.0000	9.000000
Email-Enron / spectral	0.0000	0.9822	0.8189	0.027476
com-DBLP / GCC	41.0000	100.0000	100.0000	0.000000
com-DBLP / spectral	0.2063	3.1406	3.1201	0.000853

The main higher-order panel in Table 3 reports default full-shortlist query counts; this appendix table reports targeted diagnostic counts, so the query counts should not be compared as identical settings.

Table 24: Local-support proxy load and shortlist interaction width on real tasks ($r = 24$, radius-2 footprints). The proxy columns summarize local-support decomposition size; width is a min-fill upper bound for the direct-overlap graph on the appendix-union shortlist.

Dataset	Support nodes	h	Load proxy	Interaction width
CA-HepPh	5820	4703	24	20
Email-Enron	27909	25495	24	22
com-DBLP	33251	16856	24	20

Table 25: Large-k diagnostic on five real tasks with $L = 10$. Gap is averaged after dividing each task's loss by its greedy reference magnitude. Residual is the unnormalized maximum over sampled validation subsets, then averaged across tasks. Ties use a 10^{-9} tolerance.

	$r = 24$			$r = 48$		
k	Ties / 5	Gap	Residual	Ties / 5	Gap	Residual
4	4	8.81×10^{-4}	1.64	5	0	3.02
8	4	2.31×10^{-4}	2.48	4	5.79×10^{-5}	8.86
12	5	0	4.30	3	4.52×10^{-4}	9.48
16	4	2.09×10^{-4}	5.13	3	7.18×10^{-4}	12.90
20	4	5.57×10^{-5}	6.16	2	0.0015	14.93

F.1.2 Sensitivity, Robustness, and Stress Tests

Table 26: Raw values for the synthetic scaling analysis (20 seeds).

r	Mean Λ_2	Std. Λ_2	Mean load proxy
16	150.6	5.6	2.25
32	304.7	6.0	2.45
64	614.2	9.4	2.80
128	1231.8	16.1	2.95
256	2468.2	20.0	3.00
512	4929.9	34.8	3.00
1024	9853.9	40.7	3.00
2048	19700.4	75.2	3.00
4096	39399.1	126.1	3.00

F.1.3 Validation Residuals on Public-Network Tasks

These public-network diagnostics use a separate protocol: sample r candidates without replacement from the top-$4r$ degree pool, using seeds 0–9 and radius-two footprints.

Table 27: Offline validation on the separate public-network protocol ($r = 24$, $k = 4$, $L = 10$, 10 seeds). Residual is the maximum over every subset of size at most four. Residual and regret are divided by the original GCC size, spectral radius, or unmodified IC outbreak size; the IC bank has 200 rollouts.

Task	Mean relative residual	Sample std.	Mean relative regret
CA-HepPh / GCC	2.68×10^{-5}	6.02×10^{-5}	0
CA-HepPh / spectral	5.71×10^{-5}	7.05×10^{-6}	1.18×10^{-6}
com-DBLP / GCC	9.46×10^{-7}	2.99×10^{-6}	0
com-DBLP / spectral	8.81×10^{-5}	5.34×10^{-5}	3.07×10^{-6}
Email-Enron / GCC	3.56×10^{-4}	1.95×10^{-4}	0
Email-Enron / spectral	0.0012	5.79×10^{-4}	0
Email-Eu-core / IC	4.38×10^{-4}	1.00×10^{-4}	1.51×10^{-5}

F.1.4 Exact Synthetic Recovery for General Order

Table 28: Exact synthetic recovery for general d (10 seeds, exhaustive search over $|B| \leq 4$). Queries include the empty set and all admissible cliques. Reported time includes fitting and deployment search; exact offline enumeration is timed separately.

d	r	Queries	Fit + search (ms)	Deployment regret
2	18	30.4 ± 2.5	8.7 ± 1.0	0.0000 ± 0.0000
2	24	40.1 ± 3.5	39.9 ± 9.8	0.0000 ± 0.0000
2	30	50.2 ± 3.1	154.7 ± 20.7	0.0000 ± 0.0000
3	18	31.2 ± 2.0	13.9 ± 0.6	0.0000 ± 0.0000
3	24	42.0 ± 3.2	52.2 ± 2.8	0.0000 ± 0.0000
3	30	51.9 ± 3.5	136.3 ± 18.9	0.0000 ± 0.0000

F.1.5 Noisy Synthetic Recovery with Validation Residuals

Table 29: Noisy synthetic recovery (20 seeds, $r = 20$, $k = 4$; mean $\pm$ sample standard deviation). Gaussian noise has standard deviation equal to the median absolute singleton coefficient. All errors are divided by the exact optimum magnitude; validation covers every feasible subset.

Repeats n	Mean coeff. error	$\widehat{\varepsilon}^{\mathrm{val}}_{H,2}$	Mean regret
1	0.159 ± 0.057	0.993 ± 0.343	0.210 ± 0.152
2	0.114 ± 0.036	0.700 ± 0.193	0.134 ± 0.102
4	0.086 ± 0.021	0.548 ± 0.134	0.132 ± 0.077
8	0.058 ± 0.014	0.360 ± 0.073	0.067 ± 0.058
16	0.042 ± 0.015	0.246 ± 0.069	0.037 ± 0.040

F.1.6 Decomposition Sensitivity

Table 30: Bag-shadow decomposition sensitivity on the separate public-network protocol (10 seeds per graph and shortlist size). $\Lambda_2 = \sum_t (\ell_t + \binom{\ell_t}{2})$ is a bag-load upper bound, not the count of distinct queries. Ratio averages min-degree/min-fill within each seed; gap differences are measured against the exact size-four optimum on each shortlist.

Dataset	r	Min-fill Λ_2	Min-degree Λ_2	Ratio	Gap diff. $\times 10^{-3}$
CA-GrQc	20	53937 ± 12942	55303 ± 12796	1.028	0.0
CA-GrQc	24	57262 ± 15988	59738 ± 15516	1.053	0.0
CA-HepTh	20	179594 ± 12250	165030 ± 13942	0.919	0.0
CA-HepTh	24	229966 ± 18881	209205 ± 15206	0.911	0.0
Email-Eu-core	20	148174 ± 1413	141558 ± 491	0.955	0.0
Email-Eu-core	24	214076 ± 2885	202826 ± 877	0.948	0.0
Facebook	20	259357 ± 33026	258981 ± 34939	0.998	0.0
Facebook	24	363913 ± 44314	362307 ± 52939	0.993	0.0
LastFM-Asia	20	294745 ± 18007	272951 ± 21790	0.925	0.0
LastFM-Asia	24	440434 ± 30895	404459 ± 24310	0.919	0.0
p2p-G06	20	298109 ± 14331	278924 ± 12675	0.936	0.0
p2p-G06	24	421681 ± 30072	391931 ± 24923	0.930	0.0
Power	20	1332 ± 318	1340 ± 312	1.007	0.0
Power	24	1990 ± 479	2006 ± 468	1.010	0.0
Wiki-Vote	20	1208478 ± 28392	1134516 ± 12892	0.939	0.0
Wiki-Vote	24	1721193 ± 25419	1625103 ± 17682	0.944	0.0

F.1.7 Local-Support Load Statistics on the Real Tasks

The sufficient conditions in Corollary 9 are easiest to interpret on the union of shortlisted footprints rather than on the full ambient graph. Table 31 summarizes those local-support

quantities for the three real datasets used in the main moderate benchmark. The local-support bag-shadow quantities can be large even when the induced shortlist interaction width stays modest; this is why the practical implementation is organized around shortlist-domain width rather than around a claim of globally small treewidth.

Table 31: Local-support decomposition statistics for the real moderate datasets ($r = 24$, radius-2 footprints). The load proxy is a local-support statistic; interaction width is a min-fill upper bound on the direct-overlap graph of the appendix-union shortlist.

Dataset	Support nodes	Support edges	h	Load proxy	Interaction width
CA-HepPh	5820	98996	4703	24	20
Email-Enron	27909	167328	25495	24	22
com-DBLP	33251	169647	16856	24	20

F.1.8 Wide-Width Degradation and Thin Fallback

Table 32: All-admissible-pair versus thin querying as shortlist size grows. Width is a min-fill upper bound; both call counts include deployment. "Exact" means recovery of the allowed pair coefficients.

Dataset	$\|S\|$	Width	All pairs calls / s	Thin calls / s	Thin gap
Email-Enron	24	22	296 / 9.62	36 / 0.28	0.0329
Email-Enron	48	45	1167 / 32.40	60 / 1.19	0.0000
Email-Enron	72	69	2615 / 55.84	84 / 3.73	0.0000
Email-Enron	96	93	4639 / 77.84	108 / 8.42	0.0000
ogbn-arxiv	24	23	302 / 52.75	36 / 0.20	0.0000
ogbn-arxiv	48	46	1177 / 253.91	60 / 0.84	0.0000
ogbn-arxiv	72	68	2609 / 543.26	84 / 3.05	0.0000
ogbn-arxiv	96	89	4565 / 906.87	108 / 5.92	0.0000

F.1.9 Adaptive Shortlist Expansion

Table 33: Offline expansion on known pairwise objectives over 120-node graphs ($k = 3$, 10 seeds). Stopping uses the exact full-model loss. Total calls include full-library fitting and deployments. Regrets and bounds use the recorded full-model optimum as denominator; stability measures the Jaccard overlap between the final two deployments.

Graph	Static regret	Adaptive regret	Final $\|S\|$
BA-120	0.022 ± 0.032	0.007 ± 0.015	12.8 ± 1.7
SBM-120	0.117 ± 0.125	0.006 ± 0.011	14.4 ± 2.1
Email-120	0.022 ± 0.031	0.008 ± 0.017	12.8 ± 1.7
Power-120	0.046 ± 0.060	0.002 ± 0.008	13.6 ± 2.1

Graph	Relative bound	Stability	Total calls
BA-120	0.553 ± 0.112	0.87 ± 0.28	473.2 ± 0.4
SBM-120	0.818 ± 0.161	0.61 ± 0.36	449.0 ± 15.9
Email-120	0.714 ± 0.189	0.90 ± 0.21	1676.0 ± 517.6
Power-120	0.630 ± 0.182	0.74 ± 0.35	293.0 ± 30.5

F.1.10 Certificate-Tightness Calibration on Stressed Shortlists

We form a 32-candidate pool under the union ranking, then create 24-candidate shortlists by replacing five strong singleton candidates with lower-ranked candidates. For each omission pattern

we compute the one-step pairwise proxy from the fitted pool model. The denominator is the true shortlist loss obtained by exact enumeration on the full library and on the stressed shortlist. Table 34 reports the resulting calibration ratios. The replacement patterns have the same loss and proxy within each task, so their sample standard deviations are zero.

Table 34: Offline pairwise-proxy/true-loss ratios on stressed 24-candidate shortlists within 32-candidate pools ($k = 4$, 20 patterns per task). Every pattern has positive loss. The proxy alone omits the real-objective approximation term and is not a certified real-objective bound.

Task	Positive-loss patterns	Mean ratio $\pm$ sample std.
CA-HepPh / GCC	20	1.727 ± 0.000
com-DBLP / GCC	20	1.609 ± 0.000
Email-Enron / GCC	20	1.449 ± 0.000
Email-Enron / spectral	20	2.400 ± 0.000

F.1.11 Certificate Conservatism and Local Graph Structure

To understand when the positive-part certificate is more conservative, we summarize graph-level structure for three separate small-graph families. The three families do not exhibit a clean monotone dependence on degree or clustering alone. The most robust empirical statement is narrower: all three induce almost-complete shortlist overlap graphs, and the loosest ratio occurs on the smallest family, Florentine Families. In other words, certificate conservatism appears to depend on overlap and candidate-pool granularity together rather than on a single scalar graph statistic. These measured ratios describe proxy conservatism on the sampled shortlists.

Table 35: Separate toy-graph proxy calibration ($k = 4$, 20 omission patterns per graph, shortlist size 12). Pools have 24 candidates on Karate/Les Misérables and 15 on Florentine Families. Ratios use exact finite-pool loss; the proxy omits approximation error.

Graph	Ratio mean	Degree mean	Degree p95	Average clustering	Overlap density	h	Load proxy	Width
Karate Club	1.28	4.59	13.40	0.57	0.989	29	24	22
Les Misérables	1.32	6.60	16.20	0.57	1.000	69	24	23
Florentine Families	2.50	2.67	4.60	0.16	0.971	12	15	13

F.1.12 Sensitivity to the Retained Pair Budget

Table 36: Pair-budget sensitivity on the separate public-network protocol (GCC on three graphs, 10 seeds each, $r = 24$, $k = 4$). Calls include deployment. Gap and omitted-pair tail are divided by the exact shortlist optimum; the tail is an offline coefficient diagnostic.

L	Calls	Relative gap	Relative omitted-pair tail
8	34	0.0000 ± 0.0000	0.0195 ± 0.0381
12	38	0.0000 ± 0.0000	0.0185 ± 0.0376
16	42	0.0000 ± 0.0000	0.0184 ± 0.0376
24	50	0.0000 ± 0.0000	0.0143 ± 0.0343

F.1.13 NOISY-QUERY ROBUSTNESS ON SYNTHETIC PUBLIC-NETWORK EPIDEMIC SUPPRESSION

Table 37: Email-Eu-core IC on the separate public-network protocol (10 seeds). Training uses the stated rollout count; an independent 400-rollout bank defines the exact reference. Regret is divided by the unmodified evaluation-bank outbreak size. The total call budgets are 36 for our method and GraphComBO, and 92 for the greedy reference.

Rollouts / query	Ours gap	Greedy gap	GraphComBO gap
50	0.000265 ± 0.000303	0.000281 ± 0.000326	0.002607 ± 0.001228
100	0.000101 ± 0.000217	0.000101 ± 0.000217	0.002473 ± 0.001038
200	0.000057 ± 0.000096	0.000077 ± 0.000103	0.002667 ± 0.000918
400	0.000057 ± 0.000096	0.000077 ± 0.000103	0.002346 ± 0.001246

F.1.14 CORRELATED OBSERVATION-NOISE PILOT

The independent-noise proposition is a clean theorem, but real simulators can share common random effects across queried subsets. Table 38 reports a pilot on CA-HepPh spectral suppression with additive common-random-effect noise. The intent is not to claim a monotone curve in ρ from a small pilot, but to show that once query noise is correlated, the original i.i.d. coefficient should be read as an effective-variance statement rather than reused unchanged.

Table 38: Correlated observation-noise pilot on CA-HepPh spectral suppression (repeats $= 4$, $\sigma = 1$).

ρ	Mean value	Std. value	Mean regret
0.00	4.0786	0.5518	0.7338
0.25	3.9191	0.6670	0.8933
0.50	4.1039	0.5824	0.7085

F.1.15 REAL FRONT END ROBUSTNESS DETAILS

Table 39: Offline expansion diagnostics on three real tasks and five front ends. Endpoint values average the five front ends; correlation uses all 20 front-end/shortlist-size observations per task. The pairwise outside quantity is a proxy, with setup and full-pool references charged separately.

Task	Corr(loss, cert)	Loss $\lvert S\rvert = 12$	Loss $\lvert S\rvert = 24$	Cert $\lvert S\rvert = 12$	Cert $\lvert S\rvert = 24$
CA-HepPh / GCC	0.909	24.000	0.000	35.600	4.000
CA-HepPh / spectral	0.347	0.075	0.000	4.567	0.754
Email-Enron / GCC	0.890	315.600	0.000	923.200	77.800

F.1.16 FRONT END EXPANSION OVERHEAD

Table 40: Offline front-end expansion to zero exact pool loss (mean $\pm$ sample standard deviation over three tasks). Calls include each task's 41,957–41,970 setup queries and every evaluated stage. Negative extra cost is measured against the union front end on the same task.

Front end	Rounds to zero loss	Calls to zero loss	Final $\lvert S\rvert$	Extra calls vs. union	Extra nodes vs. union
Union	1.00 ± 0.00	42013.3 ± 7.5	16.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
Degree	0.67 ± 0.58	42004.0 ± 21.0	14.7 ± 2.3	-9.3 ± 16.2	-1.3 ± 2.3
PageRank	0.67 ± 0.58	42004.0 ± 8.7	14.7 ± 2.3	-9.3 ± 16.2	-1.3 ± 2.3
Core	1.00 ± 0.00	42013.3 ± 7.5	16.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
Random-top-pool	0.67 ± 0.58	42004.0 ± 21.0	14.7 ± 2.3	-9.3 ± 16.2	-1.3 ± 2.3

F.1.17 COMPLETE CERTIFICATE PIPELINE CASE

Table 41: Offline pipeline diagnostic on Email-Enron GCC with core initialization. The policy uses ex-post pool information. Setup costs 41,970 calls for fitting and exhaustive reference values; stage counts include deployment, and cumulative counts include setup.

$\|S\|$	One-step proxy	Joint proxy	Exact loss	Value	Stage calls	Cum. calls	Stop rule
12	1760	1716	1097	953	24	41994	expand
16	524	513	0	2050	28	42022	$C_{\text{joint}} \leq 600$

F.1.18 RADIUS SENSITIVITY ON MODERATE REAL TASKS

Table 42: Radius sensitivity on four moderate tasks using this diagnostic's union shortlists ($r = 24$, $k = 4$, $L = 10$, 36 calls including deployment). Outside mass is an offline pairwise proxy; width is a min-fill upper bound. These shortlists are separate from the fixed main benchmark.

Dataset	Objective	Radius	Value	Shortlist gap	Outside proxy	Width	Support nodes
CA-HepPh	GCC	1	35.0000	0.0000	17.0000	18	1524
CA-HepPh	GCC	2	35.0000	0.0000	17.0000	20	5921
CA-HepPh	GCC	3	35.0000	0.0000	17.0000	22	9782
CA-HepPh	GCC	4	35.0000	0.0000	17.0000	23	10985
CA-HepPh	spectral	1	4.8124	0.0000	4.1278	18	1524
CA-HepPh	spectral	2	4.8124	0.0000	4.1278	20	5921
CA-HepPh	spectral	3	4.8124	0.0000	4.1278	22	9782
CA-HepPh	spectral	4	4.8124	0.0000	4.1278	23	10985
Email-Enron	GCC	1	2050.0000	0.0000	76.0000	17	12065
Email-Enron	GCC	2	2050.0000	0.0000	76.0000	22	27517
Email-Enron	GCC	3	2050.0000	0.0000	76.0000	23	32918
Email-Enron	GCC	4	2050.0000	0.0000	76.0000	23	33590
Email-Enron	spectral	1	7.9465	0.0000	0.9822	17	12065
Email-Enron	spectral	2	7.9465	0.0000	0.9822	22	27517
Email-Enron	spectral	3	7.9465	0.0000	0.9822	23	32918
Email-Enron	spectral	4	7.9465	0.0000	0.9822	23	33590

F.1.19 SIMPLE 35-CALL SEARCH CONTROLS

Table 43: Random and evolutionary search on appendix-union shortlists: mean $\pm$ sample standard deviation over three seeds. Both methods deploy their best evaluated subset within 35 calls.

Dataset / Objective	Random search	Evolutionary search
CA-HepPh / GCC	20.0000 ± 1.0000	29.0000 ± 6.0000
CA-HepPh / spectral	4.5371 ± 0.1636	4.6914 ± 0.0414
Email-Enron / GCC	1696.3333 ± 74.2181	1833.0000 ± 116.0517
Email-Enron / spectral	4.7292 ± 0.3450	6.1266 ± 1.2551
com-DBLP / GCC	59.3333 ± 14.2244	76.0000 ± 7.0000
com-DBLP / spectral	2.6762 ± 0.5605	3.9130 ± 0.5352

F.1.20 Sensitivity Sweep Summary

Table 44: The (r, k, L) sweep on 72 fixed-library configurations: $r \in \{16, 24, 32\}$, $k \in \{3, 4, 5\}$, $L \in \{8, 12, 16, 24\}$. Ours uses $r + L + 2$ calls; BO receives that many requests. Comparisons use the saved float64 observations with tolerance 10^{-9}; COMBO settings follow the disclosed protocol.

Task	Configs	Ours $\geq$ best BO	Ours $>$ best BO	Ours = Greedy	Ours $<$ Greedy
CA-HepPh / GCC	36	36	31	36	0
Email-Enron / spectral	36	34	24	29	7
Total	72	70	55	65	7

F.1.21 Synthetic 100M-Edge Stress Test

Table 45: Localized stress test on the synthetic 10^8-edge graph ($n = 5{,}000{,}000$, $|E| = 99{,}999{,}790$; evaluated region 253,889 nodes). Calls include deployment; BO receives 36 requests and invalid-cardinality COMBO proposals receive penalties without objective calls.

Method	Objective value	Calls
Ours	1429	36
Greedy	1431	91
BayesOptG	1422	36
GraphComBO	1425	36
COMBO	1394	12

G Supplementary Figures

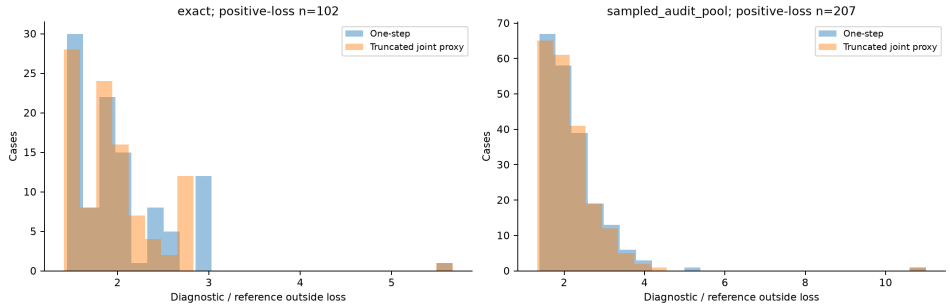


Figure 13: Certificate-ratio distribution. Histograms show 309 positive-loss cases, separated by exact and sampled reference. The 303 zero-loss cases have no finite ratio and are excluded; no kernel density estimate is fitted.

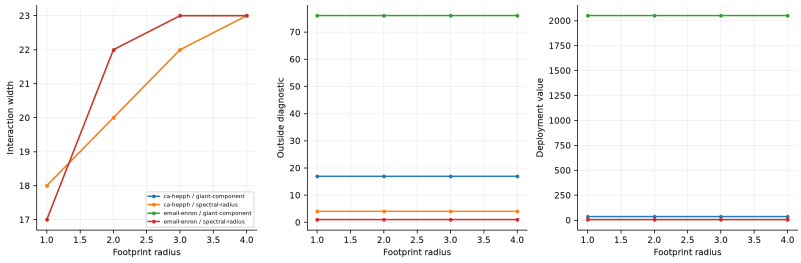


Figure 14: Footprint-radius sensitivity. Radius changes the interaction-width bound; the deployed values and outside proxies remain unchanged in these four tasks.

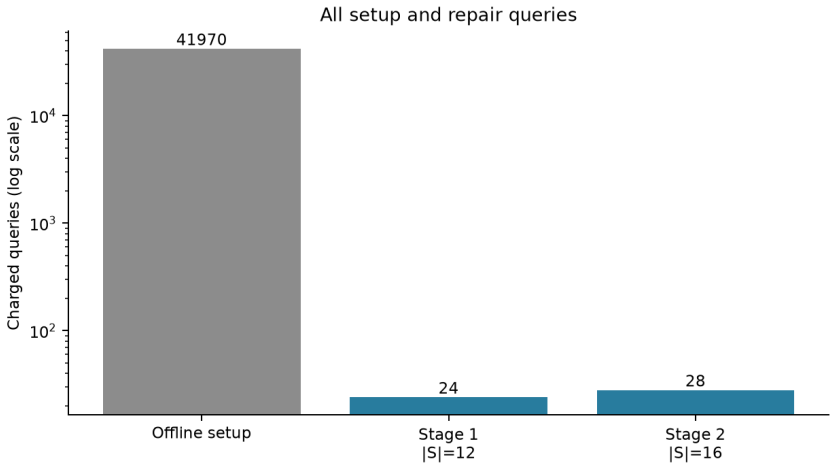


Figure 15: Total-call accounting on the Email-Enron GCC adaptive-repair trace. The offline setup includes full-pool fitting and exhaustive references; later stages include fitting and deployment.

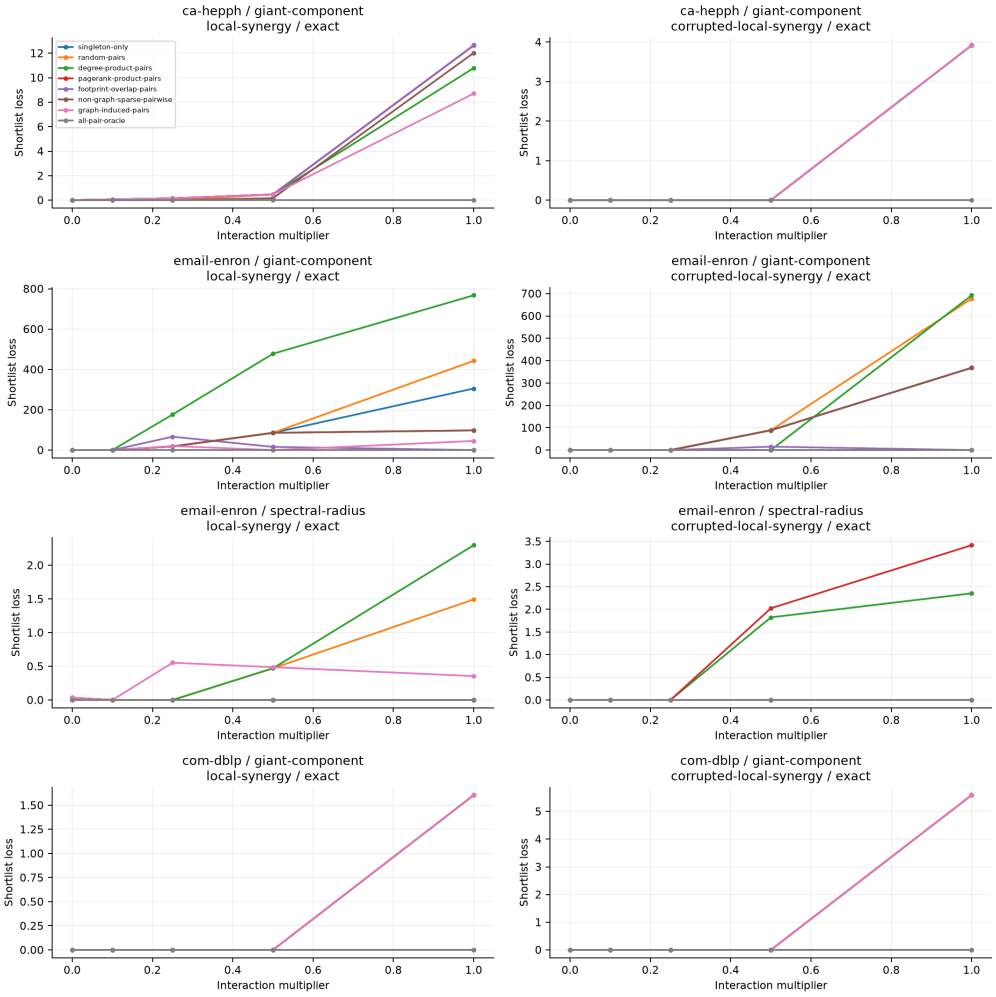


Figure 16: Task-level synergy sweep: eight selectors, four tasks, and two interaction modes. Loss is measured relative to each task's exact shortlist optimum.

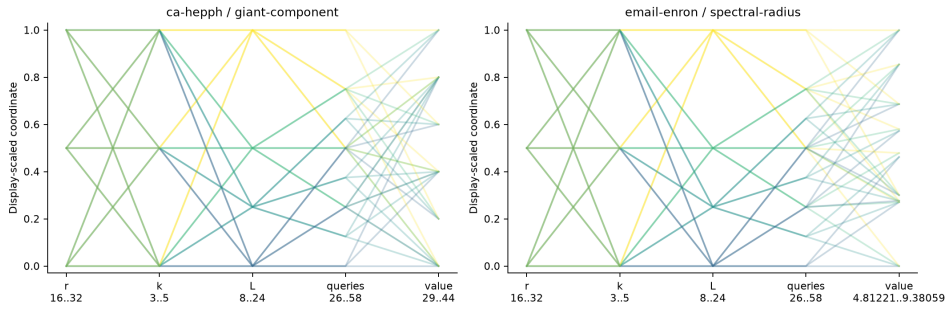


Figure 17: Parallel-coordinate view of the (r, k, L) sweep. The plot shows how shortlist size, deployment budget, retained-pair budget, calls, and value move together across the sensitivity grid.

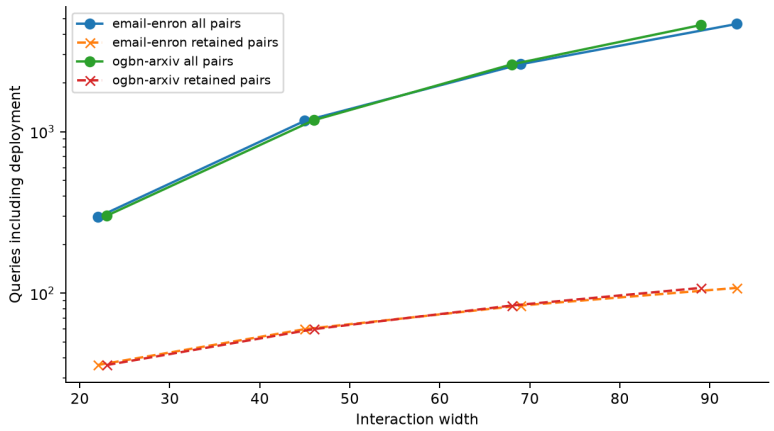


Figure 18: Width-query density for all-admissible-pair querying. Query counts include deployment; the graph width is a min-fill upper bound. Exact recovery of these pair coefficients alone gives no guarantee that the surrogate fits the underlying real objective.

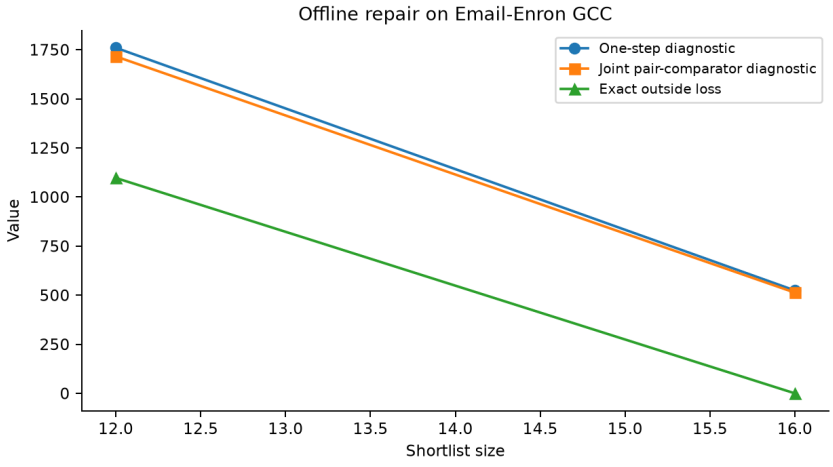


Figure 19: Offline expansion trace on Email-Enron GCC. The policy uses ex-post pool information; the fully charged online gate is evaluated separately in Appendix E.4.3.