

Monte Carlo simulation and breast cancer transcriptomic benchmarking characterize reliability limits of high-dimensional biomedical feature selection

Hao Ma

Jinzhou Medical University

Jiaxin Yin

yinjx@jzmu.edu.cn

Jinzhou Medical University

Article

Keywords: feature selection, reliability, Monte Carlo simulation, transcriptomics, support recovery, regularization

Posted Date: October 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11147390/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Reliable feature selection requires adequate signal recovery, few false selections and reproducible feature sets. We evaluated these requirements jointly using Monte Carlo simulation and breast cancer transcriptomic benchmarking. Breast cancer expression matrices supplied empirical covariates; signal supports and binary outcomes were artificially generated. Joint success required true-positive rates ≥ 0.80 and false-discovery proportions ≤ 0.10 in both members of a data pair, with Jaccard similarity ≥ 0.75 . Broad screening covered 624 configurations and nine methods, followed by independent confirmation. The principal transcriptomic benchmark compared composite minimax concave penalty (cMCP) and fixed-mixing-parameter Elastic Net across 2,000 and 22,092 genes. With low-correlation supports and 1,000 patients per side, cMCP achieved 296 to 299 successes in 300 repetitions; high-correlation supports yielded none. Conservative tuning improved reliability when recall remained sufficient, but reduced success from 93/300 to 29/300 in a 2,000-gene, 500-patient setting. Weaker and heterogeneous effects further restricted recovery. These findings characterize conditional reliability limits rather than a universal method ranking. Joint evaluation and component-level diagnostics identify when sample size, support correlation, signal distribution and regularization permit reproducible recovery of known features.

Introduction

Feature selection is used throughout high-dimensional biomedical research to reduce large measurements to a manageable set of candidate variables. For a selected set to support subsequent investigation, it should recover a substantial fraction of the relevant signals, contain relatively few irrelevant variables and recur when the analysis is repeated with new data. These requirements concern different properties of the output. Good prediction alone does not establish accurate feature recovery, and repeated selection alone does not establish that the selected features are relevant. A reliability evaluation therefore needs a defined recovery target and a way to assess these properties together.

Gene expression provides a useful setting for studying this problem. Correlated genes may substitute for one another in a prediction model, allowing similar predictive performance to coexist with substantially different selected sets. Early work on cancer expression signatures demonstrated that the resulting gene lists could depend strongly on the sampled patients¹. Comparisons of molecular signatures have consequently examined accuracy, stability and interpretability as separate outcomes². Formal work on selection stability has also shown why the properties of the similarity measure and the number of selected variables matter³. When both analyses select nothing, for example, their sets agree perfectly without recovering any signal. Evaluating stability together with recovery and false selection makes that distinction explicit.

Regularized regression offers several ways to construct sparse candidate sets. LASSO induces zero coefficients, Elastic Net combines penalties to accommodate correlated predictors, and the minimax concave penalty (MCP) changes the shrinkage applied to nonzero coefficients^{4, 5, 6}. Group penalties

incorporate structure among variables, while bi-level selection permits sparsity both between and within groups^{7,8,9}. Their recovery properties depend on the relationship between the signal and the design matrix. Support-recovery theory identifies requirements under specified model assumptions^{10,11}, but finite-sample performance under empirical dependence structures remains a question for direct evaluation. The relevant comparison is therefore between fully specified procedures operating under defined data conditions.

Tuning adds another source of variation. The minimum cross-validated deviance rule selects the penalty with the lowest estimated predictive loss. The one-standard-error rule selects a stronger penalty within an uncertainty tolerance around that minimum. The latter can remove false selections but may also discard true signals. Stability selection and complementary-pairs stability selection (CPSS) address erroneous selection through repeated subsampling under their respective assumptions^{12,13}. Those error-control targets are distinct from the probability that two fitted feature sets jointly meet specified recall, false-selection and overlap requirements. We use a joint event to evaluate that probability directly.

This study examines the reliability limits of high-dimensional feature selection through Monte Carlo simulation and empirical-covariate benchmarking. Broad diagnostic simulations identify output-size constraints and candidate recoverable settings; independent repetitions then assess sample-size patterns and tuning effects. Breast cancer expression matrices provide realistic covariate dependence, while artificially specified supports and simulated binary outcomes make the recovery target observable. The principal transcriptomic experiment evaluates 2,000 and 22,092 candidate genes using nonoverlapping patient pools. Further experiments vary signal strength, effect heterogeneity and synthetic dependence structure. Following simulation-study guidance^{14,15}, we distinguish diagnostic screening, independent confirmation and secondary analyses, and interpret reliability through the components of the joint endpoint.

The empirical-covariate component has a specific methodological role. A fully parametric generator allows direct control over a covariance model, whereas a statistical plasmode resamples observed covariates and applies a specified outcome-generating model¹⁶. The latter retains aspects of the original feature dependence while supplying known coefficients for evaluating support recovery. Its realism depends on both parts of this construction: empirical predictors do not make an artificial outcome an observed clinical endpoint. We therefore report the source matrices, reference-set calibration, artificial supports and sampling scheme as separate design elements. This separation makes the simulation target explicit while allowing the biomedical covariance structure to inform the comparison.

Results

Joint evaluation identifies recoverable conditions and output constraints

The study combined broad diagnosis with focused independent confirmation (Table 1; Fig. 1). The diagnostic stage comprised 624 configurations, nine methods and 30 paired repetitions per configuration. Of 5,616 method-configuration combinations, 76 produced at least one joint success and 17 had estimated reliability of at least 0.80. These latter combinations involved CPSS or nongrouped MCP. The diagnostic map guided subsequent investigation of sample size and regularization, while independent repetitions provided the main probability estimates. Computational failures and non-estimable results remained in the denominators throughout.

Table 1
Study design and evidence roles.

Stage	Design	Pairs per cell	Evidence role
Broad diagnostic simulation	624 configurations; 9 methods	30	Feasibility and failure mapping
Local sample-size confirmation	Selected n; CPSS or MCP	270 or 300 new	Local sample-size patterns
cMCP tuning confirmation	$p = 200$; $s = 40$; $n = 1,000$	300	Paired tuning effect
Correlation/sample-size transfer	$p = 200$; $s = 40$; $2 n \times 2 \rho$	300	4-cell confirmation
GSE96058	200 genes; 40 signals; $n = 500/1000$	300	4-cell breast cancer plasmode
GSE96058 high-dimensional extension	2,000/22,092 genes; 20 signals; $n = 500/1,000$	300	8-cell paired confirmation
Signal-realism module	2,000 genes; $AUC = 0.85$; equal/heterogeneous effects	300	4-cell secondary confirmation
GSE7390	200 genes; $s = 10/40$; $n = 50$ to 150	60	8-cell breast cancer plasmode
Alternative DGPs	$p = 1000$; $s = 10$; $3 \text{ DGP} \times 2 n \times 2 \text{ AUC}$	60 or 300	12-cell confirmation
Note. n denotes observations per synthetic side, bootstrap draws in the original 200-gene empirical experiments, and distinct patients per side in the high-dimensional experiment. Screening data are excluded from independent confirmation estimates.			

A simple output-size requirement explained some failures before model fitting. For a support containing s true features, simultaneous recall of at least 0.80 and a false-discovery proportion of at most 0.10 require the selected-set size k to satisfy $\text{ceiling}(0.80s) \leq k \leq \text{floor}(s/0.90)$. The fixed output of minimum-redundancy maximum-relevance selection violated this necessary range in 544 of 624 configurations. The maximum number of candidates available to the cluster-representative procedure was below the recall requirement in 166 configurations. Thus, a feature selector can be structurally incompatible with the evaluation target even when its fitted predictions are useful.

The joint endpoint also separated reproducibility from successful recovery. Among 165,611 normally completed diagnostic method-pair results, 50,509 were stable but did not satisfy the accuracy requirements. Most of these were stable empty sets, accounting for 49,268 results. Only 987 pairs satisfied both sides' recall and false-selection requirements, and 928 also passed the Jaccard threshold. In this diagnostic region, achieving sufficient recovery with few false selections was the main obstacle. Full failure counts, including 2,839 non-estimable results and 30 computational failures, are retained in the Supplementary Information.

Independent sample-size analyses showed that recovery could change sharply within a fixed design. For CPSS in setting A, success increased from 130/300 at $n = 325$ to 266/270 at $n = 400$; the corresponding Wilson 95% intervals were 37.8% to 49.0% and 96.3% to 99.4%. For MCP in setting B, improvement was more gradual, reaching 229/270 at $n = 1,000$. These curves used newly generated repetitions at sample sizes selected after screening (Supplementary Fig. S1). They establish local patterns for the specified signals and dependence structures, rather than a common sample-size requirement across methods.

Conservative tuning improves reliability when recall is sufficient

The dedicated cMCP tuning experiment provides a clear example of how reducing false selection can improve the joint endpoint. In setting C, the minimum-deviance rule succeeded in 233/300 pairs and the one-standard-error rule in 286/300 pairs, an increase from 77.7% to 95.3% (Fig. 2). There were 53 pairs in which only the one-standard-error rule succeeded and none with the opposite outcome. The paired difference was 17.7 percentage points, with a conservative 95% interval of 11.5 to 23.2 points. Both rules met the two-sided recall requirement in all 300 pairs.

The continuous measures explain this improvement. Stronger penalization reduced the mean number selected from 41.97 to 40.83 and mean FDP from 0.048 to 0.028. Mean TPR changed only slightly, from 0.997 to 0.991. The procedure retained enough true features for the recall threshold while removing selections that jeopardized the FDP requirement. A separate four-cell synthetic confirmation reproduced the improvement at $n = 1,000$ and within-module correlation zero: successes increased from 230/300 to 286/300, with a four-comparison interval for the difference of 11.0 to 25.5 percentage points. At correlation 0.60, the difference was smaller and its interval included zero.

The 200-gene GSE96058 experiment showed a similar mechanism under empirical expression dependence. At 1,000 draws per side, the one-standard-error rule increased successes from 251/300 to 291/300 for the high-correlation artificial support, and from 245/300 to 287/300 for the low-correlation support. The paired gains were 13.3 and 14.0 percentage points, with four-comparison intervals of 6.6 to 19.5 and 6.7 to 20.7 points. For the high-correlation support, mean TPR remained 0.992 while mean FDP decreased from 0.046 to 0.027. These results concern a 200-gene task with 40 artificial signals and bootstrap sampling; the high-dimensional benchmark below uses a different support and patient-sampling design.

High-dimensional transcriptomic reliability depends on support and sample size

The principal GSE96058 benchmark evaluated two candidate spaces, two support-correlation conditions and two sample sizes. It used 20 artificial signals, 300 paired repetitions per setting and four selection rules: minimum-deviance and one-standard-error tuning for cMCP and for Elastic Net with alpha fixed at 0.5. Each data pair contained patients sampled without replacement from separate left and right pools. The 2,000-gene and 22,092-gene conditions shared patients, generated outcomes and fold assignments within matched repetitions. This design isolates the candidate-space comparison within each support and sample-size condition (Table 2; Fig. 3).

Table 2
Joint reliability in the high-dimensional GSE96058 confirmation.

Candidate space	Support	n per side	cMCP min	cMCP 1se	Elastic Net min	Elastic Net 1se
2,000 genes	High-correlation	500	0/300 (0.0%)	0/300 (0.0%)	0/300 (0.0%)	0/300 (0.0%)
2,000 genes	High-correlation	1,000	0/300 (0.0%)	0/300 (0.0%)	0/300 (0.0%)	0/300 (0.0%)
2,000 genes	Low-correlation	500	93/300 (31.0%)	29/300 (9.7%)	0/300 (0.0%)	0/300 (0.0%)
2,000 genes	Low-correlation	1,000	297/300 (99.0%)	299/300 (99.7%)	0/300 (0.0%)	0/300 (0.0%)
22,092 genes	High-correlation	500	0/300 (0.0%)	0/300 (0.0%)	0/300 (0.0%)	0/300 (0.0%)
22,092 genes	High-correlation	1,000	0/300 (0.0%)	0/300 (0.0%)	0/300 (0.0%)	0/300 (0.0%)
22,092 genes	Low-correlation	500	0/300 (0.0%)	0/300 (0.0%)	0/300 (0.0%)	0/300 (0.0%)
22,092 genes	Low-correlation	1,000	296/300 (98.7%)	297/300 (99.0%)	0/300 (0.0%)	0/300 (0.0%)
Note. The joint endpoint requires both sides to achieve TPR at least 0.80 and FDP at most 0.10, with Jaccard at least 0.75. Exact intervals and continuous metrics are reported in Table S7. Elastic Net used prespecified alpha = 0.5.						

At 1,000 patients per side, cMCP recovered the low-correlation support reliably in both candidate spaces. Successes were 297/300 and 299/300 for the minimum-deviance and one-standard-error rules in 2,000 genes, and 296/300 and 297/300 in 22,092 genes. The corresponding paired differences between tuning rules were only 0.7 and 0.3 percentage points, with eight-comparison intervals including zero. The full-gene task had 22.092 candidates per patient, demonstrating that high dimension alone did not prevent

recovery under these particular signal conditions. Candidate dimension must be interpreted together with the support structure and the information available per signal.

The same comparison at 500 patients per side exposed a different regime. In 2,000 genes, cMCP minimum-deviance tuning succeeded in 93/300 pairs, compared with 29/300 for the one-standard-error rule. There were 67 minimum-only successes and three one-standard-error-only successes. Stronger penalization reduced reliability by 21.3 percentage points, with an eight-comparison interval of -30.1 to -11.6 points (Fig. 4). Mean TPR decreased from 0.810 to 0.701. Expanding the candidate space to 22,092 genes yielded no successes for either rule. Thus, the conditions that allowed stronger penalization to help in the earlier experiments did not hold in this lower-sample high-dimensional setting.

Support correlation also separated the recovery regimes. Neither cMCP nor the fixed-alpha Elastic Net baseline achieved a joint success for the high-correlation support at either dimension or sample size. In the low-correlation, 1,000-patient settings, Elastic Net approached complete recall but had mean FDP between 0.780 and 0.916, accounting for its zero joint successes. These component results identify different routes to the same joint outcome: inadequate recovery for some settings and excessive false selection for others. Complete proportions, exact intervals and continuous measures are supplied in Supplementary Table S7. Zero observed success is reported with its positive confidence bound, rather than treated as a zero underlying probability.

Signal distribution and dependence structure delimit tuning gains

The secondary signal experiments held the candidate space at 2,000 genes and the sample size at 1,000 per side. Under the low-correlation support and a target oracle AUC of 0.85, equal-magnitude effects gave 2/300 successes for minimum-deviance cMCP and 75/300 for the one-standard-error rule. The paired gain was 24.3 percentage points, with a four-comparison interval of 16.0 to 31.7 points. Here, reducing false selections produced a substantial relative improvement, although absolute reliability remained well below the strong-signal result. This distinction is important when comparing a favorable tuning contrast with the probability of meeting the actual recovery target.

Keeping the target AUC at 0.85 but introducing a 4:1 ratio between the largest and smallest effect magnitudes reduced one-standard-error successes to 1/300. Its difference relative to the equal-effect condition was -24.7 percentage points, with a 16-comparison interval of -33.2 to -14.9 points. Therefore, the same nominal discrimination level can conceal markedly different support-recovery difficulty. The signal distribution affects whether small coefficients survive penalization, even when the outcome model as a whole remains informative. The complete four-cell results appear in Supplementary Table S8 and Supplementary Fig. S2.

Alternative synthetic mechanisms confirmed that tuning effects depend on the signal's placement within the correlation structure. At $n = 1,000$, $p = 1,000$ and target AUC = 0.95, both cMCP rules performed well under the latent-factor/batch and module mechanisms. Successes were 300/300 for both rules in the

former and 297/300 versus 300/300 in the latter. In the AR(1) design, where the ten signals occupied adjacent correlated positions, minimum-deviance tuning achieved 286/300 successes and the one-standard-error rule 248/300. The paired difference was – 12.7 percentage points, with a 12-comparison interval of – 21.5 to – 3.2 points. Full results across all 12 settings and four methods are retained in Supplementary Table S4 and Supplementary Fig. S3.

The AR(1) component analysis explains the reversal. Stronger penalization lowered mean TPR from 0.996 to 0.941. The minimum-deviance rule had ten stable false-selection failures and four unstable results; the one-standard-error rule had one stable false-selection failure and 51 unstable results. The latter rule also failed the two-sided recall requirement in 51 pairs. Removing extra variables reduced false-selection failures but increased omissions and disagreement enough to lower joint reliability. Supplementary Fig. S4 displays these mutually exclusive categories together with the weak-signal settings.

Sensitivity analyses support a conditional interpretation

Re-evaluation under 27 combinations of recall, FDP and Jaccard thresholds preserved the main high-dimensional tuning pattern. In the 2,000-gene, low-correlation, 500-patient setting, all 27 point differences favored minimum-deviance tuning; 15 remained below zero after correction across 216 comparisons. No corrected high-dimensional threshold contrast significantly favored the one-standard-error rule. The earlier 200-gene experiment showed broader positive tuning differences at 1,000 draws per side, indicating that robustness to threshold choice was itself conditional on the task (Supplementary Figs. S5 and S6).

The smaller GSE7390 empirical benchmark and module-level analyses provided additional boundary checks. All 32 GSE7390 method-setting combinations had 0/60 successes, with a single-cell exact upper bound of 5.96%. Under the synthetic module mechanism, allowing selection of a correlated substitute modestly improved some small-sample results, while preserving the main strong-signal conclusions. These analyses are reported in full in the Supplementary Information. Together, they show why sample size, the evaluation unit and false-selection burden need to be specified when describing a procedure as reliable.

Discussion

This study evaluates high-dimensional feature selection as a joint recovery problem. The central result is a set of conditional reliability limits: some designs permit repeated recovery of a sufficiently complete and precise feature set, whereas others produce stable empty outputs, excessive false selections or inconsistent omissions. The proposed endpoint gives these outcomes a common interpretation by requiring recall and false-selection control on both sides of a data pair before accepting their agreement as reliable. Its value is diagnostic as well as comparative, because the component measures identify which part of the recovery objective remains unmet.

The high-dimensional transcriptomic benchmark provides the clearest illustration. In 22,092 genes, cMCP achieved 296 or 297 successes among 300 repetitions for a low-correlation support with 1,000 patients per side. Under a high-correlation support, the same sample size and target signal calibration produced no joint successes. Reducing the sample size also changed the effect of both candidate dimension and penalization. These comparisons indicate that a ratio such as p/n is insufficient to summarize recovery difficulty. The locations of nonzero effects, their relation to correlated alternatives and their magnitude distribution determine what information the selector must distinguish.

The tuning experiments connect those design conditions to an actionable statistical mechanism. When recall was nearly complete, the one-standard-error rule could remove enough extra variables to improve the joint endpoint without sacrificing the required signals. When recall was already close to the threshold, the same strategy caused more omissions and less agreement. The original 200-gene benchmark, the 2,000-gene smaller-sample comparison and the AR(1) confirmation therefore describe different operating regions of a single trade-off. A recommendation for stronger regularization should depend on whether false selection is the dominant remaining constraint and whether enough recall is preserved.

This interpretation also separates successful tuning from adequate absolute performance. The weaker, equal-effect experiment showed a clear gain for the one-standard-error rule, but only 75/300 repetitions met the joint target. A favorable contrast between two procedures does not establish that either is sufficiently reliable for the intended task. Conversely, when both procedures achieved almost complete success, their small tuning difference was of limited practical importance. Reporting the joint success proportion alongside paired contrasts prevents a method comparison from displacing the underlying question of how often the required output is obtained.

The results extend the motivation of expression-signature studies that considered accuracy, stability and interpretability together^{1, 2, 3}. They are also consistent with broader gene-expression benchmarks that evaluate multiple properties of a selection pipeline¹⁷. Our use of an artificial support makes missed and false selections directly measurable, while the empirical covariates preserve expression dependence. This combination supports a controlled assessment of recovery under realistic covariate structure. The biomedical role of the breast cancer datasets is to supply that structure; selected genes in these experiments derive their signal status from the specified outcome generator.

Several distinctions are important for using this framework. First, the benchmark's oracle AUC measures the discrimination of the known data-generating model, not the predictive performance of a fitted selector. Equal target AUCs do not imply equal difficulty for recovering every nonzero effect, as the heterogeneous-effect experiment demonstrates. Second, a feature-level target and a module-level target answer different questions. Allowing correlated substitutes can change recovery counts, but many irrelevant modules can still make the resulting output imprecise. Third, fixed method settings define the comparison. The Elastic Net results describe $\alpha = 0.5$ with the implemented tuning rules, rather than an exhaustive search over all possible configurations.

The one-standard-error rule itself uses uncertainty in cross-validated prediction loss; the joint support-recovery criteria do not enter its optimization. Prior work on the rule has similarly emphasized dependence on complexity, signal and design conditions¹⁸. In a known-support benchmark, one can inspect recall, FDP and overlap directly to identify an appropriate operating region. For an observed biomedical outcome, those quantities require additional assumptions or external knowledge because the true support is unknown. The present framework is therefore most directly useful for evaluating pipelines in simulations or scientifically informed benchmarks before interpreting a selected feature set in an applied study.

The empirical design has specific strengths and limits. A separate reference set fixed candidate genes, artificial supports, scaling and calibration before repeated evaluation. The high-dimensional study additionally used nonoverlapping patient pools and sampling without replacement within each side. Its paired comparisons shared patients and generated outcomes across candidate dimensions. These choices make the conditions of comparison explicit. The earlier 200-gene experiments used bootstrap draws and different supports; their results should consequently be read as complementary mechanism evidence rather than points on a continuous dimension-response curve. GSE7390 changes cohort, platform, reference-set size and sampling conditions together, so its lower recovery cannot be attributed to a single factor.

Monte Carlo precision and analysis identity also affect the claims. Broad screening used 30 pairs per configuration and served to locate patterns. Main probability estimates came from independent confirmation, with the specified denominators and multiplicity families retained. Finite repetitions leave uncertainty even at zero or complete success, and values close to the 0.80 reliability target can depend on the interval definition. The selected joint thresholds represent one explicit set of recovery requirements. Sensitivity analyses describe nearby choices without converting the most favorable threshold combination into a new primary endpoint. The original diagnostic generator's retry behavior and the corrected confirmation implementation are documented in the Supplementary Information.

The framework evaluates realized FDP within each pair and the probability of a joint event across repetitions. These quantities differ from false discovery rate averaged over repeated datasets. Guarantees associated with BH, stability selection or Model-X knockoffs apply to their own targets and assumptions^{12, 19, 20}; they do not automatically establish the joint criterion used here. Work on derandomized selection also illustrates the importance of algorithmic randomness for reproducibility²¹. The present implementation addresses comparability through shared data, controlled randomization and independent confirmation, while leaving other error-control procedures available for future evaluations under the same explicitly defined target.

In practical terms, reliability assessment should begin by checking whether a procedure can output a set of an appropriate size, then examine recall and false selection, and finally interpret agreement conditional on those requirements. The experiments support this sequence across controlled and empirical dependence structures. Sample size can improve recovery, but its benefit is conditioned by

support correlation and signal distribution; stronger shrinkage can help, but only when the retained signals remain sufficient. Joint reporting turns these dependencies into observable evaluation criteria for high-dimensional biomedical feature selection.

The joint endpoint complements, rather than replaces, the component measures. Nogueira and colleagues developed a statistical treatment of feature-selection stability with properties supporting formal comparison³. Our endpoint instead asks whether a paired recovery task passes a specified set of requirements, and retains Jaccard as one of those requirements. Its interpretation depends on the chosen cutoffs and the known support. Consequently, two procedures with identical joint proportions may still differ in how much recall, precision or agreement they achieve outside the successful region. The reported continuous measures and failure categories preserve this information and make the joint summary more useful for diagnosing a pipeline.

A further distinction concerns the unit of replication. The independent confirmation batches quantify repeated algorithm performance under a frozen design. They do not represent prospective replication of a biomarker association in new clinical cohorts. In the empirical experiments, repetitions draw on a finite patient pool and generate new outcomes from the specified model. This design answers the methodological recovery question while retaining the cohort's measured covariates. As recommended for simulation studies^{14,15}, keeping the data-generating mechanism, performance measure and repetition scheme explicit is essential to interpreting the estimated probability. The archived manifests and calibration records link each reported estimate to those choices.

Methods

Study design and endpoint

The completed study comprised broad diagnostic screening, local sample-size analyses, independent tuning confirmation, empirical-covariate benchmarks and alternative synthetic mechanisms (Table 1). Each repetition produced two datasets with the same true support S and selected sets A and B . For either set A , $TPR = |A \cap S| / |S|$ and $FDP = |A \setminus S| / \max(1, |A|)$. Jaccard similarity was $|A \cap B| / |A \cup B|$, defined as one when both sets were empty. Joint success required both $TPRs \geq 0.80$, both $FDPs \leq 0.10$ and Jaccard ≥ 0.75 . Reliability was the proportion of all K pairs satisfying this event. Computational failures and non-estimable outcomes counted as unsuccessful repetitions. The 0.80 per-side recall threshold is distinct from assessing whether the reliability probability itself exceeds 0.80.

Failures were partitioned into computational/non-estimable, unstable, stable empty, stable nonempty false selection and stable nonempty omission. Instability took precedence over the two nonempty accuracy categories; among stable nonempty outputs, an FDP violation took precedence over insufficient recall. Component failures were additionally counted without imposing exclusivity. For module-level recovery, any selected variable represented its signal module and background variables formed separate singleton modules; the same thresholds applied.

Selection procedures

The broad stage evaluated LASSO with minimum-deviance and one-standard-error tuning, Elastic Net at $\alpha=0.5$, nongrouped MCP, univariate testing with BH correction, minimum-redundancy maximum-relevance selection, CPSS, cluster-representative LASSO and Group LASSO. Focused confirmation compared Group LASSO, cMCP with both tuning rules, and nongrouped MCP. The high-dimensional benchmark compared the two cMCP and two Elastic Net rules. LASSO and Elastic Net used glmnet with binomial deviance and 50 penalty values²². Nongrouped MCP used ncvreg²³. Univariate screening used a correlation-based t approximation followed by BH correction at 0.10¹⁹. mRMR used mRMRe with $\min(p, \max(5, \text{floor}(\sqrt{p})))$ selected variables^{24, 25}.

For the earlier experiments, each data side was standardized and zero-variance features removed before fitting. Stratified cross-validation targeted three folds, reducing the fold count where required by class counts and recording failures when fitting was infeasible. Hierarchical clustering used distance $1 - |\text{Pearson correlation}|$, average linkage and $\max(2, \min(p, \text{floor}(\sqrt{p})))$ groups. Cluster-representative LASSO retained the variable with the strongest absolute outcome correlation in each group. Group LASSO and cMCP used grpreg²⁶ and the same grouping scheme. cMCP applies a composite penalty supporting both between-group and within-group selection⁹; grouping followed the data-dependent approach described above²⁷. CPSS used 20 complementary half-sample pairs, fitting one-standard-error LASSO on each of 40 halves, with selection frequency ≥ 0.80 .

Both cMCP tuning rules used the same validation path and folds. The minimum rule minimized mean validation deviance. The one-standard-error rule selected the largest penalty whose mean deviance did not exceed the minimum plus its standard error. Nonconvex selection sets need not be nested along the penalty path; results were evaluated from the saved selections. High-dimensional preprocessing and grouping differed as detailed below.

Monte Carlo designs

Broad screening covered 144 anchor and 480 Latin-hypercube configurations, including $n=50$ to 1,000, $p=200$ to 5,000, $s=5$ to 50, within-module correlation up to 0.90 and module size from one to ten. There were 30 pairs per configuration. The module generator placed each signal first in its module, drew effect signs separately on each side, scaled the linear predictor to the specified signal parameter and calibrated the intercept to the target event proportion. Setting A fixed $p=200$ and $s=10$ for CPSS sample-size evaluation; B fixed $p=213$ and $s=12$ for MCP; C fixed $p=200$ and $s=40$ for cMCP tuning. New repetitions numbered 270 at previously screened sample sizes or 300 at newly added levels. Dedicated tuning confirmation used 300 new pairs, as did each of four cells combining $n=500$ or 1,000 and correlation zero or 0.60. Full generator parameters and distinct experiment identities appear in Supplementary Sections S1 to S3.

Alternative mechanisms fixed $p=1,000$, $s=10$ and event proportion 0.30. AR(1) features had correlation 0.60 raised to their positional distance, with signals in the first ten positions. The latent-factor/batch design used three factors, loading 0.70 and a batch shift affecting every fourth feature, with dispersed signals. The module design used ten three-variable modules with within-module correlation 0.60 and 970 background variables. A 20,000-row reference sample calibrated each mechanism to oracle AUC 0.75 or 0.95 using fixed alternating effect signs; coefficients then remained fixed across sides and repetitions. Crossing mechanism, $n=300$ or 1,000 and AUC gave 12 cells. After 30-pair screening, cells with either cMCP reliability ≥ 0.20 received 300 independent confirmation pairs; other cells received 60. This yielded 5,760 method-pair results.

Breast cancer empirical-covariate benchmarks

GSE96058 RNA-sequencing and GSE7390 microarray data supplied empirical expression matrices^{28, 29}. These experiments used artificially defined nonzero coefficients and Bernoulli outcomes generated from logistic probabilities; measured clinical outcomes were not used as the selection target. This is an empirical-covariate or plasmode design¹⁶. Existing harmonization retained samples with identified ER records, reconciled identifiers and handled technical duplicates. ER availability affected upstream eligibility, while subsequent outcome generation was artificial. The harmonized matrices contained 3,189 samples and 23,065 features for GSE96058, and 198 samples and 12,343 unique genes for GSE7390.

For the 200-gene GSE96058 experiment, a fixed split assigned 637 samples to reference calibration and 2,552 to the sampling pool. The reference set determined the 200 highest-variance genes and two disjoint 40-gene supports with high or low mean absolute correlation. Alternating coefficient signs were fixed by gene identifier. Scaling, a common effect multiplier and an intercept were determined in the reference set, targeting theoretical AUC 0.95 and event proportion 0.50. Each side independently drew 500 or 1,000 samples with replacement from the sampling pool, then generated outcomes using the frozen model. Thus, n denotes draws, and original patients can recur within or across sides. Four cells each used 300 independent confirmation pairs after screening.

GSE7390 used 39 reference samples and a 159-sample pool. Candidate genes and supports were constructed within its own reference set. Eight cells combined ten signals with $n=50$, 100 or 150 and both support types, plus 40 signals at $n=150$. Each model was calibrated and frozen separately. The prespecified rule assigned 300 confirmation pairs if either cMCP screening reliability lay between 0.20 and 0.80, otherwise 60. All eight cells received 60; fitting failures remained in their denominators. The reference and sampling-pool differences are detailed in Supplementary Section S4.

High-dimensional benchmark and signal experiments

The principal GSE96058 experiment reused the reference set but deterministically hashed the 2,552 remaining patient identifiers into two disjoint pools of 1,276. Each side sampled 500 or 1,000 patients

without replacement from its own pool. Genes were ranked by reference-set variance. The top 2,000 formed the smaller space; all 22,092 reference-set positive-variance genes formed the full space. Each support contained 20 genes from the top 2,000, with at least 50 distinct reference expression values. This criterion and the supports were fixed before generating new outcomes. Both dimensions used the same support and frozen outcome model, calibrated to AUC 0.95 and event proportion 0.50.

Groups were estimated exclusively from reference covariates by average-linkage clustering of $1 - |r|$, with 44 groups for 2,000 genes and 148 for 22,092. Full-gene distances were calculated exactly in blocks. The fixed grouping was reused across methods, sides and repetitions. Within each side, patient identifiers determined three outcome-stratified cross-validation folds. Standardization and removal of zero-variance features were estimated within each training fold. cMCP used $\gamma=3$, $\tau=1/3$ and a 100-point master penalty path; tuning retained the common prefix with finite predictions in all three folds if a training path terminated early. Elastic Net used $\alpha=0.5$ and 50 penalties. Final models were refitted on the full side. Supports were not supplied to the grouping or fitting functions.

Eight main cells crossed dimension, support and sample size, each with 300 pairs. Four secondary cells fixed 2,000 genes and $n=1,000$, crossing support type with equal or heterogeneous effects at target AUC 0.85. Heterogeneous magnitudes decreased geometrically in identifier order with a 4:1 maximum-to-minimum ratio, root-mean-square normalization and alternating signs. Matched signal conditions shared patients and Bernoulli uniform random numbers. Overall, the high-dimensional and signal experiments produced 3,600 data pairs and 14,400 method results with no fitting failures or within-pair patient-isolation violations.

Statistical analysis

Confirmation proportions used two-sided exact Clopper to Pearson intervals. Local sample-size curves retained Wilson intervals for new repetitions. Bonferroni family intervals covered the specified four, eight or 12 cells within each method. For a paired tuning contrast, p_+ denoted the probability of one-standard-error-only success and p_- minimum-only success. Separate exact binomial intervals used error rate $0.05/(2m)$ for each discordance probability in a family of m contrasts. The difference interval was $[L_+ - U_-, U_+ - L_-]$, yielding conservative simultaneous coverage. Screening and confirmation denominators were not pooled for primary inference, and rates were not pooled across experiments.

High-dimensional families comprised eight cMCP tuning contrasts, 24 Elastic Net baseline contrasts and 16 dimension contrasts; the secondary module comprised four tuning and 16 signal contrasts. Sensitivity analyses crossed TPR thresholds 0.70, 0.80 and 0.90, FDP thresholds 0.05, 0.10 and 0.20, and Jaccard thresholds 0.50, 0.75 and 0.90. The 27 combinations across eight main cells formed a family of 216 cMCP contrasts. An exploratory reanalysis of the four 200-gene GSE96058 cells used 108 contrasts. These analyses reconstructed endpoints from saved selections or saved component metrics without refitting. Software and archived analysis identities are detailed in the Supplementary Information.

Declarations

Ethics

This study used publicly available, de-identified expression data and computer-generated outcomes. It involved no new participant recruitment or collection of identifiable patient information. Under the applicable institutional requirements, additional ethics approval and participant consent were not required for this secondary analysis.

Data availability

The publicly available transcriptomic datasets analysed in this study are available from the NCBI Gene Expression Omnibus under accession numbers GSE96058 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE96058>) and GSE7390 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE7390>). The numerical results supporting the findings of this study, together with the reference-set calibration information, task identifiers, and source-file checksums, are included in the Supplementary Information, Supplementary Data, and supporting archive accompanying this article.

Code availability

The custom code used to perform the analyses and reproduce the reported results, together with the reproduction instructions, is available from the corresponding author upon reasonable request.

Author contributions

H.M. and J.Y. conceived and designed the study and interpreted the results. H.M. conducted the study, developed the software, curated the data, performed the statistical analyses, prepared the visualizations and wrote the original draft. J.Y. supervised the study, reviewed the methodology and critically revised the manuscript.

Funding

This research received no specific funding.

Competing interests

The authors declare no competing interests.

References

1. Ein-Dor, L., Zuk, O. & Domany, E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc. Natl. Acad. Sci. USA* **103**, 5923–5928 (2006). <https://doi.org/10.1073/pnas.0601231103>
2. Haury, A. C., Gestraud, P. & Vert, J. P. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PLoS ONE*. **6**, e28210. <https://doi.org/10.1371/journal.pone.0028210> (2011).
3. Nogueira, S., Sechidis, K. & Brown, G. On the stability of feature selection algorithms. *J. Mach. Learn. Res.* **18**, 1–54 (2018). <https://jmlr.org/papers/v18/17-514.html>
4. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R Stat. Soc. B.* **58**, 267–288. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x> (1996).
5. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R Stat. Soc. B.* **67**, 301–320. <https://doi.org/10.1111/j.1467-9868.2005.00503.x> (2005).
6. Zhang, C. H. Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.* **38**, 894–942. <https://doi.org/10.1214/09-AOS729> (2010).
7. Yuan, M. & Lin, Y. Model selection and estimation in regression with grouped variables. *J. R Stat. Soc. B.* **68**, 49–67. <https://doi.org/10.1111/j.1467-9868.2005.00532.x> (2006).
8. Meier, L., Van De Geer, S. & Bühlmann, P. The group lasso for logistic regression. *J. R Stat. Soc. B.* **70**, 53–71. <https://doi.org/10.1111/j.1467-9868.2007.00627.x> (2008).
9. Breheny, P. & Huang, J. Penalized methods for bi-level variable selection. *Stat. Interface.* **2**, 369–380. <https://doi.org/10.4310/SII.2009.v2.n3.a10> (2009).
10. Zhao, P. & Yu, B. On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7**, 2541–2563 (2006). <https://jmlr.org/papers/v7/zhao06a.html>
11. Wainwright, M. J. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (Lasso). *IEEE Trans. Inf. Theory.* **55**, 2183–2202. <https://doi.org/10.1109/TIT.2009.2016018> (2009).
12. Meinshausen, N., Bühlmann, P. & Stability selection *J. R Stat. Soc. B* **72**, 417–473 <https://doi.org/10.1111/j.1467-9868.2010.00740.x>. (2010).
13. Shah, R. D. & Samworth, R. J. Variable selection with error control: another look at stability selection. *J. R Stat. Soc. B.* **75**, 55–80. <https://doi.org/10.1111/j.1467-9868.2011.01034.x> (2013).
14. Morris, T. P., White, I. R. & Crowther, M. J. Using simulation studies to evaluate statistical methods. *Stat. Med.* **38**, 2074–2102. <https://doi.org/10.1002/sim.8086> (2019).
15. Burton, A., Altman, D. G., Royston, P. & Holder, R. L. The design of simulation studies in medical statistics. *Stat. Med.* **25**, 4279–4292. <https://doi.org/10.1002/sim.2673> (2006).
16. Schreck, N., Slynko, A., Saadati, M. & Benner, A. Statistical plasmode simulations—potentials, challenges and recommendations. *Stat. Med.* **43**, 1804–1825. <https://doi.org/10.1002/sim.10012> (2024).

17. Bommert, A., Welchowski, T., Schmid, M. & Rahnenführer, J. Benchmark of filter methods for feature selection in high-dimensional gene expression survival data. *Brief. Bioinform.* **23**, bbab354. <https://doi.org/10.1093/bib/bbab354> (2022).
18. Chen, Y. & Yang, Y. The one standard error rule for model selection: does it work? *Stats* **4**, 868–892. <https://doi.org/10.3390/stats4040051> (2021).
19. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R Stat. Soc. B.* **57**, 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x> (1995).
20. Candès, E., Fan, Y., Janson, L. & Lv, J. Panning for gold: Model-X knockoffs for high dimensional controlled variable selection. *J. R Stat. Soc. B.* **80**, 551–577. <https://doi.org/10.1111/rssb.12265> (2018).
21. Ren, Z. & Barber, R. F. Derandomised knockoffs: leveraging e-values for false discovery rate control. *J. R Stat. Soc. B.* **86**, 122–154. <https://doi.org/10.1093/jrsssb/qkad085> (2024).
22. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22. <https://doi.org/10.18637/jss.v033.i01> (2010).
23. Breheny, P. & Huang, J. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann. Appl. Stat.* **5**, 232–253. <https://doi.org/10.1214/10-AOAS388> (2011).
24. Peng, H., Long, F. & Ding, C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**, 1226–1238. <https://doi.org/10.1109/TPAMI.2005.159> (2005).
25. De Jay, N. et al. mRMRe: an R package for parallelized mRMR ensemble feature selection. *Bioinformatics* **29**, 2365–2368. <https://doi.org/10.1093/bioinformatics/btt383> (2013).
26. Breheny, P. & Huang, J. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Stat. Comput.* **25**, 173–187. <https://doi.org/10.1007/s11222-013-9424-2> (2015).
27. Bühlmann, P., Rütimann, P., van de Geer, S. & Zhang, C. H. Correlated variables in regression: Clustering and sparse estimation. *J. Stat. Plan. Inference.* **143**, 1835–1858. <https://doi.org/10.1016/j.jspi.2013.05.019> (2013).
28. Brueffer, C. et al. Clinical value of RNA sequencing-based classifiers for prediction of the five conventional breast cancer biomarkers: a report from the population-based multicenter Sweden Cancerome Analysis Network–Breast Initiative. *JCO Precis Oncol.* **2**, 1–18. <https://doi.org/10.1200/PO.17.00135> (2018).
29. Desmedt, C. et al. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin. Cancer Res.* **13**, 3207–3214. <https://doi.org/10.1158/1078-0432.CCR-06-2765> (2007).

Figures

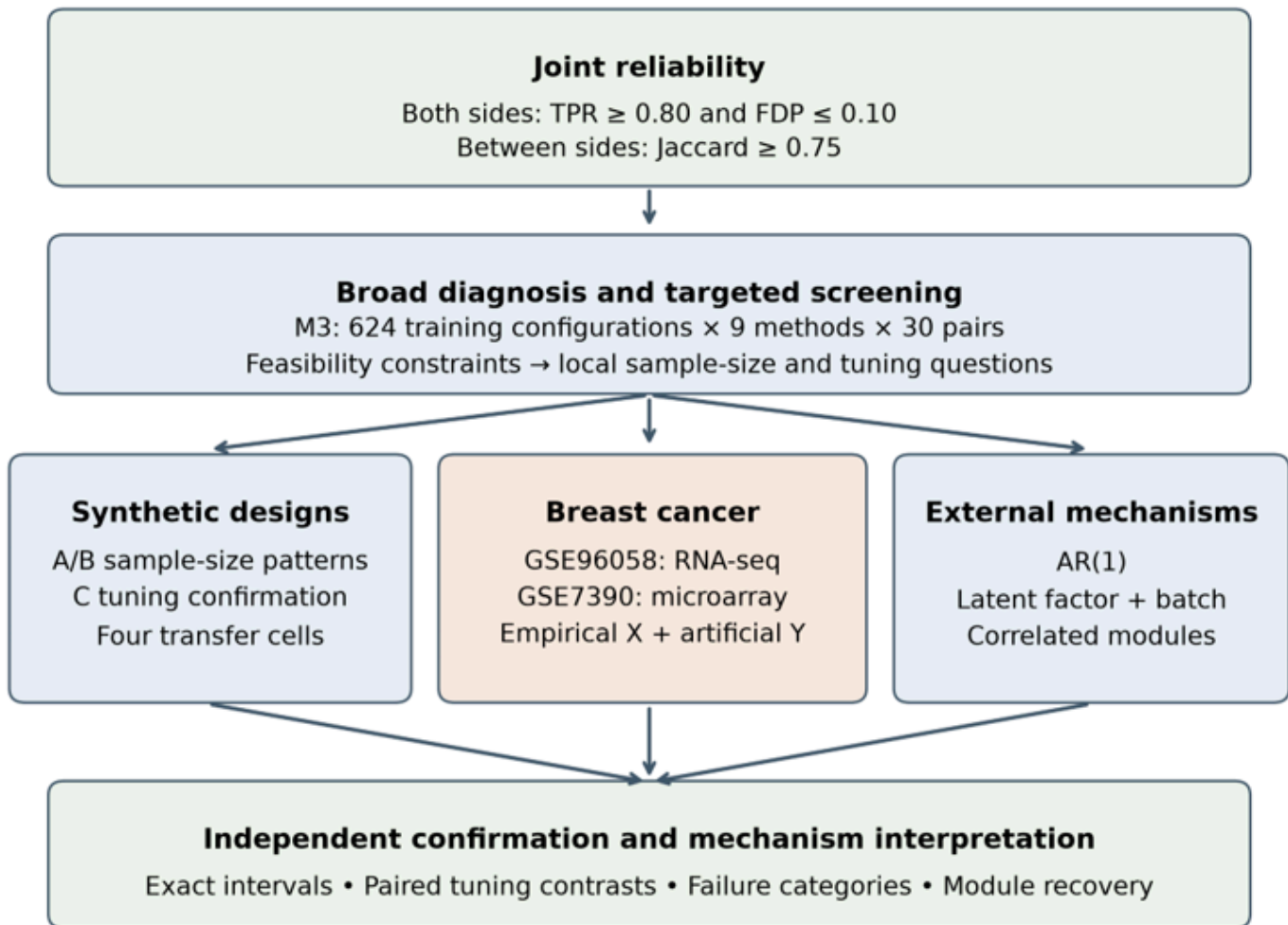


Figure 1

Study design and evidence stages. The diagram summarizes broad diagnosis, targeted screening and independent confirmation. Empirical expression matrices supply covariates; supports and binary outcomes are artificial. The high-dimensional GSE96058 extension uses 2,000 and 22,092 genes, separate patient pools and 300 pairs per cell, as specified in Table 1.

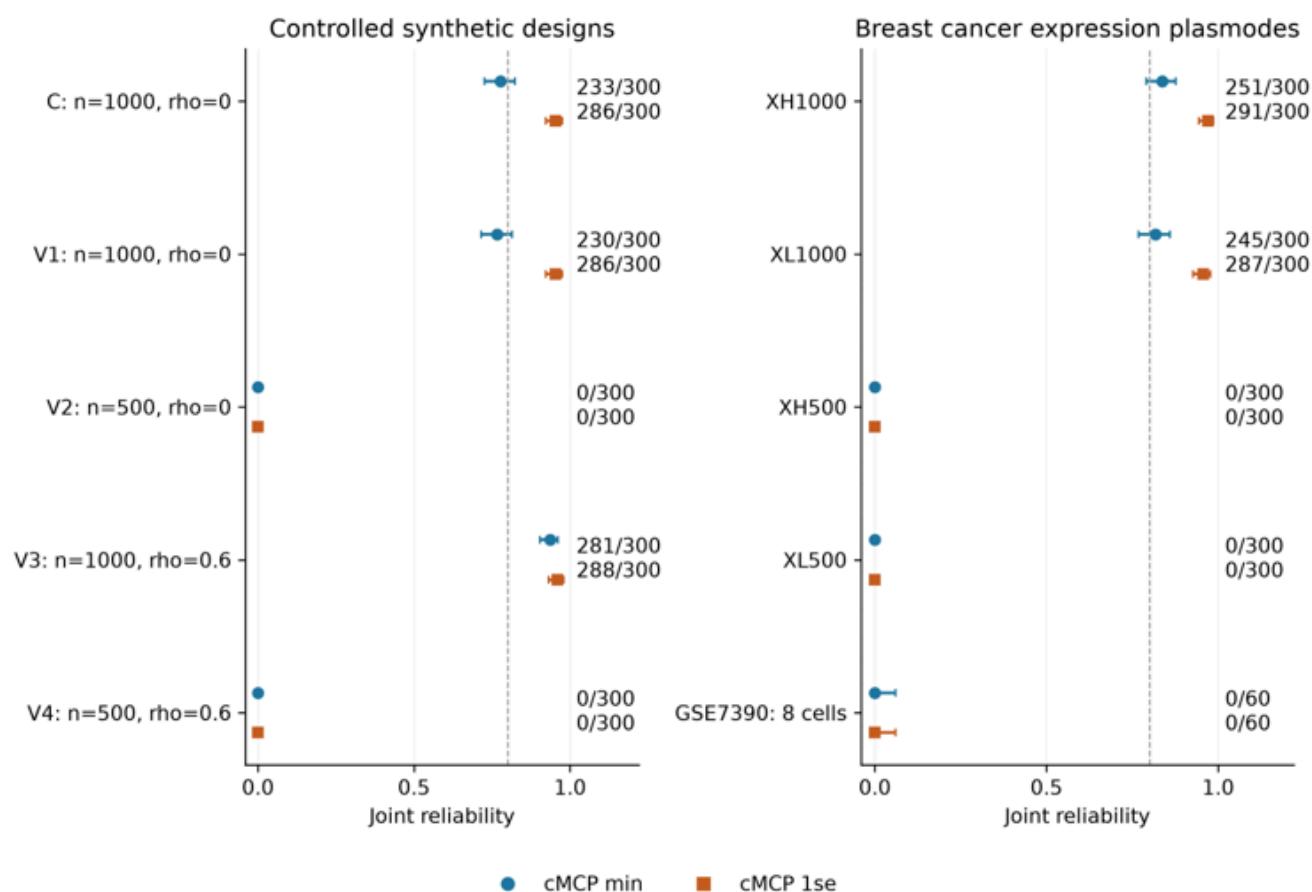


Figure 2

cMCP tuning in controlled designs and 200-gene empirical benchmarks. Points show independent confirmation proportions with unadjusted exact 95% binomial intervals. C is dedicated tuning confirmation; V1 to V4 are independent synthetic cells. XH and XL indicate high- and low-correlation supports in GSE96058. GSE7390 cells use K=60; other cells use K=300. In these empirical benchmarks, n is the number of bootstrap draws per side.

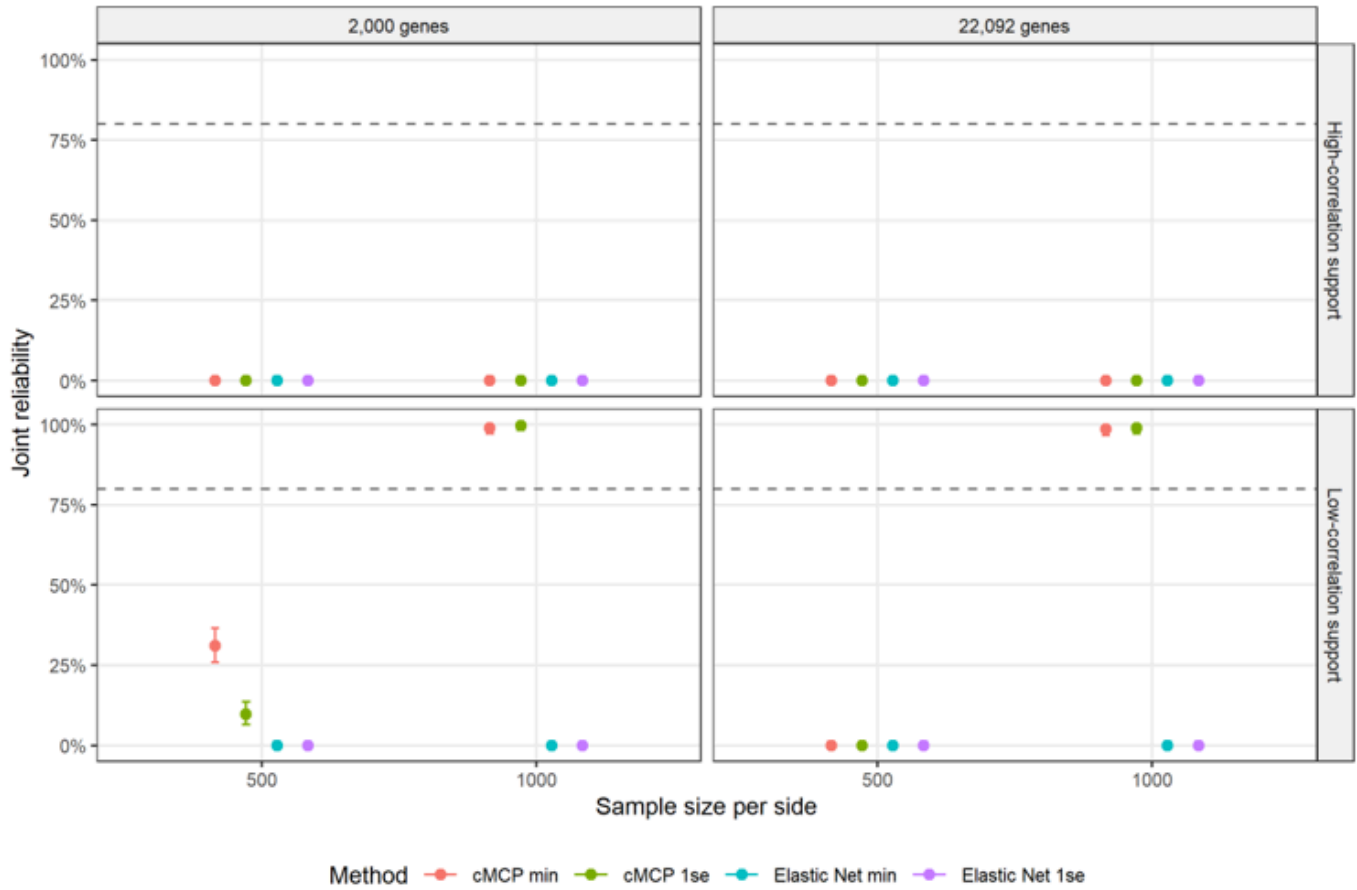


Figure 3

Joint reliability in the high-dimensional GSE96058 benchmark. Each cell contains 300 paired repetitions. Points show success proportions and bars exact 95% binomial intervals. Candidate spaces contain 2,000 or 22,092 genes, with high- or low-correlation artificial supports and 500 or 1,000 patients per side. The dashed line denotes reliability 0.80. Elastic Net uses $\alpha=0.5$.

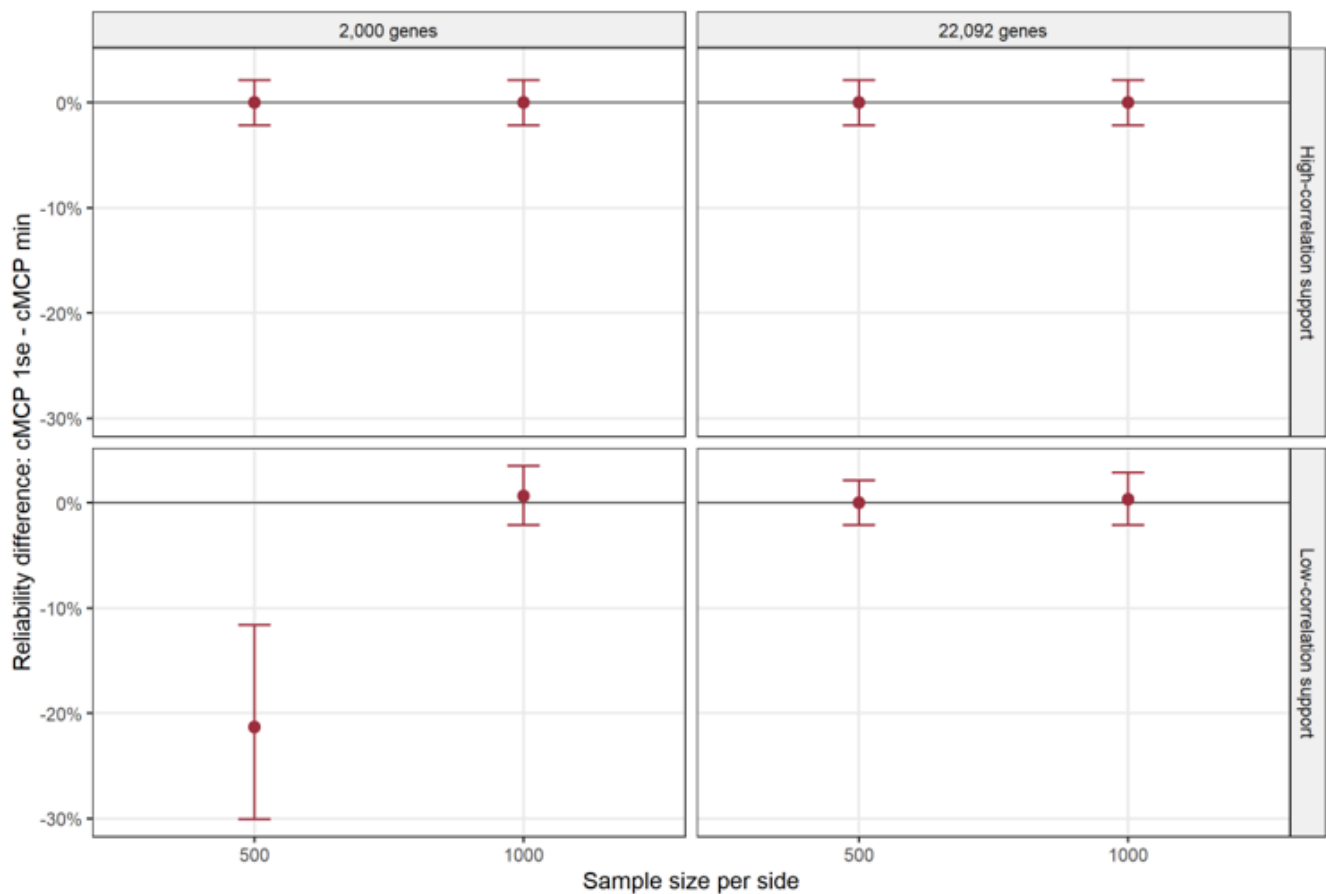


Figure 4

Paired cMCP tuning differences in the high-dimensional benchmark. Differences are one-standard-error minus minimum-deviance reliability in percentage points. Positive values favor the one-standard-error rule. Intervals control coverage across the eight main tuning contrasts. Each setting contains 300 pairs. P2K and FULL denote 2,000 and 22,092 genes; H and L denote support correlation.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryInformation.docx](#)