

Proposed simulation protocol for permutation centred mutual information

Comparisons between samples of different sizes

Nick Barua and Koji Oda | Reconstructed v0.5 | 14 October 2026

Status: draft for review. Not frozen, registered or approved for the main study.

This document consolidates the supplied v0.1 PDF and the agreed v0.2 to v0.5 changes reported in the conversation. It is a reconstruction, not an export of Claude's unavailable live document. Its purpose is to provide a single, reviewable specification before budgeting, freezing and running the new study.

The scientific question, development grid, estimator set and reported amendments are preserved. Details absent from the available record are explicitly labelled reconstruction proposals R1 to R9 in Appendix A. Those proposals are concrete defaults for review; they are not represented as previously agreed or already executed.

0 Status and prior knowledge

The original manuscript results were known before this protocol was drafted. They include the $N = 78$ stellar application, the archived Gaussian calibration, the single-coordinate quadratic experiment and its 5,000-catalogue audit, and the earlier three-coordinate quadratic experiment. The different quadratic experiments use different signal designs and do not contradict one another. They are retained as provenance.

The Pandey and Mondal preprint [1] and its cumulative-profile construction were also known before the new study was planned. The original experiments and those published claims are motivation, not prospectively generated evidence for this protocol.

Reference calculations and implementation checks are already available. At $s = 0.64$, the quadratic information is approximately 0.403627 nats per active pair and 1.210882 nats for three independent pairs. The original three-pair linear alternative at correlation 0.35 carries approximately 0.196017 nats. These values quantify an information mismatch in the earlier experiments; they do not by themselves constitute a new methodological result.

The new evidence to be generated is the calibration of raw and centred estimates over a matched-information grid, comparisons between independent samples of different sizes, cumulative profiles under constant true information, and evaluation of frozen guidance on held-out settings. No main-study estimator-performance result has been generated for this reconstruction.

AI assistance includes protocol drafting and revision, mathematical and code review, implementation, reference checks and preparation of this reconstructed document, using Claude and ChatGPT/Codex. The authors remain responsible for the design, source verification, code review, execution and interpretation. The final disclosure must describe the assistance actually used.

1 Research questions and scope

Primary question. At which sample sizes, dimensions and dependence geometries does permutation-centred mutual information preserve the mean difference between samples with known equal or unequal true information?

Secondary question. Do common- N subsampling, an analytic Gaussian correction or a different KSG neighbour order improve the specified outcomes on held-out settings?

Permutation centring subtracts an estimated null offset. It does not establish that the remaining score equals true mutual information under dependence. The study evaluates that distinction directly. It separates mean calibration and profile distortion from the uncertainty of an individual comparison.

The cumulative experiment randomises row order independently of the data, keeping the population distribution constant at every prefix. It isolates effects associated with changing N . A real radial cut can change the joint distribution as well as N ; therefore a simulated profile distortion would not, by itself, refute an astronomical inference in [1]. No new stellar catalogue or astronomical conclusion is promised by this protocol.

2 Completed verification and remaining gates

Equation (8) requires no correction. Direct inspection of the Word equation markup and rendered manuscript verified the coefficient 0.8 divided by the square root of 2. The earlier apparent “0.82” mismatch was a text-extraction error. The manuscript, Supplementary Methods and both archived code copies agree.

$$D_1 = \frac{0.8}{\sqrt{2}} (C_1^2 - 1) + 0.6 \varepsilon_1$$

Implementation package v0.2 is distinct from protocol v0.5. It retains `truth.py` unchanged, vendors `mi_audit.py` byte for byte, implements independent generators and exposes $k = 3, 4, 5, 8$ through `Prepared.ksg` rather than the archived $k = 5$ shortcut.

Check	Recorded result
Unit tests	37 tests passed using Python unittest; pytest is not required.
Archive regression	All 400 $N = 78$, correlation = 0.4, deformation = 0 rows reproduced exactly for GC and KSG5.
Deterministic truth	184 matched-information settings checked by independent numerical routes; maximum difference below 1×10^{-8} nats.
Generator reference checks	All eight designs checked at $s = 0.64$ using 10 million generated draws each; all within three Monte Carlo standard errors.
Runtime benchmark	48 rows across 24 settings; $B = 0$ and $B = 10$ measured. $B = 100$ and $B = 500$ are extrapolations.

The largest absolute generator-check z score was 2.398 for L1. All eight outcomes are retained. A passing check is not repeated merely to obtain a smaller discrepancy. Additional investigation requires a concrete concern and retains the original result. These are implementation checks, not held-out estimator-performance results.

3 Estimators and permutation centring

All information is in nats and all logarithms are natural. KSG uses the archived estimator 1 with maximum-norm distances, strict marginal neighbour counts and within-sample coordinate standardisation with denominator $N - 1$ [2]. GC uses average marginal ranks, normal scores at $(\text{rank} - 0.5)/N$, and sample covariance with denominator $N - 1$. Constant coordinates or invalid covariance calculations must be reported as failures, not silently discarded.

Code	Procedure
KSG k	Raw KSG at $k = 3, 4, 5, 8$; $k = 4$ is the primary KSG choice.
KSG k c	Raw KSG minus the corresponding mean over permuted D rows.
GC	Raw covariance score after marginal normal-score transformation.
GC c	Raw GC minus its permutation mean.
GC a	Raw GC minus the analytic Gaussian log-determinant correction below.

Use $B = 100$ whole-row permutations of D for the primary centred estimate. Each permutation preserves the paired coordinates within D and is shared by GC and all requested k values. Use $B = 500$ for the specified 10% sensitivity subset. No raw, centred or corrected score is clipped to zero. Distance correlation is excluded from the information-scale primary analysis.

$$\hat{I}_c = \hat{I} - \frac{1}{B} \sum_{b=1}^B \hat{I}_{nb}$$

For sample covariance S , the Gaussian score and its correction are defined as follows. Here ψ is the digamma function, p and q are block dimensions, and N must exceed $p + q$.

$$G = \frac{1}{2} [\ln \det(S_C) + \ln \det(S_D) - \ln \det(S_{CD})]$$

$$B_m(N) = \sum_{j=1}^m \psi\left(\frac{N-j}{2}\right) + m \ln 2 - m \ln(N-1)$$

$$\delta(N, p, q) = \frac{1}{2} [B_p(N) + B_q(N) - B_{p+q}(N)], G_a = G - \delta(N, p, q)$$

At $N = 78$, $p = 4$ and $q = 3$, $\delta = 0.08221750305196363$ nats. The correction is exact in expectation for nonsingular ordinary sample covariance from iid jointly Gaussian observations. Empirical rank-normal scores do not have that exact sampling law; GC a is a candidate correction whose residual error must be measured [3]. The implementation is checked against the official GCMI code [4], accounting for that code's conversion to bits.

At fixed N , p and q , δ is constant over permutations and cancels when centring is applied. Corrected-and-centred GC is therefore the same procedure as GC c and is counted once. GC's error against true information can include a population approximation error, especially for symmetric nonlinear dependence; it must not all be described as finite-sample bias.

4 Generating models and information truth

For each active pair, C and ϵ are independent standard normal variables. Different active pairs are independent; all unused coordinates are independent standard normals. The parameter s is a signal-variance fraction with $0 \leq s < 1$, not the archived linear correlation. For a nonnegative linear correlation, s equals its square.

$$D_j = \sqrt{s} g(C_j) + \sqrt{1-s} \epsilon_j$$

$$g_L(c) = c, g_Q(c) = \frac{c^2 - 1}{\sqrt{2}}, g_H(c) = \frac{|c| - \sqrt{\frac{2}{\pi}}}{\sqrt{1 - \frac{2}{\pi}}}$$

Design	p and q	Active pairs	Signal and role
L1	1 and 1	1	Linear Gaussian reference
L3	4 and 3	3	Three independent linear pairs
T3	4 and 3	3	L3 followed by the invertible deformation below
Q1	1 and 1	1	One quadratic pair
Q1x	4 and 3	1	One quadratic pair plus independent coordinates
Q3	4 and 3	3	Three independent quadratic pairs

Design	p and q	Active pairs	Signal and role
H1	1 and 1	1	Held-out absolute-value pair
H3	4 and 3	3	Held-out absolute-value pairs; implemented interpretation R1

For T3, replace D_2 by $D_2 + D_1^2 - 1$ after drawing L3. This is the archived deformation with coefficient 1. Its inverse has unit Jacobian, so T3 and L3 have the same true block information. The estimator receives the deformed observations.

For a scalar channel, truth is calculated from the marginal differential entropy. Multiply the scalar result by the number of independent active pairs. Additional independent coordinates contribute no information.

$$I(C; D) = h(D) - \frac{1}{2} \ln[2\pi e(1 - s)], I_{\text{linear pair}} = -\frac{1}{2} \ln(1 - s)$$

The two deterministic routes are characteristic-function FFT inversion followed by entropy integration, and independent adaptive real-space integration. They must agree on total block information within 1×10^{-8} nats. Grid resolution, transform period and integration windows are enlarged separately. Information matching is checked using the independent route. The implemented nonlinear numerical range ends at $s = 0.9999$; unsupported settings fail explicitly.

The original requirement for Monte Carlo truth cross-checks is retained: each distinct nonlinear family and signal setting used in the frozen grid receives a reference-density check with at least 10 million draws and its measured standard error. Agreement within three standard errors is the stated diagnostic. The completed $s = 0.64$ checks are representative and do not replace the remaining target-specific checks. Every generator must also be tied to the reference through actual draws from `generators.py`, with an independently written density evaluator.

Development targets are 0.05, 0.10, 0.25, 0.50 and 0.75 nats. Solve for s separately for each design so total information matches the target. E2 comparison ratios extend the maximum target to 1.50 nats. Record solver settings, method agreement and all Monte Carlo outcomes; do not replace an inconvenient seed or omit a diagnostic failure.

5 Development experiments and candidate procedures

A cell fixes the design, information target or target pair, sample size or size pair, and experiment. Replicates are independent complete catalogues, independent pairs of catalogues for E2, or independent parent catalogues for E3. The counts below are the original proposed counts before any uniform runtime-based reduction.

Experiment	Grid	Replicates per cell
E1 calibration	Six development designs $\times$ five targets $\times$ $N = 50, 78, 156, 300, 500, 1000, 2000$	1,000; 200 when $N \geq 1000$
E2 comparison	Six designs $\times$ five base targets $\times$ six size pairs below $\times$ $r = 1, 1.25, 1.5, 2$	1,000 independent A and B pairs
E3 profiles	Q1, Q1x, L1, T3 $\times$ targets 0.10 and 0.50 $\times$ ten nested prefixes below	200 parent catalogues

E1 Single sample calibration

Compute all procedures in Section 3 on each catalogue. The primary quantity is mean error against true information; report RMSE and the distribution of individual errors as secondary outcomes. For GC, use the term error rather than attributing the whole discrepancy to finite-sample bias.

E2 Two sample comparisons

The size pairs (N_A, N_B) are (78,78), (78,156), (78,500), (156,500), (78,2000) and (500,2000). Use the same generating design for both samples, with $I_B = r I_A$ and $r \geq 1$. Draw A and B independently, as made explicit by reconstruction proposal R2. Each pair is retained together during analysis and resampling.

The primary outcome is signed difference error: the estimated B-minus-A difference minus the true B-minus-A difference. Report RMSE of difference error, ordering probabilities, ties, and ratios descriptively. At $r = 1$, a 50% ordering rate is not imposed as a pass threshold. Ratios are evaluated only when both estimates are strictly positive; report the excluded proportion and do not take logarithms of signed estimates.

E3 Cumulative profiles

Draw a parent catalogue with 5,000 rows and apply an independent random row ordering. Use prefixes 100, 150, 225, 340, 500, 750, 1100, 1700, 2500 and 5000. The joint population distribution and true information remain fixed across prefixes. Prefixes within one catalogue are dependent.

KSG 4 c is the primary profile procedure. Reconstruction proposal R3 records all four KSG choices and the GC family at the same prefixes for secondary candidate comparisons. Report mean error at every prefix, the mean-profile endpoint difference and the range of the mean profile. Per-catalogue slopes and Spearman correlations are descriptive only; there is no Spearman pass threshold.

P1 Common sample size subsampling

P1 applies to E2 and E3. Use 20 row subsets without replacement within each subset; subsets may overlap. Sample C and D using the same row indices. Recalculate standardisation, ranks and permutation means within each subset. Reconstruction proposal R4 applies P1 to the same estimator family as the corresponding baseline call.

For E2, reuse the original estimate for the smaller sample A. Independently draw 20 subsets of size N_A from B, estimate each subset, then average its estimator values before forming the B-minus-A comparison. When N_A equals N_B , reuse the two baseline estimates and do not introduce redundant subset or permutation randomness.

For E3, the common size is the smallest prefix: 100 for development and 120 for the proposed held-out grid. Reuse that smallest-prefix estimate. At each larger prefix draw 20 subsets independently of the subsets at other prefixes, conditional on the same parent catalogue, and average their estimates. Catalogue-level resampling preserves the resulting dependence. Do not flatten the profile by deliberately reusing one identical subset at every prefix.

P2 Gaussian correction and P3 neighbour choice

P2 is GC a as defined in Section 3; it is evaluated wherever raw GC is evaluated. P3 selects among $k = 3, 4, 5, 8$. All four are retained in development comparisons. Any practical rule for choosing k or applying P1 must be written and frozen after development and before any held-out performance is inspected. It may use observable sample descriptors, not the unavailable true information or known simulation design label at application time.

The held-out claim is limited to that frozen rule. Exploring multiple replacement rules after seeing validation results creates exploratory follow-up work, not another confirmatory use of E4. Combinations of P1 to P3 must be named explicitly in the guidance record and evaluated against a named baseline; the permitted components do not authorise unlimited unreported searches.

Permutation sensitivity subset

Reconstruction proposal R5 selects exactly 10% of replicate IDs in every cell using a separate, outcome-independent selection stream before main outputs are available. Select whole A/B pairs for E2 and whole parent catalogues for E3. Include every baseline and P1 subset call belonging to the selected replicate. Selection is shared across estimators.

The first 100 permutations are identical for the $B = 100$ and $B = 500$ calculations. Extend selected calls with 400 additional permutations and retain both results. This extension capability is not yet in the delivered wrapper and must be implemented and tested. Report the paired change and permutation-mean MCSE; $B = 500$ is a sensitivity check, not a replacement selected after viewing the results.

6 Outcomes uncertainty and classification

The practical reporting tolerance is the larger of 0.01 nats and 10% of the reference information. For E1 and E3, the reference is the true block information in that cell. For E2, it is I_A , the smaller true information in the pair. This tolerance does not guarantee correct ranking or precise ratios for individual catalogues.

$$\tau(I_{ref}) = \max(0.01, 0.10 I_{ref}) \text{ nats}$$

E1 and E2 mean errors

$$e_r = \hat{I}_r - I^* \text{ for } E1$$

$$e_r = (\hat{I}_{B,r} - \hat{I}_{A,r}) - (I_B - I_A) \text{ for } E2$$

The parameter being classified is the expectation of e across independent replicate records. Construct a 95% interval $[L, U]$ for its mean. A unit is one fixed experiment cell and procedure, not an aggregate of incompatible sample sizes or designs.

Classification	Rule for the mean error interval
Pass	$L \geq -\tau$ and $U \leq \tau$
Fail	$U < -\tau$ or $L > \tau$
Indeterminate	Neither rule is satisfied

A pass supports an average signed error within tolerance at the simulated setting. It does not establish that every estimate or comparison is accurate. Report RMSE and ordering separately; do not convert those descriptive outcomes into unregistered pass criteria.

E3 Profile distortion

Let μ_j be the expected estimated information at prefix j , and let the sample mean over independent catalogues estimate μ_j . Define endpoint difference $\Delta = \mu_j - \mu_1$. The range H is the maximum μ_j minus the minimum μ_j across prefixes, so H is nonnegative. Report the mean error $\mu_j - I$ at every prefix separately.

Pass requires the endpoint interval to lie entirely in $[-\tau, \tau]$ and the upper bound for H to be at most τ . Fail requires the endpoint interval to lie entirely outside that band or the lower bound for H to exceed τ . Otherwise the outcome is indeterminate. Since $H \geq |\Delta|$, the range condition implies the endpoint condition under the joint construction below; the two summaries are retained for interpretation, not treated as independent evidence.

E3 measures profile distortion, not absolute calibration. A flat but uniformly biased profile can pass E3 while failing the corresponding calibration criterion. Report that distinction explicitly.

Monte Carlo intervals

Reconstruction proposal R6 fixes 95% intervals using 9,999 bootstrap resamples of complete independent replicate records. E2 retains A and B together. E3 retains every prefix and every procedure for each parent catalogue. These are Monte Carlo uncertainty intervals across generated catalogues; do not resample individual stars or prefixes for this purpose.

For E1 and E2, calculate the bootstrap distribution of the difference between the resampled mean error and the original mean error. Let c be the 95th percentile of its absolute value, and use the symmetric interval from the original mean minus c to the original mean plus c . The bootstrap quantile convention is the NumPy linear interpolation convention, recorded with the analysis code.

For E3, use simultaneous contrasts rather than a direct percentile bootstrap of the nonsmooth range. Within each resample calculate the largest absolute centred change over every pairwise mean-profile contrast. Let c be its 95th percentile. Use $[\text{estimated } \Delta - c, \text{estimated } \Delta + c]$ for the endpoint and $[\max(0, \text{estimated } H - c), \text{estimated } H + c]$ for the range. This approximates simultaneous coverage across prefixes within one profile unit; it is not simultaneous coverage over the entire experiment grid.

Report bootstrap or Wilson intervals for descriptive outcomes as appropriate. Ratio summaries and their intervals remain conditional on both estimates being positive; retain the exclusion rate. If too few valid records remain, report the summary as not estimable. Numerical failures retain their assigned replicate IDs and are investigated; they are not silently dropped and replaced. A unit with unresolved failures is indeterminate.

7 Contribution and interpretation rules

Outcome A. Report the observed operating limits of the procedures, including passing, failing and indeterminate units and how they vary with N , dimensions and geometry. A useful contribution requires an interpretable result beyond known covariance blindness or the fact that an intersection test cannot exceed its weakest component.

Outcome B. A candidate must convert at least half of its applicable failing development units to passing before being presented as promising guidance. Validation must show actual improvement on held-out settings, with no more than 10% degradation of previously passing held-out units. Report outcomes separately for E1, E2 and E3; do not pool incompatible unit types.

Reconstruction proposal R7 makes “actual improvement” operational: within each experiment family claimed to benefit, require at least one baseline-fail to candidate-pass conversion and a positive net increase in the number of passing applicable units. Count any baseline-pass unit that becomes fail or indeterminate as degraded. Divide degradation by all applicable baseline-pass units. If that denominator is zero, report the degradation check as not evaluable and do not claim that this full criterion has been met. Baseline-indeterminate units are not counted as failing-unit conversions. These are descriptive grid-level rules; individual unit intervals do not establish a grid-wide confirmatory significance level.

Outcome C. If no failure is resolved, report the tested settings and remaining uncertainty honestly. There is no automatic “no failure means short note” rule. Paper form, title and venue depend on usefulness and novelty after the results are known. A general statistical finding may support a statistics submission; an astronomy claim requires evidence that addresses astronomical inference. None of the implementation checks guarantees publication.

8 Held out validation grid

The available record specifies H1, H3, L1 and L3, information targets 0.15 and 0.40 nats, sample sizes 120, 250 and 750, and a separate seed base. It reports that a fuller grid was written in an unavailable revision but does not give its values. The complete grid below is therefore reconstruction proposal R1, not a recovered quotation.

Validation analogue	Proposed grid	Replicates
E4 calibration	4 designs $\times$ 2 targets $\times$ $N = 120, 250, 750$	1,000 per cell
E4 comparison	4 designs $\times$ 2 base targets $\times$ 6 pairs below $\times$ $r = 1, 1.25, 1.5, 2$	1,000 independent pairs per cell
E4 profiles	4 designs $\times$ 2 targets; prefixes 120, 250, 750; parent $N = 750$	200 parent catalogues per cell

Held-out size pairs are (120,120), (120,250), (120,750), (250,250), (250,750) and (750,750). H3 is interpreted as three independent active absolute-value pairs, consistent with the implemented generator. For comparisons, the same design is used for A and B and the larger true information is r times the smaller. The largest held-out target is therefore 0.80 nats.

Apply the same estimator definitions, P1 implementation, 10% permutation extension, outcome definitions and interval rules. Full procedure values may be recorded for transparent reporting, but the guidance rule must be frozen before E4 begins and cannot be retuned using those values. Report the entire held-out grid whatever it shows. A bug-fix rerun retains the original outputs and is documented in the deviation log.

The held-out reference and generator checks already performed establish implementation consistency only. They have not evaluated finite-sample estimator performance on E4. L1 and L3 at new N and information targets assess transfer in settings; H1 and H3 add a new dependence family.

9 Reproducibility and random streams

Retain development seed base 202611010000 and held-out seed base 202612010000. Use explicit NumPy SeedSequence objects with separate streams for data, permutations, ordering, subset selection and bootstrap resampling. SeedSequence's entropy argument is a seed input, unrelated to entropy measured in nats [5].

The delivered seeding.py uses the spawn-key layout (schema version, stage ID, role ID, stable task coordinates). Reconstruction proposal R8 reserves distinct task namespaces for experiment and sensitivity selection, and appends fixed integer indices for design, target, ratio, size or pair, replicate, prefix and P1 subset as applicable. Use a separate ordering role when extending the role registry; increment the schema version if the mapping changes. Save the complete registry before the main run.

Never derive stream keys from floating-point hashes, Python hash values, worker numbers or completion order. A task's stream identity must be independent of scheduling. Store seed base, full spawn key, pool size and bit-generator name with each output record. Generator validation also records batching because C/noise draw interleaving changes when batch boundaries change.

Implementation-validation and benchmark seed bases are separate from the study bases. Do not replace the study seeds with the seeds used for the already completed checks. For selected $B = 500$ extensions, reconstruct the same permutation stream and continue after permutation 100; the extension must reproduce the first 100 values exactly.

Pin Python 3.12, NumPy 2.3.5 and SciPy 1.17.0 for the reference run. The original draft also named pandas 2.2.3; the delivered core package does not require pandas. Pin it only if a future runner or analysis script uses it. Record exact Python patch version, platform, worker count, native thread counts and all added dependencies. Parallel workers use one numerical-library thread each unless an explicitly measured configuration is frozen.

Keep mi_audit.py unchanged and record every code and configuration hash. Main-study execution must refuse a mismatch with the frozen manifest. Record dependencies and versions as well as file hashes; a hash alone does not preserve an environment. Preserve every assigned replicate, its results or error status, and its provenance.

10 Implementation specification and validation gates

Component	Current status and required completion
truth.py	Implemented and previously validated; complete any remaining target-specific reference checks for the frozen grid.
generators.py	All eight mapped designs implemented; independent formula and reference-density checks completed at $s = 0.64$.

Component	Current status and required completion
estimators.py	Raw, centred and corrected outputs implemented with shared permutations. P1 and continuation from 100 to 500 remain to be added.
seeding.py	Stable task streams implemented; extend and freeze the full runner task registry.
Experiment drivers	E1, E2, E3 and E4 drivers, checkpointing, profile ordering and result tables remain to be implemented.
Analysis and freeze	Catalogue bootstrap, classification, guidance record, budget calculator and freeze/hash gate remain to be implemented.

Run the existing 37 unittest checks before integration. Add focused checks for P1 row pairing and marginal preservation, equal-N identity, correct nested prefixes, $B = 500$ continuation, stable task IDs under reordered workers, interval classification and numerical-failure retention. Use analytic examples or small deterministic fixtures to test these mechanics. Do not inspect development or held-out performance to tune the implementation.

The archived regression is stronger than a comparison of means: all 400 selected GC and KSG5 rows already match exactly. Preserve that fixture and the unchanged archive-code hash. If a change breaks an existing equality, diagnose it before running the study rather than relaxing tolerances without a stated reason.

11 Reporting and deviations

Report every experimental cell, including indeterminate units and numerical failures. For each procedure provide true information, mean estimate, mean error, RMSE, Monte Carlo interval, permutation count, sample size and dimensions. E2 additionally reports difference error, ordering probabilities, ties, positive-only ratio summaries and excluded shares. E3 reports every mean prefix estimate, mean error, endpoint difference and range.

Store a dated deviation log with version, setting changed, original and revised values, reason, code/configuration hashes, and whether any relevant main-study results had been viewed. Separate a bug correction from a design change. Preserve superseded results with their status; never silently replace them.

State that the earlier manuscript experiments followed known stellar results, that this protocol was written later, and exactly which new analyses were frozen before their outcomes were inspected. The existing implementation checks and runtime measurements belong in the pre-freeze knowledge record. Keep the AI-assistance statement accurate and verify manuscript references against their sources.

12 Compute budget and order of work

The runtime benchmark is provisional: three representative designs, four N values, two k sets and two repeats on a shared host. $B = 100$ and $B = 500$ costs are extrapolated from $B = 0$ and $B = 10$. Full budgeting must include baseline calls, P1 subset calls, the additional 400 permutations for the 10% subset, truth checks, bootstrap analysis, checkpointing and measured parallel efficiency. Do not multiply every replicate by a full $B = 500$ cost.

Use explicit interpolation or a conservative dimension/geometry proxy for unmeasured settings and label that assumption. Validate major cost contributors with timing-only measurements. No score, bias, power or ranking result from the main study may guide a replication cut. Replacing four k values by $k = 4$ changes the planned comparison and is not a neutral timing adjustment.

Reconstruction proposal R9 applies one common reduction factor γ to every originally specified development and held-out replication count. Round down to a multiple of 10 so the 10% sensitivity subset remains exact: $R_{\text{new}} = 10 \times \text{floor}(\gamma R_{\text{original}} / 10)$. Choose $0 < \gamma \leq 1$ so all cells retain at least 10 replicates. Record γ , the target compute budget and the resulting counts before the freeze. No selective cell cuts or dropped cells are allowed under this rule. Low precision is reported as indeterminate under Section 6; a reduction does not justify weakening the criteria.

No reduction is applied in this draft: the displayed counts use $\gamma = 1$ for planning. The full budget has not yet been measured or approved. Appendix B shows why the 20-subset procedure must be included in the call count.

1. Review the reconstruction proposals in Appendix A and adopt or amend them before freezing.
2. Complete the remaining truth, wrapper, P1, driver and analysis checks without running main-study performance experiments.
3. Calculate estimator calls from the final grid, measure the dominant costs, and set the runtime-only replication factor and machine allocation.
4. Freeze the protocol, configuration, stream registry, software environment and hashes; obtain the authors' review and make a timestamped deposit before the main run.
5. Run the development experiments E1 to E3, retaining every record and deviation.
6. Write and freeze the practical guidance to evaluate, using only the permitted candidate components and development results.
7. Run E4 once under that guidance; retain and explain any bug-fix rerun.
8. Apply the reporting and contribution rules, assess novelty and scope, and then revise the manuscript and choose the journal.

Creating this document is not a public deposit, author approval or freeze. No external publication action has been taken.

13 Change history

Version	Source and substantive change
v0.1	Supplied PDF. Initial question, six development designs, matched targets and experiment grid. Included an erroneous Equation (8) warning and criteria later withdrawn.
v0.2	Reported in chat. Removed the equation warning; changed E2 to signed difference error; replaced the profile Spearman gate; added exact GC correction and revised truth precision.
v0.3	Reported in chat. Added prior-knowledge record, mean-profile range, descriptive E2 outcomes, procedure coverage, independent streams and a runtime gate.
v0.4	Reported in chat. Clarified nats and seed inputs, I_A reference, whole-catalogue uncertainty, runtime-only cuts and order of work.
v0.5	Reported in chat. Named FFT and adaptive real-space truth methods and required generator-to-reference checks for every design.
Reconstructed v0.5	This document. Consolidates available sources, adds completed implementation status, and labels unrecovered operational details R1 to R9. It is not a verified copy of the unavailable live version.

Appendix A Reconstruction proposals for review

These defaults make this draft concrete without falsely attributing missing settings to an earlier document. They remain open to documented changes before freezing. All other numerical grid values are taken from the supplied PDF or explicit subsequent changes.

ID	Detail not recoverable from the record	Proposed default in this document
R1	Full held-out grid and H3 active-pair count	Section 8 grid; three active H3 pairs; 1,000 calibration/comparison replicates and 200 profiles.

ID	Detail not recoverable from the record	Proposed default in this document
R2	Whether E2 samples share observations	Independent catalogues A and B; reuse the paired record only for outcome analysis and bootstrap.
R3	Secondary estimator coverage in profiles	KSG 4 c remains primary; all four k values and GC family are recorded for candidate comparisons.
R4	P1 profile common N, coupling and estimator coverage	Smallest prefix as common N; 20 subsets; all estimator families; reuse equal-N baseline; independent subset draws between prefixes.
R5	Which 10% receive B = 500 and how	Outcome-independent selection within each cell; whole replicate records and every associated P1 call; extend shared first 100 by 400.
R6	Bootstrap count and precise interval construction	9,999 resamples, 95% centred intervals; simultaneous pairwise mean-profile contrast bands for E3.
R7	Held-out improvement and degradation denominators	Per-family positive net pass increase plus at least one fail-to-pass conversion; count pass-to-indeterminate as degradation; zero denominator is not evaluable.
R8	Complete task and role registry	Extend the validated SeedSequence scheme with fixed integer experiment, prefix, subset and selection identifiers; freeze the mapping with code.
R9	Exact uniform-reduction rounding rule	One common γ across all original counts; round down to multiples of 10; retain every cell and at least 10 replicates.

Appendix B Preliminary estimator call counts

The following counts describe the complete proposed grid at $\gamma = 1$, including R1 to R5. One estimate call returns all four KSG choices and the GC family on one dataset. Do not multiply its cost by the number of returned procedures again. Raw scores, centred scores and GC a share that call. No result reuse across different cells is assumed.

Experiment	Cells	Replicate records	Baseline calls	Extra P1 calls
E1 development	210	162,000	162,000	0
E2 development	720	720,000 pairs	1,440,000	12,000,000
E3 development	8	1,600 catalogues	16,000	288,000
E4 calibration	24	24,000	24,000	0
E4 comparison	192	192,000 pairs	384,000	1,920,000
E4 profiles	8	1,600 catalogues	4,800	64,000
Total	1,162	1,101,200 records	2,030,800	14,272,000

The proposal therefore contains 16,302,800 baseline-plus-subset estimate calls. Exactly 1,630,280 of these calls receive a 400-permutation extension under the whole-replicate 10% rule. This is not an additional 1,630,280 complete B = 500 calls if continuation is implemented as specified. The arithmetic counts calls; it does not imply that this full grid has been budgeted or will be run at these replication counts.

For E2 development, five of six size pairs are unequal, giving $600 \text{ cells} \times 1,000 \text{ pairs} \times 20 \text{ additional B subsets}$. For E3 development, each of 1,600 catalogues has ten baseline prefixes and nine larger prefixes with 20 subsets each. The held-out comparison has three unequal and three equal size pairs; its profiles have three baseline prefixes and two larger prefixes.

Measured benchmark examples at $N = 5,000$ with all four k values were 13.70 seconds at $B = 10$ for Q1x and 12.91 seconds for T3. Extrapolated $B = 100$ costs were approximately 126.83 and 120.72 seconds. These costs do not apply to every P1 call: P1 uses the smaller common N . A budget must weight each call by its actual subset size, design and permutation count. Shared-host timing and missing-grid proxies remain material uncertainty sources.

References and source records

- [1] Pandey B and Mondal A. Stellar Halo Memory: A New Observable for Galactic Archaeology. arXiv:2607.10863v1, 2026. <https://doi.org/10.48550/arXiv.2607.10863>
- [2] Kraskov A, Stögbauer H and Grassberger P. Estimating mutual information. Physical Review E 69, 066138, 2004. <https://doi.org/10.1103/PhysRevE.69.066138>
- [3] Ince RAA, Giordano BL, Kayser C, Rousselet GA, Gross J and Schyns PG. A statistical framework for neuroimaging data analysis based on mutual information estimated via a Gaussian copula. Human Brain Mapping 38, 1541–1573, 2017. <https://doi.org/10.1002/hbm.23471>
- [4] Ince RAA. Official GCMI implementation. Reviewed Python source blob f0bb30affdb6f990a6c7e87cd7b3a14d0fee780e. https://github.com/robinince/gcmi/blob/master/python/src/gcmi/_core.py
- [5] NumPy documentation. Parallel random number generation. <https://numpy.org/doc/stable/reference/random/parallel.html>

Reconstruction sources: supplied “Pre-registered simulation protocol permutation-centred mutual information” PDF, draft v0.1; the v0.2 to v0.5 change descriptions supplied in this conversation; MI_Truth_Validation_v0_1.zip; and MI_Estimator_Validation_v0_2.zip. The latter ZIP has SHA-256 94ce5214023565cf7e26ae95dba0fb818d893f02f9b2a77fb5b01d438bdcfc93. These sources are distinguished from an unavailable original v0.5 export.