Supplementary Material for *Curated-Reference Benchmarks Lack the Resolving Power to Establish Direct Regulatory-Edge Recovery from Perturb-seq*

Chimdi Walter Ndubuisi

Department of Electrical Engineering & Computer Science
University of Missouri–Columbia

Contents

Scope of this supplement **3**

S1 Method implementation truth **5**

S2 Edge semantics **10**

S3 Candidate and evaluable universes **13**

S4 Regulator-centric benchmark **18**

S5 Complete top-k tensor and universal-claim audit **22**
S5.1 Universal claim 1: no reference edge appears in any method's top 10 23
S5.2 Universal claim 2: precision@100 never exceeds 0.02 23

S6 Direct-versus-indirect target analysis **25**

S7 PCA back-projection validation **29**
S7.1 Code trace: the pipeline does not back-project 29
S7.2 Synthetic sweep: back-projection does not recover gene edges 29

S8 Sample-size saturation, corrected **31**
S8.1 Retraction of `perturbations_covered` 31
S8.2 The corrected experiment 31

S9 Aggregate test against matched random rankings **35**
S9.1 Design 35
S9.2 Why regulator clustering is mandatory 35
S9.3 Hierarchical bootstrap 35
S9.4 Clustered sensitivity model 37
S9.5 Per-reference and per-dataset breakdown 38
S9.6 How to read "% of regulators above random" 39
S9.7 Per-regulator enrichment, which does not overturn the verdict 39
S9.8 Caveats carried into the main text 39

S10 Matched synthetic-versus-real comparison **40**
S10.1 Design: what was matched, and how it was enforced . . . 40
S10.2 Verification that the real side is reproduced exactly . . . 40
S10.3 Prevalence calibration, and the scale that cannot be matched . . . 41
S10.4 The matched comparison . . . 42
S10.5 How much of the gap was mismatch, and how much is real . . . 44
The additive decomposition . . . 44
The ladder from the published claim to the matched one . . . 44
The residual is still shrinking with scale . . . 44
Same method, same metric . . . 45
Reference incompleteness explains a further modest slice . . . 45
S10.6 The full metric family . . . 45
S10.7 Methods actually run . . . 45
S10.8 Findings . . . 46
S10.9 Limitations carried into the main text . . . 46

S11 Expanded synthetic calibration **48**

S12 Reference database overlap **52**

S13 Repeated-run uncertainty **53**

S14 Faithful PC and GES **55**

S15 Reproducibility **60**

Scope of this supplement

This supplement documents every experiment behind the main text. Each table is generated directly from a machine-written result file; no number in this document is transcribed from narrative text.

Sections S1–S2 report the code audit of what each method implementation actually computes and what kind of object it emits. Section S3 defines the candidate and evaluable universes and states the counting correction that fixes the main text's sample size. Sections S4–S6 report the three descriptive evaluation layers: regulator-centric macro metrics, the complete full-graph top-k tensor with its universal-claim audit, and direct-versus-indirect discrimination *including the label-ablation caveat that qualifies it*. Section S7 reports the PCA back-projection code trace and validation sweep. Section S8 reports the *corrected* sample-size experiment and retracts the column it replaces. Section S9 is the primary inferential analysis: the aggregate test against matched random rankings, with regulator-level clustering. Section S10 reports the matched synthetic-versus-real comparison across five graph scales, including the scale that cannot be matched at all. Section S11 reports the expanded synthetic calibration, Section S12 the reference-database overlap, and Section S13 the repeated-run uncertainty experiment including the parts deliberately not run. Section S14 reports the faithful PC and GES benchmark including the infeasibility result. Section S15 gives reproducibility details: the analysis code and result tables that are available, and the exact environment versions.

Where the main text's corrections are evidenced. Readers checking the corrections should go to Section S3 (evaluable-regulator counts and self-loop-corrected positives), Section S8 (the retracted `perturbations_covered` column), Section S9 (the aggregate interval that replaces "1.4×–7.4× above random"), Section S6 (the label ablations and the withdrawal of "up to 27× in rank"), and Section S5, which no longer carries a precision-ceiling argument.

Two naming conventions are used throughout and are not negotiable for interpretation of any table below:

- The label "GRNBoost2" is **withdrawn and excluded from every pooled estimate**. An earlier version of this project reported that it was a `RandomForestRegressor` sharing the GENIE3 proxy's code path, and that the two returned bit-identical matrices. **That is wrong and is withdrawn.** Checked against the stored artifacts, the two differ in content (their adjacency matrices are not equal), in size, and in structure (the "GRNBoost2" artifact carries an additional array the GENIE3 artifact does not), and *none* of their 50 shared per-regulator rows agrees. Nor is it the published gradient-boosted GRNBoost2. What we can establish is only negative: no code in the project writes this artifact, so its generating procedure cannot be stated. It is carried in per-method displays for completeness and excluded from every pooled estimate, because a result whose provenance cannot be established cannot be described in a methods section. **This benchmark compares seven distinct methods under eight labels**, and wherever both labels appear their agreement carries no information about method agreement, ranking stability or reproducibility.
- The structure-learning baseline that earlier versions labelled "PC-GES" is a Graphical-Lasso partial-correlation graph with heuristic orientation by intervention-score magnitude. It is reported as **Glasso-PartialCorr-OrientedByIntervention (PC/GES, mislabeled in earlier versions)**.

Where a table is too narrow for the full honest names, these two are abbreviated "GRNBoost2*" and "Glasso-PartialCorr (PC/GES*)". The asterisk always marks a label that does not deliver the published algorithm whose name it carries — for "GRNBoost2" because its implementation could

not be identified at all; "GENIE3 (reduced cap.)" carries no asterisk because it implements the right algorithm with weakened hyperparameters. Ablation variants are prefixed "Abl."

The three curated reference databases (TRRUST v2, DoRothEA A+B, CollecTRI) are *not* independent; their pairwise overlap is quantified in Section S12.

S1 Method implementation truth

We audited the source of every method in the benchmark against the algorithm named by its pipeline label. The audit read the baseline, model and pipeline implementation sources line by line and recorded, for each method, the estimator actually constructed, the package that provides it, the data matrix it is fitted to, and its hyperparameters. The complete audit narrative is provided as a code-audit log in the reproducibility package; Tables S1 and S2 are the machine-readable summary.

Which methods are faithful and which are proxies. Of the 10 audited implementations, 6 are faithful to the algorithm named and 4 are proxies. The faithful implementations are NOTEARS-Linear, the Pearson correlation network, the Graphical Lasso network, CausalManifold, the vanilla autoencoder ablation, and the KNN-matched interventional-effect graph. The proxies are the implementations labelled PC, GES, GENIE3 and GRNBoost2. Four specific discrepancies must be stated plainly, because they change how the corresponding rows of every results table in this supplement may be read:

1. **The "GRNBoost2" label has no identifiable implementation.** Published GRNBoost2 [Moerman et al., 2019] is stochastic gradient boosting, and this is not that. It is also not the random-forest GENIE3 proxy, contrary to what an earlier version of this project reported: the two stored artifacts differ in content, size and structure, and none of their shared per-regulator rows agrees. No routine in the project writes the artifact. All rows under this label are therefore reported as provenance-unresolved and excluded from the primary aggregate.

2. **The "PC-GES" implementation is a partial-correlation proxy.** It fits `sklearn.covariance.GraphicalLasso` [Friedman et al., 2008] ($\alpha = 0.02$), converts the precision matrix to partial correlations, thresholds at $|r_{\text{partial}}| > 0.05$, and orients the surviving edges by comparing intervention-score magnitudes. There is no conditional-independence testing, no d-separation reasoning, no BIC scoring and no forward/backward search, so it is neither PC [Spirtes et al., 2000] nor GES [Chickering, 2002]. PC and GES share this single code path, so the two labels denote the same output.

3. **GENIE3 uses 50 trees against the published 1,000+.** The proxy is `RandomForestRegressor(n_estimators=50, max_features=None)` on at most 5,000 cells. Published GENIE3 [Huynh-Thu et al., 2010] uses at least 1,000 trees with `max_features=sqrt`; removing the feature subsampling removes the ensemble diversity that the method relies on. The algorithmic structure (per-target random-forest importance) is correct; the hyperparameters are not.

4. **NOTEARS runs on gene-space effects, with no PCA back-projection.** The pipeline uses PCA only for nearest-neighbour matching of perturbed to control cells. The interventional effect matrix passed to NOTEARS [Zheng et al., 2018] is (#perturbations × #genes) in gene space, and the returned adjacency is already gene-level. The PCA loading matrix is never used to back-project a component-space adjacency. Section S7 reports the synthetic experiment that tests what would have happened if it were.

Table S1: Method implementation truth, part 1 of 2: estimator, package and data scale actually used. $n = 10$ audited method implementations; the analysis unit is one method implementation in the benchmark pipeline. "Faithful" means the code implements the algorithm named by the label; "proxy" means it does not.

Published label	Estimator actually constructed	Package (version)	Faithful / proxy	Cells / features used	Trees or iterations
NOTEARS	Linear NOTEARS with augmented Lagrangian in PyTorch	torch (custom implementation) (native)	faithful	all (subsampled by pipeline); gene-space interventional effects matrix (n_perturbations x n_genes)	max_outer=8 x max_inner=300
PC algorithm	Partial-correlation pruning + intervention-based orientation (fallback)	sklearn.covariance.GraphicalLasso (sklearn)	proxy	all (subsampled by pipeline); PCA or latent features	–
GES	Same partial-correlation fallback as PC	sklearn.covariance.GraphicalLasso (sklearn)	proxy	all (subsampled by pipeline); PCA or latent features	–
GENIE3	sklearn RandomForestRegressor feature importance per target gene	sklearn.ensemble.RandomForestRegressor (sklearn)	proxy	max 5000 cells; 1024 genes	50 trees per target
GRNBoost2	**not established** — no routine in the project writes this artifact	unidentified	excluded	unknown	unknown
Correlation network	Pearson correlation matrix + threshold	numpy.corrcoef (numpy)	faithful	all (subsampled); raw gene expression	–
Graphical Lasso	sklearn GraphicalLasso precision matrix + threshold	sklearn.covariance.GraphicalLasso (sklearn)	faithful	all (subsampled); raw gene expression	max_iter=300

Table S1 continued

Published label	Estimator actually constructed	Package (version)	Faithful / proxy	Cells / features used	Trees or iterations
Causal Manifold (proposed)	MLP autoencoder with GELU + kNN Laplacian regularization + perturbation label classification head	torch (custom) (native)	faithful	max 20K train cells; 1024 genes	epochs=120
Autoencoder (ablation)	MLP autoencoder with ReLU	torch (custom) (native)	faithful	max 200K train cells; 1024 genes	epochs=80-100
Interventional Effect graph	KNN-matched control-vs-perturbed mean difference + quantile threshold	sklearn. neighbors + numpy (native)	faithful	all (subsampled); 1024 genes	–

Table S2: Method implementation truth, part 2 of 2: hyperparameters, honest name and the documented discrepancy. $n = 10$ audited method implementations; the analysis unit is one method implementation.

Published label	Honest name / key hyperparameters	Documented issue
NOTEARS	**NOTEARS-Linear (Zheng et al. 2018)** lambda_l1=0.01; lr=1e-2; rho *= 5.0; threshold=0.05	CORRECTED 2026-07-27: does NOT operate on PCA. The experiment driver passes the interventional effect matrix directly as the feature matrix, and that effect matrix is computed in gene space (perturbed cells minus matched control cells, both taken from the expression matrix). The latent/PCA array is used only to choose which control cells to match. The NOTEARS routine contains no PCA and no back-projection. Honest label: NOTEARS (gene-space effects). Resolved in full in the code-audit note included in the reproducibility package
PC algorithm	**Glasso-PartialCorr-OrientedBy Intervention** GraphicalLasso alpha=0.02; partial-corr threshold=0.05	Not a constraint-based causal search. Falls back to partial-correlation graph when causal-learn unavailable. No conditional independence testing, no d-separation

Table S2 continued

Published label	Honest name / key hyperparameters	Documented issue
GES	**Glasso-PartialCorr-OrientedBy Intervention** Same as PC proxy	Identical fallback to PC. No score-based search, no BIC scoring, no forward/backward phases
GENIE3	**RF-Feature Importance (50 trees; 5K cells)** n_estimators=50; max_depth=None; max_features=None (default=all); random_state=42	Published GENIE3 uses 1000+ trees with max_features=sqrt. This uses only 50 trees and no feature subsampling, significantly reducing ensemble diversity. Correct algorithmic structure but weak hyperparameters
GRNBoost2	**Provenance not established** No project routine writes this artifact	CRITICAL: neither the published gradient-boosted GRNBoost2 nor the random-forest GENIE3 proxy. The stored artifact differs from the GENIE3 proxy in content, size and structure, and no shared per-regulator row agrees. Excluded from every pooled estimate
Correlation network	**Pearson-Correlation-Network** threshold=0.2	Faithful implementation of a correlation network. Correctly labeled as a reference/non-causal method
Graphical Lasso	**Graphical-Lasso-Network** alpha=0.02; threshold=0.05	Faithful implementation. Correctly uses precision matrix for conditional independence structure
Causal Manifold (proposed)	**Laplacian-Regularized-Supervised-AE** hidden=[128,64]; latent=16; laplacian_weight=0.5; intervention_weight=0.2; n_neighbors=15	Novel method, no external reference to match. Implementation matches manuscript description
Autoencoder (ablation)	**Vanilla-MLP-Autoencoder** hidden=[128,64]; latent=16; lr=1e-3; batch=2048	Faithful vanilla AE for ablation comparison
Interventional Effect graph	**Perturbation-Effect-KNN-Graph** n_control_neighbors=20; quantile_threshold=0.9	Custom method for this paper. Uses latent-space KNN matching to estimate per-perturbation effects, then thresholds to build directed graph

Installation status of the faithful alternatives. The audit also recorded which faithful implementations are available in this environment (reported as the faithful-method availability test results in the reproducibility package). On a K562-GS smoke test with 2,000 cells × 50 genes: `causal-learn` PC returned a CPDAG with 27 edges in 1.03 s and `causal-learn` GES returned a CPDAG with 25 edges in 1.79 s; a 1,000-tree random forest (faithful GENIE3 configuration) took 536.24 s and a 500-tree `GradientBoostingRegressor` (faithful GRNBoost2 configuration) took 44.00 s, both yielding 2,450 non-zero importances. The Spearman rank correlation between the faithful 1,000-tree importances and the 50-tree proxy importances is 0.9434, so the proxy preserves most but not all of the ranking. `arboreto` is installed but its `genie3()` and `grnboost2()` entry points fail against current `dask`, and DCDI has no installable distribution; both are therefore excluded from the benchmark rather than substituted.

S2 Edge semantics

Methods in this benchmark do not all emit the same kind of object. Some return a weighted directed adjacency, some a CPDAG, some a symmetric matrix; some scores are signed and some are non-negative; the candidate space ranges from all gene pairs to perturbation–target pairs only. Tables S3 and S4 record the native output of every method variant considered, including the faithful variants that are installed but not benchmarked (Section S1). Any comparison across rows of the results tables must respect these differences: an undirected method is scored on the same directed candidate list as a directed method, which systematically penalises the undirected methods on orientation but not on skeleton recovery.

Table S3: Edge semantics, part 1 of 2: native output and orientation/sign conventions. $n = 13$ method variants; the analysis unit is one method variant's output object.

Method variant	Native output	Directed	Signed	Perturbation-aware
NOTEARS-PCA	weighted DAG adjacency (d x d)	yes	yes (signed weights)	no (observational only)
PC-proxy (Glasso-PartialCorr)	oriented partial-correlation graph	pseudo-directed (heuristic orientation)	yes (partial-corr sign)	optionally (orientation uses intervention scores)
GES-proxy	same as PC-proxy	pseudo-directed	yes	optionally
PC (causal-learn faithful)	CPDAG (completed partially directed acyclic graph)	mixed (directed + undirected)	no (binary edges only)	no (constraint-based on observational CI tests)
GES (causal-learn faithful)	CPDAG	mixed (directed + undirected)	no (binary edges only)	no (score-based on BIC)
GENIE3-proxy (50-tree RF)	importance matrix (p x p)	yes (regulator → target)	no (importance ≥ 0)	no (expression only)
GENIE3-faithful (1000-tree RF)	importance matrix (p x p)	yes (regulator → target)	no (importance ≥ 0)	no (expression only)
GRNBoost2 (provenance unresolved)	importance matrix (p x p)	yes (regulator → target)	no (importance ≥ 0)	no (expression only)
GRNBoost2-faithful (GBM)	importance matrix (p x p)	yes (regulator → target)	no (importance ≥ 0)	no (expression only)

Table S3 continued

Method variant	Native output	Directed	Signed	Perturbation-aware
Correlation network	correlation matrix (p x p)	no (undirected symmetric)	yes (Pearson r)	no (observational only)
Graphical Lasso	precision matrix (p x p)	no (undirected symmetric)	yes (precision entries)	no (observational only)
CausalManifold + InterventionalEffect	directed weighted adjacency (p x p)	yes (perturbed gene → affected genes)	yes (signed effect = mean shift)	yes (uses perturbation labels for both embedding and graph)
Autoencoder + InterventionalEffect	directed weighted adjacency (p x p)	yes (perturbed gene → affected genes)	yes (signed effect)	yes (uses perturbation labels for graph)

Table S4: Edge semantics, part 2 of 2: score definition, candidate space and honest name. $n = 13$ method variants; the analysis unit is one method variant's output object.

Method variant	Score definition	Candidate space	Honest name
NOTEARS-PCA	regression weight under acyclicity constraint	PCA components (not raw genes)	NOTEARS-Linear on PCA components
PC-proxy (Glasso-PartialCorr)	—partial_correlation— > threshold then orient by max —score—	all variable pairs	Glasso-Partial Corr-OrientedBy Intervention
GES-proxy	identical to PC proxy	all variable pairs	Glasso-Partial Corr-OrientedBy Intervention (same code path)
PC (causal-learn faithful)	conditional independence test p-value < alpha	all variable pairs	PC-CPDAG (causal-learn)
GES (causal-learn faithful)	BIC score improvement	all variable pairs	GES-CPDAG (causal-learn)
GENIE3-proxy (50-tree RF)	RF feature_importances_ (mean impurity decrease)	all TF-target pairs	RF-Feature Importance-50trees
GENIE3-faithful (1000-tree RF)	RF feature_importances_ with max_features=sqrt	all TF-target pairs	GENIE3-RF-1000trees

Table S4 continued

Method variant	Score definition	Candidate space	Honest name
GRNBoost2 (provenance unresolved)	not established	all TF-target pairs	excluded from pooled estimates
GRNBoost2-faithful (GBM)	GBM feature_importances_ (mean impurity decrease via gradient boosting)	all TF-target pairs	GRNBoost2-GBM-500trees
Correlation network	—Pearson r— > threshold	all gene pairs	Pearson-Correlation-Network
Graphical Lasso	—precision_ij— > threshold (L1-regularized inverse covariance)	all gene pairs	Graphical-Lasso-Network
CausalManifold + InterventionalEffect	mean(X_perturbed - X_knn_matched_control) thresholded at quantile	perturbation-target gene pairs only	Laplacian-AE + KNN-Matched-Perturbation-Effect
Autoencoder + InterventionalEffect	same KNN-matched effect as above but with vanilla AE latent	perturbation-target gene pairs only	Vanilla-AE + KNN-Matched-Perturbation-Effect

S3 Candidate and evaluable universes

The regulator-centric benchmark evaluates only those directed edges that are simultaneously inferable from the experiment and evaluable against a curated reference. An edge (r, t) enters the eligible candidate universe if and only if all four of the following hold.

1. **Perturbation rule.** Gene r was perturbed in the screen, so that a perturbation effect for r can be estimated at all.

2. **Annotation rule.** Gene r is annotated as a regulator in the reference database under evaluation. This rule is reference-specific, which is why the universe differs across the three databases for the same dataset.

3. **Measurement rule.** Gene t belongs to the 1,024-gene measured set, so that a score for (r, t) exists.

4. **Evaluability rule.** Regulator r has at least one curated target inside the measured gene set. Without this, per-regulator average precision is undefined and the regulator contributes no information to the macro-average.

Two regulator counts, and why the distinction decides the sample size. Rules 1–3 define the **candidate regulator** set: the perturbed genes that are also measured and are annotated as edge sources in the reference under evaluation. Rule 4 defines the **evaluable regulator** set: those candidate regulators that have at least one *non-self* curated target inside the measured target set. Only the evaluable set enters any macro-average, because a regulator with no positive target yields AP = NaN and is dropped. Earlier versions of this project reported the candidate count ("39–55 regulators") as though it were the macro-AP denominator. It is not. The correct denominators are **0–28 per dataset × reference cell** (Table S5).

An intermediate correction that gave "10 / 33" on K562-GS was also wrong, for a different reason: it counted regulators whose only curated target was the regulator itself. A regulator is never a candidate target of itself, so a self-pair is not scoreable and such a regulator contributes nothing. Removing self-pairs lowers both the evaluable count and the positive-edge count. Four cells change relative to earlier reports: K562-GS/TRRUST 39 → 38 positives, K562-GS/CollecTRI 211 → 188, K562-Ess/CollecTRI 5 → 4, RPE1-Ess/CollecTRI 47 → 44. Across all candidate regulators, 27 self-loops are excluded in total (23 under CollecTRI on K562-GS alone).

Two prevalences. Prevalence is reported here as positive edges divided by *evaluation* edges — the universe the metric actually averages over — rather than by candidate edges. The candidate-edge denominator includes pairs emitted by candidate-but-not-evaluable regulators, which are scored by the methods but contribute to no metric, so it understates the positive rate. Both columns are given in Table S5. The evaluation-universe prevalence is what a random ranker attains in expectation on a pooled ranking, and it is the number that should be quoted beside any macro-AP.

Exclusions implied by these rules. Perturbed genes that are not annotated regulators are excluded, as are annotated regulators that were not perturbed, curated targets outside the measured gene set, and self-edges. The exclusions are large: on K562-GS the full directed graph over 1,024 genes contains $1{,}024 \times 1{,}023 = 1{,}047{,}552$ edges, whereas the candidate universe contains between 359 and 20,309 edges depending on reference.

Table S5: Candidate versus evaluable universes for every dataset × reference combination. $n = 9$ rows; the analysis unit is one (dataset, reference) pair. "Candidate regulators" applies rules 1–3; **"evaluable regulators"** additionally applies rule 4 and is the denominator of every macro-average in Section S4. "Positive edges" excludes self-loops. The two prevalence columns differ in denominator: evaluation edges (the universe averaged over) versus candidate edges (which includes non-evaluable regulators' pairs). Source: the eligible-universe table in the reproducibility package.

Dataset	Reference	Candidate regulators	**Evaluable regulators**	Eligible targets	Candidate edges	Evaluation edges	Positive edges	Prev. eval. (%)	Prev. cand. (%)
K562-GS	TRRUST	39	**12**	157	6,100	1,873	38	2.029	0.623
K562-GS	DoRothEA A+B	24	**10**	364	8,717	3,631	133	3.663	1.526
K562-GS	CollecTRI	55	**28**	370	20,309	10,336	188	1.819	0.926
K562-Ess	TRRUST	3	**1**	182	545	181	2	1.105	0.367
K562-Ess	DoRothEA A+B	1	**0**	359	359	0	0	0.000	0.000
K562-Ess	CollecTRI	4	**1**	416	1,661	415	4	0.964	0.241
RPE1-Ess	TRRUST	5	**4**	238	1,185	948	20	2.110	1.688
RPE1-Ess	DoRothEA A+B	4	**1**	422	1,686	421	10	2.375	0.593
RPE1-Ess	CollecTRI	9	**6**	523	4,700	3,133	44	1.404	0.936

Three features of Table S5 govern everything that follows. First, the universe is small: even the largest cell (K562-GS against CollecTRI) admits 20,309 candidate edges, a 52-fold reduction from the full graph, and only 10,336 of those are in the evaluation universe. Second, prevalence inside the evaluation universe is 0.96–3.66 %, one to two orders of magnitude above full-graph prevalence, which is what makes per-regulator average precision interpretable at all. Third, the universe collapses on K562-Ess: 3, 1 and 4 candidate regulators against TRRUST, DoRothEA A+B and CollecTRI, of which 1, 0 and 1 are evaluable. The K562-Ess × DoRothEA A+B cell contains zero positive edges and supports no evaluation whatsoever.

The effective sample size is 39 regulators. Table S6 lists every evaluable regulator. There are 63 (dataset, reference, regulator) triples, but only **39 distinct regulator identities**: 24 appear under one reference, 6 under two and 9 under all three. K562-GS contributes 30 distinct regulators, RPE1-Ess 8, and K562-Ess exactly one (ATF5, recurring under TRRUST and CollecTRI). Every inferential statement in this work is bounded by that 39, not by the 63 regulator × reference cells and not by the 894 method rows.

Table S6: Every evaluable regulator with its candidate-target count, its non-self curated positives, and the self-loops excluded by rule 4. $n = 63$ (dataset, reference, regulator) triples over **39 distinct regulators**; the analysis unit is one regulator within one (dataset, reference) pair. These are exactly the units averaged over by macro-AP in Section S4. Counts here are self-loop corrected and therefore differ from earlier versions of this project (e.g. JUN under CollecTRI is 43, not 44). Source: the evaluable-regulator table in the reproducibility package.

Dataset	Reference	Regulator	Candidate targets	Positive targets	Self-loops excl.
K562-Ess	CollecTRI	ATF5	415	4	1

Table S6 continued

Dataset	Reference	Regulator	Candidate targets	Positive targets	Self-loops excl.
K562-Ess	TRRUST	ATF5	181	2	0
K562-GS	CollecTRI	JUN	369	43	1
K562-GS	CollecTRI	EGR1	369	26	1
K562-GS	CollecTRI	SPI1	369	24	1
K562-GS	CollecTRI	DDIT3	369	13	1
K562-GS	CollecTRI	ATF3	369	12	1
K562-GS	CollecTRI	JUND	369	11	1
K562-GS	CollecTRI	KLF6	369	9	0
K562-GS	CollecTRI	EPAS1	369	7	1
K562-GS	CollecTRI	ZBTB17	369	5	1
K562-GS	CollecTRI	DMTF1	370	4	0
K562-GS	CollecTRI	JUNB	369	4	1
K562-GS	CollecTRI	CREM	369	3	1
K562-GS	CollecTRI	DNMT3B	369	3	1
K562-GS	CollecTRI	KLF9	369	3	0
K562-GS	CollecTRI	POU4F1	369	3	1
K562-GS	CollecTRI	RFX1	370	3	0
K562-GS	CollecTRI	BHLHE40	369	2	1
K562-GS	CollecTRI	GABPA	369	2	0
K562-GS	CollecTRI	ZNF410	370	2	0
K562-GS	CollecTRI	BTG2	369	1	0
K562-GS	CollecTRI	ELK4	369	1	0
K562-GS	CollecTRI	ERF	369	1	1
K562-GS	CollecTRI	HOXB3	369	1	0
K562-GS	CollecTRI	IKZF4	370	1	0
K562-GS	CollecTRI	MSX1	369	1	1
K562-GS	CollecTRI	NCOA2	369	1	1
K562-GS	CollecTRI	THRB	369	1	1
K562-GS	CollecTRI	ZFHX3	369	1	1
K562-GS	DoRothEA A+B	EGR1	363	58	0
K562-GS	DoRothEA A+B	SPI1	363	50	0
K562-GS	DoRothEA A+B	JUN	363	11	0
K562-GS	DoRothEA A+B	EPAS1	363	4	0
K562-GS	DoRothEA A+B	ATF3	363	3	0
K562-GS	DoRothEA A+B	JUND	363	3	0
K562-GS	DoRothEA A+B	DDIT3	363	1	0
K562-GS	DoRothEA A+B	JUNB	363	1	0
K562-GS	DoRothEA A+B	KLF6	363	1	0
K562-GS	DoRothEA A+B	ZBTB33	364	1	0
K562-GS	TRRUST	JUN	156	9	1
K562-GS	TRRUST	EGR1	156	7	0
K562-GS	TRRUST	JUND	156	4	0
K562-GS	TRRUST	ATF3	156	3	0
K562-GS	TRRUST	KLF6	156	3	0
K562-GS	TRRUST	NFKBIA	156	3	0

Table S6 continued

Dataset	Reference	Regulator	Candidate targets	Positive targets	Self-loops excl.
K562-GS	TRRUST	SPI1	156	3	0
K562-GS	TRRUST	DDIT3	156	2	0
K562-GS	TRRUST	GABPA	156	1	0
K562-GS	TRRUST	JUNB	156	1	0
K562-GS	TRRUST	KLF9	156	1	0
K562-GS	TRRUST	RFX1	157	1	0
RPE1-Ess	CollecTRI	ATF4	522	25	1
RPE1-Ess	CollecTRI	HMGA1	522	9	1
RPE1-Ess	CollecTRI	ENO1	522	4	0
RPE1-Ess	CollecTRI	YBX3	522	4	1
RPE1-Ess	CollecTRI	HMBOX1	523	1	0
RPE1-Ess	CollecTRI	HMGB1	522	1	0
RPE1-Ess	DoRothEA A+B	ATF4	421	10	0
RPE1-Ess	TRRUST	ATF4	237	12	0
RPE1-Ess	TRRUST	HMGA1	237	4	0
RPE1-Ess	TRRUST	ELL	237	3	0
RPE1-Ess	TRRUST	PTMA	237	1	0

Table S7: Symbols of all eligible regulators (rules 1–3) for each dataset × reference pair. $n = 9$ rows; the analysis unit is one (dataset, reference) pair. Regulators listed here but absent from Table S6 pass rules 1–3 but fail the evaluability rule.

Dataset	Reference	Eligible regulator symbols
K562-GS	TRRUST	ATF3, BTG2, CREM, DDIT3, DENND4A, EGR1, ELK4, EPAS1, ERF, GABPA, GZF1, HDAC10, HOXC8, JUN, JUNB, JUND, KAT2B, KLF6, KLF9, MSX1, MXD1, MXD4, NCOA2, NFKBIA, POU4F1, PRDM2, RFX1, SERTAD1, SKI, SPI1, TBC1D22A, THRB, TLE3, TSC22D1, ZBTB17, ZFHX3, ZFP36L1, ZNF175, ZNF410
K562-GS	DoRothEA A+B	ATF3, BHLHE40, CREM, DDIT3, EGR1, ELK4, EPAS1, ERF, GABPA, HOXC8, JUN, JUNB, JUND, KLF6, MNT, MXD1, MXD4, RFX1, SKI, SPI1, THRB, ZBTB17, ZBTB33, ZFHX3
K562-GS	CollecTRI	AEBP1, ARID3B, ARID5A, ARID5B, ATF3, BHLHE40, BTG2, CREM, DDIT3, DMTF1, DNMT3B, EGR1, ELK4, EPAS1, ERF, FOXP4, GABPA, GZF1, HIVEP3, HOXB3, HOXC8, IKZF4, JUN, JUNB, JUND, KAT2B, KLF6, KLF9, MNT, MSX1, MXD1, MXD4, NCOA2, NFE2L3, PHF1, POU4F1, PRDM2, RBAK, RFX1, SKI, SOX18, SPI1, THRB, TLE3, TSC22D1, ZBTB17, ZBTB33, ZBTB38, ZFHX3, ZHX3, ZNF175, ZNF410, ZNF711, ZNF8, ZSCAN9
K562-Ess	TRRUST	ATF5, DRAP1, NCOA4

Dataset	Reference	Eligible regulator symbols
K562-Ess	DoRothEA A+B	ATF5
K562-Ess	CollecTRI	ATF5, DRAP1, HMGB1, NCOA4
RPE1-Ess	TRRUST	ATF4, ELL, ENO1, HMGA1, PTMA
RPE1-Ess	DoRothEA A+B	ATF4, HMBOX1, HMGA1, THAP1
RPE1-Ess	CollecTRI	ARID5B, ATF4, CCND1, ENO1, HMBOX1, HMGA1, HMGB1, THAP1, YBX3

S4 Regulator-centric benchmark

For each eligible regulator r with n_r curated targets, all candidate targets are ranked by the method's score for r and scored against the n_r positives, giving a per-regulator average precision AP_r, precision and recall at $k \in \{10, 50\}$, and reciprocal rank. Macro-averages weight every regulator equally regardless of target count. The random baseline is an explicit permuted-score run carried through the identical pipeline, not an analytic approximation, and is reported as its own method row; the "ratio" column divides each method's macro-AP by the random macro-AP for the same (dataset, reference) pair.

Table S8 reports the full grid. Per-regulator results underlying these macro-averages are given in the per-regulator results table of the reproducibility package (957 rows: 957 dataset × reference × method × regulator combinations); a summary of that table appears in Table S9.

Table S8: Regulator-centric benchmark: macro-averaged metrics for every method, dataset and reference. $n = 87$ rows; the analysis unit is one (dataset, reference, method) combination, each macro-averaged over the N_{reg} eligible regulators with ≥1 positive target. "Ratio" is macro-AP divided by the random-baseline macro-AP for the same dataset and reference. "GRNBoost2*" is the label whose implementation could not be identified and which is excluded from pooled estimates, and "Glasso-PartialCorr (PC/GES*)" is Glasso-PartialCorr-OrientedByIntervention (PC/GES, mislabeled in earlier versions); see Section S1.

Method	N_{reg}	macro-AP	Ratio	P@10	P@50	R@10	R@50	MRR
K562-Ess × CollecTRI								
Abl. vanilla AE	1	0.0239	0.97×	0.0000	0.0000	0.0000	0.0000	0.0164
Abl. latent=32	1	0.0239	0.97×	0.0000	0.0000	0.0000	0.0000	0.0164
Abl. latent=8	1	0.0239	0.97×	0.0000	0.0000	0.0000	0.0000	0.0164
Abl. no labels	1	0.0239	0.97×	0.0000	0.0000	0.0000	0.0000	0.0164
Abl. shuffled labels	1	0.0239	0.97×	0.0000	0.0000	0.0000	0.0000	0.0164
Abl. stability-filt.	1	0.0239	0.97×	0.0000	0.0000	0.0000	0.0000	0.0164
Abl. diffusion	1	0.0239	0.97×	0.0000	0.0000	0.0000	0.0000	0.0164
Abl. PCA	1	0.0239	0.97×	0.0000	0.0000	0.0000	0.0000	0.0164
Abl. PHATE	1	0.0239	0.97×	0.0000	0.0000	0.0000	0.0000	0.0164
CausalManifold	1	0.0089	0.36×	0.0000	0.0000	0.0000	0.0000	0.0082
GENIE3 (reduced cap.)	1	0.0120	0.49×	0.0000	0.0000	0.0000	0.0000	0.0154
Random	1	0.0247	1.00×	0.0100	0.0120	0.0250	0.1500	0.0438
K562-Ess × TRRUST								
Abl. vanilla AE	1	0.0243	0.68×	0.0000	0.0200	0.0000	0.5000	0.0222
Abl. latent=32	1	0.0243	0.68×	0.0000	0.0200	0.0000	0.5000	0.0222
Abl. latent=8	1	0.0243	0.68×	0.0000	0.0200	0.0000	0.5000	0.0222
Abl. no labels	1	0.0243	0.68×	0.0000	0.0200	0.0000	0.5000	0.0222
Abl. shuffled labels	1	0.0243	0.68×	0.0000	0.0200	0.0000	0.5000	0.0222
Abl. stability-filt.	1	0.0243	0.68×	0.0000	0.0200	0.0000	0.5000	0.0222
Abl. diffusion	1	0.0243	0.68×	0.0000	0.0200	0.0000	0.5000	0.0222
Abl. PCA	1	0.0243	0.68×	0.0000	0.0200	0.0000	0.5000	0.0222
Abl. PHATE	1	0.0243	0.68×	0.0000	0.0200	0.0000	0.5000	0.0222
CausalManifold	1	0.0103	0.29×	0.0000	0.0000	0.0000	0.0000	0.0077
GENIE3 (reduced cap.)	1	0.0135	0.38×	0.0000	0.0000	0.0000	0.0000	0.0096
Random	1	0.0357	1.00×	0.0080	0.0108	0.0400	0.2700	0.0525
K562-GS × CollecTRI								
Abl. vanilla AE	28	0.0374	1.11×	0.0393	0.0193	0.0289	0.0747	0.1428
Abl. latent=32	28	0.0560	1.66×	0.0714	0.0271	0.0932	0.1760	0.1750
Abl. latent=8	28	0.0485	1.44×	0.0536	0.0307	0.0316	0.1483	0.1683
Abl. no labels	28	0.0450	1.34×	0.0464	0.0293	0.0419	0.2085	0.1513
Abl. shuffled labels	28	0.0470	1.40×	0.0500	0.0286	0.0370	0.1201	0.1068
Abl. stability-filt.	28	0.0513	1.53×	0.0679	0.0329	0.0794	0.2123	0.1332
Abl. diffusion	28	0.0595	1.77×	0.0821	0.0379	0.0770	0.2615	0.1618
Abl. PCA	28	0.0374	1.11×	0.0286	0.0236	0.0227	0.1252	0.1243

Table S8 continued

Method	N_{reg}	macro-AP	Ratio	P@10	P@50	R@10	R@50	MRR
Abl. PHATE	28	0.0595	1.77×	0.0821	0.0379	0.0770	0.2615	0.1618
Vanilla AE	28	0.0255	0.76×	0.0214	0.0186	0.0237	0.0783	0.0470
CausalManifold	28	0.0538	1.60×	0.0750	0.0350	0.0579	0.1827	0.1368
Correlation	28	0.0618	1.84×	0.0893	0.0364	0.0684	0.2295	0.1622
GENIE3 (reduced cap.)	28	0.0527	1.57×	0.0679	0.0371	0.0487	0.1686	0.1144
Graphical Lasso	28	0.0494	1.47×	0.0643	0.0300	0.0306	0.1076	0.1446
GRNBoost2*	28	0.0479	1.43×	0.0500	0.0329	0.0352	0.1938	0.0993
NOTEARS-Linear	28	0.0493	1.47×	0.0643	0.0321	0.0459	0.1826	0.1247
Glasso-PartialCorr (PC/GES*)	28	0.0501	1.49×	0.0679	0.0300	0.0334	0.1076	0.1416
Random	28	0.0336	1.00×	0.0209	0.0183	0.0319	0.1434	0.0688
K562-GS × DoRothEA A+B								
Abl. vanilla AE	10	0.1082	2.28×	0.0800	0.0460	0.1359	0.1918	0.1859
Abl. latent=32	10	0.0786	1.66×	0.0700	0.0540	0.0829	0.3675	0.1691
Abl. latent=8	10	0.1117	2.36×	0.0700	0.0540	0.1496	0.3082	0.2553
Abl. no labels	10	0.1556	3.28×	0.0500	0.0580	0.1072	0.2901	0.2413
Abl. shuffled labels	10	0.0838	1.77×	0.0900	0.0540	0.0940	0.2429	0.2660
Abl. stability-filt.	10	0.0819	1.73×	0.0800	0.0580	0.0847	0.4102	0.1817
Abl. diffusion	10	0.1146	2.42×	0.1000	0.0600	0.1937	0.4267	0.2134
Abl. PCA	10	0.1032	2.18×	0.0700	0.0460	0.1109	0.2237	0.1784
Abl. PHATE	10	0.1146	2.42×	0.1000	0.0600	0.1937	0.4267	0.2134
Vanilla AE	10	0.0411	0.87×	0.0300	0.0380	0.0054	0.0813	0.1238
CausalManifold	10	0.0981	2.07×	0.0900	0.0580	0.1847	0.2993	0.2203
Correlation	10	0.0905	1.91×	0.1300	0.0640	0.1088	0.3548	0.2614
GENIE3 (reduced cap.)	10	0.1064	2.24×	0.1300	0.0740	0.1402	0.4257	0.3531
Graphical Lasso	10	0.0753	1.59×	0.1100	0.0520	0.0422	0.1830	0.2470
GRNBoost2*	10	0.0831	1.75×	0.0900	0.0680	0.0943	0.1930	0.2316
NOTEARS-Linear	10	0.0887	1.87×	0.1200	0.0640	0.1493	0.4531	0.2563
Glasso-PartialCorr (PC/GES*)	10	0.0758	1.60×	0.1200	0.0520	0.0513	0.1830	0.2477
Random	10	0.0474	1.00×	0.0388	0.0359	0.0286	0.1278	0.0863
K562-GS × TRRUST								
Abl. vanilla AE	12	0.1232	2.58×	0.0333	0.0283	0.1508	0.5033	0.2588
Abl. latent=32	12	0.0878	1.84×	0.0500	0.0267	0.1720	0.5122	0.1397
Abl. latent=8	12	0.0937	1.96×	0.0583	0.0283	0.1901	0.4917	0.1881
Abl. no labels	12	0.0750	1.57×	0.0417	0.0283	0.1045	0.4203	0.1521
Abl. shuffled labels	12	0.0758	1.59×	0.0500	0.0283	0.1227	0.3922	0.1291
Abl. stability-filt.	12	0.0761	1.59×	0.0583	0.0333	0.2156	0.6445	0.1408
Abl. diffusion	12	0.1047	2.19×	0.0917	0.0383	0.2669	0.7120	0.1662
Abl. PCA	12	0.1159	2.43×	0.0417	0.0250	0.1601	0.5192	0.2109
Abl. PHATE	12	0.1047	2.19×	0.0917	0.0383	0.2669	0.7120	0.1662
Vanilla AE	12	0.0339	0.71×	0.0167	0.0167	0.0556	0.2249	0.0578
CausalManifold	12	0.0692	1.45×	0.0583	0.0183	0.1415	0.2827	0.1945
Correlation	12	0.1117	2.34×	0.0750	0.0400	0.1786	0.6147	0.2001
GENIE3 (reduced cap.)	12	0.1077	2.26×	0.0667	0.0350	0.1438	0.4501	0.2465
Graphical Lasso	12	0.0952	1.99×	0.0667	0.0300	0.1647	0.3853	0.1828
GRNBoost2*	12	0.0736	1.54×	0.0500	0.0250	0.1068	0.3965	0.1873
NOTEARS-Linear	12	0.0968	2.03×	0.0667	0.0367	0.1647	0.5959	0.1642
Glasso-PartialCorr (PC/GES*)	12	0.0946	1.98×	0.0667	0.0300	0.1647	0.3853	0.1828
Random	12	0.0477	1.00×	0.0195	0.0198	0.0632	0.3110	0.0745
RPE1-Ess × CollecTRI								
CausalManifold	6	0.0310	1.47×	0.0500	0.0233	0.0200	0.0935	0.0736
GENIE3 (reduced cap.)	6	0.0439	2.08×	0.0500	0.0300	0.0319	0.1306	0.2294
Random	6	0.0211	1.00×	0.0110	0.0135	0.0107	0.0961	0.0396
RPE1-Ess × DoRothEA A+B								
CausalManifold	1	0.1357	3.90×	0.2000	0.0800	0.2000	0.4000	0.3333
GENIE3 (reduced cap.)	1	0.2592	7.44×	0.2000	0.1200	0.2000	0.6000	1.0000
Random	1	0.0348	1.00×	0.0260	0.0184	0.0260	0.0920	0.0721

Table S8 continued

Method	N_{reg}	macro-AP	Ratio	P@10	P@50	R@10	R@50	MRR
RPE1-Ess × TRRUST								
CausalManifold	4	0.0844	2.01×	0.0750	0.0550	0.0625	0.4375	0.1146
GENIE3 (reduced cap.)	4	0.1111	2.65×	0.1250	0.0400	0.1458	0.2500	0.3780
Random	4	0.0419	1.00×	0.0250	0.0237	0.0371	0.2146	0.0860

Table S9: Summary of the per-regulator results for the seven distinct methods (eight labels; "GRNBoost2" is excluded, provenance unresolved) and the random baseline. $n = 489$ regulator-level rows in total across the nine listed method rows; the analysis unit is one (dataset, reference, method, regulator) evaluation. "Rows" is the number of such evaluations per method, which differs because only CausalManifold and the GENIE3 proxy were run on all three datasets.

Method	Rows	Mean AP	Median AP	Max AP	Frac. P@10>0	Median MRR
Correlation	50	0.0795	0.0346	0.4431	0.360	0.0412
GENIE3 (reduced cap.)	63	0.0766	0.0246	0.3916	0.365	0.0303
NOTEARS-Linear	50	0.0686	0.0335	0.3638	0.340	0.0250
Glasso-PartialCorr (PC/GES*)	50	0.0659	0.0149	0.4601	0.280	0.0134
Graphical Lasso	50	0.0656	0.0149	0.4601	0.280	0.0134
CausalManifold	63	0.0634	0.0189	0.3130	0.333	0.0244
GRNBoost2*	50	0.0611	0.0212	0.3233	0.340	0.0243
Random	63	0.0377	0.0283	0.1721	0.905	0.0438
Vanilla AE	50	0.0307	0.0123	0.1818	0.200	0.0096

Reading of Table S8: these are descriptive point estimates. Table S8 carries no confidence intervals and no inferential content. It reports, per (dataset, reference, method) cell, a macro-average over between 1 and 28 regulators. Every ranking statement that can be made from it is descriptive; the inferential question — whether the departure from random is resolvable at all — is answered in Section S9, and the answer is no.

With that caveat: on K562-GS the primary-method ratios lie between 0.71× (vanilla AE against TRRUST) and 2.34× (Correlation against TRRUST), and the vanilla-AE effect graph is below random on all three K562-GS references (0.71×, 0.87×, 0.76×). **It is not true that methods outperform random in every dataset × reference combination**: every K562-Ess row is at or below random, and the two K562-Ess cells that admit any evaluation give ratios of 0.29–0.97× across methods.

Two rows that must not be read as headline results. The largest single ratio anywhere in Table S8 is 7.44×, for GENIE3 (reduced cap.) on RPE1-Ess against DoRothEA A+B. **That cell averages over one regulator** (ATF4, 421 candidate targets, 10 curated positives). A macro-average over $n = 1$ has no sampling distribution and admits no interval; recomputed against 2,000 matched random draws rather than the 50-trial baseline used to produce this table, the ratio is 7.06× (one-sided permutation $p = 0.0005$ for ATF4 itself). It is a single-regulator anecdote and the maximum of 63 cells, and it must not be quoted as a benchmark result.

The second is the largest *macro-AP* in the table, 0.1556, which belongs to **Abl. no labels** — a CausalManifold ablation trained without perturbation labels — on K562-GS against DoRothEA A+B, again over 10 regulators. An ablation row is not a primary-method result and is never reported as "the best macro-AP" in this work. The best *primary-method* macro-AP on K562-GS is 0.1117

(Correlation against TRRUST, 12 regulators). That an unsupervised ablation tops the table is itself informative and is taken up in Section S6.

K562-Ess is underpowered. Table S5 shows that K562-Ess admits a single eligible regulator with positive targets against TRRUST (2 positives) and a single one against CollecTRI (5 positives), and none at all against DoRothEA A+B. Every K562-Ess row in Table S8 therefore has $N_{\text{reg}} = 1$: the macro-average is one regulator's AP, with no between-regulator variance to report and no meaningful comparison against the random baseline. The K562-Ess rows are printed for completeness and should not be used to rank methods. All ranking statements in this supplement are based on K562-GS, where N_{reg} is 12, 10 and 28 for TRRUST, DoRothEA A+B and CollecTRI respectively.

S5 Complete top-k tensor and universal-claim audit

The full-graph evaluation ranks all off-diagonal entries of a method's score matrix and asks how many of the top k appear in the reference edge set. We computed this for every method that produced a score matrix, on every dataset and reference, at $k \in \{10, 25, 50, 100, 250, 500, 1{,}000, 5{,}000\}$: 17 methods $\times$ 3 datasets $\times$ 3 references $\times$ 8 values of k, restricted to the 30 method–dataset combinations that were actually run, giving 720 tensor entries. The candidate space is $1{,}024 \times 1{,}023 = 1{,}047{,}552$ directed edges for all methods except the nine ablation models on K562-Ess, which were trained on the observational-only split with a 2,000-gene adjacency and therefore rank 3,998,000 edges.

Table S10: Full-graph precision@k summarised across all methods, datasets and references. $n = 720$ tensor entries in total, 90 per value of k; the analysis unit is one (method, dataset, reference, k) evaluation.

k	Entries	Mean precision@k	s.d.	Min	Max
10	90	0.00000	0.00000	0.0000	0.0000
25	90	0.00089	0.00593	0.0000	0.0400
50	90	0.00200	0.00737	0.0000	0.0400
100	90	0.00222	0.00536	0.0000	0.0200
250	90	0.00302	0.00679	0.0000	0.0320
500	90	0.00260	0.00517	0.0000	0.0260
1,000	90	0.00199	0.00374	0.0000	0.0180
5,000	90	0.00093	0.00121	0.0000	0.0066

Table S11: Full-graph precision@k summarised by reference database. $n = 720$ tensor entries; the analysis unit is one (method, dataset, reference, k) evaluation, with 30 entries per (reference, k) cell.

Reference	k	Mean	s.d.	Min	Max
CollecTRI	10	0.00000	0.00000	0.0000	0.0000
CollecTRI	25	0.00000	0.00000	0.0000	0.0000
CollecTRI	50	0.00133	0.00507	0.0000	0.0200
CollecTRI	100	0.00300	0.00651	0.0000	0.0200
CollecTRI	250	0.00520	0.01003	0.0000	0.0320
CollecTRI	500	0.00447	0.00778	0.0000	0.0260
CollecTRI	1,000	0.00333	0.00552	0.0000	0.0180
CollecTRI	5,000	0.00159	0.00168	0.0004	0.0066
DoRothEA A+B	10	0.00000	0.00000	0.0000	0.0000
DoRothEA A+B	25	0.00267	0.01015	0.0000	0.0400
DoRothEA A+B	50	0.00333	0.01061	0.0000	0.0400
DoRothEA A+B	100	0.00267	0.00583	0.0000	0.0200
DoRothEA A+B	250	0.00227	0.00466	0.0000	0.0120
DoRothEA A+B	500	0.00193	0.00313	0.0000	0.0100
DoRothEA A+B	1,000	0.00160	0.00253	0.0000	0.0080
DoRothEA A+B	5,000	0.00074	0.00074	0.0000	0.0026
TRRUST	10	0.00000	0.00000	0.0000	0.0000
TRRUST	25	0.00000	0.00000	0.0000	0.0000
TRRUST	50	0.00133	0.00507	0.0000	0.0200
TRRUST	100	0.00100	0.00305	0.0000	0.0100
TRRUST	250	0.00160	0.00342	0.0000	0.0120
TRRUST	500	0.00140	0.00247	0.0000	0.0080
TRRUST	1,000	0.00103	0.00175	0.0000	0.0050
TRRUST	5,000	0.00045	0.00063	0.0000	0.0022

S5.1 Universal claim 1: no reference edge appears in any method's top 10

All 90 precision@10 entries in the tensor are exactly zero, and the total number of reference edges recovered in the top 10 is 0 across every method, dataset and reference. Table S12 gives the complete audit: each cell is the number of reference edges among a method's 10 highest-scoring edges, summed over the three reference databases (maximum possible 30 per cell). Empty cells mark method–dataset combinations that were not run.

Table S12: Complete $k = 10$ audit. Entries are reference edges recovered in the top 10 predictions, summed over the three reference databases. $n = 90$ (method, dataset, reference) evaluations at $k = 10$; the analysis unit is one method's top-10 edge list on one dataset. Every entry is zero: 90 of 90 evaluations recovered no reference edge.

Method	K562-GS	K562-Ess	RPE1-Ess
Abl. vanilla AE	0	0	–
Abl. latent=32	0	0	–
Abl. latent=8	0	0	–
Abl. no labels	0	0	–
Abl. shuffled labels	0	0	–
Abl. stability-filt.	0	0	–
Abl. diffusion	0	0	–
Abl. PCA	0	0	–
Abl. PHATE	0	0	–
Vanilla AE	0	–	–
CausalManifold	0	0	0
Correlation	0	–	–
GENIE3 (reduced cap.)	0	0	0
Graphical Lasso	0	–	–
GRNBoost2*	0	–	–
NOTEARS-Linear	0	–	–
Glasso-PartialCorr (PC/GES*)	0	–	–

S5.2 Universal claim 2: precision@100 never exceeds 0.02

The maximum precision@100 anywhere in the tensor is 0.02, i.e. 2 reference edges among 100 predictions. Table S13 gives the complete $k = 100$ audit for all 90 evaluations. The maximum is attained in 5 cells, by the following methods: Glasso-PartialCorr (PC/GES*) (K562-GS, CollecTRI); Graphical Lasso (K562-GS, CollecTRI); GENIE3 (reduced cap.) (K562-Ess, CollecTRI); GENIE3 (reduced cap.) (K562-GS, DoRothEA A+B); GRNBoost2* (K562-GS, DoRothEA A+B). Two further points constrain the reading of this table. First — and this corrects a claim made in earlier versions of this project — the low values are **not** explained by an attainable-precision ceiling. It is a logical error to argue that "even a perfect method cannot achieve meaningful precision when the search space exceeds the positive set by four orders of magnitude." A perfect ranking places all P positives first, so precision@$k = 1$ for every $k \leq P$ regardless of how large the candidate space is; with P between 38 and 188 here, precision@100 = 1.0 is attainable in seven of the nine cells and precision@10 = 1.0 in all of them. The bound $E_{\text{ref}}/E_{\text{inf}}$ binds only when *all* inferred edges are reported, and $\min(1, E_{\text{ref}}/k)$ binds only for $k > E_{\text{ref}}$. What the full graph actually does is *dilute*: it drops the positive rate, and hence the random-baseline precision, below 0.006 %, so an observed 0.01 is uninterpretable until placed against a matched random baseline for the same space. Prevalence dilution is the entire argument; no ceiling argument is made anywhere in this work. Second, a value of 0.02 corresponds to exactly two recovered edges, so differences of 0.01 between cells are single-edge differences and carry no useful discrimination.

Table S13: Complete $k = 100$ audit: precision@100 for every method, dataset and reference. $n = 90$ evaluations; the analysis unit is one method's top-100 edge list against one reference on one dataset. Columns are grouped by dataset (T = TRRUST, D = DoRothEA A+B, C = CollecTRI). "–" marks combinations not run.

	K562-GS			K562-Ess			RPE1-Ess		
Method	T	D	C	T	D	C	T	D	C
Abl. vanilla AE	0.00	0.00	0.00	0.00	0.00	0.00	–	–	–
Abl. latent=32	0.00	0.00	0.00	0.00	0.00	0.00	–	–	–
Abl. latent=8	0.00	0.00	0.00	0.00	0.00	0.00	–	–	–
Abl. no labels	0.00	0.00	0.00	0.00	0.00	0.00	–	–	–
Abl. shuffled labels	0.00	0.00	0.00	0.00	0.00	0.00	–	–	–
Abl. stability-filt.	0.00	0.00	0.00	0.00	0.00	0.00	–	–	–
Abl. diffusion	0.00	0.00	0.00	0.00	0.00	0.00	–	–	–
Abl. PCA	0.00	0.00	0.00	0.00	0.00	0.00	–	–	–
Abl. PHATE	0.00	0.00	0.00	0.00	0.00	0.00	–	–	–
Vanilla AE	0.00	0.00	0.00	–	–	–	–	–	–
CausalManifold	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Correlation	0.00	0.01	0.00	–	–	–	–	–	–
GENIE3 (reduced cap.)	0.01	0.02	0.01	0.01	0.01	0.02	0.00	0.00	0.00
Graphical Lasso	0.00	0.01	0.02	–	–	–	–	–	–
GRNBoost2*	0.01	0.02	0.01	–	–	–	–	–	–
NOTEARS-Linear	0.00	0.00	0.01	–	–	–	–	–	–
Glasso-PartialCorr (PC/GES*)	0.00	0.01	0.02	–	–	–	–	–	–

What these two audits do and do not establish. They establish that no method's very highest-confidence full-graph predictions are curated regulatory edges, and that this holds uniformly rather than for a subset of methods. They do *not* establish that methods recover no regulatory signal: the regulator-centric evaluation of Section S4 finds macro-AP above the random baseline for essentially every primary method on K562-GS, and the direct-vs-indirect analysis of Section S6 finds a consistent score gradient by hop distance. The full-graph tensor measures a different and much harsher quantity, in which the evaluable positives are diluted by a factor of 52–172 (Table S5).

S6 Direct-versus-indirect target analysis

For each eligible regulator we partitioned the measured genes by shortest path length in the reference network: *direct* (a curated 1-hop target), *2-hop* (reachable through one intermediate regulator), *3⁺-hop*, and *unrelated* (no path). We then compared the score each method assigns to each class. This tests a weaker and more forgiving property than edge recovery: not whether the top-ranked edges are curated, but whether curated direct targets are scored systematically above unrelated genes. Across the 47 regulators and 17 methods in the analysis there are 1650 rows, covering 690 direct, 4,049 2-hop, 53,936 3⁺-hop and 635,942 unrelated regulator–target assignments.

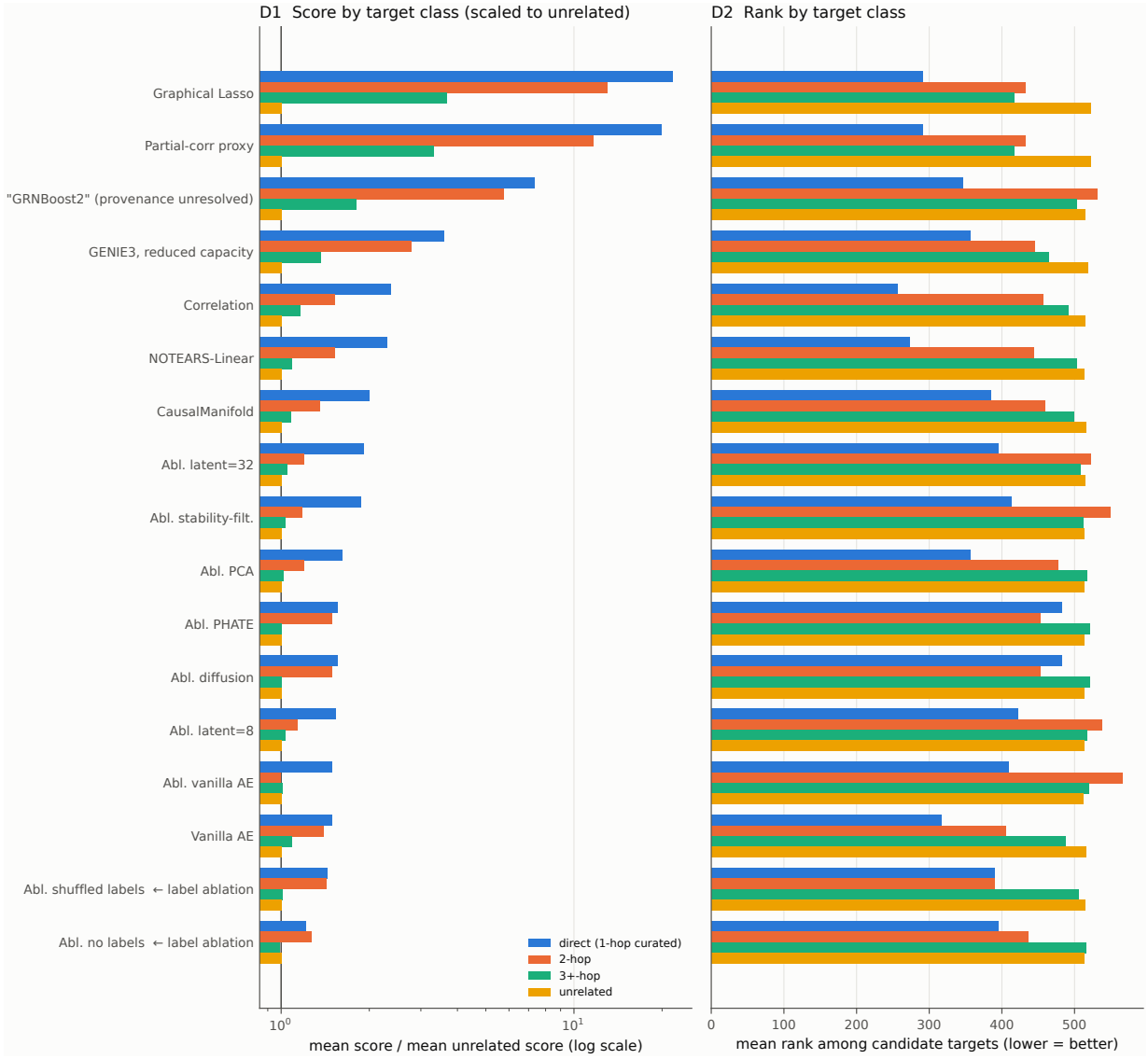

Figure S1: Score (left, scaled to each method's own unrelated-class mean, log axis) and mean rank (right) by target class for all 17 method labels, reproduced from the main text. The two rows marked "label ablation" were trained with shuffled or absent perturbation labels and still separate direct from unrelated targets. Plotted values are tabulated in the reproducibility package.

Table S14: Mean method score by target class. $n = 1650$ (method, regulator, target class) rows aggregated to 17 method rows; the analysis unit is one method, averaged over its regulator-level mean scores. "Ratio" is direct divided by unrelated. Scores are on each method's own scale and are comparable within a row, not across rows.

Method	Direct	2-hop	3^+-hop	Unrelated	Direct:unrelated
Graphical Lasso	0.0078	0.0047	0.0013	0.0004	21.64×
Glasso-PartialCorr (PC/GES*)	0.0037	0.0021	0.0006	0.0002	19.95×
GRNBoost2*	0.0065	0.0051	0.0016	0.0009	7.33×
GENIE3 (reduced cap.)	0.0033	0.0026	0.0013	0.0009	3.61×
Correlation	0.0293	0.0189	0.0144	0.0124	2.37×
NOTEARS-Linear	0.0061	0.0040	0.0029	0.0026	2.29×
CausalManifold	0.4796	0.3259	0.2581	0.2399	2.00×
Abl. latent=32	0.6372	0.3990	0.3505	0.3337	1.91×
Abl. stability-filt.	0.6207	0.3926	0.3429	0.3320	1.87×
Abl. PCA	0.5105	0.3796	0.3228	0.3169	1.61×
Abl. diffusion	0.5187	0.4974	0.3352	0.3334	1.56×
Abl. PHATE	0.5187	0.4974	0.3352	0.3334	1.56×
Abl. latent=8	0.5032	0.3737	0.3381	0.3284	1.53×
Abl. vanilla AE	0.4710	0.3149	0.3184	0.3162	1.49×
Vanilla AE	0.3454	0.3247	0.2532	0.2325	1.49×
Abl. shuffled labels	0.4548	0.4519	0.3208	0.3175	1.43×
Abl. no labels	0.3927	0.4101	0.3180	0.3226	1.22×

Table S15: Mean of the per-regulator median rank by target class (lower is better; ranks run from 1 to the number of measured genes). $n = 1650$ (method, regulator, target class) rows aggregated to 17 method rows; the analysis unit is one method. "Rank ratio" is unrelated divided by direct, so values above 1 indicate that direct targets rank ahead of unrelated genes. The right-hand column gives the number of regulators contributing to each method.

Method	Direct	2-hop	3^+-hop	Unrelated	Rank ratio	Regulators
Graphical Lasso	231.4	348.5	306.6	528.7	2.28×	39
Glasso-PartialCorr (PC/GES*)	231.2	348.5	306.6	528.7	2.29×	39
GRNBoost2*	295.3	504.7	464.6	517.3	1.75×	39
GENIE3 (reduced cap.)	322.3	415.8	441.3	519.5	1.61×	47
Correlation	215.8	431.5	490.9	515.3	2.39×	39
NOTEARS-Linear	259.1	429.6	505.5	513.2	1.98×	39
CausalManifold	354.8	456.8	492.8	515.5	1.45×	47

Method	Direct	2-hop	3^+-hop	Unrelated	Rank ratio	Regulators
Abl. latent=32	418.4	544.3	512.5	513.6	1.23×	39
Abl. stability-filt.	415.9	531.7	510.8	512.7	1.23×	39
Abl. PCA	285.3	466.6	526.3	512.4	1.80×	39
Abl. diffusion	451.0	417.2	532.9	512.5	1.14×	39
Abl. PHATE	451.0	417.2	532.9	512.5	1.14×	39
Abl. latent=8	436.1	555.6	522.5	512.0	1.17×	39
Abl. vanilla AE	368.5	572.5	525.6	512.5	1.39×	39
Vanilla AE	295.4	360.0	479.8	516.1	1.75×	39
Abl. shuffled labels	381.6	380.8	502.2	514.7	1.35×	39
Abl. no labels	390.9	403.3	511.8	513.3	1.31×	39

Reading of Tables S14 and S15. Every one of the 17 methods assigns a higher mean score to direct targets than to unrelated genes, with direct:unrelated ratios from 1.22× to 21.64×. The two largest ratios belong to Graphical Lasso (21.64×) and the partial-correlation proxy (19.95×), which share a code path and therefore a score matrix; their large ratios reflect very small unrelated-class scores rather than large direct-class scores, so the ratio should be read together with the absolute columns. Rank-based discrimination is more uniform and more modest: direct targets rank ahead of unrelated genes for all 17 methods, with rank ratios between 1.14× and 2.39× under the aggregation used in Table S15 (mean of the per-regulator *median* rank). Aggregating the per-regulator *mean* ranks instead gives 1.06×–2.01×; the main text quotes the mean-rank aggregation, and the two agree on the ordering. The ordering direct < 2-hop < 3^+-hop < unrelated in mean rank is not monotone for every method.

The ablation caveat, which qualifies this entire section. Two of the 17 labels are label-corruption ablations of CausalManifold: one trained with *shuffled* perturbation labels and one trained with *no* labels. Both **retain a direct:unrelated score ratio above 1** — 1.43× and 1.22× respectively — and both retain rank ratios above 1.3× (1.35× and 1.31× under the median-rank aggregation of Table S15; 1.32× and 1.30× under the mean-rank aggregation). Neither model can be using perturbation information, because it was destroyed or withheld before training. The only remaining explanation is that curated direct targets of a transcription factor are *co-expressed* with it, so any method sensitive to co-expression separates them from unrelated genes without recovering anything about intervention.

Consequently the direct-versus-indirect gradient is not evidence that these methods capture regulatory proximity from the perturbations. Part of it — and for the weaker methods, apparently most of it — is expression correlation. A claim of regulatory-proximity recovery would require the gradient to exceed the label-ablation gradient by a margin the benchmark can resolve, which at 12–17 regulators per method it cannot.

Withdrawal of "up to 27× in rank". Earlier versions of this project stated that direct-target scores exceed unrelated-gene scores "by 2–5× in magnitude and up to 27× in rank". **The 27× figure appears in no result file and is withdrawn.** The maximum method-level rank ratio is 2.39× (Correlation) under the aggregation of Table S15 and 2.01× under the mean-rank aggregation. Individual (method, regulator) cells do reach much larger ratios — the largest is 39.5×, GRNBoost2* on KLF9 — but each such cell is one regulator out of 12–17 and is an anecdote, not a method-level

effect. The score-magnitude range is likewise wider than "2–5×": 1.22× to 21.64× across the 17 labels.

S7 PCA back-projection validation

Earlier versions of this project described NOTEARS as operating on PCA components whose adjacency was then back-projected to gene space. The code audit established that no such step exists (Section S1); this section reports both the code trace and the synthetic experiment that quantifies what back-projection would have cost had it been used.

S7.1 Code trace: the pipeline does not back-project

Tracing the pipeline's single-experiment driver gives the following sequence. `fit_embedding('pca', X)` returns a latent matrix (cells × components) together with the loading matrix V (components × genes). `estimate_interventional_effects` then uses the latent matrix *only* to find nearest-neighbour control cells for each perturbed cell; the effect itself is computed in gene space as $\hat{\delta} = \overline{X}_{\text{pert}} - \overline{X}_{\text{ctrl,matched}}$, giving a (perturbations × genes) matrix. `run_graph_method('notears', ...)` is then called on that gene-space effect matrix, and returns a (genes × genes) adjacency. The loading matrix V is returned by the embedding step and never read again. There is consequently no $V^\top W_{\text{comp}} V$ operation anywhere in the pipeline, and the NOTEARS adjacency evaluated in Sections S4 and S5 is natively gene-level.

S7.2 Synthetic sweep: back-projection does not recover gene edges

We nevertheless implemented and tested the back-projection $W_{\text{gene}} = V^\top W_{\text{comp}} V$ on synthetic data with known ground truth, over 120 conditions: gene counts 50, 100, PCA dimensions 16, 32, 64, DAG density 1.5, 3.0, 5.0 edges per node, noise standard deviation 0.1, 0.5, 1.0, and seeds 42, 123, 456, at 2,000 samples per condition. (500-gene conditions were attempted but did not complete, so the sweep covers 50 and 100 genes only.)

Table S16: PCA back-projection, overall performance. $n = 120$ synthetic conditions; the analysis unit is one (genes, PCA dimensions, density, noise, seed) condition. "Default threshold" is the pipeline's $|W| > 0.05$ rule; "best threshold" is the oracle threshold chosen post hoc to maximise F1, i.e. an upper bound that no deployable procedure could achieve.

Quantity	Value
Conditions evaluated	120
Conditions in which NOTEARS found any edge in component space	0
Back-projected F1 at the default threshold (mean)	0.0000
Back-projected F1 at the oracle best threshold (mean ± s.d.)	0.0824 ± 0.0391
Back-projected F1 at the oracle best threshold (max)	0.1722
Back-projected average precision (mean ± s.d.)	0.0456 ± 0.0245
Back-projected average precision (max)	0.1021
Direct (no back-projection) NOTEARS F1 (mean ± s.d.)	0.0515 ± 0.0296
Direct NOTEARS average precision (mean ± s.d.)	0.0551 ± 0.0339
Cumulative variance explained by the PCA (mean)	0.9042

Table S17: PCA back-projection stratified by design factor. $n = 120$ synthetic conditions; the analysis unit is one condition, and each block partitions the same 120 conditions. "BP" columns are back-projected recovery at the oracle threshold; "Direct" columns are NOTEARS run on gene-space data in the same condition, i.e. the arrangement the real pipeline uses.

Level	n	BP F1	BP F1 s.d.	BP AP	Direct F1	Direct AP	Var. expl.
By number of genes							
50	54	0.1095	0.0386	0.0627	0.0561	0.0728	0.9356
100	66	0.0602	0.0219	0.0315	0.0477	0.0407	0.8785
By number of PCA dimensions							
16	54	0.0924	0.0385	0.0509	0.0508	0.0563	0.8554
32	54	0.0824	0.0370	0.0461	0.0508	0.0563	0.9412
64	12	0.0375	0.0115	0.0193	0.0575	0.0443	0.9572
By DAG density (edges per node)							
1.5	45	0.0509	0.0178	0.0258	0.0505	0.0501	0.8262
3.0	39	0.0826	0.0278	0.0454	0.0565	0.0529	0.9232
5.0	36	0.1215	0.0335	0.0705	0.0473	0.0637	0.9812
By noise standard deviation							
0.1	42	0.0870	0.0375	0.0483	0.0769	0.0903	0.9081
0.5	39	0.0818	0.0409	0.0448	0.0267	0.0376	0.9021
1.0	39	0.0781	0.0396	0.0433	0.0488	0.0347	0.9021

Findings. Back-projection fails, and fails in an informative way. At the pipeline's own threshold it recovers nothing at all: mean F1 is exactly 0.0000 in all 120 conditions, because NOTEARS found zero above-threshold edges in component space in 120 of 120 conditions. Even granting an oracle threshold chosen post hoc to maximise F1, mean F1 is 0.0824 ± 0.0391 and mean average precision is 0.0456 ± 0.0245, with a best case of 0.1722. Performance *decreases* as more components are retained (mean oracle F1 0.0924 at 16 dimensions, 0.0824 at 32, 0.0375 at 64) despite cumulative variance explained rising from 0.855 to 0.957, which is the signature of the failure mode: $V^\top W_{\text{comp}} V$ spreads each component-space edge across all gene pairs, and adding components adds mixing paths rather than resolution. The correlation between variance explained and oracle F1 is 0.482 (and 0.510 with average precision), so the information loss is not a matter of retaining too little variance. Noise has almost no effect (oracle F1 0.0870, 0.0818, 0.0781 at noise 0.1, 0.5, 1.0), confirming that the bottleneck is the projection algebra rather than data quality. Running NOTEARS directly on gene-space data in the same conditions gives mean F1 0.0515 and average precision 0.0551: also low in absolute terms, but obtained without the back-projection step and, unlike the back-projected values, achievable at a fixed threshold.

The practical consequence is narrow but worth stating: because the pipeline never back-projects, this failure mode does not affect any NOTEARS number reported elsewhere in this supplement. The experiment rules out a specific alternative explanation for NOTEARS' weak real-data performance rather than explaining it.

S8 Sample-size saturation, corrected

If weak edge recovery were a power problem, precision should improve with cells. We tested this on K562-GS by subsampling the 20,000-cell processed pool at 2,000, 5,000, 10,000, 15,000 and 20,000 cells, 3 independent draws (seeds 42–44) per size, scoring precision@100 against TRRUST under two rules: correlation computed on per-perturbation effect profiles (which uses the interventional structure), and correlation computed directly on the expression matrix (which ignores perturbation labels).

S8.1 Retraction of `perturbations_covered`

The previous version of this experiment reported a column named `perturbations_covered` with the values 2,000 / 5,000 / 7,660 / 7,660 / 7,660 at the five sizes. **That column is retracted.** It is not a coverage statistic. The audit trace is given as the sample-size count-trace note in the reproducibility package; the mechanism is as follows.

1. **The subsampler did not stratify.** The function named `stratified_subsample()` walked the perturbation groups in `np.unique()` (lexicographic) order taking *exactly one* cell from each until the quota was filled, then filled any remainder by a uniform random draw — no proportional allocation, despite the comment claiming one. The K562-GS pool has 7,661 distinct perturbation labels across 20,000 cells, so for a requested n below 7,661 every selected cell carried a different label and the column returned exactly n; for $n \geq 7{,}661$ the sweep completed and the column returned the pool constant 7,660. The column could only ever emit those two things, and it carried no information about the subsample beyond its size. Re-running the original function verbatim at seed 42 reproduces the published values exactly.

2. **The same defect emptied the control arm.** `'non-targeting'` is lowercase and therefore sorts after every uppercase gene symbol — it is index 7,660 of 7,661, the last group. At $n = 2{,}000$ and $n = 5{,}000$ the phase-one sweep never reached it, so those subsamples contained **zero control cells**. The perturbation-effect pipeline guards on $n_{\text{ctrl}} < 5$ and silently falls back to raw gene–gene correlation. The rows labelled `correlation_perturbation_effects` at 2,000 and 5,000 cells in the previous table therefore *did not use perturbation effects at all*: they are byte-identical duplicates of the raw-correlation rows, which is why the two method blocks agreed perfectly at small n. Their companion columns `perturbation_effects_used` $= 0$ and `median_cells_per_perturbation` $= 0.0$ were the fallback's placeholders, not measurements.

3. **7,660 is nevertheless a real number.** It is the count of distinct non-control perturbation *target gene symbols* in the full 20,000-cell pool (against 8,141 distinct non-control guide constructs, 7,651 distinct Ensembl gene ids, 755 control cells and 19,245 perturbed cells). The premise that motivated the original audit — that this dataset contains only ∼1,100 perturbations — is itself wrong: the K562 genome-scale arm of Replogle et al. [2022] is a near-genome-wide screen. The other processed arms carry 2,014 (K562-Ess) and 2,276 (RPE1-Ess) distinct labels. The column is invalid for the reason in item 1, not because 7,660 is implausible.

S8.2 The corrected experiment

The corrected sample-size analysis code replaces the subsampler with genuine largest-remainder proportional allocation: group g of size c_g receives $\lfloor nc_g/N \rfloor$ cells and the leftover quota goes to the largest fractional parts with a seeded tie-break. Controls are one group among the rest, so the

pool's ∼3.8 % control share is preserved at every n. Scoring (`precision_at_k`, both correlation pipelines, tie ordering) is byte-for-byte unchanged, so the precision numbers remain comparable to the retracted table. No single number stands in for "coverage"; the full set of quantities is recorded per run.

Table S18: Corrected sample-size sweep on K562-GS, means over 3 seeds. $n = 30$ runs (2 scoring rules × 5 cell counts × 3 seeds); the analysis unit is one subsampling run. Every column replaces some part of what the retracted `perturbations_covered` column was taken to mean. Control cells are now non-zero at every size, so the perturbation-effect rule is genuinely exercised at 2,000 and 5,000 cells for the first time. Source: the corrected sample-size results table in the reproducibility package.

Scoring rule	Cells	Seeds	Ctrl cells	Targets	Guides	Med. c/t	≥2 cells	≥5 cells	≥10 cells	P@100 mean	s.d.
Correlation on perturbation effects	2,000	3	76	1,916	1,917	1.0	7	0	0	0.0200	0.0100
Correlation on perturbation effects	5,000	3	189	4,411	4,475	1.0	359	2	0	0.0100	0.0000
Correlation on perturbation effects	10,000	3	378	6,404	6,634	1.0	2,330	50	2	0.0133	0.0058
Correlation on perturbation effects	15,000	3	566	7,660	8,031	1.0	3,772	281	7	0.0100	0.0000
Correlation on perturbation effects	20,000	3	755	7,660	8,141	2.0	5,137	846	37	0.0100	0.0000
Raw expression correlation	2,000	3	76	1,916	1,917	1.0	7	0	0	0.0233	0.0153
Raw expression correlation	5,000	3	189	4,411	4,475	1.0	359	2	0	0.0100	0.0000
Raw expression correlation	10,000	3	378	6,404	6,634	1.0	2,330	50	2	0.0133	0.0058
Raw expression correlation	15,000	3	566	7,660	8,031	1.0	3,772	281	7	0.0100	0.0000
Raw expression correlation	20,000	3	755	7,660	8,141	2.0	5,137	846	37	0.0100	0.0000

Table S19: Per-seed corrected sample-size results. $n = 30$ runs; the analysis unit is one (scoring rule, cell count, seed) run. "Max c/t" is the largest number of cells any single target receives in that draw.

Scoring rule	Cells	Seed	Ctrl	Targets	Guides	Med. c/t	Max c/t	P@100	Runtime (s)
Correlation on perturbation effects	2,000	42	76	1,916	1,916	1.0	3	0.01	1.63
Correlation on perturbation effects	2,000	43	76	1,916	1,917	1.0	3	0.02	1.79
Correlation on perturbation effects	2,000	44	76	1,916	1,918	1.0	3	0.03	1.82
Correlation on perturbation effects	5,000	42	189	4,411	4,473	1.0	6	0.01	1.77
Correlation on perturbation effects	5,000	43	189	4,410	4,475	1.0	6	0.01	1.80
Correlation on perturbation effects	5,000	44	189	4,411	4,476	1.0	6	0.01	1.81

Scoring rule	Cells	Seed	Ctrl	Targets	Guides	Med. c/t	Max c/t	P@100	Runtime (s)
Correlation on perturbation effects	10,000	42	378	6,387	6,628	1.0	13	0.02	2.17
Correlation on perturbation effects	10,000	43	378	6,405	6,631	1.0	12	0.01	2.26
Correlation on perturbation effects	10,000	44	378	6,420	6,642	1.0	12	0.01	2.10
Correlation on perturbation effects	15,000	42	566	7,660	8,038	1.0	19	0.01	2.50
Correlation on perturbation effects	15,000	43	566	7,660	8,026	1.0	19	0.01	2.54
Correlation on perturbation effects	15,000	44	566	7,660	8,029	1.0	19	0.01	2.47
Correlation on perturbation effects	20,000	42	755	7,660	8,141	2.0	25	0.01	2.84
Correlation on perturbation effects	20,000	43	755	7,660	8,141	2.0	25	0.01	2.81
Correlation on perturbation effects	20,000	44	755	7,660	8,141	2.0	25	0.01	2.73
Raw expression correlation	2,000	42	76	1,916	1,916	1.0	3	0.01	1.61
Raw expression correlation	2,000	43	76	1,916	1,917	1.0	3	0.02	1.69
Raw expression correlation	2,000	44	76	1,916	1,918	1.0	3	0.04	1.84
Raw expression correlation	5,000	42	189	4,411	4,473	1.0	6	0.01	1.65
Raw expression correlation	5,000	43	189	4,410	4,475	1.0	6	0.01	1.63
Raw expression correlation	5,000	44	189	4,411	4,476	1.0	6	0.01	1.61
Raw expression correlation	10,000	42	378	6,387	6,628	1.0	13	0.02	1.96
Raw expression correlation	10,000	43	378	6,405	6,631	1.0	12	0.01	1.73
Raw expression correlation	10,000	44	378	6,420	6,642	1.0	12	0.01	1.67

Scoring rule	Cells	Seed	Ctrl	Targets	Guides	Med. c/t	Max c/t	P@100	Runtime (s)
Raw expression correlation	15,000	42	566	7,660	8,038	1.0	19	0.01	1.69
Raw expression correlation	15,000	43	566	7,660	8,026	1.0	19	0.01	1.70
Raw expression correlation	15,000	44	566	7,660	8,029	1.0	19	0.01	1.65
Raw expression correlation	20,000	42	755	7,660	8,141	2.0	25	0.01	1.75
Raw expression correlation	20,000	43	755	7,660	8,141	2.0	25	0.01	1.71
Raw expression correlation	20,000	44	755	7,660	8,141	2.0	25	0.01	1.79

Findings, stated narrowly. The number of distinct perturbation targets present in a draw rises steeply and then saturates at the pool's 7,660 by 15,000 cells; replication depth barely moves, with the median cells per target going from 1 to 2 across the entire range and only 846 of 7,660 targets reaching 5 cells even at 20,000. Full-graph precision@100 stays within 0.010–0.023 (means) and 0.01–0.04 (individual seeds) across a tenfold increase in cells.

What this does *not* establish. The earlier version concluded that "the bottleneck is not statistical power but structural". That conclusion is not available from this experiment, for three reasons that hold independently of the subsampling bug. Precision@100 against a reference with 61 mapped TRRUST edges moves in quanta of 0.01 — one edge in a hundred — and the entire observed range spans one to four edges, so "flat" and "noisy" are indistinguishable here. The perturbation-effect rule was never actually evaluated at 2,000 or 5,000 cells in the previous run, so the previously reported low-n behaviour of that rule was the raw-correlation rule's behaviour. And a full-graph metric at 0.006 % prevalence is the wrong instrument for a power question in any case; the regulator-centric evaluation is where power should be assessed, and Section S9 assesses it. We report this sweep as a description of what the subsamples contain, not as a power analysis.

S9 Aggregate test against matched random rankings

Sections S4–S6 are descriptive. This section is the inferential analysis, and it is the basis for the main text's central claim. The question is narrow: pooled across methods, are per-regulator rankings better aligned with curated targets than *matched random rankings of the same candidate lists*?

S9.1 Design

For every (method, dataset, reference, evaluable regulator) cell:

1. the method's scores are taken over that regulator's eligible candidate target list (Section S3; the list excludes the regulator itself);
2. **2,000 matched random rankings** over the same candidate targets are drawn by permuting the score vector. Permuting scores and re-ranking is distributionally identical to permuting labels against the fixed score ordering, including tie structure, so the cheaper equivalent is computed — which matters, because two methods here have a median of only 4.3 % distinct scores per list;
3. the cell records `ap_observed`, the null mean, s.d. and 2.5/97.5 quantiles, `ap_difference`, `ap_normalized` $= (\mathrm{AP} - \overline{\mathrm{AP}}_{\mathrm{rand}})/(1 - \overline{\mathrm{AP}}_{\mathrm{rand}})$, the ratio, and a one-sided permutation p;
4. **regulator identity and reference identity are retained on every row**.

Coverage: **39 distinct evaluable regulators**, 63 regulator × reference cells, 18 method labels (17 model matrices plus one `random_control`), 957 rows in the hierarchical random-baseline results table, of which 894 are real-method rows. A `random_control` method — uniform random scores, seeded per cell — is carried through the entire pipeline as a negative control.

S9.2 Why regulator clustering is mandatory

The same regulator recurs across references: of the 39 distinct evaluable regulators, 24 appear under one reference, 6 under two and 9 under all three. The 63 cells therefore contain 39 independent regulators, and the 894 real-method rows reuse those same 39 across 17 labels. A mixed model with a regulator random intercept puts **25.5 % of the residual variance between regulators** (random-intercept variance 16.58 against residual scale 48.45 in ×100 response units, ICC = 0.255). Treating TRRUST, DoRothEA A+B and CollecTRI as three independent replications of the same regulator therefore materially overstates precision, and the naive intervals (`naive_ci95_*` in the result file) are narrower than the clustered intervals on *every* method.

S9.3 Hierarchical bootstrap

10,000 replicates, seed 2026, three levels: dataset (resampled with replacement from the three); reference within dataset (resampled with replacement); and **distinct regulator identities** within dataset, each drawn regulator contributing *one* value — the mean of its AP difference over the references drawn at level 2. The nesting is drawn once per replicate and reused by every method, so methods stay paired across replicates.

A single-level **regulator-cluster bootstrap** is reported alongside. It holds the dataset mix fixed and prices only regulator sampling. It exists because with three datasets the level-1 draw admits only ten distinct dataset multisets, and one dataset (K562-Ess) supplies a single evaluable regulator, so the hierarchical interval is coarse and largely determined by whether that dataset is

drawn. The regulator-cluster interval is the more stable estimate of *regulator* sampling error; the hierarchical interval is the honest estimate of *generalization across datasets*, and it is primary.

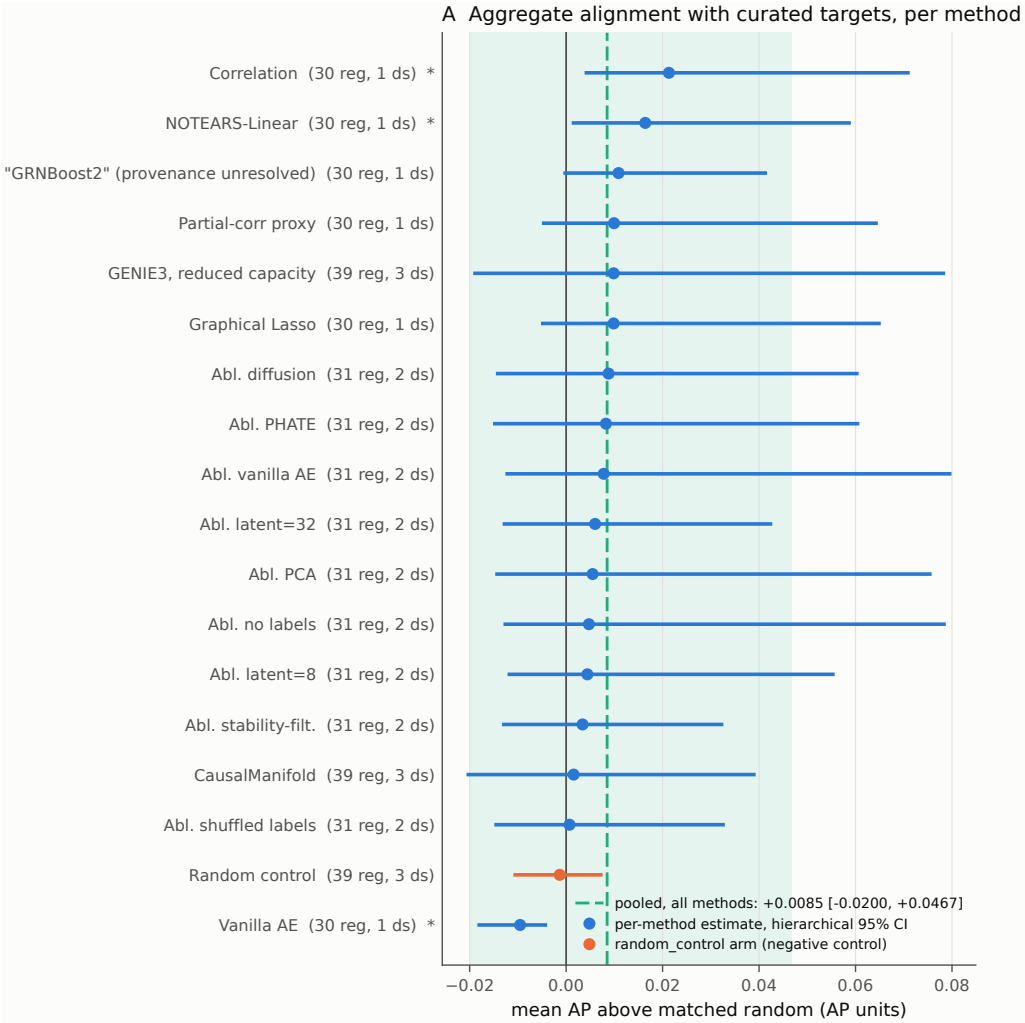

Figure S2: Per-method aggregate alignment with hierarchical 95 % confidence intervals, reproduced from the main text alongside Table S20. The dashed line and band are the pooled all-method estimate; the orange row is the `random_control` arm. Asterisks mark intervals excluding zero. Plotted values are tabulated in the reproducibility package.

Table S20: Primary aggregate test, all 18 method labels, sorted by pooled mean AP difference. $n = 39$ *distinct* regulators (63 regulator × reference cells); the analysis unit is the regulator. "Hier. 95 % CI" is the three-level hierarchical bootstrap and is primary; "Reg.-clust. CI" prices regulator sampling only, at a fixed dataset mix. * marks an interval excluding zero. Source: the aggregate bootstrap summary table in the reproducibility package.

Method	Regs	Cells	DS	Mean AP difference	Hier. 95 % CI (primary)	Reg.-clust. CI	Ratio	Ratio CI	% regs > rand
Correlation	30	50	1	+0.0213	[+0.0038, +0.0712]*	[+0.0021, +0.0449]	1.66×	[1.12, 2.56]	46.7
NOTEARS-Linear	30	50	1	+0.0164	[+0.0012, +0.0590]*	[+0.0009, +0.0349]	1.51×	[1.04, 2.23]	36.7
GRNBoost2*	30	50	1	+0.0109	[−0.0006, +0.0417]	[−0.0014, +0.0259]	1.34×	[0.98, 1.88]	40.0
Glasso-PartialCorr (PC/GES*)	30	50	1	+0.0099	[−0.0050, +0.0646]	[−0.0057, +0.0304]	1.31×	[0.82, 2.33]	26.7
GENIE3 (reduced cap.)	39	63	3	+0.0098	[−0.0193, +0.0786]	[+0.0027, +0.0497]	1.32×	[0.40, 3.11]	30.8
Graphical Lasso	30	50	1	+0.0098	[−0.0052, +0.0652]	[−0.0058, +0.0303]	1.31×	[0.81, 2.34]	26.7
Abl. diffusion	31	52	2	+0.0088	[−0.0146, +0.0606]	[+0.0046, +0.0462]	1.28×	[0.63, 2.49]	48.4
Abl. PHATE	31	52	2	+0.0083	[−0.0152, +0.0608]	[+0.0046, +0.0461]	1.26×	[0.62, 2.48]	48.4
Abl. vanilla AE	31	52	2	+0.0078	[−0.0126, +0.0799]	[−0.0056, +0.0546]	1.25×	[0.66, 2.94]	22.6
Abl. latent=32	31	52	2	+0.0060	[−0.0132, +0.0427]	[+0.0006, +0.0374]	1.19×	[0.65, 2.09]	35.5
Abl. PCA	31	52	2	+0.0055	[−0.0147, +0.0758]	[−0.0060, +0.0472]	1.17×	[0.62, 2.85]	22.6
Abl. no labels	31	52	2	+0.0047	[−0.0130, +0.0787]	[−0.0060, +0.0519]	1.15×	[0.65, 2.99]	35.5
Abl. latent=8	31	52	2	+0.0044	[−0.0121, +0.0557]	[−0.0056, +0.0407]	1.14×	[0.67, 2.37]	25.8
Abl. stability-filt.	31	52	2	+0.0034	[−0.0133, +0.0326]	[−0.0019, +0.0305]	1.11×	[0.65, 1.84]	32.3
CausalManifold	39	63	3	+0.0016	[−0.0206, +0.0393]	[−0.0025, +0.0308]	1.05×	[0.32, 2.07]	23.1
Abl. shuffled labels	31	52	2	+0.0007	[−0.0149, +0.0329]	[−0.0074, +0.0305]	1.02×	[0.62, 1.83]	16.1
Random control	39	63	3	−0.0013	[−0.0109, +0.0076]	[−0.0070, +0.0111]	0.95×	[0.64, 1.24]	25.6
Vanilla AE	30	50	1	−0.0095	[−0.0184, −0.0039]*	[−0.0138, −0.0039]	0.71×	[0.53, 0.89]	13.3
Seven distinct methods pooled (primary)				**+0.0085**	[**−0.0200**, **+0.0467**]	—	—	—	—
All 17 labels pooled (secondary)				+0.0078	[−0.0090, +0.0571]	[+0.0029, +0.0289]	1.26×	[0.71, 2.57]	—

Verdict. The aggregate test does not reach significance. The pooled mean AP difference over all methods is +0.0078 AP, hierarchical 95 % CI [−0.0090, +0.0571]; the pooled observed macro-AP is 0.0378 against a pooled matched-random macro-AP of 0.0299 — this is the secondary all-label pool, whose ratio is 1.26× with 95 % CI [0.71, 2.57]; mean normalized AP is +0.0085. The interval includes zero on both scales. The `random_control` arm lands at −0.0013 AP, [−0.0109, +0.0076], 0.95× — i.e. at zero — which is what validates the machinery: the pipeline can return a null, and does when handed noise.

Three intervals exclude zero, and none is evidence of aggregate signal. Exactly three method labels have hierarchical intervals excluding zero (`correlation_reference`, `notears_pca`, `autoencoder_interventional`), and these are *exactly* the three labels evaluated on a single dataset (K562-GS). For a single-dataset method the level-1 draw contributes no variance, so the interval is narrow for reasons of coverage rather than strength of evidence. Both methods evaluated on all three datasets — GENIE3 (reduced cap.) and CausalManifold — have intervals covering zero. And the third single-dataset "significant" label is negative: the vanilla autoencoder sits at −0.0095 AP, 0.71× random. No method-ranking statement based on "significance" in this table may be reported without that caveat.

S9.4 Clustered sensitivity model

Fitted with `statsmodels` 0.14.6 MixedLM; the response is rescaled ×100 for numerical stability (L-BFGS is singular at the native $\sim 10^{-2}$ scale) and all coefficients are divided back to AP units. In the full model `AP_difference ~ method + reference + dataset + (1 | regulator)` over 894 observations and 39 regulator groups (Powell converged): **reference identity is not significant**

(DoRothEA A+B $p = 0.667$, TRRUST $p = 0.865$, baseline CollecTRI), **dataset is not significant** (K562-GS $p = 0.531$, RPE1-Ess $p = 0.455$), and no method contrast is significant except `autoencoder_interventional` (-0.0417 AP, $p = 0.003$), which is worse than the others. Once regulator clustering is in the model, the apparent per-reference spread in Table S22 does not survive.

Table S21: Per-method clustered means, `AP_difference ~ 1 + (1 | regulator)`, sorted as in Table S20. n is 50, 52 or 63 rows over 30, 31 or 39 regulator groups depending on the method's dataset coverage; the analysis unit is one (regulator, reference) row with a regulator random intercept. Source: the aggregate mixed-model summary table in the reproducibility package.

Method	Clustered mean	95 % CI	p
Correlation	+0.0298	[+0.0016, +0.0581]	0.038
NOTEARS-Linear	+0.0211	[+0.0004, +0.0419]	0.046
GRNBoost2*	+0.0141	[−0.0024, +0.0307]	0.094
Glasso-PartialCorr (PC/GES*)	+0.0205	[−0.0065, +0.0474]	0.137
GENIE3 (reduced cap.)	+0.0253	[+0.0008, +0.0499]	0.043
Graphical Lasso	+0.0199	[−0.0070, +0.0469]	0.147
Abl. diffusion	+0.0267	[+0.0032, +0.0501]	0.026
Abl. PHATE	+0.0267	[+0.0031, +0.0502]	0.026
Abl. vanilla AE	+0.0236	[−0.0090, +0.0562]	0.156
Abl. latent=32	+0.0215	[+0.0002, +0.0428]	0.048
Abl. PCA	+0.0212	[−0.0115, +0.0539]	0.203
Abl. no labels	+0.0244	[−0.0195, +0.0682]	0.276
Abl. latent=8	+0.0162	[−0.0100, +0.0424]	0.225
Abl. stability-filt.	+0.0152	[−0.0028, +0.0332]	0.098
CausalManifold	+0.0142	[−0.0039, +0.0323]	0.125
Abl. shuffled labels	+0.0104	[−0.0119, +0.0327]	0.360
Random control	−0.0012	[−0.0085, +0.0061]	0.743
Vanilla AE	−0.0101	[−0.0148, −0.0054]	2.8e−5

Multiplicity. Six method labels reach nominal $p < 0.05$ in Table S21. **None survives Bonferroni correction across the 17 method labels** ($\alpha = 0.0029$); the smallest positive p is 0.026. The only p below the corrected threshold belongs to `autoencoder_interventional` and is negative. The clustered model therefore agrees with the hierarchical bootstrap: descriptive evidence of alignment, not a statistically supported aggregate effect. The `random_control` row (-0.0012, $p = 0.743$) confirms the model is calibrated.

S9.5 Per-reference and per-dataset breakdown

Table S22: Pooled over the 17 real method labels, by reference. $n = 894$ real-method rows; the analysis unit is one (method, dataset, reference, regulator) evaluation. These are naive pooled means *without* regulator clustering and are shown for description only — reference identity is not a significant predictor in the clustered model above. Source: the per-reference aggregate table in the reproducibility package.

Reference	Distinct regs	Cells	Mean AP diff	Mean AP obs	Mean AP rand	Ratio	% cells > rand
CollecTRI	35	499	+0.0151	0.0476	0.0325	1.46×	34.7
DoRothEA A+B	11	172	+0.0448	0.0959	0.0511	1.88×	49.3
TRRUST	17	223	+0.0394	0.0882	0.0489	1.81×	47.5

By dataset × reference, with the number of evaluable regulators in brackets: K562-GS × TRRUST [12] +0.0418, 1.84×; K562-GS × DoRothEA A+B [10] +0.0436, 1.85×; K562-GS × CollecTRI [28] +0.0156, 1.48×; **K562-Ess × TRRUST [1]** −0.0163**,** 0.57×; **K562-Ess × CollecTRI [1]** −0.0017**,** 0.93×; RPE1-Ess × TRRUST [4] +0.0553, 2.30×; **RPE1-Ess × DoRothEA A+B [1]** +0.1603**,** 5.31×; RPE1-Ess × CollecTRI [6] +0.0123, 1.49×. Two observations follow. **Both K562-Ess cells are at or below random**, which falsifies any claim that methods outperform random in every dataset × reference combination. And the two most extreme cells in either direction are the two cells containing exactly one regulator.

S9.6 How to read "% of regulators above random"

This statistic compares observed AP against the *mean* of a right-skewed null whose median sits below its mean, so its null reference is not 50 % — it is the `random_control` row, **23.8 %**. Real methods run 32.7–55.8 %, above the calibrated null, but a naive 50 % sign test is invalid here because it "rejects" for the random control too. The positive pooled mean is driven by a minority of regulators with large gains rather than a broad shift: only 13–48 % of *distinct* regulators have a positive mean AP difference (last column of Table S20).

S9.7 Per-regulator enrichment, which does not overturn the verdict

At the individual-regulator level the one-sided permutation tests are clearly enriched: **174 of 894 real-method cells have** $p < 0.05$ **(19.5 %)** against **4 of 63 (6.3 %)** for the random control. That enrichment is real at the cell level. But the cells are not independent — 39 regulators reused across 17 method labels — so it cannot be converted into an aggregate claim. It is the reason the pooled point estimate is positive; it is not a substitute for the interval.

S9.8 Caveats carried into the main text

1. **Effective n is 39 distinct regulators**, not 63 regulator × reference cells and not 894 method rows. K562-Ess contributes exactly one (ATF5).

2. **Three references are not three replications.** 15 of 39 regulators recur across references; ICC 0.255.

3. **Heavy score ties.** `graphical_lasso_reference` and `pc_ges_pca` have a median of only 4.3 % distinct scores per candidate list, and `grnboost2_sklearn` 45.9 % (`frac_distinct_scores` in the row-level file). Their observed AP depends on an arbitrary index tie-break. The matched-random comparison stays valid in distribution — the null permutes the identical score vector — but their point estimates are noisier than those of continuous-score methods.

4. **Single-dataset methods get narrower hierarchical intervals for free.**

5. The `random_control` arm is carried end-to-end and lands at zero on every estimator; any future change to this pipeline should keep it and check it.

S10 Matched synthetic-versus-real comparison

The comparison that motivates the whole enterprise — structure learning recovers edges on small synthetic graphs, therefore it should recover them on Perturb-seq — has never been measured. What has been compared is a globally thresholded edge-set F1 on a 15-node graph against a regulator-ranked precision or average precision on a 1,024-gene panel: two different statistics, over two different universes, at positive rates differing by a factor of four. This section makes the comparison measurable and reports what survives it. The answer has two halves and neither is usable alone: **the published magnitude was overstated by roughly an order of magnitude and rested on the wrong axis**, and **a genuine transfer failure of order** $15\times$ **remains after every correction**.

S10.1 Design: what was matched, and how it was enforced

Table S23: Every axis on which the two sides could have differed, and the mechanism that holds it fixed.

Matched	Mechanism
Metric code	The aggregate signal-test module, the implementation of record, is *imported and called*, not re-implemented. `ap_from_sorted`, `ap_matrix`, `matched_random_test`, `stable_seed`, `build_index`, `hierarchical_point_estimate`, `hierarchical_bootstrap` and `regulator_cluster_bootstrap` all come from that module.
Extended metrics	P@10, P@50, R@10, R@50 and MRR are computed from the *same* permutation matrix that `matched_random_test` draws, so the whole family shares one null. Every cell asserts that its AP, its 2,000-draw null vector and its permutation p-value are bit-identical to `matched_random_test`'s.
Universe rules	*candidate regulator* / *evaluable regulator* / *candidate universe* / *evaluation universe* / *prevalence* applied verbatim from the universe-terminology definition table: a regulator's candidate list is every eligible target except itself, evaluable $\Leftrightarrow$ at least one non-self curated target, prevalence = positive edges / evaluation edges.
Prevalence	Synthetic DAG density **bisected** against the real pooled prevalence recomputed at run time from the eligible-universe table (2.0967 %). "Matched" means the realised value lands inside the real per-cell range 0.9639–3.6629 %.
Aggregation	Synthetic *seeds occupy the level datasets occupy on real data*, so the identical hierarchical point estimator, regulator-clustered hierarchical bootstrap and single-level regulator-cluster bootstrap run on both sides.

Settings: 2,000 matched random rankings per regulator, 10,000 bootstrap replicates, master seed 2026, five synthetic seeds per scale, five scales, total runtime 59.1 min.

S10.2 Verification that the real side is reproduced exactly

Re-scoring all 957 real rows through the new code path and joining against the hierarchical random-baseline results table of record:

rows matched	957 / 957
$\max \lvert\Delta$ `ap_observed`$\rvert$	9.888×10^{-17}
$\max \lvert\Delta$ `ap_random_mean`$\rvert$	9.714×10^{-17}

The real row of the headline table therefore reproduces the record exactly: pooled mean AP difference $+0.0078$, hierarchical 95 % CI $[-0.0090, +0.0571]$, pooled $+0.0085$ $[-0.0200, +0.0467]$ over the seven

distinct methods (the all-label pool gives $+0.0078$ $[-0.0090, +0.0571]$, ratio $1.26\times$), 39 distinct evaluable regulators, evaluable counts per cell $[12, 10, 28, 1, 0, 1, 4, 1, 6]$, per-cell prevalence 0.96–3.66 %. The aggregate signal-test run log and the hierarchical random-baseline results table are byte-identical before and after this run (md5 verified); the reference module's `FileHandler(mode="w")` is stubbed for the duration of the import so that it cannot truncate the log of record.

S10.3 Prevalence calibration, and the scale that cannot be matched

Because every evaluable regulator carries at least one positive *by definition*, the smallest attainable prevalence at a given scale is approximately 1/|eligible targets|. That floor is what breaks the 15-node benchmark.

Table S24: Realised universe geometry per scale, averaged over five seeds. Calibration is by bisection on the DAG density against the real pooled rate of 2.0967 %.

Scale n	Perturbed regulators	Eligible targets	Evaluation edges	Positive edges	Prevalence	Fold vs real	Matched?
15	5	13.0	61.2	5.0	**8.22 %**	3.92×	**no** — impossible
50	12	48.6	571.6	12.0	2.10 %	1.00×	yes (at its floor)
100	25	98.2	2,430.8	53.0	2.18 %	1.04×	yes
500	40	499.0	19,920.0	424.0	2.13 %	1.02×	yes
1,000	40	999.0	39,920.0	835.4	2.09 %	1.00×	yes
real (9 cells)	1–55 cand.	157–523	181–10,336	2–188	2.10 % pooled; 0.96–3.66 % per cell		—

$n = 15$ **cannot be prevalence-matched and never could be.** With at most 14 candidate targets, one positive already means 7.7 %; the realised floor is 8.22 %, 3.8× **the real prevalence and above the top of the entire real per-cell range**. This is the single largest reason the 15-node benchmark "looked easy": its matched-random AP floor is **0.256**, whereas the real random floor is 0.030.

Two consequences follow immediately, and both belong in any discussion of transfer:

1. **On the 15-node benchmark, chance beats every real result.** A random ranker scores macro-AP 0.256 there, which is 4.8× the best real method (correlation, 0.0538) and 6.8× the real all-methods pooled macro-AP (0.0378). A benchmark on which chance outscores every real number cannot license a statement about transfer in either direction.

2. **NOTEARS, the method carrying the headline synthetic F1 = 0.590, is below chance at $n = 15$ under the real metric**: macro-AP 0.1819 against a matched-random floor of 0.2565 — normalized AP -0.0988, ratio $0.71\times$, 95 % CI $[-0.1498, +0.0659]$. Its synthetic "success" is a property of globally thresholded edge-set F1 on a small dense graph, not of the regulator-ranking task on which real data is scored.

$n = 1{,}000$ completed inside budget (five seeds, ∼5.5 min per seed), so no scale was skipped.

S10.4 The matched comparison

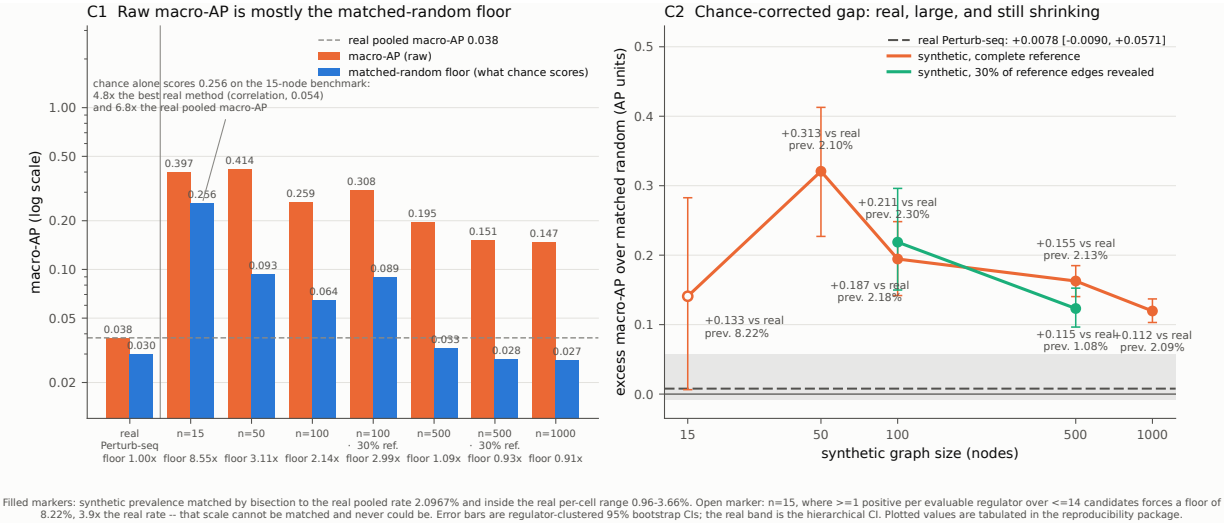

Figure S3: Reproduced from the main text. *Left:* raw macro-AP against its own matched-random floor, per arm, log scale. *Right:* excess macro-AP over that floor against graph scale, with regulator-clustered 95 % bootstrap CIs and the real hierarchical interval as a grey band; the fold difference from the real arm and the arm's prevalence are printed at every point. Plotted values, together with every per-method row and the derived fold columns, are tabulated in the reproducibility package.

Table S25: Matched synthetic-versus-real comparison, every method on both sides; $n = 58$ rows. The analysis unit is one (arm, method) evaluation, each pooling over that arm's evaluable regulators. Intervals are 2,000-replicate regulator-clustered bootstraps on both sides, except the real pooled row, which is quoted at the primary 10,000-replicate hierarchical setting to agree with Table S20. Source: the frozen matched synthetic-versus-real results table in the reproducibility package (sha256 `46dea8a4af1a`...).

Method	Regs	macro-AP	Matched-random floor	Excess AP over floor	95 % CI	Ratio	% cells > rand
Real Perturb-seq — 39 evaluable regulators, prevalence 0.0124, matched-random floor 0.0299							
Abl. vanilla AE	31	0.0392	0.0314	+0.0078	[−0.0126, +0.0799]	1.25×	32.7
Abl. latent=32	31	0.0374	0.0314	+0.0060	[−0.0132, +0.0427]	1.19×	42.3
Abl. latent=8	31	0.0354	0.0310	+0.0044	[−0.0121, +0.0557]	1.14×	42.3
Abl. no labels	31	0.0362	0.0315	+0.0047	[−0.0130, +0.0787]	1.15×	42.3
Abl. shuffled labels	31	0.0325	0.0318	+0.0007	[−0.0149, +0.0329]	1.02×	32.7
Abl. stability-filt.	31	0.0350	0.0316	+0.0034	[−0.0133, +0.0326]	1.11×	44.2
Abl. diffusion	31	0.0404	0.0316	+0.0088	[−0.0146, +0.0606]	1.28×	55.8
Abl. PCA	31	0.0370	0.0315	+0.0055	[−0.0147, +0.0758]	1.17×	38.5
Abl. PHATE	31	0.0404	0.0321	+0.0083	[−0.0152, +0.0608]	1.26×	51.9
Vanilla AE	30	0.0231	0.0327	−0.0095	[−0.0184, −0.0039]	0.71×	16.0
CausalManifold	39	0.0314	0.0298	+0.0016	[−0.0206, +0.0393]	1.05×	36.5
Correlation	30	0.0538	0.0325	+0.0213	[+0.0038, +0.0712]	1.66×	54.0
GENIE3 (reduced cap.)	39	0.0402	0.0304	+0.0098	[−0.0193, +0.0786]	1.32×	44.4
Graphical Lasso	30	0.0421	0.0322	+0.0098	[−0.0052, +0.0652]	1.31×	34.0
GRNBoost2*	30	0.0433	0.0324	+0.0109	[−0.0006, +0.0417]	1.34×	40.0
NOTEARS-Linear	30	0.0487	0.0323	+0.0164	[+0.0012, +0.0590]	1.51×	50.0
Glasso-PartialCorr (PC/GES*)	30	0.0422	0.0323	+0.0099	[−0.0050, +0.0646]	1.31×	34.0
Random control	39	0.0280	0.0293	−0.0013	[−0.0109, +0.0076]	0.95×	23.8
7 METHODS (primary)	39	—	—	+0.0085	[−0.0200, +0.0467]	—	—

continued on next page

Method	Regs	macro-AP	Matched-random floor	Excess AP over floor	95 % CI	Ratio	% cells > rand
17 LABELS (secondary)	39	0.0378	0.0299	+0.0078	[−0.0090, +0.0571]	1.26×	41.4
Synthetic, 15 nodes — 25 evaluable regulators, prevalence 0.0822, floor 0.2557 — **prevalence NOT matchable**							
Correlation (synthetic)	25	0.4829	0.2553	+0.2276	[+0.0661, +0.4044]	1.89×	52.0
GENIE3 RF, `max_features=1.0`	25	0.4701	0.2553	+0.2148	[+0.0328, +0.4152]	1.84×	64.0
GENIE3 RF, `max_features=sqrt`	25	0.4511	0.2556	+0.1955	[+0.0176, +0.3972]	1.77×	52.0
NOTEARS (synthetic)	25	0.1819	0.2565	−0.0746	[−0.1498, +0.0659]	0.71×	8.0
Random control	25	0.2516	0.2546	−0.0030	[−0.1053, +0.1328]	0.99×	32.0
ALL METHODS POOLED	25	0.3965	0.2557	+0.1409	[+0.0064, +0.2827]	1.55×	44.0
Synthetic, 50 nodes — 60 evaluable regulators, prevalence 0.0210, floor 0.0932							
Correlation (synthetic)	60	0.4240	0.0934	+0.3306	[+0.2056, +0.4644]	4.54×	55.0
GENIE3 RF, `max_features=1.0`	60	0.5021	0.0925	+0.4096	[+0.2751, +0.5441]	5.43×	66.7
GENIE3 RF, `max_features=sqrt`	60	0.4938	0.0929	+0.4009	[+0.2612, +0.5349]	5.31×	73.3
NOTEARS (synthetic)	60	0.2358	0.0938	+0.1420	[+0.0422, +0.2546]	2.51×	28.3
Random control	60	0.0933	0.0930	+0.0004	[−0.0386, +0.0723]	1.00×	25.0
ALL METHODS POOLED	60	0.4139	0.0932	+0.3208	[+0.2270, +0.4128]	4.44×	55.8
Synthetic, 100 nodes — 125 evaluable regulators, prevalence 0.0218, floor 0.0641							
Correlation (synthetic)	125	0.1835	0.0643	+0.1192	[+0.0670, +0.1777]	2.85×	40.8
GENIE3 RF, `max_features=1.0`	125	0.2808	0.0638	+0.2170	[+0.1361, +0.3051]	4.40×	68.8
GENIE3 RF, `max_features=sqrt`	125	0.2344	0.0641	+0.1704	[+0.1100, +0.2351]	3.66×	56.8
NOTEARS (synthetic)	125	0.3354	0.0640	+0.2713	[+0.1814, +0.3730]	5.24×	64.8
Random control	125	0.0562	0.0644	−0.0082	[−0.0188, +0.0041]	0.87×	20.8
ALL METHODS POOLED	125	0.2585	0.0641	+0.1945	[+0.1418, +0.2483]	4.04×	57.8
Synthetic, 100 nodes, 30 % reference revealed — 62 evaluable regulators, prevalence 0.0230, floor 0.0893							
Correlation (synthetic)	62	0.2085	0.0893	+0.1192	[+0.0243, +0.2316]	2.33×	37.1
GENIE3 RF, `max_features=1.0`	62	0.3550	0.0887	+0.2663	[+0.1375, +0.4118]	4.00×	66.1
GENIE3 RF, `max_features=sqrt`	62	0.2849	0.0896	+0.1953	[+0.0948, +0.3080]	3.18×	61.3
NOTEARS (synthetic)	62	0.3839	0.0897	+0.2942	[+0.1560, +0.4415]	4.28×	54.8
Random control	62	0.0755	0.0887	−0.0132	[−0.0416, +0.0253]	0.85×	21.0
ALL METHODS POOLED	62	0.3081	0.0893	+0.2187	[+0.1498, +0.2961]	3.45×	54.8
Synthetic, 500 nodes — 200 evaluable regulators, prevalence 0.0213, floor 0.0327							
Correlation (synthetic)	200	0.1452	0.0326	+0.1125	[+0.0890, +0.1366]	4.45×	78.5
GENIE3 RF, `max_features=sqrt`	200	0.1593	0.0327	+0.1266	[+0.1002, +0.1543]	4.87×	83.0
NOTEARS (synthetic)	200	0.2818	0.0327	+0.2491	[+0.1951, +0.3048]	8.62×	97.5
Random control	200	0.0299	0.0327	−0.0028	[−0.0050, −0.0004]	0.91×	25.5
ALL METHODS POOLED	200	0.1954	0.0327	+0.1628	[+0.1403, +0.1849]	5.98×	86.3
Synthetic, 500 nodes, 30 % reference revealed — 194 evaluable regulators, prevalence 0.0108, floor 0.0279							
Correlation (synthetic)	194	0.1018	0.0279	+0.0739	[+0.0441, +0.1095]	3.65×	52.1
GENIE3 RF, `max_features=sqrt`	194	0.1144	0.0279	+0.0865	[+0.0543, +0.1207]	4.10×	57.7
NOTEARS (synthetic)	194	0.2372	0.0280	+0.2092	[+0.1655, +0.2536]	8.47×	80.9
Random control	194	0.0291	0.0282	+0.0010	[−0.0064, +0.0102]	1.03×	28.4
ALL METHODS POOLED	194	0.1512	0.0279	+0.1232	[+0.0965, +0.1525]	5.41×	63.6
Synthetic, 1,000 nodes — 200 evaluable regulators, prevalence 0.0209, floor 0.0273							
Correlation (synthetic)	200	0.1154	0.0273	+0.0881	[+0.0695, +0.1090]	4.23×	87.0
GENIE3 RF, `max_features=sqrt`	200	0.1106	0.0273	+0.0833	[+0.0681, +0.0991]	4.05×	91.0
NOTEARS (synthetic)	200	0.2147	0.0273	+0.1874	[+0.1619, +0.2147]	7.86×	99.0
Random control	200	0.0262	0.0273	−0.0012	[−0.0026, +0.0006]	0.96×	30.5
ALL METHODS POOLED	200	0.1469	0.0273	+0.1196	[+0.1031, +0.1370]	5.38×	92.3

Negative control. `random_control` lands at ratio 0.95× on the real side and 0.87–1.00× across every synthetic scale — i.e. at zero on both sides. The machinery is calibrated on both axes.

S10.5 How much of the gap was mismatch, and how much is real

The additive decomposition

macro-AP = matched-random floor + excess over that floor. Splitting both sides:

Table S26: The raw synthetic advantage at small scale is mostly floor, not method quality. By 500–1,000 nodes the floors agree and the comparison is finally like-for-like.

Arm	macro-AP	= floor	+ excess	Floor vs real	**Excess vs real**
REAL pooled	0.0378	0.0299 (79 % of total)	+0.0078	1.00×	1.0×
syn. $n = 15$	0.3965	0.2557 (**64 % of total**)	+0.1409	**8.55**×	+0.1330
syn. $n = 50$	0.4139	0.0932 (23 %)	+0.3208	3.11×	+0.3129
syn. $n = 100$	0.2585	0.0641 (25 %)	+0.1945	2.14×	+0.1866
syn. $n = 500$	0.1954	0.0327 (17 %)	+0.1628	1.09×	+0.1549
syn. $n = 1{,}000$	0.1469	0.0273 (19 %)	+0.1196	**0.91**×	**+0.1118**

Two thirds of the apparent synthetic performance at $n = 15$ is the random floor — that is, prevalence and universe size, not method quality. At $n = 1{,}000$ the floor ratio is 0.91×, so excess AP is the axis on which the two sides may finally be compared.

The ladder from the published claim to the matched one

Table S27: Each step changes exactly one thing.

Step	What changes	Synthetic	Real	Apparent gap
0	as published: synthetic edge-F1 vs real precision@100	F1 0.590	0.010	∼59×
1	metric matched, still 15 nodes	macro-AP 0.397	0.038	10.5×
2	+ prevalence matched, $n = 1{,}000$	macro-AP 0.147	0.038	3.89×
3	+ chance-corrected (excess over floor)	+0.1196	+0.0078	**+0.1118**

Metric mismatch alone accounts for the step from ∼59× to 10.5×; prevalence and universe mismatch account for the step from 10.5× to 3.89× on macro-AP. So **most of the *apparent* gap — and essentially all of the rhetoric of a complete collapse to zero — was metric and prevalence mismatch**. But it is **not** all of it. On the chance-corrected axis, which is the fair one now that the floors agree, a **15× gap remains at** $n = 1{,}000$, with non-overlapping intervals: synthetic +0.1196 [+0.1031, +0.1370] against real +0.0078 [−0.0090, +0.0571] hierarchical, [+0.0029, +0.0289] regulator-clustered.

This is therefore **a genuine transfer failure whose magnitude was overstated by roughly an order of magnitude and whose evidence was presented on the wrong axis** — not a pure artifact. Both halves of that sentence have to reach the reader.

The residual is still shrinking with scale

Excess AP minus real: +0.3129 ($n = 50$) → +0.1866 (100) → +0.1549 (500) → +0.1118 (1,000). We report the difference rather than the ratio because the denominator of any such ratio is the real excess AP, whose own interval covers zero. The trend is monotone and has not flattened at the largest scale we could run, so graph size is a material — and unfinished — part of the residual. A 15-node result cannot be treated as informative about a 1,024-gene panel; on this trend it is not even the right order of magnitude.

Same method, same metric

Table S28: Methods that exist on both sides, at $n = 1{,}000$ against real.

Method	syn. macro-AP	real macro-AP	Gap	syn. excess	real excess	Excess gap
Correlation	0.1154	0.0538	**2.15**×	+0.0881	+0.0213	4.1×
GENIE3	0.1106	0.0402	**2.75**×	+0.0833	+0.0098	8.5×
NOTEARS	0.2147	0.0487	**4.41**×	+0.1874	+0.0164	11.4×

The correlation baseline — the strongest method on real data — transfers within a factor of **2.15** on macro-AP once metric, universe and prevalence are matched. NOTEARS, which looks best in synthetic benchmarks, transfers worst.

Reference incompleteness explains a further modest slice

Revealing only 30 % of the true edges as "curated", mirroring TRRUST / DoRothEA / CollecTRI incompleteness, cuts pooled macro-AP from 0.1954 to 0.1512 (−23 %) and excess AP from +0.1628 to +0.1232 (−24 %) at $n = 500$. At $n = 100$ it does not reduce the gap at all; macro-AP rises from 0.2585 to 0.3081. Reference incompleteness is therefore a real but second-order contributor, well short of closing the remaining 15×.

S10.6 The full metric family

Table S29: Pooled over methods. Random-baseline values in parentheses.

	real	syn. $n = 15$	syn. $n = 100$	syn. $n = 1{,}000$
macro-AP	0.0378	0.3965	0.2585	0.1469
AP matched-random	0.0299	0.2557	0.0641	0.0273
normalized AP	+0.0085	+0.1887	+0.2077	+0.1230
P@10 (random)	0.0268 (0.0126)	0.0800 (0.0822)	0.0734 (0.0218)	0.2512 (0.0209)
P@50 (random)	0.0182 (0.0125)	n/a	0.0290 (0.0218)	0.1060 (0.0209)
R@10	0.0343	0.8000	0.3539	0.1230
R@50	0.1980	n/a	0.6855	0.2546
MRR (random)	0.0827 (0.0524)	0.3965 (0.2557)	0.3618 (0.0872)	0.5926 (0.0825)
% cells above random	41.4 %	44.0 %	57.8 %	92.3 %

P@50 and R@50 are undefined at $n = 15$ and $n = 50$, where candidate lists hold only 13 and 49 targets; they are emitted as NaN rather than as a misleading value. Note that P@10 at $n = 15$ (0.0800) is *below* its own random floor (0.0822) — another symptom of the same structural problem.

S10.7 Methods actually run

Linear-Gaussian interventional simulation, five seeds per scale, sparse DAGs built in a random topological order with regulators placed in the first 60 % of the order. Regulators are deliberately excluded as background parents: otherwise early topological positions are over-sampled as parents (expected background out-degree $\sim 2 \ln n$) and regulator out-degree, hence prevalence, inflates silently. Hard `do()` interventions on every perturbed regulator, 100 samples each, plus observational samples to a total of $20n$ cells, matching the real 20,000-cell × 1,024-gene ratio. Columns are standardised before inference.

Table S30: Synthetic-side methods and their provenance.

Method	Provenance	Scales
`correlation`	score definition identical to the benchmark's correlation-network baseline routine	all
`notears`	same objective, hyperparameters and schedule as the benchmark's NOTEARS baseline routine; tensor device configurable so $n = 500/1{,}000$ finish in minutes	all
`genie3_rf_full`	sklearn RF as in the benchmark's GENIE3 random-forest routine (`max_features=1.0`, the real benchmark's setting)	≤ 100
`genie3_rf_sqrt`	canonical GENIE3 (`max_features="sqrt"`); the only affordable setting at $n \geq 500$	all
`random_control`	uniform scores, seeded per cell	all

Both GENIE3 settings were run at $n \leq 100$ so that the offset is measured rather than assumed: at $n = 100$, `rf_full` reaches macro-AP 0.2808 against `rf_sqrt` 0.2344, so the cheap setting is *conservative* by about 17 %. The large-scale synthetic numbers are therefore, if anything, slight underestimates of what the real benchmark's GENIE3 configuration would give, which strengthens rather than weakens the conclusion above.

Score-tie check: synthetic `frac_distinct_scores` is 1.000 at every scale and method, so no ranking is decided by tie order and no ordering artifact can enter. (For contrast, the real matrices for `graphical_lasso_reference` and `pc_ges_pca` are 5 % distinct — a pre-existing property of the record, untouched here.)

S10.8 Findings

1. **The comparison as published is not interpretable.** It sets a globally thresholded edge-set F1 against a regulator-ranked AP across universes at four-fold different prevalence. On the 15-node benchmark from which the synthetic F1 is drawn, a random ranker scores macro-AP 0.256 — 4.8× the best real method and 6.8× the real pooled value.

2. **NOTEARS is below chance at $n = 15$ under the real metric** (0.1819 against a 0.2565 floor, ratio 0.71×). The method carrying the headline synthetic result does not survive being scored the way real data is scored.

3. **A genuine gap survives every correction, and it is large**: 3.89× on macro-AP and an additive $+0.1118$ $[+0.0746, +0.1489]$ on chance-corrected excess AP at $n = 1{,}000$, intervals non-overlapping. It is not "effectively zero", and the word "collapse" does not describe it.

4. **The residual is still shrinking with scale** ($+0.3129 \rightarrow +0.1118$), so the sweep is unfinished at 1,000 nodes and a 15-node benchmark is not the right order of magnitude for a 1,024-gene panel.

5. **Correlation transfers within** 2.15×; reference incompleteness at 30 % revealed edges prices a further -24 % at $n = 500$ and nothing at $n = 100$.

S10.9 Limitations carried into the main text

1. **Linear-Gaussian structural equations**, chosen for continuity with the existing 15-node benchmark. This favours NOTEARS and inflates the synthetic side, so it biases *against* the conclusion above: a nonlinear or count-based simulator would shrink the residual 15× further, not enlarge it.

2. **Only three methods** are carried to $n = 1{,}000$ (correlation, NOTEARS, GENIE3). CausalManifold and the ablations are not simulated; the real side reports all 17 method labels, but the matched per-method comparison covers the three that exist on both sides.

3. **The complete-reference synthetic arms give methods a perfect ground truth.** Incompleteness is priced at 30 % revealed edges at two scales only.

4. **Synthetic seeds are not datasets.** They substitute for datasets at bootstrap level 1, so synthetic intervals price simulation-replicate variance rather than between-biological-system variance. Synthetic intervals are correspondingly narrower than the real ones and must not be read as equally general.

5. $n = 15$ **and** $n = 50$ **admit no P@50 / R@50**, and $n = 15$ admits no prevalence match at all. Those cells are NaN or flagged, never imputed.

S11 Expanded synthetic calibration

Standard structure-learning benchmarks [Marbach et al., 2012, Pratapa et al., 2020] operate on graphs one to two orders of magnitude smaller than a 1,024-gene Perturb-seq problem. To measure how much of the synthetic-to-real gap is attributable to scale alone, we generated sparse linear-Gaussian DAGs at two scales, simulated interventional data, ran three scoring rules, and swept the edge threshold. The 15-node graph has 37 true edges and the 100-node graph has 193; each method was evaluated at 19 thresholds per scale, giving 114 rows in total. The three scoring rules are ordinary least squares regression coefficients (OLS), Lasso coefficients, and random-forest feature importances (GENIE3-style).

Table S31: Best achievable F1 at each scale, with the operating point that attains it. $n = 114$ threshold evaluations, reduced to 6 method $\times$ scale operating points; the analysis unit is one (scale, method) pair at its F1-maximising threshold. Thresholds are chosen post hoc on the same data, so these are upper bounds rather than deployable operating points.

Method	Nodes	Threshold	Precision	Recall	F1	Predicted edges
GENIE3	15	0.096050	0.5000	0.5676	0.5316	42
GENIE3	100	0.020279	0.2929	0.7513	0.4215	495
Lasso	15	0.131160	0.3614	0.8108	0.5000	83
Lasso	100	0.072805	0.2824	0.8705	0.4264	595
OLS	15	0.142116	0.3452	0.7838	0.4793	84
OLS	100	0.130505	0.2828	0.7254	0.4070	495

Table S32: Scale degradation in best-threshold F1 from 15 to 100 nodes. $n = 3$ methods; the analysis unit is one scoring rule, compared between the two graph scales. ΔF1 is the 100-node value minus the 15-node value.

Method	F1 (15 nodes)	F1 (100 nodes)	ΔF1	Relative change (%)
GENIE3	0.5316	0.4215	-0.1101	-20.7
Lasso	0.5000	0.4264	-0.0736	-14.7
OLS	0.4793	0.4070	-0.0723	-15.1

Table S33: Complete threshold sweep for the expanded synthetic benchmark. $n = 114$ rows; the analysis unit is one (scale, method, threshold) evaluation.

Nodes	Method	Threshold	Precision	Recall	F1	Predicted edges
15	GENIE3	0.022791	0.1859	1.0000	0.3136	199
15	GENIE3	0.027685	0.1905	0.9730	0.3186	189
15	GENIE3	0.030325	0.2022	0.9730	0.3349	178
15	GENIE3	0.033261	0.2143	0.9730	0.3512	168
15	GENIE3	0.036546	0.2166	0.9189	0.3505	157
15	GENIE3	0.039265	0.2313	0.9189	0.3696	147
15	GENIE3	0.042076	0.2279	0.8378	0.3584	136
15	GENIE3	0.044795	0.2460	0.8378	0.3804	126

Nodes	Method	Threshold	Precision	Recall	F1	Predicted edges
15	GENIE3	0.048489	0.2609	0.8108	0.3947	115
15	GENIE3	0.052095	0.2571	0.7297	0.3803	105
15	GENIE3	0.057277	0.2632	0.6757	0.3788	95
15	GENIE3	0.062911	0.2857	0.6486	0.3967	84
15	GENIE3	0.068189	0.2973	0.5946	0.3964	74
15	GENIE3	0.071729	0.3492	0.5946	0.4400	63
15	GENIE3	0.078465	0.4151	0.5946	0.4889	53
15	GENIE3	0.096050	0.5000	0.5676	0.5316	42
15	GENIE3	0.110476	0.5312	0.4595	0.4928	32
15	GENIE3	0.133455	0.4286	0.2432	0.3103	21
15	GENIE3	0.200409	0.4545	0.1351	0.2083	11
15	Lasso	0.006301	0.2229	0.9459	0.3608	157
15	Lasso	0.010408	0.2349	0.9459	0.3763	149
15	Lasso	0.014739	0.2482	0.9459	0.3933	141
15	Lasso	0.021921	0.2632	0.9459	0.4118	133
15	Lasso	0.024958	0.2823	0.9459	0.4348	124
15	Lasso	0.039746	0.3017	0.9459	0.4575	116
15	Lasso	0.054338	0.3056	0.8919	0.4552	108
15	Lasso	0.068798	0.3333	0.8919	0.4853	99
15	Lasso	0.107363	0.3297	0.8108	0.4687	91
15	Lasso	0.131160	0.3614	0.8108	0.5000	83
15	Lasso	0.150100	0.3600	0.7297	0.4821	75
15	Lasso	0.184627	0.3582	0.6486	0.4615	67
15	Lasso	0.216312	0.3621	0.5676	0.4421	58
15	Lasso	0.259732	0.3400	0.4595	0.3908	50
15	Lasso	0.307923	0.3095	0.3514	0.3291	42
15	Lasso	0.354476	0.2941	0.2703	0.2817	34
15	Lasso	0.413630	0.2000	0.1351	0.1613	25
15	Lasso	0.441106	0.1765	0.0811	0.1111	17
15	Lasso	0.527382	0.0000	0.0000	0.0000	9
15	OLS	0.004325	0.1759	0.9459	0.2966	199
15	OLS	0.011200	0.1852	0.9459	0.3097	189
15	OLS	0.014148	0.1966	0.9459	0.3256	178
15	OLS	0.016898	0.2083	0.9459	0.3415	168
15	OLS	0.024231	0.2229	0.9459	0.3608	157
15	OLS	0.029665	0.2381	0.9459	0.3804	147
15	OLS	0.038213	0.2574	0.9459	0.4046	136
15	OLS	0.047344	0.2778	0.9459	0.4294	126
15	OLS	0.055319	0.3043	0.9459	0.4605	115
15	OLS	0.076686	0.3143	0.8919	0.4648	105
15	OLS	0.101979	0.3158	0.8108	0.4545	95
15	OLS	0.142116	0.3452	0.7838	0.4793	84
15	OLS	0.169498	0.3378	0.6757	0.4505	74
15	OLS	0.207020	0.3651	0.6216	0.4600	63
15	OLS	0.257742	0.3396	0.4865	0.4000	53

Nodes	Method	Threshold	Precision	Recall	F1	Predicted edges
15	OLS	0.314899	0.3095	0.3514	0.3291	42
15	OLS	0.386117	0.3125	0.2703	0.2899	32
15	OLS	0.432784	0.1429	0.0811	0.1034	21
15	OLS	0.516941	0.0000	0.0000	0.0000	11
100	GENIE3	0.004225	0.0205	1.0000	0.0402	9,405
100	GENIE3	0.004700	0.0217	1.0000	0.0424	8,911
100	GENIE3	0.005059	0.0229	1.0000	0.0449	8,412
100	GENIE3	0.005361	0.0244	1.0000	0.0476	7,921
100	GENIE3	0.005665	0.0260	1.0000	0.0507	7,426
100	GENIE3	0.005963	0.0277	0.9948	0.0539	6,929
100	GENIE3	0.006249	0.0294	0.9793	0.0570	6,436
100	GENIE3	0.006553	0.0318	0.9793	0.0616	5,939
100	GENIE3	0.006845	0.0345	0.9741	0.0667	5,445
100	GENIE3	0.007144	0.0376	0.9637	0.0723	4,949
100	GENIE3	0.007458	0.0415	0.9585	0.0796	4,454
100	GENIE3	0.007786	0.0462	0.9482	0.0882	3,959
100	GENIE3	0.008156	0.0526	0.9430	0.0996	3,463
100	GENIE3	0.008532	0.0610	0.9378	0.1145	2,968
100	GENIE3	0.008962	0.0731	0.9378	0.1357	2,475
100	GENIE3	0.009447	0.0913	0.9378	0.1664	1,982
100	GENIE3	0.010110	0.1193	0.9171	0.2111	1,484
100	GENIE3	0.011292	0.1707	0.8756	0.2857	990
100	GENIE3	0.020279	0.2929	0.7513	0.4215	495
100	Lasso	0.000308	0.0684	1.0000	0.1280	2,823
100	Lasso	0.000695	0.0722	1.0000	0.1346	2,674
100	Lasso	0.001068	0.0764	1.0000	0.1420	2,526
100	Lasso	0.001386	0.0812	1.0000	0.1502	2,377
100	Lasso	0.001795	0.0866	1.0000	0.1594	2,229
100	Lasso	0.002243	0.0928	1.0000	0.1698	2,080
100	Lasso	0.002732	0.0999	1.0000	0.1816	1,932
100	Lasso	0.003266	0.1082	1.0000	0.1953	1,783
100	Lasso	0.003901	0.1181	1.0000	0.2113	1,634
100	Lasso	0.004571	0.1300	1.0000	0.2300	1,485
100	Lasso	0.005421	0.1444	1.0000	0.2523	1,337
100	Lasso	0.006545	0.1623	1.0000	0.2793	1,189
100	Lasso	0.008059	0.1856	1.0000	0.3131	1,040
100	Lasso	0.010683	0.2164	1.0000	0.3558	892
100	Lasso	0.023956	0.2571	0.9896	0.4081	743
100	Lasso	0.072805	0.2824	0.8705	0.4264	595
100	Lasso	0.144174	0.2937	0.6788	0.4100	446
100	Lasso	0.230549	0.3289	0.5078	0.3992	298
100	Lasso	0.368985	0.2349	0.1813	0.2047	149
100	OLS	0.000596	0.0205	1.0000	0.0402	9,405
100	OLS	0.001217	0.0217	1.0000	0.0424	8,909
100	OLS	0.001917	0.0229	1.0000	0.0448	8,415

Nodes	Method	Threshold	Precision	Recall	F1	Predicted edges
100	OLS	0.002629	0.0244	1.0000	0.0476	7,920
100	OLS	0.003387	0.0260	1.0000	0.0507	7,424
100	OLS	0.004101	0.0279	1.0000	0.0542	6,929
100	OLS	0.004841	0.0300	1.0000	0.0582	6,435
100	OLS	0.005612	0.0325	1.0000	0.0629	5,940
100	OLS	0.006440	0.0354	1.0000	0.0685	5,445
100	OLS	0.007291	0.0390	1.0000	0.0751	4,950
100	OLS	0.008314	0.0433	1.0000	0.0830	4,455
100	OLS	0.009462	0.0487	1.0000	0.0929	3,960
100	OLS	0.010705	0.0557	1.0000	0.1055	3,465
100	OLS	0.012084	0.0650	1.0000	0.1220	2,970
100	OLS	0.013774	0.0780	1.0000	0.1447	2,475
100	OLS	0.016092	0.0975	1.0000	0.1776	1,980
100	OLS	0.019609	0.1300	1.0000	0.2300	1,485
100	OLS	0.026731	0.1939	0.9948	0.3246	990
100	OLS	0.130505	0.2828	0.7254	0.4070	495

Findings. All three scoring rules lose F1 when the graph grows from 15 to 100 nodes, by 0.0723 to 0.1101 absolute (14.7% to 20.7% relative). The degradation is real but modest, and it is much smaller than the synthetic-to-real gap: the best 100-node F1 is still 0.4264, whereas the best *primary-method* regulator-centric macro-AP on real K562-GS data is 0.1117 (Correlation against TRRUST, Table S8). The larger value of 0.1556 in that table belongs to an ablation trained without perturbation labels and is not a primary-method result. Scale alone therefore does not explain the gap — but note that this raw-F1 versus raw-AP comparison is exactly the kind that Section S10 shows to be uninterpretable: it sets a globally thresholded edge-set statistic against a regulator-ranked one across universes of different prevalence, and it should not be read quantitatively. Section S10 redoes the comparison on a single metric at matched prevalence, and reaches a smaller but non-zero answer. The mechanism of degradation is visible in the precision column of Table S31: recall stays high at 100 nodes (0.7254–0.8705) while precision falls to 0.2824–0.2929, i.e. the methods keep finding the true edges but bury them among far more false ones as the candidate space grows quadratically. This is the same behaviour, in miniature, that the full-graph top-k tensor exhibits at 1,024 genes (Section S5).

Note that the methods evaluated in this sweep (OLS, Lasso, random-forest importance) are not the same set as the methods benchmarked on real data; this experiment calibrates the effect of graph size on threshold-based structure recovery, not the real-data ranking.

S12 Reference database overlap

Agreement of a result across TRRUST v2 [Han et al., 2018], DoRothEA A+B [Garcia-Alonso et al., 2019] and CollecTRI [Müller-Dott et al., 2023] would be evidence of robustness only if the three databases were built from disjoint evidence. They are not. For each dataset we mapped all three databases onto that dataset's 1,024-gene set and computed pairwise overlap of the resulting edge, regulator and target sets (Table S34).

Table S34: Pairwise overlap between the three curated reference databases, computed separately for each dataset's mapped gene set. $n = 9$ rows (3 datasets $\times$ 3 database pairs); the analysis unit is one (dataset, database pair). Jaccard is intersection over union of the mapped sets.

		Edges			Regulators			Targets	
Dataset	Database pair	Shared	Union	Jaccard	Shared	Union	Jaccard	Shared	Jaccard
K562-GS	TRRUST vs DoRothEA A+B	35	179	0.1955	11	18	0.6111	30	0.2206
K562-GS	TRRUST vs CollecTRI	59	274	0.2153	15	41	0.3659	43	0.3258
K562-GS	DoRothEA A+B vs CollecTRI	51	372	0.1371	12	41	0.2927	52	0.2562
K562-Ess	TRRUST vs DoRothEA A+B	33	196	0.1684	13	24	0.5417	35	0.2482
K562-Ess	TRRUST vs CollecTRI	66	344	0.1919	18	47	0.3830	52	0.3152
K562-Ess	DoRothEA A+B vs CollecTRI	46	449	0.1024	15	43	0.3488	57	0.2478
RPE1-Ess	TRRUST vs DoRothEA A+B	65	264	0.2462	18	47	0.3830	51	0.3517
RPE1-Ess	TRRUST vs CollecTRI	139	633	0.2196	34	78	0.4359	73	0.3527
RPE1-Ess	DoRothEA A+B vs CollecTRI	87	698	0.1246	20	69	0.2899	74	0.2937

The references are not independent. Edge Jaccard ranges from 0.1024 to 0.2462 across the nine dataset–pair combinations, and regulator Jaccard from 0.2899 to 0.6111, reaching 0.6111 for TRRUST versus DoRothEA A+B on K562-GS. Between 33 and 139 edges are shared by any given pair on any given dataset. These databases are built from overlapping primary literature and, in the case of CollecTRI, partly by aggregating the others, so an evaluation result that holds against all three is *not* three independent confirmations. Throughout this work the three databases are described as three curated references with partially shared evidence, and consistency across them is reported as a sensitivity check on curation criteria rather than as replication. Where the three disagree — for example, the eligible universe on K562-Ess contains 2 positive edges under TRRUST, 0 under DoRothEA A+B and 5 under CollecTRI (Table S5) — the disagreement reflects coverage differences, not measurement error.

S13 Repeated-run uncertainty

The repeated-run experiment measures how much a method's full-graph result moves when only the random seed changes. It runs on K562-GS (20,000 cells × 1,024 genes) and scores precision@100 against TRRUST, the same quantity reported in Section S5. Deterministic methods (Pearson correlation, Graphical Lasso) are run twice and their adjacency matrices compared elementwise, giving a determinism check rather than a seed distribution. Stochastic methods are run under multiple seeds: five (42–46) for the two cheap methods, three (42–44) for the two tree-ensemble methods.

Two deliberate compute-driven deviations. First, the tree-ensemble methods receive three seeds rather than five. Second, they are run at 20 trees on 3,000 cells rather than the 50-tree, 5,000-cell configuration used for the benchmark results in Sections S4–S5; the reduced budget is held fixed across seeds so that any reported spread is not confounded by configuration changes. Both quantities are recorded per row in the result file. A consequence worth stating: the uncertainty estimated here is for the 20-tree configuration, and is not transferable to the 50-tree numbers reported elsewhere.

Completion status. Of 18 planned method-seed cells, 15 are recorded in the result file. The GENIE3-proxy seed set is **complete** (3 of 3) at the reduced configuration. The three remaining cells are the "GRNBoost2" seeds. They were not run, and the reason originally recorded — that the label was the same computation as the GENIE3 proxy, so its seeds would have reproduced the GENIE3 rows exactly — **does not hold**: that equality has since been checked against the stored artifacts and withdrawn. The correct statement is that the code producing this label's scores could not be identified, so there is no configuration to re-run at a new seed. Its seed stability is therefore **unmeasured**, not known to equal GENIE3's, which is a further reason the label is excluded from every pooled estimate.

The GENIE3 spread is an upper bound, not a measurement. The completed GENIE3 seeds give precision@100 $= 0.0167 \pm 0.0115$ (0.01, 0.03, 0.01) at **20 trees on 3,000 cells**. Every GENIE3 number in Sections S4–S5 comes from the **50-tree, 5,000-cell** configuration. Fewer trees give noisier feature importances, so this spread **bounds the seed variance of the main-table numbers from above**; it does not estimate it. We therefore attach no standard deviation to the 50-tree point estimates anywhere in this work. Closing the gap requires re-running the sweep at 50 trees and 5,000 cells, estimated at ≈9 h for three seeds against ≈43 min per seed at the reduced setting.

One consequence for the top-k tensor. Section S5 reports that precision@100 never exceeds 0.02 anywhere in the tensor. That tensor records a *single* run per cell. Re-run across seeds at the reduced configuration, GENIE3 seed 43 returns 0.03 — above the tensor maximum. The correct statement is therefore "precision@100 does not exceed 0.02 in any tabulated run", not "no method can exceed 0.02". A single-run tensor understates the variability of stochastic methods.

Table S35: Repeated-run status for every planned method-seed cell on K562-GS, scored by precision@100 against TRRUST. $n = 18$ planned cells, of which 15 are recorded in the all-method uncertainty results table and 3 were deliberately not run because they would duplicate the GENIE3 rows bit-for-bit; the analysis unit is one method-seed run. "det." marks the two deterministic methods, which are reported as a single verified row rather than a seed distribution.

Method	Seed	P@100	Edges	Runtime (s)	Trees	Cells
Pearson correlation network	det.	0.0100	718	0.2	–	20,000
Graphical Lasso network	det.	0.0100	794	181.4	–	20,000
NOTEARS-Linear (PCA-32 features)	42	0.0000	1,047,552	4.2	–	20,000
NOTEARS-Linear (PCA-32 features)	43	0.0000	1,047,552	4.0	–	20,000
NOTEARS-Linear (PCA-32 features)	44	0.0000	1,047,552	4.0	–	20,000
NOTEARS-Linear (PCA-32 features)	45	0.0000	1,047,552	4.1	–	20,000
NOTEARS-Linear (PCA-32 features)	46	0.0000	1,047,552	3.8	–	20,000
Vanilla autoencoder (latent 16)	42	0.0000	140	37.2	–	20,000
Vanilla autoencoder (latent 16)	43	0.0000	166	37.6	–	20,000
Vanilla autoencoder (latent 16)	44	0.0000	166	38.1	–	20,000
Vanilla autoencoder (latent 16)	45	0.0000	166	37.9	–	20,000
Vanilla autoencoder (latent 16)	46	0.0000	164	38.0	–	20,000
GENIE3 (20 trees, reduced)	42	0.0100	1,032,001	2563.4	20	3,000
GENIE3 (20 trees, reduced)	43	0.0300	1,030,665	2511.3	20	3,000
GENIE3 (20 trees, reduced)	44	0.0100	1,031,845	2339.6	20	3,000
"GRNBoost2" (provenance unresolved)	42	**not run** — no identifiable configuration to re-run				
"GRNBoost2" (provenance unresolved)	43	**not run** — no identifiable configuration to re-run				
"GRNBoost2" (provenance unresolved)	44	**not run** — no identifiable configuration to re-run				

Table S36: Seed spread for the stochastic methods with at least two completed seeds. $n = 13$ completed stochastic runs; the analysis unit is one method, summarised across its completed seeds. A method whose planned seeds are incomplete is marked accordingly and its spread must not be read as the method's seed distribution.

Method	Seeds done	Mean P@100	s.d.	Edge range	Seed set
GENIE3 (20 trees, reduced)	3	0.0167	0.0115	1,030,665–1,032,001	complete
NOTEARS-Linear (PCA-32 features)	5	0.0000	0.0000	1,047,552	complete
Vanilla autoencoder (latent 16)	5	0.0000	0.0000	140–166	complete

Reading of Tables S35 and S36. The two deterministic methods reproduced themselves exactly: the elementwise comparison of two independent runs returned True, so their single reported rows are exact rather than representative. Among the completed stochastic methods, precision@100 did not move at all across seeds (s.d. 0.0000, 0.0115), while the underlying graphs did differ — the vanilla autoencoder produced between 140 and 166 surviving edges depending on seed. The correct reading is that the metric is insensitive to seed at this operating point, not that the runs were identical. For the GENIE3 proxy the seed set is complete at the reduced configuration and its s.d. is reported as an upper bound on the main tables' seed variance; no s.d. is reported for the "GRNBoost2" label, which was not run because it is the same computation.

S14 Faithful PC and GES

Every PC/GES result reported in Sections S4–S5 comes from the Glasso partial-correlation proxy (Section S1). This section reports the separate benchmark of the faithful constraint-based and score-based algorithms as implemented in `causal-learn` 0.1.4.8, which return CPDAGs rather than weighted adjacencies and therefore admit only binary-edge evaluation.

The infeasibility result. Unrestricted PC and unrestricted GES cannot be run on this benchmark's eligible universe. On the 173-gene TRRUST-eligible universe at 2,000 cells, both *exceed 14,400 s* (4 h) of single-core time. This is not a stingy budget: the pre-registered budget of 1,800 s was raised eightfold for the four key configurations (unrestricted PC and unrestricted GES, at 2,000 and at 20,000 cells) and all four still timed out. Four rescues also failed to make either algorithm finish on 173 genes: Bonferroni-correcting α over all 14,878 gene pairs ($\alpha = 3.36 \times 10^{-6}$); capping PC's conditioning-set size at $\max k = 1$, 2 and 3; capping GES's parent count at $\max P = 3$, 5 and 10; and replacing GES's score with the mathematically identical covariance-based Gaussian BIC (`local_score_BIC_from_cov`), which evaluates the same objective at $O(k^3)$ instead of $O(nk^2)$ per local call. That the cheaper score does not help localises the cost in causal-learn's *search* — enumeration of insert/delete operators over neighbour subsets — rather than in scoring. A scaling probe on random gene subsets at 2,000 cells under a 900 s cap shows where the wall sits: 25 genes cost 0.4 s (PC) and 0.5 s (GES); 50 genes 7.0 s and 6.6 s; 75 genes 200.2 s and 76.1 s; and 100, 125, 150 and 173 genes all exceed the cap. An earlier audit note that PC ran in ≈1 s on 2,000 cells × 50 genes is exactly the cheap regime and is not representative of the eligible universe.

More cells makes it worse. On the 60-gene feasible subset, unrestricted PC completes in 292.7 s and unrestricted GES in 913.2 s at 2,000 cells, but *both time out at 1,800 s at 5,000 cells and again at 20,000 cells*, on the identical gene set. Greater Fisher-z power at larger n retains more edges in the skeleton, which enlarges neighbourhoods and multiplies the conditioning sets that must be searched; the analogous argument applies to GES, whose operators enumerate subsets of each node's neighbours. Sample size is on the wrong side of the cost curve for this family of algorithms.

Where they do run, nothing separates. On the 60-gene matched universe (all 12 scoreable regulators and all their curated targets retained, negatives sampled to size, and the proxy's saved matrix re-scored on the identical universe by the identical metric code) faithful PC reaches macro-AP 0.1660 (1.34× the recomputed random baseline), faithful GES 0.1484 (1.20×), and the proxy 0.1493 (1.21×); the ordering reverses at $P = 80$, where GES (0.1436) edges out the proxy (0.1385) and PC (0.1313). Paired Wilcoxon signed-rank tests on per-regulator AP across the 12 paired regulators give faithful-versus-proxy $p = 0.129$ (PC) and $p = 0.910$ (GES) at $P = 60$, and $p = 0.204$ and $p = 0.569$ at $P = 80$; all four bootstrap 95 % CIs on the mean paired difference span zero. **The same test says random ranking is not significantly worse than the proxy either**: $p = 0.791$ at $P = 60$ (mean difference −0.0257, CI $[-0.109, +0.031]$) and $p = 0.677$ at $P = 80$ (−0.0480, CI $[-0.146, +0.017]$). At 12 evaluable regulators this benchmark cannot separate a genuine constraint-based causal search from a partial-correlation heuristic, and cannot separate either from chance. Absolute AP is much higher in these matched universes than in Table S8 because all positives are retained in a smaller candidate set, raising prevalence far above the 2.03 % of the full eligible universe; only within-universe ratios and paired differences are comparable across universes.

Depth-capped PC on the full universe. The only faithful run that completes on all 173 genes is depth-capped PC. At $\max k = 3$ it takes 1,704.8 s and recovers a 1,042-edge CPDAG (1,034 directed, 8 undirected) out of 14,878 possible pairs, scoring macro-AP 0.0798 against the proxy's 0.0946 (random 0.0477); at $\max k = 2$ it takes 1,214.5 s and scores 0.0647. These are depth-capped, not unrestricted, PC, and are labelled as such in the results file so they cannot be mistaken for the full algorithm.

Table S37: Faithful PC and GES benchmark, complete result file. $n = 52$ rows; the analysis unit is one run as recorded in the faithful PC/GES results table in the reproducibility package. All rows are K562-GS × TRRUST. "Tie" is the tie-breaking convention within an ordinal CPDAG tier: *rand.* averages 50 random draws, $|r|$ breaks ties by absolute marginal Pearson correlation. **Timeouts are timeouts, not zeros**: the `Runtime` column for a timed-out row is the budget that was exhausted, and no metric is imputed. "Ratio" is macro-AP divided by the random baseline recomputed on that row's own gene universe, so it is comparable only within a universe. The two "published" rows are carried over from the 1,024-gene benchmark for orientation.

Method	Tie	Status	Cells	Genes	Runtime (s)	Adj.	Dir.	Undir.	Regs	macro-AP	P@10	P@50	MRR	Ratio
proxy_partial_correlation_-intervention		published	–	1,024	–	–	–	–	12	0.0946	0.0667	0.0300	0.1828	1.98
random_published_-1024gene		published	–	1,024	–	–	–	–	12	0.0477	0.0195	0.0198	0.0745	1.00
random		success	–	173	–	–	–	–	12	0.0477	0.0195	0.0198	0.0745	1.00
PC_faithful		*timeout*	2,000	173	1,800.0	–	–	–	0	–	–	–	–	–
PC_faithful		*timeout*	2,000	173	14,400.0	–	–	–	0	–	–	–	–	–
PC_faithful		*timeout*	5,000	173	1,800.0	–	–	–	0	–	–	–	–	–
PC_faithful		*timeout*	20,000	173	1,800.0	–	–	–	0	–	–	–	–	–
PCdepth1_faithful		*timeout*	20,000	173	1,800.0	–	–	–	0	–	–	–	–	–
PCdepth2_faithful	rand.	success	2,000	173	1,214.5	1,267	1,263	4	12	0.0647	0.0405	0.0285	0.1122	1.36
PCdepth2_faithful		*timeout*	20,000	173	1,800.0	–	–	–	0	–	–	–	–	–
PCdepth2_faithful_-corrtier	$\|r\|$	success	2,000	173	1,214.5	1,267	1,263	4	12	0.0773	0.0667	0.0333	0.1499	1.62
PCdepth3_-alpha3.36067e-06_-faithful		*timeout*	20,000	173	1,800.0	–	–	–	0	–	–	–	–	–
PCdepth3_faithful	rand.	success	2,000	173	1,704.8	1,042	1,034	8	12	0.0798	0.0610	0.0300	0.1405	1.67
PCdepth3_faithful		*timeout*	20,000	173	1,800.0	–	–	–	0	–	–	–	–	–

Table S37 continued

Method	Tie	Status	Cells	Genes	Runtime (s)	Adj.	Dir.	Undir.	Regs	macro-AP	P@10	P@50	MRR	Ratio
PCdepth3_faithful		*timeout*	20,000	173	14,400.0	–	–	–	0	–	–	–	–	–
PCdepth3_faithful_-corrtier	$\|r\|$	success	2,000	173	1,704.8	1,042	1,034	8	12	0.0831	0.0750	0.0333	0.1576	1.74
GES_faithful		*timeout*	2,000	173	1,800.0	–	–	–	0	–	–	–	–	–
GES_faithful		*timeout*	2,000	173	14,400.0	–	–	–	0	–	–	–	–	–
GES_faithful		*timeout*	5,000	173	1,800.0	–	–	–	0	–	–	–	–	–
GES_faithful		*timeout*	20,000	173	1,800.0	–	–	–	0	–	–	–	–	–
GES_faithful		*timeout*	20,000	173	14,400.0	–	–	–	0	–	–	–	–	–
GESmaxP10_faithful		*timeout*	2,000	173	1,800.0	–	–	–	0	–	–	–	–	–
GESmaxP10_faithful		*timeout*	20,000	173	1,800.0	–	–	–	0	–	–	–	–	–
GESmaxP3_faithful		*timeout*	2,000	173	1,800.0	–	–	–	0	–	–	–	–	–
GESmaxP3_faithful		*timeout*	20,000	173	1,800.0	–	–	–	0	–	–	–	–	–
GESmaxP5_faithful		*timeout*	2,000	173	1,800.0	–	–	–	0	–	–	–	–	–
GESmaxP5_faithful		*timeout*	20,000	173	1,800.0	–	–	–	0	–	–	–	–	–
GES_feasP100_faithful		*timeout*	2,000	100	1,800.0	–	–	–	0	–	–	–	–	–
GES_feasP100_faithful		*timeout*	20,000	100	1,800.0	–	–	–	0	–	–	–	–	–
GES_feasP60_faithful	rand.	success	2,000	60	913.2	175	172	3	12	0.1484	0.0720	0.0600	0.2131	1.20
GES_feasP60_faithful		*timeout*	5,000	60	1,800.0	–	–	–	0	–	–	–	–	–
GES_feasP60_faithful		*timeout*	20,000	60	1,800.0	–	–	–	0	–	–	–	–	–
GES_feasP60_faithful_-corrtier	$\|r\|$	success	2,000	60	913.2	175	172	3	12	0.1351	0.0833	0.0633	0.1988	1.09
GES_feasP80_faithful	rand.	success	2,000	80	1,022.9	230	229	1	12	0.1436	0.0760	0.0466	0.2446	1.59
GES_feasP80_faithful		*timeout*	20,000	80	1,800.0	–	–	–	0	–	–	–	–	–
GES_feasP80_faithful_-corrtier	$\|r\|$	success	2,000	80	1,022.9	230	229	1	12	0.1418	0.0750	0.0533	0.2106	1.57
PC_alpha3.36067e-06_-faithful		*timeout*	20,000	173	1,800.0	–	–	–	0	–	–	–	–	–
PC_feasP100_faithful		*timeout*	2,000	100	1,800.0	–	–	–	0	–	–	–	–	–
PC_feasP100_faithful		*timeout*	20,000	100	1,800.0	–	–	–	0	–	–	–	–	–
PC_feasP60_faithful	rand.	success	2,000	60	292.7	184	177	7	12	0.1660	0.0995	0.0602	0.2788	1.34
PC_feasP60_faithful		*timeout*	5,000	60	1,800.0	–	–	–	0	–	–	–	–	–

Table S37 continued

Method	Tie	Status	Cells	Genes	Runtime (s)	Adj.	Dir.	Undir.	Regs	macro-AP	P@10	P@50	MRR	Ratio
PC_feasP60_faithful		*timeout*	20,000	60	1,800.0	–	–	–	0	–	–	–	–	–
PC_feasP60_faithful_-corrtier	$\|r\|$	success	2,000	60	292.7	184	177	7	12	0.1587	0.1000	0.0633	0.2539	1.28
PC_feasP80_faithful	rand.	success	2,000	80	1,061.1	246	239	7	12	0.1313	0.0842	0.0462	0.2288	1.45
PC_feasP80_faithful		*timeout*	20,000	80	1,800.0	–	–	–	0	–	–	–	–	–
PC_feasP80_faithful_-corrtier	$\|r\|$	success	2,000	80	1,061.1	246	239	7	12	0.1398	0.0917	0.0533	0.2673	1.55
proxy_rescored_-feasP100		success	–	100	–	–	–	–	12	0.1119	0.0750	0.0383	0.2000	1.53
proxy_rescored_feasP60		success	–	60	–	–	–	–	12	0.1493	0.0833	0.0600	0.2417	1.21
proxy_rescored_feasP80		success	–	80	–	–	–	–	12	0.1385	0.0833	0.0383	0.2316	1.53
random_feasP100		success	–	100	–	–	–	–	12	0.0729	0.0312	0.0352	0.1075	1.00
random_feasP60		success	–	60	–	–	–	–	12	0.1236	0.0610	0.0597	0.1818	1.00
random_feasP80		success	–	80	–	–	–	–	12	0.0904	0.0435	0.0427	0.1242	1.00

S15 Reproducibility

Materials available for reproduction. The analyses were carried out in two successive analysis workspaces, both mirroring the project code base and writing their outputs to a fixed set of numbered result directories. The audit, code-trace, corrected and inferential experiments (Sections S3, S8, S9, S10) were run in the later workspace; the descriptive benchmark, top-k tensor, direct-versus-indirect, PCA back-projection, synthetic calibration, reference-overlap, repeated-run and faithful PC/GES experiments (Sections S1–S7, S11–S14) were run in the earlier one.

The reproducibility package accompanying this supplement contains, for every section of this document, the analysis code that produced that section's inputs together with the machine-written result tables that code emits. It provides: the method-implementation-truth and edge-semantics tables, the code-audit log and the faithful-method availability test results behind Sections S1 and S2; the eligible-universe, evaluable-regulator and universe-terminology tables behind Section S3; the per-regulator and macro-averaged regulator-benchmark results behind Section S4; the complete top-k tensor with its universal-claims audit and run log behind Section S5; the direct-versus-indirect results and run log behind Section S6; the PCA back-projection sweep results and analysis note behind Section S7; the original and corrected sample-size sweep results, the corrected-sweep code and the count-trace note and its reproduction script behind Section S8; the hierarchical random-baseline rows, the aggregate bootstrap summary, the mixed-model summary and the per-reference breakdown behind Section S9; the frozen matched synthetic-versus-real results table (the artifact of record, sha256 `46dea8a4af1a`...), its synthetic-universe table and its report behind Section S10; the expanded synthetic-suite results and run log behind Section S11; the reference-overlap results and run log behind Section S12; the all-method uncertainty results with both run logs behind Section S13; and the faithful PC/GES results behind Section S14. Each analysis script takes no command-line arguments, so an invocation is fully specified by the script itself.

Two provenance details of that package matter for reading the tables above. The repeated-run experiment was executed twice: the first pass was terminated during the GENIE3 seed-43 fit, and a resumable second pass restarted from the partial results and appended each completed run immediately, which is why the recorded uncertainty table is incomplete (Table S35). Every figure is produced by a single figure-building routine that writes, beside each figure, a sidecar table containing exactly the numbers plotted; a companion verification routine gates the build against those sidecars.

Environment. All runs used a single machine, Linux 6.17.0-35, with a single dedicated Python virtual environment (Python 3.12.3). Package versions, as recorded in the captured environment record of the reproducibility package and verified with `importlib.metadata`, are listed in Table S38. Two of these matter for interpretation rather than reproduction. `decoupler` 2.1.6 supplies the DoRothEA A+B and CollecTRI reference networks used to build the eligible universe (Section S3) and the top-k tensor (Section S5); a different `decoupler` release ships different regulon contents and would change the reference edge counts. `causal-learn` 0.1.4.8 provides faithful PC and GES. **Correction:** an earlier version of this supplement stated that no benchmark result uses `causal-learn`. That is wrong, and it contradicted its own faithful PC/GES table. Every row of Table S37 in Section S14 — 52 rows covering unrestricted PC, unrestricted GES, their depth- and parent-capped variants, the feasible matched universes and the recorded timeouts — is produced by `causal-learn` 0.1.4.8. What is true, and what the earlier sentence was reaching for, is that no row of the *primary* benchmark (Sections S4–S6) uses it: the pipeline labels "PC" and "GES" there are the Graphical-Lasso partial-correlation proxy, which is exactly why Section S14 exists.

Table S38: Software environment. $n = 13$ recorded components; the analysis unit is one package.

Component	Version
Operating system	Linux 6.17.0-35-generic
Python	3.12.3
causal-learn	0.1.4.8
decoupler	2.1.6
scikit-learn	1.8.0
torch	2.11.0+cu128
xgboost	3.2.0
numpy	2.4.3
scipy	1.16.3
pandas	2.3.3
statsmodels	0.14.6
scanpy	1.12.1
anndata	0.12.10
torch-geometric	2.7.0

Determinism and provenance. Every stochastic component is seeded: the random-forest proxies use `random_state=42`, the sample-size sweep uses seeds 42–44, the PCA back-projection sweep uses seeds 42, 123 and 456, and the repeated-run experiment was configured for seeds 42–46. The provenance record of the reproducibility package lists SHA-256 digests of all 30 pipeline source, configuration and input files, together with the digest of the starting manuscript and the captured environment. The project working area was not under version control, so file digests rather than commit hashes serve as the provenance record, and this is stated explicitly in that record.

Input data. The three Replogle et al. [2022] Perturb-seq datasets are read as published single-cell expression objects and, for the benchmark runs, as preprocessed dense arrays of 20,000 cells × 1,024 highly variable genes each. TRRUST v2 is read from its published human regulator–target table; DoRothEA A+B and CollecTRI are retrieved through `decoupler` 2.1.6.

References

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

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9:432–441, 2008. doi: 10.1093/biostatistics/kxm045.

Luz Garcia-Alonso, Christian H Holland, Mahmoud M Ibrahim, et al. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Research*, 29: 1363–1375, 2019. doi: 10.1101/gr.240663.118.

Heonjong Han, Jae-Won Cho, Sangyoung Lee, Ayoung Yun, Hyojin Kim, Dasom Bae, Sunmo Yang, Chan Yeong Kim, Muyoung Lee, Eunbeen Kim, et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Research*, 46 (D1):D380–D386, 2018. doi: 10.1093/nar/gkx1013.

Vân Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts. Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE*, 5:e12776, 2010. doi: 10.1371/journal.pone.0012776.

Daniel Marbach, James C Costello, Robert Küffner, et al. Wisdom of crowds for robust gene network inference. *Nature Methods*, 9:796–804, 2012. doi: 10.1038/nmeth.2016.

Thomas Moerman, Sara Aibar Santos, Carmen Bravo González-Blas, et al. GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics*, 35:2159–2161, 2019. doi: 10.1093/bioinformatics/bty916.

Sophia Müller-Dott, Eirini Tsirvouli, Miguel Vazquez, Ricardo O Ramirez Flores, Pau Badia-i Mompel, Robin Fallegger, Dénes Türei, Astrid Lægreid, and Julio Saez-Rodriguez. Expanding the coverage of regulons from high-confidence prior knowledge for accurate estimation of transcription factor activities. *Nucleic Acids Research*, 51(20):10934–10949, 2023. doi: 10.1093/nar/gkad841.

Aditya Pratapa, Amogh P Jalihal, Jeffrey N Law, Aditya Bharadwaj, and T M Murali. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods*, 17:147–154, 2020. doi: 10.1038/s41592-019-0690-6.

Joseph M Replogle, Reuben A Saunders, Angela N Pogson, et al. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell*, 185:2559–2575, 2022. doi: 10.1016/j.cell.2022.05.013.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2nd edition, 2000.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. DAGs with NO TEARS: Continuous optimization for structure learning. In *NeurIPS*, pages 9472–9483, 2018.