Resampling Robustness Does Not Establish Comparator Adequacy: A Leakage-Free Hyperspectral Benchmark

Abstract

Empirical machine-learning benchmarks quantify uncertainty over seeds, splits and sampling masks, but these estimates are conditional on a fixed benchmark specification: a preprocessing pipeline that must respect the hold-out, and a comparator set that defines the reported margin. We study both assumptions on identical stored runs from a hyperspectral reconstruction benchmark. First, we formulate hold-out invariance and show that smoothing spectra before the hold-out is chosen violates it: a 7-point Savitzky–Golay filter contaminates 99.6 % of nominally observed channels under an interleaved split and 37.5 % under a binned variant. In a fixed-target control this contamination produces heterogeneous error reductions, with no median effect under a spatial hold-out the spectral smoother cannot contaminate and reductions as large as 46 % in individual method-and-cube cells. Second, we define comparator-set sensitivity and measure it from stored per-run errors. An additive Gaussian field scored against two comparator baselines, linear interpolation and PCA, is best on 11 of 12 cubes by 12.5 % (7.7–17.1 % hierarchical bootstrap interval), yet against the full evaluated pool of 15 comparators it trails the strongest method by 24.4 % (16.8–33.4 %). The same reversal occurs for SIREN, whereas Soft-Impute remains sign-stable. Across comparator subsets, variation can cross the scientific decision boundary even where repeated sampling yields a narrow interval conditional on the restricted suite. We therefore propose a Comparator-Set Robustness Audit that reports how a claim depends on the evaluated comparator pool, while treating subset agreement as a descriptive sensitivity analysis rather than a probability over future benchmark choices.

1 Introduction

Empirical machine learning often treats robustness as a sampling problem: rerun seeds, perturb splits, report intervals, and ask whether a conclusion survives resampling (Henderson et al., 2018; Bouthillier et al., 2021). These checks are necessary, but they condition on a fixed benchmark specification. In particular, they assume that preprocessing respects the hold-out and that the chosen comparator set is adequate for the claim being made. Additional resampling can reduce uncertainty conditional on those choices; it cannot test choices that remain fixed. We therefore distinguish sampling uncertainty from benchmark-specification sensitivity and study both on the same stored runs.

Target leakage is a preprocessing operator whose output at an observed coordinate depends on a value that will be withheld. Comparator-set sensitivity is the extent to which a reported margin changes when the evaluated comparator subset changes. The first is a property of the pipeline and can be tested directly; the second is a property of the comparison and is usually argued qualitatively.

What is new here. That weak or incomplete baselines can exaggerate apparent progress is established (Ferrari Dacrema et al., 2019; Musgrave et al., 2020; Lipton & Steinhardt, 2019), and we do not claim otherwise. Our contribution is to make comparator dependence an explicit, measurable sensitivity axis of the benchmark estimand. Because a reported margin is defined relative to the best error in a comparator set, the quantity being estimated is indexed by that set. Ordinary resampling quantifies uncertainty conditional on the set and

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

therefore cannot reveal how the claim changes when different comparators are admitted. We measure this dependence directly from stored per-run errors and report sign stability, the location of the originally reported set among equal-size alternatives, influential comparators and families, and approximations for pools too large to enumerate exhaustively.

Why hyperspectral reconstruction is a good case study. The benchmark contains several independently variable design choices. Preprocessing acts along the same spectral axis that can be withheld, making leakage directly testable. Multiple method families, interpolation, geostatistics, low-rank factorisation, tensor methods, coordinate networks and explicit Gaussian fields, solve related reconstruction tasks under different assumptions. Different missingness geometries reward different inductive biases, and every reported error can be recomputed from stored artifacts. The additive Gaussian field, the natural adaptation of 3D Gaussian Splatting (Kerbl et al., 2023) to a spectrum image, is used here as a controlled method under test. The field is not the contribution of this paper; it is the controlled case through which two benchmark failure modes become measurable.

What we find. Spectral smoothing applied before the hold-out is chosen carries withheld values into the training view: with a 7-point Savitzky–Golay filter, withholding every 5th raw channel contaminates 99.6 % of nominally observed channels, and the binned variant 37.5 %. Under a fixed-target control, where both pipelines are scored against the same raw withheld values, the effect is heterogeneous rather than a uniform discount: the median method-and-cube cell improves by 0.4 % and the largest by 46 %, while a spatial hold-out the smoother cannot contaminate has no median leakage advantage. We therefore treat the contaminated fraction as a geometric exposure measure, not as a linear predictor of gain (§4).

The result we did not expect. A restricted benchmark can be both statistically precise and substantively wrong. Against two comparator baselines, linear interpolation and PCA, plus a diagnostic constant predictor, the field is best on 11 of 12 cubes by 12.5 %, with an interval excluding zero that remains positive across 20 masks per cube, threshold sweeps and a permutation test. Against the full evaluated pool it is best on 0 of 12 and trails the strongest comparator by 24.4 %. Robust central summaries of the comparator axis are of comparable scale to that sampling interval; the substantive distinction is that comparator changes cross the scientific decision boundary whereas the restricted sampling interval does not. The same qualitative reversal occurs when SIREN is the method under test, while Soft-Impute remains sign-stable (§5.3–5.4).

What we are not claiming. We do not interpret the fraction of enumerated comparator subsets that agree with the full pool as a probability that a future benchmark will reach the same conclusion. The subsets form a finite sensitivity lattice, not a probability distribution over scientific practice. We also do not claim that our comparator pool is exhaustive, that a fixed number or family count guarantees adequacy, or that an audit can certify methods that were never evaluated. The audit is retrospective: it measures how conditional a reported claim is on the evaluated pool.

Contributions. (i) Hold-out invariance. We formalise a black-box tamper test for whether withheld targets can influence the training view, and use fixed-target controls to separate contamination from the easier-target effect introduced by smoothing (§4). (ii) Benchmark uncertainty. We separate ordinary sampling uncertainty from comparator-set sensitivity, and evaluate methods only on missingness geometries for which their parameterisation is structurally applicable (§4.1, §5.2). (iii) Comparator-Set Robustness Audit. From stored per-run errors we report sign stability over the evaluated comparator lattice, the position of the originally reported set among equal-size alternatives, leave-one-family-out sensitivity, and influential comparators (Appendix N). (iv) Empirical case study. We show that a restricted comparator suite can yield a narrow, repeatable positive conclusion that reverses under a broader evaluated pool, examine the effect for 3 methods under test across 5 geometries, and stress-test the direction on public FIB-EDX data (§5.2–5.5).

2 Related work

Explicit primitives and neural fields. 3D Gaussian Splatting (Kerbl et al., 2023) established explicit anisotropic primitives as a scene representation, and hyperspectral variants followed (Thirgood et al., 2025; Sinha et al., 2025); those methods synthesise novel views from calibrated cameras, so the representation transfers to a spectrum image but the task does not. GaussianImage (Zhang et al., 2024) drops alpha compositing for accumulation for the same reason we do; our third axis is a physical spectral coordinate rather than a second spatial one, which is what makes the spectral hold-outs of §4 meaningful. Coordinate networks are the other continuous option (Mildenhall et al., 2021; Müller et al., 2022; Chen et al., 2024); both families are in our pool, a Fourier-feature network (Tancik et al., 2020) and a SIREN (Sitzmann et al., 2020).

Factorisations and completion, which are the real competition. Spectrum images have been compressed and unmixed by bilinear and multilinear factorisation for decades (Bioucas-Dias et al., 2012): truncated SVD, non-negative matrix factorisation (Lee & Seung, 1999), MCR-ALS (Jaumot et al., 2015), which adds the constraints that make components interpretable as chemical species, and tensor decompositions that exploit the third axis directly (Kolda & Bader, 2009; Kossaifi et al., 2019). Scattered missingness is matrix completion (Candès & Recht, 2009; Liu et al., 2013), where the reference method shrinks singular values rather than truncating them (Mazumder et al., 2010); that distinction turns out to matter here, because a truncated factorisation and a shrunken one from the same family land on opposite sides of our conclusion. RamanSPy (Georgiev et al., 2024b) standardises the preprocessing pipeline such comparisons depend on (Georgiev et al., 2024a).

Scattered-data interpolation and geostatistics. A Gaussian field fitted to scattered samples is a radial basis function expansion with fitted centres, so the classical methods built on the same idea are the comparison it most needs: multiquadric and thin-plate interpolation (Hardy, 1971; Wendland, 2004), and kriging (Matheron, 1963; Cressie, 1990), which is Gaussian process regression with a covariance fitted to the data (Rasmussen & Williams, 2005) and models exactly the spatial autocorrelation our own predictor is built on. Both were absent from the comparison that first convinced us, which is the subject of §5.3.

Evaluation: four literatures and the gap between them. Benchmark variance is the most audited source of error: seeds, splits and sampling all move a reported number (Henderson et al., 2018; Bouthillier et al., 2021). Leakage is the familiar hazard of estimating anything from the full data before splitting it (Kaufman et al., 2012), widespread in machine-learning-based science (Kapoor & Narayanan, 2023); it appears here in a form specific to spectroscopy, where smoothing along the axis one is about to withhold from is standard practice. Weak baselines are documented to manufacture progress (Ferrari Dacrema et al., 2019; Musgrave et al., 2020; Lipton & Steinhardt, 2019). What is largely absent is sensitivity analysis of the benchmark specification itself: the comparator set is rarely treated as an argument of the estimand, baseline criticism is usually qualitative and retrospective, and variation across comparator sets is rarely put on the same scale as run-level resampling. That is the gap this paper measures, and it reports the measurement as a procedure rather than a verdict.

3 An additive spectral Gaussian field

The field is the controlled method under test, not the contribution; Figure 2 shows one fitted cell and how it fails.

3.1 Representation

A measured volume is represented by a set of anisotropic Gaussian primitives evaluated directly on the sampling grid,

$$\hat{V}(\mathbf{p}) \;=\; \sum_i a_i \exp\!\big(-\tfrac{1}{2}(\mathbf{p}-\boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\mathbf{p}-\boldsymbol{\mu}_i)\big)\,, \qquad \boldsymbol{\Sigma}_i = \mathbf{R}(\mathbf{q}_i)\,\mathbf{S}_i\mathbf{S}_i^\top\mathbf{R}(\mathbf{q}_i)^\top, \qquad (1)$$

with $\mathbf{p} = (x, y, \nu)$ the two spatial coordinates and the spectral one. Three things differ from scene-rendering Gaussian splatting, and all follow from the measurement. Primitives accumulate additively, because there is no camera and nothing to occlude, which is the correct forward model for incoherent emission in the optically thin limit. Amplitudes are non-negative, as $a_i = \exp(\alpha_i)$, so no primitive can subtract. The axes share one frame: micrometres and cm^{-1} are incommensurable, so all three are mapped to $[0, 1]$ and the amplitude divided by a high percentile of the masked cube, without which first-order methods stall.

The rotation is a full three-dimensional one, so a primitive's band centre may drift across the specimen. The fitted fields use that freedom (49 % of primitives above 0.3) but do not need it: constraining the rotation axis to ν changes held-out PSNR by $+0.24\,\mathrm{dB}$ (-0.26 to $+0.86\,\mathrm{dB}$).

3.2 Fitting

Each cube is fitted independently by first-order descent on closed-form gradients of

$$\mathcal{L} \;=\; \tfrac{1}{2} \sum_{\mathbf{p} \,\in\, \text{observed}} \big(\hat{V}(\mathbf{p}) - V(\mathbf{p})\big)^2 \;+\; \lambda_a \sum_i a_i \;+\; \lambda_s \sum_i \|\log \mathbf{s}_i\|^2, \tag{2}$$

entirely on a CPU. The sum runs over observed entries only, so the hold-out enters the fit nowhere. The loss is unweighted, which after baseline removal and normalisation is defensible for Raman but not for a photon-counting modality, where a variance-stabilising transform is required (§5.5); the remaining implementation details are in Appendix B.

Comparing a Gaussian field with a truncated SVD or a coordinate network is only meaningful at equal stored size, so every method is measured by what it must write to reconstruct the cube: the field stores 11 scalars per primitive at 32-bit precision, uncompressed, which is the pessimistic accounting and the honest one to report. Budgets are quoted as a fraction of the cube's own scalar count (Appendix H).

4 Benchmark validity: leakage and comparator-set dependence

Reconstruction quality is assessed by withholding part of the measurement and predicting it. Two properties of that procedure decide whether the numbers mean anything, and repeating it tests neither: whether the training view is free of the values being predicted, and whether the comparison contains the methods that suit the task.

4.1 Two things a benchmark holds fixed

A benchmark specification is a tuple $B = (P, H, S, A)$: a preprocessing pipeline, a hold-out generator, a comparator set and an aggregation. Resampling varies the realisations drawn under a fixed B; the two failures above are failures of P and S (Figure 10).

Definition 1 (Hold-out invariance). P is hold-out invariant for H if replacing the withheld raw entries by arbitrary values leaves every observed entry of the training view, and every statistic fitted from it, unchanged. §4.2 is a test of this property, not an argument about it.

For a method under test T, comparator set S and run $r = (c, m)$, a cube and a mask, the reported margin is

$$M(T, r, S) \;=\; \frac{E_{\min}(S, r) - E_T(r)}{E_{\min}(S, r)} \;=\; 1 - \frac{E_T(r)}{E_{\min}(S, r)}, \qquad E_{\min}(S, r) = \min_{j \in S} E_j(r), \tag{3}$$

with all errors positive, and the claim is made about its aggregate $\bar{M}(T, S) = A\big[M(T, r, S)\big]$ over runs. Positive means the method under test beat every member of S.

Definition 2 (Comparator robustness). A claim is comparator robust over an evaluated pool $\mathcal{P}$ if it holds under every non-empty $S \subseteq \mathcal{P}$. Robustness is claim-relative: the same pool can support "T is competitive" and fail to support "T is best".

What this measures is sensitivity within $\mathcal{P}$, which is an operational diagnostic of comparator adequacy and not a certificate of it: $\mathcal{P}$ itself may omit the method that matters. We therefore use the full evaluated pool as a reference rather than as ground truth, the most inclusive comparison this study can make, against which restricted subsets are scored.

Two properties of (3) explain why the two failures behave differently under replication. Both are elementary; their value is that they make explicit why resampling and comparator-set sensitivity answer different questions, and the contribution is operationalising that consequence.

Proposition 1 (Comparator monotonicity). If $S_1 \subseteq S_2$ then $M(T, r, S_2) \leq M(T, r, S_1)$ for every run.

Proof. $E_{\min}(S_2, r) \leq E_{\min}(S_1, r)$, and $M = 1 - E_T/E_{\min}$ is non-decreasing in $E_{\min}$. □

A restricted benchmark is therefore not one noisy estimate of a fixed quantity: it is an upper envelope on the margin, and the sign flips exactly when the added comparator beats T on the aggregate. Proposition 1 makes one aspect of comparator expansion trivial: when a claim is defined against the best available comparator, adding methods can only reduce or preserve the margin. The empirical question is therefore not whether the margin decreases, but whether realistic or reported restrictions on the comparator pool cross the decision boundary, and whether that dependence remains hidden under otherwise convincing resampling checks.

Proposition 2 (Conditionality of resampling). Any procedure that varies runs r with S held fixed estimates uncertainty in $\bar{M}(T, S)$, a quantity indexed by S. More seeds, masks or cubes may shrink that uncertainty; none of it identifies $\bar{M}(T, S')$ for $S' \neq S$.

The two axes are therefore different arguments of the same functional: the comparator set is not a nuisance variable outside the claim, it enters through $E_{\min}$, so changing it changes the estimand and not merely its variance.

4.2 Preprocessing carries the answer into the question

Any preprocessing operator whose output at an observed coordinate depends on a value later assigned to the hold-out violates Definition 1; spectral smoothing is the instance we can quantify. Raman spectra are smoothed before almost anything else is done to them, and a Savitzky–Golay filter is the standard choice (Savitzky & Golay, 1964). A centred filter of window w makes output channel k a function of channels $k \pm \lfloor w/2 \rfloor$, so if the hold-out is chosen after smoothing, every observed channel within $\lfloor w/2 \rfloor$ of a withheld one carries its value.

Interleaved spectral withholding is a plausible protocol whose variants appear in prior hyperspectral evaluation, and under a centred smoother its leakage depends on the spacing and binning rule: in our two variants 99.6 % and 37.5 % of nominally observed channels lie within the support of a withheld value, against 2.2 % for a contiguous band and none for a spatial hold-out (Figure 3).

Our protocol seals the targets before any parameter is estimated, fits the baseline and the normalisation from observed entries only, and omits spectral smoothing, since a smoother is itself a reconstruction of the missing values. The gate is the tamper test of Definition 1, which needs no knowledge of the operator; the conventional pipeline fails it and ours passes for all hold-out families.

What the contamination costs, at a fixed target. A contaminated fraction is a property of the stencil, and the obvious comparison confounds two things, since a pipeline that smooths before splitting also predicts a smoother target. We therefore ran both pipelines on the same cubes, mask, seed, budget and methods and scored both against one fixed target, the raw withheld values of the untouched cube, each prediction mapped back to raw units by its own baseline and scale, so the arms differ only in whether withheld values could reach the training view. Over 96 method-and-cube cells the effect is heavy-tailed rather than a uniform discount: the median cell gains 0.4 % and the mean 10 %, the upper quartile exceeds 22 % and the largest reaches 46 %. Scored instead on each pipeline's own target, as a study

using it would, the same runs report a flat 14 % median reduction, so most of that apparent across-the-board gain is the easier target. Contamination advantages some methods and not others, which is how it distorts a comparison.

What the control says the effect is not. The controls establish the direction of the effect without implying a linear dose–response (Appendix E). Under the interleaved split, contamination is nearly complete at 99.7 % and the median fixed-target reduction is 0.83 %; under the contiguous block it is 0.04 %; under the spatial hold-out, which the spectral smoother cannot contaminate, it is 0.00 %. Individual cells are heavy-tailed, so the contaminated fraction should be read as exposure to leakage rather than as a proportional predictor of the reduction. Rank changes should not be attributed causally to leakage alone: ordering changes under both the contaminated and the uncontaminated condition (4 and 5 of 45 pairs), which shows that smoothing itself can alter rankings even where withheld values cannot enter the training view. We therefore make a causal claim only about the leakage-induced error advantage.

Smoothing also disqualifies the usual claim that a reconstruction has reached the measurement limit: the noise estimate that claim rests on stops measuring the instrument once the spectra are filtered, so we report a within-pipeline residual scale instead (Appendix F).

4.3 Hold-outs that separate the abilities being claimed

A single hold-out cannot distinguish spectral interpolation from spatial generalisation. We use six families, each one boolean mask over the volume: H0 every nth channel; H1 one contiguous channel run; H2 spatial patches behind a guard ring; H3 a channel run over a spatial subset; R5 scattered voxels; and R7 complete spectra at a random subset of stage positions, the sparse survey an instrument can actually run (Appendix D).

H2 and a regular super-resolution variant do not discriminate, all 10 methods landing within 1.4 % of one another, so we report them but never rank them (Appendix D).

5 Experiments

5.1 Data, comparators and fairness

The corpus is 23 Raman acquisitions of carbonaceous specimens, which hashing reduces to 18 distinct cubes. The cube is the unit of replication, so margins are averaged over the masks of a cube and then over cubes, and every interval is a hierarchical bootstrap (Efron & Tibshirani, 1994) over 10 000 draws resampling cubes and then masks: the voxel-wise results are 12 cubes measured 20 times, not 240 independent samples.

The evaluated pool contains 15 comparators in 6 families. 8 are representation methods matched to the field's stored-scalar budget: PCA, Soft-Impute, NMF, MCR-ALS, Tucker, CP, a Fourier-feature network and SIREN. 7 are interpolation or geostatistical references evaluated under their natural observation-retention model: linear interpolation, spatial interpolation, three radial-basis interpolants and two kriging models. We do not describe the latter as storage-matched because they retain the observations they interpolate. The constant predictor is a diagnostic floor and is never counted as a comparator baseline. The field receives at least as much tuning freedom as anything it meets, a full coordinate search frozen on validation masks, and nothing is tuned on a test mask (Appendix C). Every budget is 2 % of the cube's own scalar count.

5.2 The field across missingness geometries

Among the methods that can use an observed entry for the task (Figure 5 and Table 6), the field places 8th of 9 on the contiguous spectral block and 16th of 16 on the joint block, 10th of 16 under scattered voxels, and 9th of 9 on the sparse survey; on the public field it is 15th of 16, and on the joint block it ranks below the diagnostic floor itself. Among methods structurally applicable to each geometry, the field therefore ranks near the bottom of the

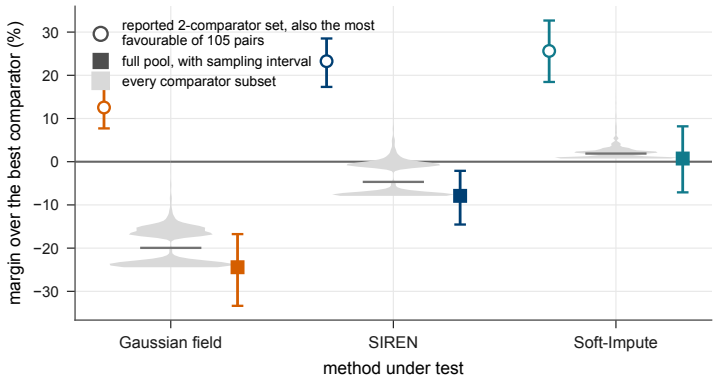


Figure 1: Sampling uncertainty and comparator-set sensitivity on the same stored runs. For each method under test, the open circle is the margin under the 2-comparator set the corpus study actually reported, which the audit afterwards showed to be the most favourable of the 105 such pairs, with its hierarchical bootstrap interval, the grey distribution the exact margins over all non-empty subsets of the evaluated pool, and the filled square the full-pool margin with its interval. Subset distributions are descriptive sensitivity summaries, not probability distributions over future benchmark choices. The nontrivial result is the crossing of the zero-margin decision boundary under comparator changes that ordinary resampling does not interrogate.

evaluated pool across the principal tasks. No single comparator dominates every geometry either, which is why we report task-specific eligible method sets rather than one universal ranking: kriging (Matérn) takes the spectral block, Soft-Impute the joint block, SIREN scattered voxels and RBF (Gaussian) the survey. Nominal parameter count alone does not explain the gap (Appendices H and R).

5.3 Sampling robustness does not establish comparator adequacy

A restricted comparison gives a robust positive result, conditional on that comparison. Scored against two comparator baselines, linear interpolation and PCA, with the diagnostic constant shown separately, the same runs make the field best on 11 of 12 cubes by 12.5 % (7.7–17.1 %), excluding zero. The result remains positive across 20 masks per cube and every region-of-interest threshold from 0 to 500 pixels (88 to 93 % of cubes won), and a spatial-autocorrelation association is observed under the restricted suite ($r = 0.76$, permutation test $p = 0.013$, Bonferroni over 6 predictors). A leave-one-cube-out model predicts the sign of that margin on 12 of 12 cubes against 11 for an always-positive rule, so sign accuracy is secondary; its role is to show that a coherent mechanistic explanation can emerge inside a comparator-conditioned benchmark. These checks establish that the restricted estimate is repeatable under the sampled data; they do not test whether the comparator set is adequate.

The comparator set is what the checks above hold fixed. Every one of them conditions on the same restricted suite. Against the full evaluated pool the field is best on 0 of 12 cubes and trails by 24.4 % (16.8–33.4 %), and the restricted spatial-autocorrelation association changes sign, from $r = 0.76$ to $r = -0.46$; rerun there, the same checks report the field behind at every threshold (0 % of cubes at best) and the association no longer significant ($p = 0.61$ corrected). For the field, whose reported pair the audit afterwards showed to be the most favourable of 105 (Figure 1), the restricted sampling interval is 9.4 points wide. Across the finite lattice of evaluated comparator subsets the full margin range is 39.5 points, the interquartile range is 7.9 points, and the central 90 % band is of comparable scale to the sampling interval; only 31 of the 32 767 subsets support the positive claim. The important result is therefore not that comparator variation is always numerically larger than sampling variation, but that changing the comparator set can cross the scientific decision boundary where repeated sampling around the restricted estimate does not. Increasing within-cube draws narrows that interval, from 7.4–17.8 % at 3 draws to 7.7–17.1 % at 20,

but still conditions on the same comparator set. Throughout, a share of subsets agreeing is the exact descriptive fraction of non-empty subsets whose aggregate margin has the same sign as the full-pool margin, not a probability that a future study would choose such a set.

The reversal is not unique to one method. When SIREN is treated as the method under test and the rest of the pool as its comparators, the restricted comparator set yields a positive margin of 23.3 % (17.3–28.5 %) whereas the full evaluated pool yields −7.9 %, with 4 623 of the 32 767 subsets agreeing with the restricted verdict against the field's 31. Soft-Impute provides a sign-stable control case: its margin varies across comparator subsets, by 26.9 points, but its conclusion remains positive throughout the evaluated lattice (32 767 of 32 767). The audit therefore distinguishes claims with different comparator sensitivity rather than uniformly penalising every method, and which claim is being made matters: SIREN is best in 14 % of subsets but within 5 % of the best comparator in 51 %, so one pool can support a non-inferiority claim and refuse a winner-takes-all one.

The audit. We package these analyses as the Comparator-Set Robustness Audit (Appendix N), a retrospective sensitivity analysis computed from stored per-run errors: the claim under the reported set, run-level uncertainty conditional on it, the same claim over every non-empty subset of the pool, the location of the reported set among equal-size alternatives, leave-one-family-out sensitivity, and the comparators whose inclusion most changes the margin. It measures dependence within the evaluated pool and cannot certify untested methods. Its outputs locate our own failure: of the 105 comparator pairs, 89 % reach the full pool's conclusion and ours ranks 1 by the margin it reports, its two comparators ranking 14th and 15th of 15. Nor does the conclusion rest on one reference: the field trails by −24.4 % against the full pool, −23.7 % without its strongest comparator, and at worst −17.0 % with a whole family removed.

5.4 Where comparator sensitivity bites, and where it does not

Comparator sensitivity is a property of a particular claim under a particular regime, not a generic penalty attached to a method. We repeated the audit on every geometry for each method under test: 13 cells over 5 geometries and 3 methods, all from stored errors. Some claims are sign-stable over the whole subset lattice while others change sign under substantial fractions of it, from 60 % of subsets agreeing with the full pool for Soft-Impute on H3 to 100 % elsewhere (Figure 9). A method far from the best evaluated comparator is often insensitive because no subset within the pool rescues it; a method close to the frontier can be much more sensitive to comparator choice. The comparator range exceeds the sampling width in 11 of 13 cells; in the cells where it does not, on the block hold-outs with one mask per cube, additional replication is the more useful thing to add. In our stored-run simulation, broader comparator coverage more often recovers the full-pool sign than additional mask repetitions at the same number of evaluation cells, a benchmark-specific sensitivity result rather than a compute-allocation rule, since methods differ in cost (Appendix P).

5.5 External stress test on a public FIB-EDX field

The corpus cannot be released, so we repeat the comparison on a public FIB-EDX map of a battery electrode (Donoghue et al., 2025) with the same protocol, budget and frozen configuration. The 4 binnings are 4 resolutions of one independently acquired field, spanning autocorrelation $r = -0.02$ to 0.80, not independent replicates. The archive carries no machine-readable specification of its raw layout, so we document the inferred byte order, storage order and geometry in Appendix T and check the recovered dimensions against the released element maps. The channel axis is not independently energy-calibrated, so we use the channel index only as an ordered spectral coordinate and make no chemical-line claim: this is a public count-field stress test of the reconstruction ranking, not a calibrated EDX analysis. The field wins 0 of 40 runs, trailing by 45 to 216 % and most where autocorrelation is highest (correlation −0.98), the direction of the private full-pool association and the opposite of the restricted one. To test whether the ranking is an artefact of fitting heteroscedastic count data with unweighted squared error, we rerun the full suite after an

Anscombe variance-stabilising transform (Anscombe, 1948): the raw-count and transformed rankings have rank correlation 0.98 and the qualitative conclusion is unchanged. That check uses 5 seeds per binning rather than 10, being a ranking-robustness check rather than a second primary benchmark (Appendix T).

6 Limitations

The audit is a sensitivity analysis over an evaluated comparator pool, not a certificate of benchmark validity. The subset lattice is finite and carries no natural probability distribution over the comparator sets researchers are likely to choose, so subset-agreement fractions are descriptive. The full evaluated pool is a reference, not ground truth, and comparator strength, being read from full-pool results, explains which sets failed rather than predicting which would succeed. The public FIB-EDX experiment uses one acquired field at several binnings and an inferred raw-file layout without independently validated energy calibration, so it is an external reconstruction stress test rather than an independent chemically calibrated replication. Robustness is claim-specific: a winner-takes-all claim, a non-inferiority claim and a ranking claim can have different sensitivity to the same pool.

No finite pool establishes universal comparator adequacy. Graph-regularised completion, learned cross-cube priors, diffusion-based inpainting and other methods remain outside the present pool; their absence does not invalidate the audit, it defines its scope. A stable conclusion is likewise not a valid one: if every evaluated comparator is inappropriate for the task, the audit reports agreement and is of no help. The audit is also conditional on the metric, the aggregation, the region of interest and the missingness generator, the other arguments of B, and extending it across metrics is future work.

Scale. The 39.5-point range is a measurement on this benchmark, not a rate to expect elsewhere. Scalability. Exhaustive enumeration is feasible to about twenty comparators; beyond that the audit's quantities need sampling (Appendix N). Data and method. One private corpus and one public specimen test the direction of the result rather than its size, a compressive instrument samples along a geometry we do not test, and the comparators were not tuned per cube while the field received a full search.

7 Conclusion

Benchmark robustness is conditional on what the benchmark holds fixed. Resampling estimates uncertainty in a quantity defined under one preprocessing pipeline, one missingness generator, one metric and one comparator set; it makes that estimate more precise, but cannot reveal a target leak introduced before the split, or a method never allowed to compete.

In our case study a restricted comparator suite yields a narrow positive estimate for an additive Gaussian field, and the conclusion reverses under the broader evaluated pool; the same reversal occurs for SIREN, whereas Soft-Impute remains sign-stable. The audit is not a probability over future benchmark choices: its purpose is to reveal how conditional a claim is on the methods already evaluated.

Together with hold-out invariance this suggests a reporting standard: verify that withheld targets cannot change the training view, report run-level uncertainty, name the comparator set, and audit whether the claim survives changes within the evaluated pool. A narrow confidence interval establishes precision conditional on a benchmark specification; it does not establish that the preprocessing or comparator set defining that benchmark was adequate.

Reproducibility statement

Every number in this paper is generated from result artifacts rather than transcribed: each experiment writes an immutable JSON record carrying its dataset, method, hold-out specification, storage budget, seed, hyperparameters, runtime and metrics, and the manuscript reads those through generated macros, so the prose cannot drift from the measurements. The anonymised supplement contains the representation and every baseline, the hold-out

generators, the leakage-free split with its tamper test, the storage accounting, the hyperparameter searches with their full traces, the comparator-subset analysis, and the scripts that reproduce each figure and table from the stored records. Seeds are fixed and reported per run; the searches and sweeps are deterministic given them. Section 3 gives the field and its objective in full, Section 4 the protocol, and the appendix the gradient derivations and the per-cube results behind the aggregates.

The Raman corpus is a private collection and cannot be redistributed, so the reproducibility boundary is worth stating precisely. Without the raw cubes a reader can still reproduce every derivation, every comparator-set analysis from the per-run records, every reported aggregate, every figure and table built from stored results, and the public FIB-EDX experiment end to end; only the fitting of the methods on the private cubes needs the cubes. The supplement therefore ships the complete pipeline together with those records. To make the central comparison independently checkable, Section 5.5 repeats it on a public FIB-EDX map (Donoghue et al., 2025), and the supplement runs that validation end to end from the public download, with the parsing assumptions and validation checks used for the public archive, whose raw spectral-cube layout is not accompanied by a machine-readable format specification, in Appendix T.

Ethics statement

This work involves no human subjects, no animal subjects and no personally identifying data. The Raman corpus was acquired for materials characterisation and was anonymised before analysis: specimen identifiers, material designations, instrument configurations, operator names and dated session folders were replaced with opaque codes by an automated stage whose output is checked by a gate that fails on any identifier leak. Two public reference standards are named because they are standards rather than study identifiers. No proprietary vendor data or file-format documentation is redistributed.

AI use statement

Generative AI tools were used at the following task levels in this project: implementation and debugging of the representation, the comparator suite, the evaluation pipeline and the analysis code; drafting and editing of manuscript text, including clarity, concision and phrasing; figure design; literature discovery; and discussion of how to organise the manuscript and present the research questions. AI assistance also contributed to experimental design choices, including the hold-out families, the comparator families added to the pool and the fixed-target leakage control, and to the identification of defects in our own earlier methodology, including the preprocessing leakage, the instability of the baseline fit under sparse observation and the duplicate acquisitions in the corpus.

The authors made all final research, methodological and experimental decisions. All executable analyses were run by the authors from version-controlled code; every numerical claim in the manuscript is generated from the immutable experiment records rather than transcribed, and each was checked against the record that produced it, as were the citations. The authors take responsibility for the complete contents of the submission.

References

F. J. Anscombe. The transformation of Poisson, binomial and negative-binomial data. Biometrika, 35(3–4):246–254, 1948. doi: 10.1093/biomet/35.3-4.246.

José M. Bioucas-Dias, Antonio Plaza, Nicolas Dobigeon, Mario Parente, Qian Du, Paul Gader, and Jocelyn Chanussot. Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 5(2):354–379, 2012. doi: 10.1109/JSTARS.2012.2194696.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Naz Sepah, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Dmitriy Serdyuk, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. In Proceedings of Machine Learning and Systems, volume 3, pp. 747–769, 2021.

Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. Foundations of Computational Mathematics, 9(6):717–772, 2009. doi: 10.1007/s10208-009-9045-5.

Junqi Chen, Xinjie Zhang, Zhe Gao, Tongda Xu, Yan Wang, and Jun Zhang. HINER: Neural representation for hyperspectral image. In ACM International Conference on Multimedia (MM), 2024.

Noel Cressie. The origins of kriging. Mathematical Geology, 22(3):239–252, 1990. doi: 10.1007/BF00889887.

Ronan Docherty, Sam Riley, John D. Morley, et al. Battery imaging library: Multi-length scale and multi-modal synchrotron and laboratory battery imaging data for all. ChemRxiv, 2026.

J. Donoghue, A. Vamvakeros, R. Docherty, and S. J. Cooper. Battery imaging library: Laboratory 2D FIB EDX NMC811 electrode (slow) (raw data). Zenodo dataset, 2025. CC-BY-4.0.

Bradley Efron and R.J. Tibshirani. An Introduction to the Bootstrap. Chapman and Hall/CRC, 1994. doi: 10.1201/9780429246593.

A. C. Ferrari and J. Robertson. Interpretation of Raman spectra of disordered and amorphous carbon. Physical Review B, 61(20):14095–14107, 2000. doi: 10.1103/PhysRevB.61.14095.

Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. Are we really making much progress? a worrying analysis of recent neural recommendation approaches. In Proceedings of the 13th ACM Conference on Recommender Systems, pp. 101–109, 2019. doi: 10.1145/3298689.3347058.

D. Georgiev, A. Fernández-Galiana, S. V. Pedersen, G. Papadopoulos, R. Xie, M. M. Stevens, and M. Barahona. Hyperspectral unmixing for raman spectroscopy via physics-constrained autoencoders. arXiv preprint arXiv:2403.04526, 2024a.

Dimitar Georgiev, Álvaro Fernández-Galiana, Simon Vilá Pedrosa, Georgios Papadopoulos, Ruoxiao Xie, Molly M. Stevens, and Mauricio Barahona. RamanSPy: An open-source python package for integrated raman spectroscopy data analysis. Analytical Chemistry, 96(21):8492–8500, 2024b. doi: 10.1021/acs.analchem.4c00383.

Rolland L. Hardy. Multiquadric equations of topography and other irregular surfaces. Journal of Geophysical Research, 76(8):1905–1915, 1971. doi: 10.1029/JB076i008p01905.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1), 2018. doi: 10.1609/aaai.v32i1.11694.

Joaquim Jaumot, Anna de Juan, and Romà Tauler. MCR-ALS GUI 2.0: New features and applications. Chemometrics and Intelligent Laboratory Systems, 140:1–12, 2015.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. Patterns, 4(9):100804, 2023. doi: 10.1016/j.patter.2023.100804.

Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. Leakage in data mining: Formulation, detection, and avoidance. ACM Transactions on Knowledge Discovery from Data, 6(4):1–21, 2012. doi: 10.1145/2382577.2382579.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics, 42(4), 2023. doi: 10.1145/3592433.

Tamara G. Kolda and Brett W. Bader. Tensor decompositions and applications. SIAM Review, 51(3):455–500, 2009.

Jean Kossaifi, Yannis Panagakis, Anima Anandkumar, and Maja Pantic. TensorLy: Tensor learning in Python. Journal of Machine Learning Research, 20(26):1–6, 2019.

Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755):788–791, 1999. doi: 10.1038/44565.

Zachary C. Lipton and Jacob Steinhardt. Troubling trends in machine learning scholarship. Queue, 17(1):45–77, 2019. doi: 10.1145/3317287.3328534.

Ji Liu, Przemyslaw Musialski, Peter Wonka, and Jieping Ye. Tensor completion for estimating missing values in visual data. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(1):208–220, 2013. doi: 10.1109/TPAMI.2012.39.

Georges Matheron. Principles of geostatistics. Economic Geology, 58(8):1246–1266, 1963. doi: 10.2113/gsecongeo.58.8.1246.

Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. Journal of Machine Learning Research, 11(80): 2287–2322, 2010.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM, 65(1):99–106, 2021. doi: 10.1145/3503250.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. 2020. doi: 10.1007/978-3-030-58595-2_41.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics, 41 (4):1–15, 2022. doi: 10.1145/3528223.3530127.

Carl Edward Rasmussen and Christopher K. I. Williams. Gaussian Processes for Machine Learning. The MIT Press, 2005. doi: 10.7551/mitpress/3206.001.0001.

A. Savitzky and M. J. E. Golay. Smoothing and differentiation of data by simplified least squares procedures. Analytical Chemistry, 36:1627–1639, 1964.

Saptarshi Neil Sinha, Holger Graf, and Michael Weinmann. SpectralGaussians: Semantic, spectral 3D Gaussian splatting for multi-spectral scene representation, visualization and analysis. ISPRS Journal of Photogrammetry and Remote Sensing, 227:789–803, 2025. doi: 10.1016/j.isprsjprs.2025.06.008.

Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In Advances in Neural Information Processing Systems (NeurIPS), 2020.

Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In Advances in Neural Information Processing Systems, volume 33, 2020.

Christopher Thirgood, Oscar Mendez, Erin Ling, Jon Storey, and Simon Hadfield. HyperGS: Hyperspectral 3D Gaussian splatting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5970–5979, 2025.

F. Tuinstra and J. L. Koenig. Raman spectrum of graphite. The Journal of Chemical Physics, 53(3):1126–1130, 1970. doi: 10.1063/1.1674108.

Holger Wendland. Scattered Data Approximation. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2004. doi: 10.1017/CBO9780511617539.

Xinjie Zhang, Xingtong Ge, Tongda Xu, Dailan He, Yan Wang, Hongwei Qin, Guo Lu, Jing Geng, and Jun Zhang. GaussianImage: 1000 FPS image representation and compression by 2D Gaussian splatting. In European Conference on Computer Vision (ECCV), 2024.

A The corpus and its provenance

The Raman corpus comprises 23 hyperspectral acquisitions of carbonaceous specimens, collected over several months on a confocal Raman microscope under 11 acquisition recipes. Cubes span 1000–1800 cm^{-1}, the window containing the D and G bands whose relative intensities are read as a measure of structural disorder (Tuinstra & Koenig, 1970; Ferrari & Robertson, 2000), at 292 channels for most cubes (133 to 293 overall) and spatial extents from 10×14 to 144×153 pixels.

Distinct cubes. Hashing each array reduces 23 acquisition identifiers to 18 distinct cubes: the archive contains 5 groups of byte-identical duplicates, the same measurement saved twice under different identifiers. Every statistic in the paper is computed after collapsing these, because counting a duplicate as a second cube inflates the sample and narrows confidence intervals on a replication that did not occur. The duplicate groups are detected automatically and listed in the supplementary manifest.

Anonymisation. Specimen identifiers, material designations, instrument configurations, operator names and dated session folders are replaced with opaque codes during ingestion, and the forward map is written to a file excluded from every shareable artifact. A gate scans all outputs for identifier patterns and fails on a hit; it exists because two real leaks occurred during development, one from a derived column that re-introduced material names the path scrubber had already removed, and one from a five-digit specimen pattern matching inside a content hash.

Region of interest. Each cube carries a particle mask separating specimen from substrate, derived from an intensity threshold on integrated counts. All errors are computed inside it. Cubes whose mask falls below 100 pixels are reported but excluded from aggregate statistics, on the grounds that a region of a dozen spectra is not a map; the threshold was fixed once and not tuned against results.

B The field and its gradients

Writing $\mathbf{d}_i = \mathbf{p} - \boldsymbol{\mu}_i$ and $\mathbf{P}_i = \boldsymbol{\Sigma}_i^{-1} = \mathbf{R}(\mathbf{q}_i)\,\mathrm{diag}(\mathbf{s}_i^{-2})\mathbf{R}(\mathbf{q}_i)^\top$, each primitive contributes $\phi_i(\mathbf{p}) = a_i \exp(-\tfrac{1}{2}\mathbf{d}_i^\top \mathbf{P}_i \mathbf{d}_i)$, and with residual $\varepsilon(\mathbf{p}) = \hat{V}(\mathbf{p}) - V(\mathbf{p})$ restricted to observed entries the gradients of (2) are

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}_i} = \sum_{\mathbf{p}} \varepsilon\, \phi_i\, \mathbf{P}_i \mathbf{d}_i, \qquad \frac{\partial \mathcal{L}}{\partial \alpha_i} = \sum_{\mathbf{p}} \varepsilon\, \phi_i \;+\; \lambda_a a_i, \tag{4}$$

$$\frac{\partial \mathcal{L}}{\partial \log \mathbf{s}_i} = \sum_{\mathbf{p}} \varepsilon\, \phi_i \left(\mathbf{R}_i^\top \mathbf{d}_i\right)^2 \oslash \mathbf{s}_i^2 \;+\; 2\lambda_s \log \mathbf{s}_i, \qquad \frac{\partial \mathcal{L}}{\partial \mathbf{q}_i} = -\tfrac{1}{2} \sum_{\mathbf{p}} \varepsilon\, \phi_i\, \mathbf{d}_i^\top \frac{\partial \mathbf{P}_i}{\partial \mathbf{q}_i} \mathbf{d}_i, \tag{5}$$

where $\oslash$ is elementwise division, amplitudes are parameterised as $a_i = e^{\alpha_i}$ so that the non-negativity constraint needs no projection, and $\partial \mathbf{P}_i / \partial \mathbf{q}_i$ follows from differentiating the quaternion rotation matrix and normalising the quaternion on use.

Every gradient is checked against central finite differences in the test suite, to a relative tolerance tight enough that a sign error or a missing term in any one of the four blocks fails the test. The field is evaluated on a truncated support: a primitive contributes only inside

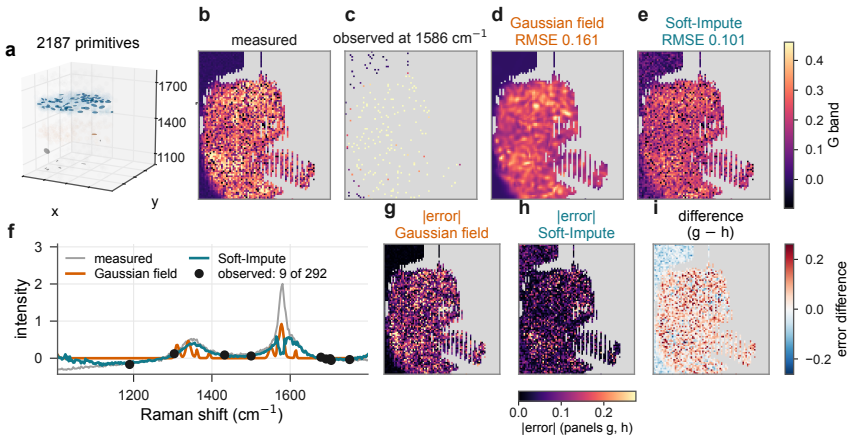


Figure 2: The field at work, and why it loses. One benchmark cell (acq-0059, 5 % observed, 2 % storage), leakage-free: (a) the 2187 fitted primitives; (b) the G-band image; (c) pixels observed at the G peak; reconstructions by (d) the field and (e) the Soft-Impute; (f) one spectrum; (g, h) absolute error on one scale and (i) its difference. The figure is diagnostic, showing how the field fails; the comparator argument is §5.3's.

a box of three standard deviations, sized from the widest primitive in the field and applied uniformly, which gives an $\mathcal{O}(N\sigma_{\max}^3)$ cost that grows with the widest primitive whatever the distribution of widths and caps our search grid at 4.5 voxels. This kept the largest cube in the corpus out of some sweeps and bounds the configurations a hyperparameter search can reach; per-primitive boxes, or the tiled rasteriser the computer-vision implementations use, would lift it.

C The comparator suite

Table 1 is the nomenclature the paper uses. "Baseline" means a comparator; the Gaussian field is the method under test and is never counted among them, and the constant predictor is a floor rather than a baseline. Storage-matched means the method is sized so that what it must store to reconstruct the cube is within a few percent of the field's budget; an interpolant is not matched, because it keeps the observed entries themselves, which at 5 % observed is more than the budget.

What each family contributes. Local interpolation interpolates along wavenumber per pixel (linear) or across the image plane per channel on the Delaunay triangulation of the observed positions (spatial). Scattered-data interpolation solves a kernel interpolant on each withheld voxel's nearest observed voxels, with a Gaussian, multiquadric or thin-plate kernel (Hardy, 1971; Wendland, 2004); the distance metric needs one modelling choice, how many channels count as one pixel, and that ratio and the kernel width are chosen on a held-back tenth of the observed entries. Kriging is the same local solve with a covariance instead of a kernel: an anisotropic Matérn-3/2 or squared-exponential covariance with a nugget, its length scales fitted by maximum marginal likelihood on compact neighbourhoods of observed voxels (Matheron, 1963; Rasmussen & Williams, 2005). Low-rank is PCA fitted by iterative hard-thresholded imputation and Soft-Impute, which soft-thresholds the singular values instead, the reference method for completing a matrix from few entries (Mazumder et al., 2010); its penalty is chosen on held-back observed entries along a decreasing path with warm starts. One pathology is worth naming: the thin-plate kernel is the only one in the suite with no shape parameter to damp its growth away from the data, and on a channel run withheld at every pixel it extrapolates catastrophically, reaching errors of order 10^{19} on one cube. It is ranked accordingly in the H1 column and is well behaved wherever the task is interpolation rather than extrapolation. Non-negative factorisation is masked multiplicative NMF and MCR-ALS with closure; tensor decomposition is Tucker and CP

Table 1: Every method in the comparison: its family, what is tuned, the data the choice is made on, the scope of the resulting configuration, and whether its stored size is matched to the field's budget. No hyperparameter, rank, threshold or model-selection decision anywhere in this paper is chosen using a test mask. "Budget" means the value follows from the storage budget alone; "global" means one configuration is frozen and reused on every cube and geometry.

method	family	tuned	selected on	scope	matched
interpolation	local interpolation	none	–	fixed	no
spatial interp.	local interpolation	none	–	fixed	no
RBF (Gaussian)	scattered-data interpolation	width, axis ratio	held-back observed	per cube	no
RBF (multiquadric)	scattered-data interpolation	width, axis ratio	held-back observed	per cube	no
RBF (thin plate)	scattered-data interpolation	axis ratio	held-back observed	per cube	no
kriging (Matérn)	scattered-data interpolation	scales, nugget	observed voxels	per cube	no
kriging (SE)	scattered-data interpolation	scales, nugget	observed voxels	per cube	no
PCA	low-rank	rank	budget	per cube	yes
Soft-Impute	low-rank	rank, penalty	budget; held-back observed	per cube	yes
NMF	non-negative factorisation	rank	budget	per cube	yes
MCR-ALS	non-negative factorisation	rank	budget	per cube	yes
Tucker	tensor decomposition	ranks	budget	per cube	yes
CP	tensor decomposition	rank	budget	per cube	yes
Fourier network	coordinate network	width	budget	per cube	yes
SIREN	coordinate network	width; freq., step	budget; validation masks	global	yes
Gaussian field	the method under test	size; full search	budget; validation masks	global	yes
constant	diagnostic floor	none	–	fixed	no

fitted by expectation-maximisation imputation. Coordinate networks are a Fourier-feature ReLU network and a SIREN, both two hidden layers wide enough to reach the budget, trained with Adam on observed voxels; the SIREN's first-layer frequency and step size were chosen by a grid on validation masks of one cube and then frozen, as the field's configuration was.

D Hold-out specifications

Each family is a boolean mask over the volume, generated before any preprocessing parameter is estimated.

Contamination, defined. With H the withheld raw channels and O the observed ones, $j \in O$ is contaminated if $\min_{k \in H} |j - k| \leq \lfloor w/2 \rfloor$, and the contaminated fraction is $|\{j \in O :$

Table 2: Which methods are ranked on which task. A method is structurally excluded when the entries it must predict have no observed value anywhere in the sub-array its parameterisation ties them to, so what it would return is a fallback rather than the method: a factorisation is asked to fill an entirely unobserved slice, spectral interpolation a position with no observed channel, spatial interpolation a channel observed at no position. Excluded methods are run and reported, but never ranked, and Figure 5 shows them as "–".

method	H1	H3	R5	R7	EDX	EDX-A	excluded because
interpolation	•	•	•	–	•	•	unobserved spectrum
spatial interp.	–	•	•	•	•	•	unobserved channel
RBF (Gaussian)	•	•	•	•	•	•	
RBF (multiquadric)	•	•	•	•	•	•	
RBF (thin plate)	•	•	•	•	•	•	
kriging (Matérn)	•	•	•	•	•	•	
kriging (SE)	•	•	•	•	•	•	
PCA	–	•	•	–	•	•	unobserved slice
Soft-Impute	–	•	•	–	•	•	unobserved slice
NMF	–	•	•	–	•	•	unobserved slice
MCR-ALS	–	•	•	–	•	•	unobserved slice
Tucker	–	•	•	–	•	•	unobserved slice
CP	–	•	•	–	•	•	unobserved slice
Fourier network	•	•	•	•	•	•	
SIREN	•	•	•	•	•	•	
Gaussian field	•	•	•	•	•	•	

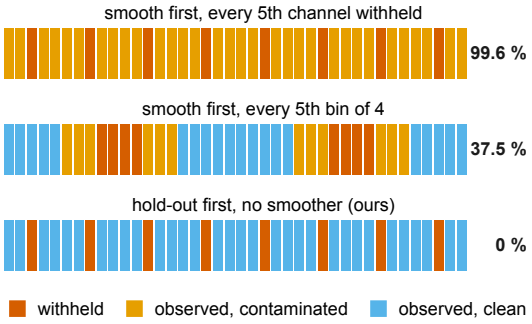


Figure 3: How much of the answer reaches the training view depends on the hold-out geometry. A run of the spectral axis under three protocols, with the channels a 7-point smoother contaminates and the contaminated fraction on the real 292-channel axis. The same codebase, filter and nominal hold-out fraction range from an almost meaningless evaluation to an essentially clean one.

contaminated}|/|O|, computed on the real 292-channel axis with $w = 7$ and no boundary special-casing. A tamper test confirms the analytical count: setting one withheld channel to a large value and smoothing changes exactly the outputs the stencil predicts.

The hold-outs that do not discriminate. H2 and a regular super-resolution variant withhold pixels with no observed channel at all, and there all 10 methods we ran land within 1.4 % of one another: a pixel never measured carries no information beyond spatial smoothness, and at this sampling density the material is not smooth. R7 asks the same question at densities an instrument would choose and does discriminate (Appendix Q).

H0, interleaved: channel k is withheld when $k \bmod n = \lfloor n/2 \rfloor$, at every pixel; $n = 5$ throughout. H1, contiguous spectral: a run of $\lceil fK \rceil$ consecutive channels at every pixel, positioned by a parameter that slides the window from the low-wavenumber end to the high without running off the axis. H2, spatial: square patches of side b on a lattice with spacing ℓb, each surrounded by a guard ring of width g excluded from training so that no withheld

pixel is adjacent to an observed one; lattice placement rather than scattering is what keeps the guard rings from consuming the training set. H3, joint: the intersection of an H2 pixel set with an H1 channel run, so that inside the patches the spectrum is missing only over the interval. R5, irregular: a random fraction of voxels, drawn independently per channel so that no channel becomes an unobserved column. R7, sparse survey: a random fraction of positions, with every channel of an unvisited position withheld, which is what an instrument that visits fewer stage positions produces.

The leakage gate applies to all of them identically: the withheld entries of the raw cube are replaced with arbitrary values, the training view is rebuilt, and every observed entry and every fitted constant is required to be unchanged. The conventional pipeline, which smooths before masking, fails this for H0 and H1 and passes for H2, which is the pattern reported in §4.

E The leakage control in full

Both arms are scored against the same fixed target throughout: the raw withheld values of the untouched cube, with each pipeline's prediction mapped back to raw units by its own fitted baseline and scale. Mask, seed, budget, method configuration, metric, region of interest and aggregation are identical between arms, so the only thing that differs is whether the observed training view could see the withheld values.

Table 3: The same contrast under three hold-out geometries, on 3 cubes and 6 methods. "Median drop" is the median reduction in reported error in the contaminated arm and "largest drop" the most extreme method-and-cube cell. A spatial hold-out cannot be contaminated by a filter that acts along wavenumber, and is the control.

hold-out	contaminated	median drop	largest drop	reversals
interleaved (H0)	99.7 %	0.83 %	34 %	4 of 45
contiguous block (H1)	3.0 %	0.04 %	21 %	3 of 45
spatial patches (H2)	0.0 %	0.00 %	14 %	5 of 45

The median reduction follows the contaminated fraction down to zero, which is the causal statement the fixed-target design supports. Rank reversals do not: they are as common under the uncontaminated control as under full contamination, which says that replacing the training input with a smoothed version reorders methods for reasons unrelated to leakage. We therefore report the reordering as an observation about smoothing rather than as a cost of leakage, and the error reduction as the cost of leakage.

Table 4: Dose-response over the smoother's width, on 3 cubes and 6 methods under the interleaved hold-out. Contamination saturates almost immediately, so the informative contrast is between the narrowest filter and the rest.

window	contaminated	median drop	largest drop	reversals
3	49.7 %	0.00 %	0 %	0 of 45
5	99.4 %	0.44 %	50 %	3 of 45
7	99.7 %	0.83 %	34 %	4 of 45
9	100.0 %	0.46 %	30 %	3 of 45
11	100.0 %	0.57 %	27 %	3 of 45

Widening the filter widens the stencil, and the contaminated fraction saturates by 3 points at 49.7 %. Below that, at a 3-point window, the effect is not measurable at all; above it the median effect is small and the tail is long, reaching 27 % on individual cells. The dose axis available on a real spectral axis is therefore short, which is itself worth knowing: a filter wide enough to smooth anything is already in the saturated regime, so the contaminated fraction separates protocols rather than filter widths.

F Residual scale, not a noise floor

It is common to normalise a residual by an estimate of the acquisition's own noise, taken from second differences along wavenumber, and to read a ratio near one as evidence that a representation has reached the measurement limit. That estimator is sound on raw channels and stops measuring the instrument once the spectra have been smoothed: the lag-one autocorrelation of first differences, -0.5 for white noise, has median -0.21 across our cubes before smoothing and $+0.74$ after it, so what survives to be measured is the high-frequency residue of the filter. We therefore call the quantity a within-pipeline residual scale rather than a noise floor: a residual below it means only that the reconstruction is smoother than the filter's output, which is a statement about the pipeline and not about the specimen.

G Per-cube results

Table 5 gives the per-cube result behind Figure 6, under the narrow comparison (linear interpolation, PCA and the constant) that §5.3 reports and then withdraws. Margin is the fractional improvement over the best of interpolation, PCA and the constant predictor, averaged over 20 seeds, with the standard deviation across those seeds. We include it because the robustness of a superseded result is the evidence for the methodological point: on every cube at least 16 of the 20 draws agree in sign with the cube's mean, on 9 cubes all of them do, and none of that was protective.

Table 5: Voxel-wise missingness at 5 % observed, per cube, for the 12 cubes whose region of interest exceeds 100 pixels, scored against the narrow comparator set. Positive margin means the Gaussian field beat all three. Against the full suite of 15 baselines the same cubes give 0 (§5.3).

map	r	ROI px	margin	s.d.
0011	0.97	135	+0.238	0.046
0016	0.14	2539	-0.057	0.081
0026	0.80	576	+0.052	0.023
0028	0.97	1696	+0.091	0.019
0037	0.96	4514	+0.236	0.016
0039	0.91	3966	+0.143	0.010
0042	0.97	1588	+0.246	0.019
0045	0.65	145	+0.108	0.021
0054	0.42	379	+0.076	0.035
0057	0.92	1662	+0.113	0.015
0059	0.94	4120	+0.143	0.009
0061	0.97	1087	+0.117	0.015

H Storage budget and rate–distortion

Every method is measured by what it must write to reconstruct the cube: centres, scales, rotations and amplitudes for the field; mean, basis and coefficients for PCA and Soft-Impute; factors and core for Tucker; all weights, biases and learned encodings for the networks. Each is then sized to within a few percent of the target, and budgets are quoted as a fraction of the cube's own scalar count. The field is best in 0 of the 12 settings of budget and hold-out on this cube.

I Ranks on every task

J Comparator coverage

Effect sizes. Against the full evaluated pool the field's margin over the best comparator has mean -24.4 % and median -18.8 % across the 12 cubes, with a hierarchical 95 % interval of

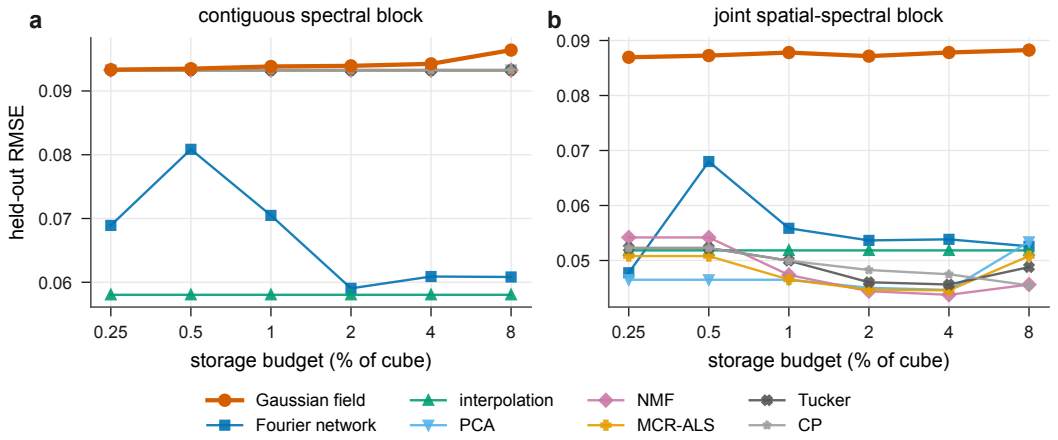


Figure 4: Held-out error against storage budget on one cube. acq-0055, the 8 methods of the first suite at 6 budgets from 0.25 % to 8 % of the cube. (a) Contiguous spectral block: both methods that can extrapolate along wavenumber beat the field at every budget, and the factorisations, which see nothing in the withheld channels, collapse onto the constant predictor beneath it. (b) Joint block: the field is last at every budget and its curve is flat in budget throughout (Appendix R). The families added later were run at the headline budget only and appear in the rank table of Appendix I rather than here.

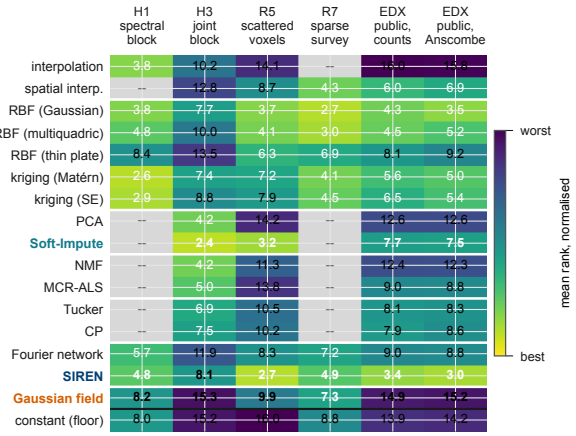


Figure 5: Every method, every missingness geometry, as ranks. Mean rank within a task (1 is best), averaged over the masks of a cube and then over cubes. Structurally excluded methods are shown as "–" rather than ranked (Appendix C); the constant predictor is the diagnostic floor, placed among the ranked methods without displacing them. Columns: the spectral block (H1), the joint block (H3), scattered voxels (R5), the sparse survey (R7) and the public FIB-EDX map in counts and after variance stabilisation. Counts: Table 6.

−33.4 to −16.8 % and 0 cubes won; against the narrow set the same runs give mean 12.5 %, median 11.5 %, interval 7.7 to 17.1 % and 11 cubes won. Against the first suite we ran, 7 comparators without scattered-data interpolation, kriging, completion or a second network, the field trails by 8.1 % (4.7 to 11.3 %) and wins 1, and the method taking most runs is Fourier-feature network rather than Soft-Impute.

Size, coverage, or strength. Three properties of a subset rank the ones that reach the full evaluated pool's conclusion above the ones that do not: its size (AUC 0.992), the number of families it covers (0.961), and its strength, the best full-pool mean rank among the

Table 6: Mean rank of each method (1 is best) within each task, the numbers behind Figure 5. Runs are ranked one at a time, averaged over the masks of a cube and then over cubes.

method	H1	H3	R5	R7	EDX	EDX-A
interpolation	3.8	10.2	14.1	–	16.0	15.8
spatial interp.	–	12.8	8.7	4.3	6.0	6.9
RBF (Gaussian)	3.8	7.7	3.7	2.7	4.3	3.5
RBF (multiquadric)	4.8	10.0	4.1	3.0	4.5	5.2
RBF (thin plate)	8.4	13.5	6.3	6.9	8.1	9.2
kriging (Matérn)	2.6	7.4	7.2	4.1	5.6	5.0
kriging (SE)	2.9	8.8	7.9	4.5	6.5	5.4
PCA	–	4.2	14.2	–	12.6	12.6
Soft-Impute	–	2.4	3.2	–	7.7	7.5
NMF	–	4.2	11.3	–	12.4	12.3
MCR-ALS	–	5.0	13.8	–	9.0	8.8
Tucker	–	6.9	10.5	–	8.1	8.3
CP	–	7.5	10.2	–	7.9	8.6
Fourier network	5.7	11.9	8.3	7.2	9.0	8.8
SIREN	4.8	8.1	2.7	4.9	3.4	3.0
Gaussian field	8.2	15.3	9.9	7.3	14.9	15.2
constant	8.0	15.2	16.0	8.8	13.9	14.2
methods ranked	9	16	16	9	16	16
cubes (or binnings)	12	12	12	56	4	4
runs	12	12	240	56	40	20

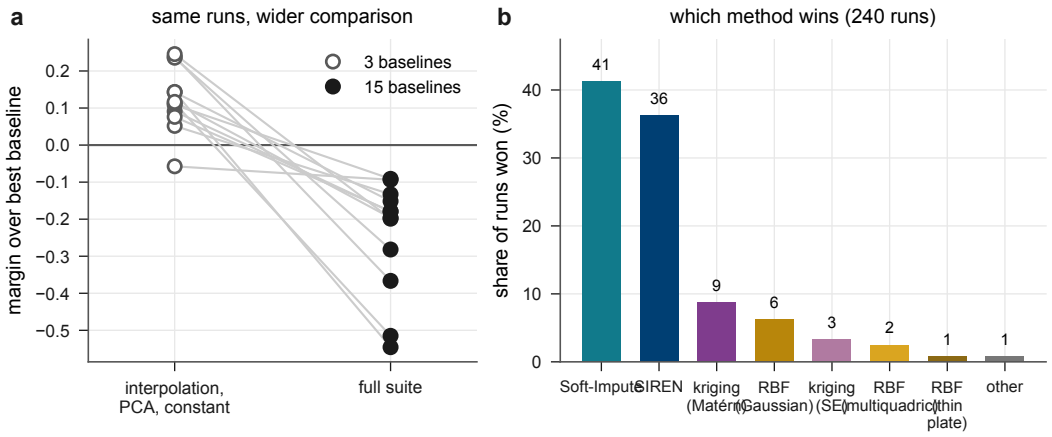


Figure 6: The same runs under two reference sets. (a) Each of the 12 cubes at 5 % of voxels observed, scored first against interpolation, PCA and a constant, then against the full evaluated pool; positive means the field beat everything it was compared with. (b) Which method wins each of the 240 runs.

methods it contains. Strength is the one that tracks it here: a retrospective threshold on full-pool comparator strength classifies the sign of all but 16 of the 32 767 subsets, so in this benchmark the outcome is largely determined by whether the set contains at least one comparator above that threshold, whatever else is in it. That is an empirical property of this pool, not a general rule. The set that first convinced us fails a strength test and not a size test: its two comparators rank 14th and 15th of 15 in the full pool, the weakest available. The threshold is retrospective by construction and is not an oracle: strength is read off the pool's own results, so the statistic explains which sets failed rather than certifying which sets would succeed. This is the same conclusion the family-balanced subsets reach from the other direction, since a family contributes to coverage only through the representative that stands for it.

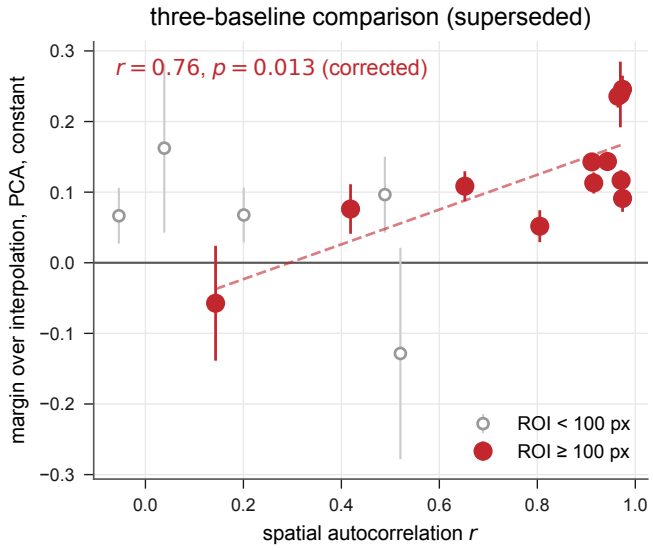


Figure 7: The association that reversed. Each cube's margin over the narrow comparator set against the spatial autocorrelation of its G-band image, measured before any model is fitted. This is the relationship that survived a permutation test, partial correlation against the collinear spectral predictor, estimation from observed entries only and leave-one-cube-out prediction, and that changes sign when the comparator set is widened (Figure 14 plots the widened version, for these cubes and for the public ladder).

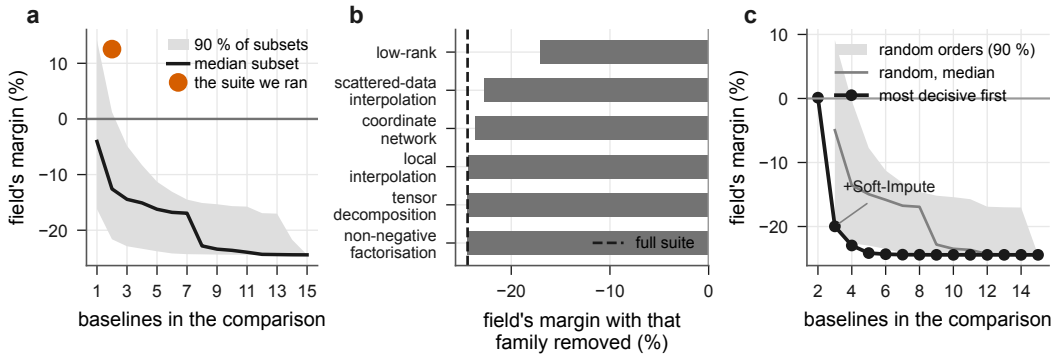


Figure 8: The comparator axis in full. (a) The margin the field reports against every non-empty comparator subset, grouped by subset size, with the set we first ran marked; the band holds 90 % of subsets and the line is the median. Positive margin means the field beat every comparator in the set. (b) The margin with one method family removed, against the full pool. (c) Adding comparators to the narrow set in the order that moves the conclusion fastest, against random orders: the conclusion flips one addition greedily and one addition in a random order.

Family presence against method multiplicity. Enumerating subsets of methods lets a family with many implementations dominate the population, so we repeated the analysis over the 1457 subsets that hold exactly one method per family (Table 7). Family coverage then predicts the conclusion well: every subset covering four or more families reaches it, and 91 % of all of them do. Which method represents a family still matters, and by about one family's worth of coverage: taking each family's strongest representative, 94 % reach the full conclusion and four families always suffice, while taking the weakest gives 78 % and needs five. A comparison is therefore not made adequate by counting methods; it is made

adequate by covering the families and representing each one with a method that is actually competitive.

Table 7: Comparator subsets holding exactly one method per family, by how many families they cover and how each family's representative is chosen. "Median margin" is over the subsets with an arbitrary representative; positive means the field wins.

families	subsets	median margin	any rep.	strongest rep.	weakest rep.
1	15	−4.0 %	60 %	67 %	33 %
2	90	−11.8 %	88 %	100 %	60 %
3	280	−14.0 %	97 %	100 %	80 %
4	480	−15.5 %	100 %	100 %	93 %
5	432	−16.4 %	100 %	100 %	100 %
6	160	−19.4 %	100 %	100 %	100 %

Table 8: How often a comparator subset of each size reaches the full evaluated pool's conclusion, over all 32 767 non-empty subsets, with a hierarchical bootstrap interval over 2000 draws, and separated by whether the subset contains a method from the family that wins the regime (low-rank). Reaching the conclusion means agreeing in sign with the full evaluated pool, where the field trails.

baselines	subsets	all	with low-rank	without
1	15	60 % [47, 80]	50 %	62 %
2	105	89 % [84, 97]	81 %	91 %
3	455	98 % [96, 100]	96 %	99 %
4	1365	100 % [99, 100]	100 %	100 %
5	3003	100 % [100, 100]	100 %	100 %
6	5005	100 % [100, 100]	100 %	100 %
7	6435	100 % [100, 100]	100 %	100 %
8	6435	100 % [100, 100]	100 %	100 %
9	5005	100 % [100, 100]	100 %	100 %
10	3003	100 % [100, 100]	100 %	100 %
11	1365	100 % [100, 100]	100 %	100 %
12	455	100 % [100, 100]	100 %	100 %
13	105	100 % [100, 100]	100 %	100 %
14	15	100 % [100, 100]	100 %	–
15	1	100 % [100, 100]	100 %	–

Table 9: Leave one method family out. Margin is the field's mean margin over the best remaining comparator, averaged over masks within a cube and then over cubes; positive means the field wins.

comparator set	margin	won	most wins	corr. r
full suite	−24.4 %	0/12	Soft-Impute (41 %)	−0.46
without low-rank	−17.0 %	0/12	SIREN (68 %)	−0.49
without non-negative factorisation	−24.4 %	0/12	Soft-Impute (41 %)	−0.46
without tensor decomposition	−24.4 %	0/12	Soft-Impute (41 %)	−0.46
without coordinate network	−23.6 %	0/12	Soft-Impute (52 %)	−0.43
without local interpolation	−24.4 %	0/12	Soft-Impute (41 %)	−0.46
without scattered-data interpolation	−22.7 %	0/12	Soft-Impute (53 %)	−0.43

K A checklist before trusting a benchmark conclusion

The audit answers one of these questions; the rest are the context that makes its answer worth having. We state them as the checks we wish we had applied to our own earlier result.

Table 10: Adding comparators to linear interpolation, PCA and the constant, in the order that moves the conclusion fastest towards the full evaluated pool's. Positive margin means the field beat every comparator in the set.

step	added	family	margin	cubes won
1	Soft-Impute	low-rank	−20.0 %	0/12
2	kriging (Matérn)	scattered-data interpolation	−23.0 %	0/12
3	SIREN	coordinate network	−24.1 %	0/12
4	RBF (Gaussian)	scattered-data interpolation	−24.3 %	0/12
5	RBF (multiquadric)	scattered-data interpolation	−24.4 %	0/12
6	kriging (SE)	scattered-data interpolation	−24.4 %	0/12
7	RBF (thin plate)	scattered-data interpolation	−24.4 %	0/12
8	NMF	non-negative factorisation	−24.4 %	0/12
9	MCR-ALS	non-negative factorisation	−24.4 %	0/12
10	Tucker	tensor decomposition	−24.4 %	0/12
11	CP	tensor decomposition	−24.4 %	0/12
12	Fourier network	coordinate network	−24.4 %	0/12
13	spatial interp.	local interpolation	−24.4 %	0/12

Preprocessing. Was the hold-out sealed before any operation that could use a withheld value? Does a tamper test leave the whole observed training view unchanged? Are baseline and normalisation estimated from observed entries alone?

Missingness. Does the hold-out geometry match the scientific claim? Are spectral, spatial, joint and sparse-acquisition cases being reported separately rather than conflated?

Comparators. Are several relevant method families represented, each by a competitive member? Is the proposed method compared against methods whose inductive bias suits the task? Does the sign of the claim survive comparator subsets? Does removing a family change it? Is one comparator controlling the result?

Statistics. Is the true experimental unit the thing being resampled? Are repeated masks treated as nested within it rather than as independent datasets? Is sampling uncertainty reported separately from comparator sensitivity, in the same units?

L The audit on every geometry

Table 11: The comparator audit on every geometry. "Margin" is the mean margin against the full evaluated pool, in points, positive when the method under test beats every comparator; "sampling" is the width of its hierarchical bootstrap interval and "comparator" the range of the same margin over all non-empty subsets of that geometry's eligible pool; "sign stable" is the share of those subsets that agree with the full pool and "positive" the number that report the method under test ahead. Block hold-outs have one mask per cube, so their sampling interval resamples cubes alone.

geometry	method	margin	sampling	comparator	sign stable	positive
H1	Gaussian field	−184.9	176.3	158.5	100 %	0 of 255
	SIREN	−41.2	49.5	80.0	97 %	7 of 255
H3	Gaussian field	−215.5	218.8	99.4	100 %	0 of 32767
	SIREN	−21.4	29.7	70.6	100 %	23 of 32767
	Soft-Impute	−3.6	6.4	57.1	60 %	13071 of 32767
R5	Gaussian field	−24.4	16.6	39.5	100 %	31 of 32767
	SIREN	−7.9	12.5	33.3	86 %	4623 of 32767
	Soft-Impute	+0.7	15.1	26.9	100 %	32767 of 32767
R7	Gaussian field	−69.7	26.4	49.7	100 %	0 of 255
	SIREN	−21.2	20.9	40.5	98 %	5 of 255
EDX	Gaussian field	−129.7	116.2	126.5	100 %	0 of 32767
	SIREN	−4.3	10.6	50.4	84 %	5311 of 32767
	Soft-Impute	−46.6	63.5	79.0	100 %	31 of 32767

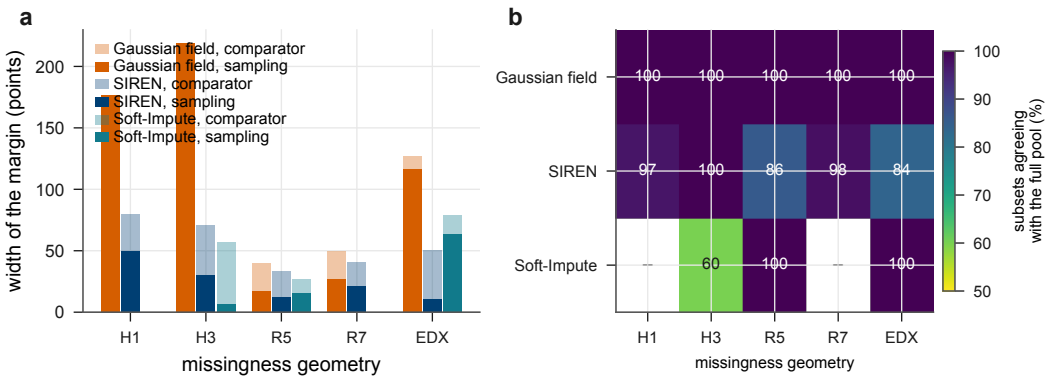


Figure 9: Claim sensitivity across geometries and methods under test. (a) For each method under test, the width of the hierarchical bootstrap interval around the full-pool margin (solid) against the range of that margin over every comparator subset (pale), in points of margin; positive margin means the method beat the best comparator in the set. (b) Each cell is the exact fraction of non-empty subsets of the evaluated pool whose aggregate margin has the same sign as the full-pool margin; "–" marks a structurally inapplicable method. The fraction describes sensitivity within the finite evaluated pool and is not the probability that a future benchmark reaches the same conclusion.

M Two kinds of variation in one conclusion

benchmark conclusion under $B = (P, H, S, A)$

run-level variation	specification variation
seeds, masks, cubes drawn under a fixed B	the preprocessing P and the comparator set S themselves
estimated by a hierarchical bootstrap	tested by hold-out invariance and the comparator audit

Figure 10: What a benchmark holds fixed, and what each kind of check varies. Resampling explores the left branch; Definition 1 and the audit of Algorithm 1 explore the right one.

N The Comparator-Set Robustness Audit

The audit in §5.3 needs no extra runs, only the per-run errors a benchmark already computes, and it is cheap: the lattice is one elementwise minimum per subset, 32 767 subsets over 240 runs in seconds. We state it as a procedure, the Comparator-Set Robustness Audit, because it transfers to any benchmark that reports a margin over a comparator set. Its inputs are the method under test T, the evaluated pool $\mathcal{P}$, the stored per-run errors $E \in \mathbb{R}^{R \times K}$ and e_T, the aggregation A the benchmark already uses, the replication hierarchy it resamples, the comparator set S_0 actually reported, the claim being made, and a subset budget when $\mathcal{P}$ is too large to enumerate. Its outputs are the sign-stability rate, the comparator-set margin distribution with robust summaries, the percentile of S_0 among equal-size sets, the influential comparators and families, and the ratio of comparator variation to sampling uncertainty.

In words, and as the numbered checks a reader can apply:

1. Store the errors, not the verdict. Keep every method's error on every run. A stored margin cannot be rescored against a different comparator set; stored errors can.
2. Enumerate the subsets. Build the margin for every non-empty subset by adding one method at a time, and report the distribution of the margin and the share of subsets that reach the conclusion being claimed.

Algorithm 1 Comparator-Set Robustness Audit

compute the headline margin $\bar{M}(T, S_0)$ under the reported set
bootstrap the replication hierarchy with S_0 held fixed
for all $S \subseteq \mathcal{P}$, $S \neq \varnothing$ do
    $E_{\min}(S, r) \leftarrow \min_{j \in S} E_{rj}$ for every run r
    $M(T, r, S) \leftarrow 1 - e_T(r)/E_{\min}(S, r)$
    $\bar{M}(T, S) \leftarrow$ the benchmark's own aggregate over runs
end for
compare the two distributions on one scale
Stability$(T) \leftarrow \left|\{S \in \mathcal{L} : \operatorname{sign} \bar{M}(T, S) = \operatorname{sign} \bar{M}(T, \mathcal{P})\}\right| / |\mathcal{L}|$ ▷ $\mathcal{L}$: the non-empty subsets of $\mathcal{P}$
rank S_0 among the subsets of its own cardinality
repeat over subsets holding one method per family
remove each family in turn and recompute
recompute every margin-dependent analysis under $\mathcal{P}$
report the comparators whose addition moves $\bar{M}$ most

3. Report both axes on one scale. Put the resampling interval and the across-subset spread in the same units, as in Figure 1. A resampling interval that is narrow relative to the subset spread means the benchmark is comparator-limited, and more seeds will not help.
4. Locate your own set. Report where the comparator set actually used ranks among the sets of its size by the margin it reports. A set at the favourable extreme is the case this audit exists to catch.
5. Remove one family at a time. Leave-one-family-out says which kind of method the conclusion depends on, which is what a reader needs in order to know what is missing.
6. Rescore the supporting analyses. Any predictor, ablation or threshold sweep defined against the margin inherits the comparator set. Rerun them against the full pool and report whether they survive; ours does not.
7. Say what would change the conclusion. Name the family whose addition moves the margin most, so that the next comparator to run is identified rather than left to chance.

The audit is descriptive, not a test with a threshold: it reports how a conclusion depends on the set it was computed against, and that dependence is a property of the claim, not a defect to be certified away. It also cannot certify what was never run: it measures sensitivity within the evaluated pool, and a method outside it remains outside it.

What the audit costs. The expense of a benchmark is fitting the comparators, not rescoring them. Enumerating all 32 767 subsets and computing every statistic in this appendix took 40 seconds on one CPU core; the 1 680 comparator fits whose errors it reads took 29 hours. The audit is therefore something a benchmark can be retrofitted with, provided it stored per-run errors rather than only the verdict.

Robust summaries of the comparator axis. A range is an extreme-value statistic, so we report it alongside less fragile summaries (Table 12). For the field the full range is 39.5 points while the interquartile range is 7.9; the conclusion does not depend on which is used, since 100 % of subsets agree with the full pool in sign either way, but a reader is entitled to know that the range is driven by the tail.

Robustness is relative to the claim. A margin over the best comparator states that the method under test is best. Weaker claims are common and behave differently under the same pool (Table 13): SIREN is best in 14 % of subsets but within 5 % of the best in 51 %, and within ten percent in all of them. The field fails all three. A pool can therefore be

Table 12: The comparator axis for each method under test, summarised six ways. Positive margin means the method beat every comparator in the set.

method under test	median	IQR	90 % band	range	positive	agree
Gaussian field	−19.9	−23.6 to −15.7	−24.4 to −14.1	39.5	0 %	100 %
SIREN	−4.7	−7.6 to −0.6	−7.9 to 1.6	33.3	14 %	86 %
Soft-Impute	1.9	0.9 to 2.4	0.8 to 4.5	26.9	100 %	100 %

adequate for a non-inferiority claim and inadequate for a winner-takes-all one, which is why the audit takes the claim as an input.

Table 13: The share of comparator subsets supporting three claims about each method under test: that it is best, and that it is within a tolerance of the best comparator in the set.

method under test	best	within 5 %	within 10 %
Gaussian field	0 %	0 %	1 %
SIREN	14 %	51 %	100 %
Soft-Impute	100 %	100 %	100 %

Choosing the pool before anything is run. The audit is retrospective by construction, which invites the fair question of what a practitioner should do beforehand. We cannot offer an oracle, but the failures we observed all have the same shape, a family present only through a weak representative, and that suggests a construction rule. Include, at minimum, a trivial diagnostic floor; the conventional baseline of the field, whatever its merits; one competitive representative of each inductive-bias family the task admits, chosen on validation data rather than convenience; at least one method whose assumptions match the missingness geometry specifically; and at least one comparator selected by someone with no stake in the proposed method. Then run the audit and report what it says. The rule is a heuristic, not a guarantee: its value is that every item on it is checkable before any result is known.

Which comparators drive the sensitivity. Step 13 is a marginal influence over the lattice. For comparator j, $I_j = \mathbb{E}_{S \not\ni j}\big[\bar{M}(T, S) - \bar{M}(T, S \cup \{j\})\big]$, which Proposition 1 makes non-negative, together with the number of subsets whose sign j changes. On our runs the most influential comparator is Soft-Impute, worth 8.1 % of margin on average and flipping 32 of the 16384 subsets it can be added to (Table 14). This is what makes the audit actionable: it names the comparator to run next rather than only reporting that one is missing.

Table 14: The 6 most influential comparators: the mean change in the field's margin when each is added to a subset that lacks it, and how many of those subsets change sign. Positive influence means adding the comparator lowers the reported margin.

comparator	family	influence	subsets whose sign it flips
Soft-Impute	low-rank	8.1 %	32 of 16384
SIREN	coordinate network	1.6 %	32 of 16384
RBF (multiquadric)	scattered-data interpolation	0.7 %	32 of 16384
RBF (Gaussian)	scattered-data interpolation	0.7 %	32 of 16384
kriging (Matérn)	scattered-data interpolation	0.6 %	32 of 16384
kriging (SE)	scattered-data interpolation	0.5 %	32 of 16384

Beyond exhaustive enumeration. The lattice costs $O(2^K)$, which is fine to about twenty comparators and not beyond. Leave-one-family-out is $O(F)$ and the greedy path $O(K^2)$, so those scale regardless; for the subset statistics, sampling subsets uniformly is enough, and because our pool is small enough to enumerate we can score the approximation against the exact answer. Over 20 repeats, 50 sampled subsets already put the sign-stability rate within 0.2 points of the exhaustive value and 2500 within 0.1. The range does not behave as well,

as an extreme-value statistic should not: it is still 8.3 points out at 2500 subsets, against 23.9 at 50. A sampled audit should therefore report sign stability and a central band of margins, which converge, rather than the minimum and maximum, which do not. What does not scale at all is running the comparators themselves, which is the cost the audit exists to direct.

O What more replication buys

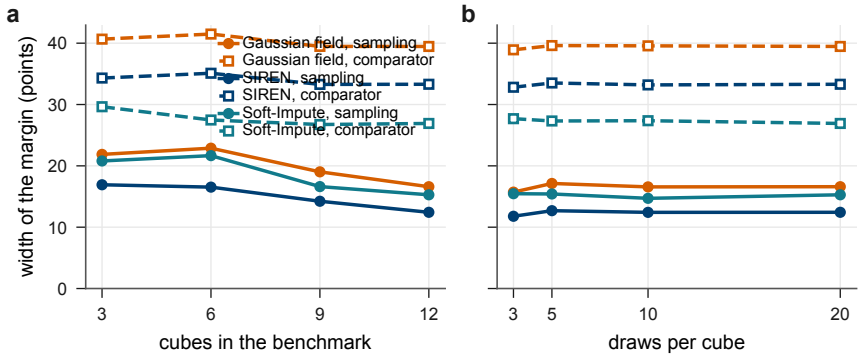


Figure 11: Replication narrows one axis and leaves the other where it was. Width of the hierarchical bootstrap interval around the full-pool margin (solid) and the range of that margin over every comparator subset (dashed), against (a) the number of cubes and (b) the number of draws per cube. Cube counts below the full 12 average over 20 random cube subsets.

Proposition 2 says replication cannot identify the comparator axis; Figure 11 is the experiment. Going from 3 to 12 cubes narrows the sampling interval from 21.9 to 16.6 points while the comparator range moves from 40.7 to 39.5. Draws per cube do not narrow the interval at all (15.7 points at 3 draws, 16.6 at 20), because the cube is the unit of replication and masks within a cube are not independent replicates of it: the axis that buys precision is more cubes, and neither axis touches the comparator range.

P Where evaluation compute is best spent

The replication curve has a practical corollary. Suppose a study can afford B evaluation cells per cube, one cell being one comparator fitted on one mask, and must choose between k comparators measured n times with $kn = B$. Drawing 400 random comparator sets and mask sets for each split and asking how often the resulting conclusion agrees in sign with the full evaluated pool's gives Table 15. At the smallest budget, 20 cells, the narrowest split (1 comparator, every mask) reaches the pool's conclusion between 7 and 100 % of the time depending on the method under test, and the broadest (10 comparators, a fifth of the masks) between 94 and 100 %. Broader coverage recovers the full-pool sign more often at every budget we tried, and the gap is largest exactly where the claim is contested. Cells are a normalised count and not wall-clock, since methods differ in cost, so this is a benchmark-specific sensitivity result about stored runs rather than a general compute-allocation rule; the direction is the one the two axes predict.

Q The sparse survey

Scattered voxels are generic completion, not an acquisition a spectrometer can perform: an instrument saves time by visiting fewer stage positions and recording the whole spectrum at each. R7 is that geometry, at 2.5 to 40 % of positions, over 56 cube and sampling-fraction cells of the reported cohort. The ranking does not transfer: the field places 9th of 9, the task is won by RBF (Gaussian) rather than by the winner under scattered voxels, and its mean

Table 15: A fixed evaluation budget spent on comparators or on repeats. "Reaches the pool's conclusion" is the share of random draws whose margin agrees in sign with the full evaluated pool's.

budget	method under test	comparators × draws	reaches the pool's conclusion
20	Gaussian field	1 × 20	61 %
		2 × 10	90 %
		4 × 5	100 %
		5 × 4	100 %
		10 × 2	100 %
	SIREN	1 × 20	7 %
		2 × 10	20 %
		4 × 5	53 %
		5 × 4	63 %
		10 × 2	96 %
	Soft-Impute	1 × 20	100 %
		2 × 10	100 %
		4 × 5	100 %
		5 × 4	100 %
		10 × 2	94 %
40	Gaussian field	2 × 20	91 %
		4 × 10	100 %
		5 × 8	100 %
		8 × 5	100 %
		10 × 4	100 %
	SIREN	2 × 20	21 %
		4 × 10	53 %
		5 × 8	64 %
		8 × 5	91 %
		10 × 4	98 %
	Soft-Impute	2 × 20	100 %
		4 × 10	100 %
		5 × 8	100 %
		8 × 5	100 %
		10 × 4	98 %
60	Gaussian field	3 × 20	98 %
		4 × 15	100 %
		5 × 12	100 %
		6 × 10	100 %
		10 × 6	100 %
		12 × 5	100 %
		15 × 4	100 %
	SIREN	3 × 20	34 %
		4 × 15	52 %
		5 × 12	64 %
		6 × 10	76 %
		10 × 6	99 %
		12 × 5	100 %
		15 × 4	100 %
	Soft-Impute	3 × 20	100 %
		4 × 15	100 %
		5 × 12	100 %
		6 × 10	100 %
		10 × 6	100 %
		12 × 5	99 %
		15 × 4	97 %

rank moves from 13.3 at the sparsest sampling to 12.7 at the densest. Nothing about the geometry is generous to any method: at 2.5 % of positions the best method improves on the constant predictor by 19 %, against 19 % at 40 %. Our voxel-wise finding therefore concerns

generic incomplete cubes, not accelerated Raman mapping, and neither result should be read as a claim about the other.

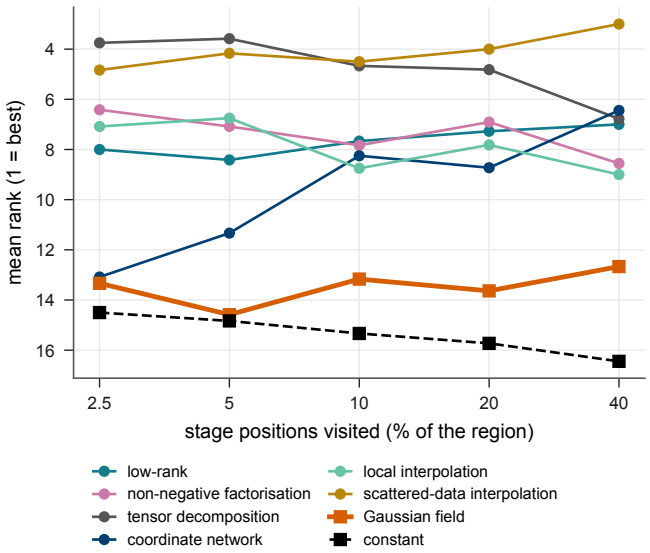


Figure 12: The sparse survey, as a function of how much of the specimen is visited. Mean rank of each family's best member, the Gaussian field and the constant predictor, over the cubes of the reported cohort. Whole spectra are recorded at the visited positions, so the task is spatial and the methods that interpolate across the image plane lead. The Gaussian field and the constant predictor are drawn individually for reference; the per-method numbers are in Table 16.

Table 16: Mean rank on the sparse survey (R7) by the share of stage positions visited, over the cubes of the reported cohort at the 2 % budget. Whole spectra are recorded at the visited positions and nothing at the others, so the task is spatial: methods that interpolate across the image plane lead, and the gap to the constant predictor grows from 19 % at 2.5 % of positions to 19 % at 40 %.

method	2.5 %	5 %	10 %	20 %	40 %
interpolation	11.3	12.1	12.8	13.7	14.2
spatial interp.	7.1	6.8	8.8	7.8	9.0
RBF (Gaussian)	6.2	4.2	4.5	4.0	3.0
RBF (multiquadric)	4.8	4.2	5.0	6.3	5.6
RBF (thin plate)	13.9	13.9	13.0	11.1	10.3
kriging (Matérn)	8.3	8.6	7.5	9.1	6.3
kriging (SE)	8.9	7.9	8.8	9.3	6.7
PCA	8.0	8.4	8.2	8.5	9.4
Soft-Impute	8.9	8.7	7.7	7.3	7.0
NMF	6.5	7.1	7.8	6.9	8.6
MCR-ALS	6.4	7.9	8.8	7.8	10.4
Tucker	4.0	3.8	4.7	4.8	6.8
CP	3.8	3.6	5.2	5.3	7.7
Fourier network	13.8	15.1	13.6	13.1	12.4
SIREN	13.1	11.3	8.2	8.7	6.4
Gaussian field	13.3	14.6	13.2	13.6	12.7
constant	14.5	14.8	15.3	15.7	16.4

R Nominal parameter count does not explain the gap

The flat curves admit two readings: too few parameters, or the wrong ones. Increasing the primitive budget 16-fold, with adaptive densification enabled, reduces training error by 5 % and leaves held-out error unchanged, plateauing at 0.52 of the data's own standard deviation: the field cannot fit data it can see. A factorised variant, a spatial Gaussian carrying a spectral signature over a shared basis, escapes that plateau and fits the observed entries better, but its held-out error is worse and grows with its spectral capacity. The gap is not explained by parameter count alone and survives the obvious repair, which implicates the representation and its optimisation together rather than the budget.

S Baseline instability under sparse sampling

Fitting the baseline on observed entries alone, as hold-out invariance requires, costs something: under sparse sampling a polynomial fitted to a spectrum's few remaining channels extrapolates wildly, swinging the constant predictor's error by a factor of 36 between draws of the same hold-out until the order is dropped with the number of observations and the fit is ridged.

Under voxel-wise missingness at 5 % a spectrum keeps a handful of scattered channels, and a fixed-order polynomial baseline fitted to those is barely determined. On acq-0028, three draws of the same hold-out gave fitted baselines reaching 551, 5 809 and 70 632, and constant-predictor errors spanning a factor of 36: the errors being compared were baseline blow-up, not reconstruction quality. Dropping the polynomial order to keep four observations per free coefficient, and ridging the fit, brings the factor to 1.0. Both use observed channels only, so the tamper test still passes. The pre-fix fit is reproduced by the supplementary code through the same split function, with only the baseline swapped.

T External validation on the public field

Data. The Battery Imaging Library (Docherty et al., 2026) publishes two acquisitions of the same FIB-EDX field of an NMC811 electrode as separate records. We use the longer-dwell one (Zenodo 17073546, CC-BY-4.0; Donoghue et al., 2025), which averages about five counts per pixel against a fraction of one for the other. The archive is undocumented. The raw file is unsigned 8-bit and band-sequential, every pixel of one channel followed by the next; read per pixel it looks like noise, and read per band the summed spectrum resolves into peaks at more than a hundred times its median. Geometry was recovered from the pixel count of the shipped element maps and the file size. The energy calibration does not map cleanly onto known NMC811 lines, so channel index serves as the spectral coordinate, which is all the reconstruction task needs.

The autocorrelation ladder. At native resolution a voxel holds a fraction of one count and a band image has no measurable spatial structure. Summing counts over $b \times b$ pixel blocks, which keeps the data Poisson, raises the autocorrelation monotonically; channels are summed into 32 contiguous bands. Every rung is then centre-cropped to the pixel extent of the coarsest one, so that all 4 cubes have the same voxel count and the same storage budget. Without the crop, coarser binning would change the autocorrelation and the cube size together and the two effects could not be separated. The rungs are four resolutions of one acquired field: they are not four independent replicates, and we report them as external validation of the direction of the result.

Protocol. Voxel-wise missingness at 5 % observed, 2 % storage budget, the same 15 baselines and constant predictor, and the Gaussian field's configuration frozen from the private search with only its two voxel-unit width bounds dropped, since those are specific to a cube's resolution. Nothing was tuned on this data.

Counts and the squared loss. The cube holds photon counts, whose variance grows with the mean, while every method minimises an unweighted squared error. We therefore reran the

entire suite on $z = 2\sqrt{x + 3/8}$ (Anscombe, 1948), which makes the variance approximately constant, and compared the two rankings (Figure 13e): the rank correlation between them is 0.98, the field wins 0 of 20 runs against 0 of 40 on raw counts, and SIREN leads either way.

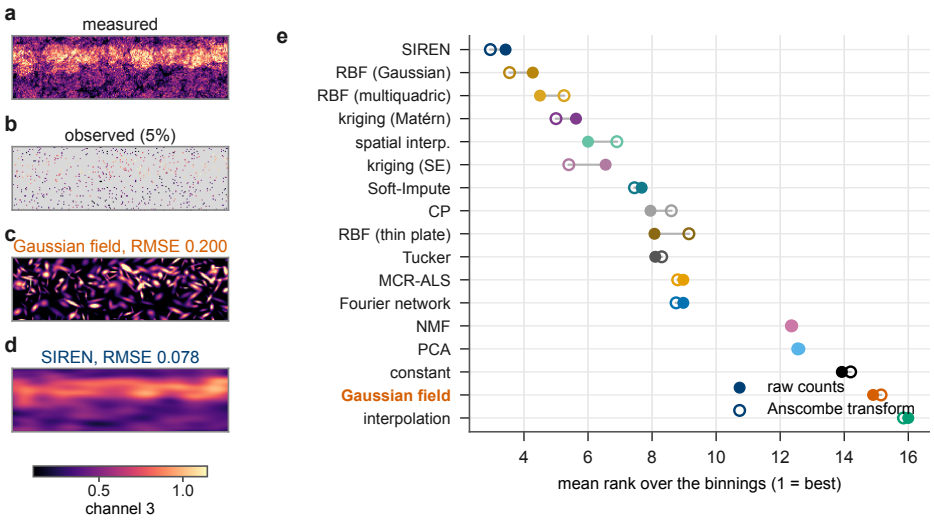


Figure 13: The same comparison on public data. (a–d) One cell of the public FIB-EDX map (Donoghue et al., 2025) at spatial binning 4, 5 % of voxels observed and the same 2 % budget: the measured channel, what the methods see, and the reconstructions of the field (RMSE 0.200) and of SIREN (0.078). (e) Mean rank of every method over the 4 binnings, on raw counts and after Anscombe's variance-stabilising transform (Anscombe, 1948); the ranking is what the transform is there to test.

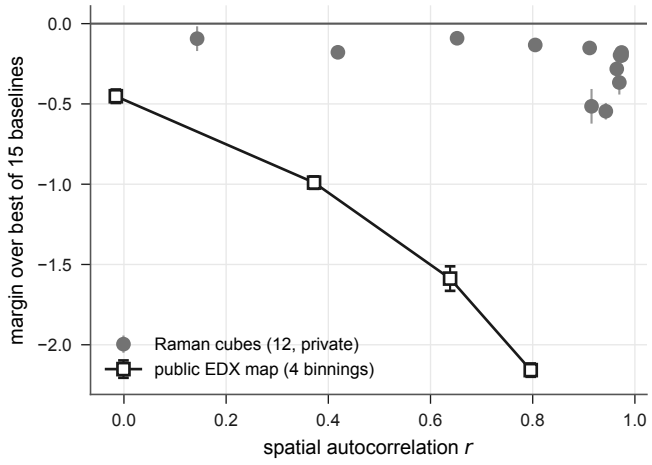


Figure 14: Margin of the Gaussian field over the best of the full evaluated pool against spatial autocorrelation, for the private Raman cubes and for the public EDX ladder, at 5 % of voxels observed. Error bars are the standard deviation over seeds. Negative means the field lost.

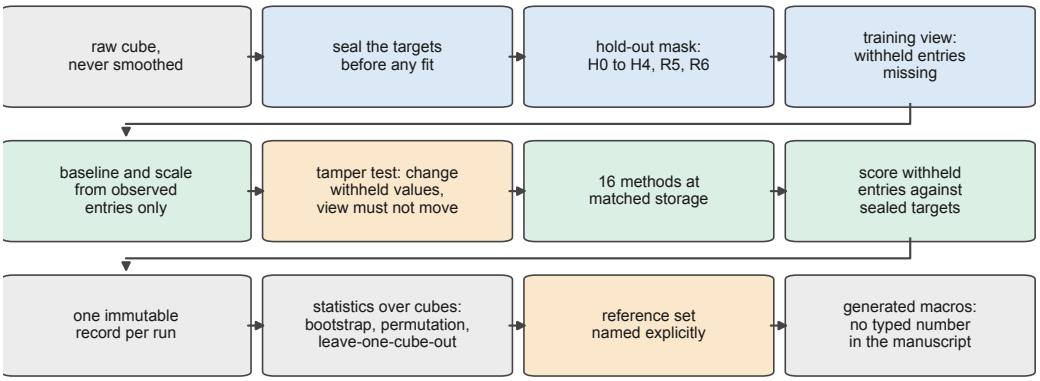


Figure 15: The leakage-free benchmark, stage by stage. Targets are sealed before anything is fitted; the baseline and the normalisation are estimated from observed entries only; and a tamper test, which replaces every withheld value and requires the training view not to move, gates every hold-out family. Every run writes one record; statistics are computed over cubes, never pixels; the reference set a margin is scored against is named explicitly; and every number in the text is a generated macro.

U The benchmark pipeline

V The field on every fitted map

W How far the hold-out geometry moves the reported quality

X The cubes the benchmark runs on

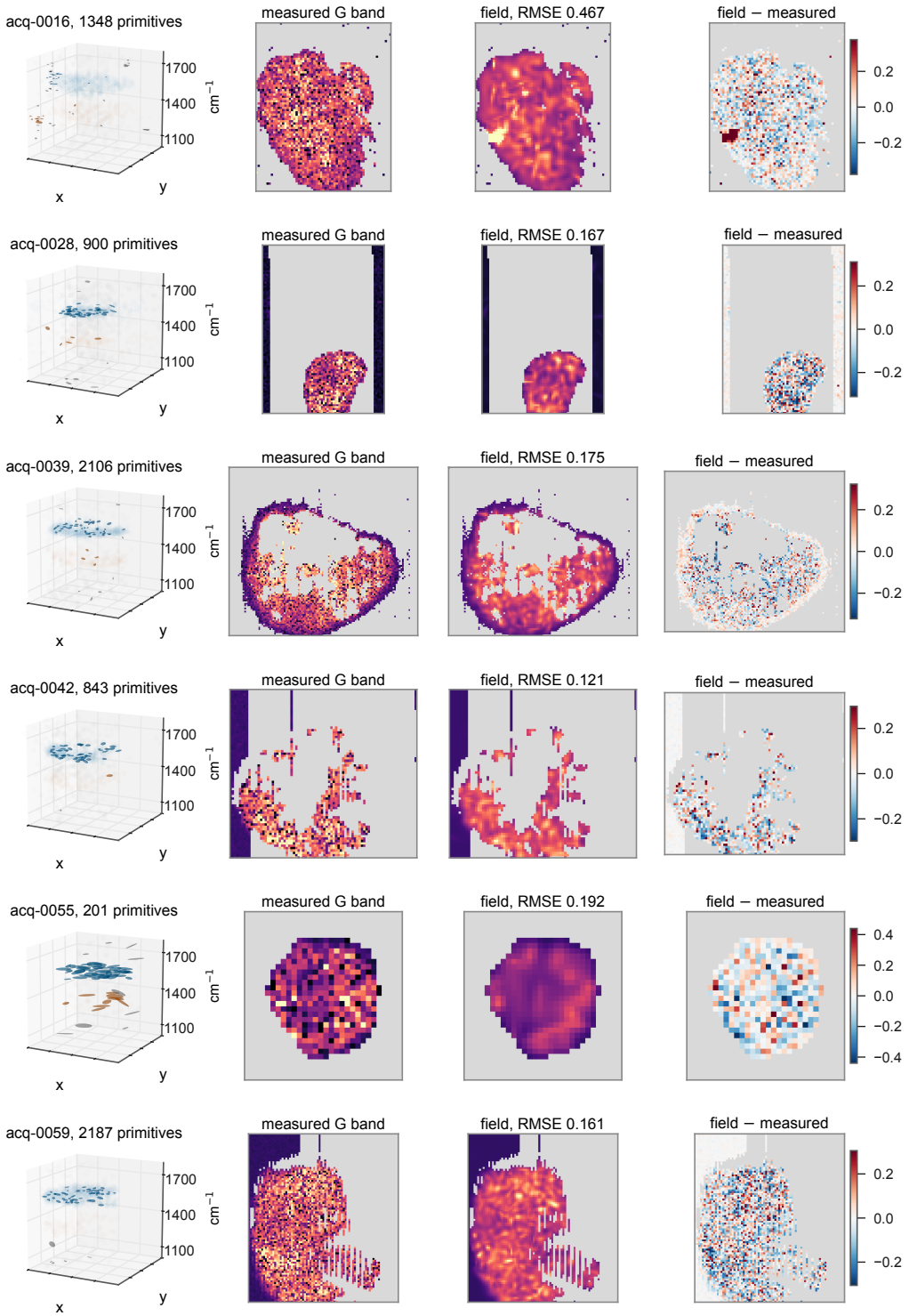


Figure 16: The same benchmark cell as Figure 2, on each of the six maps the corpus study fitted. Voxel-wise missingness at 5 % observed, 2 % storage budget, seed 0, the frozen configuration. Left to right: the fitted primitives, the measured G band, the field's reconstruction of it from the observed voxels with the withheld-voxel error, and the residual. Every error matches the sweep's stored record for the same cell. The bright artefact on acq-0016 is a primitive placed where the observed voxels do not constrain it; that cube is also the one with the lowest spatial autocorrelation in the cohort.

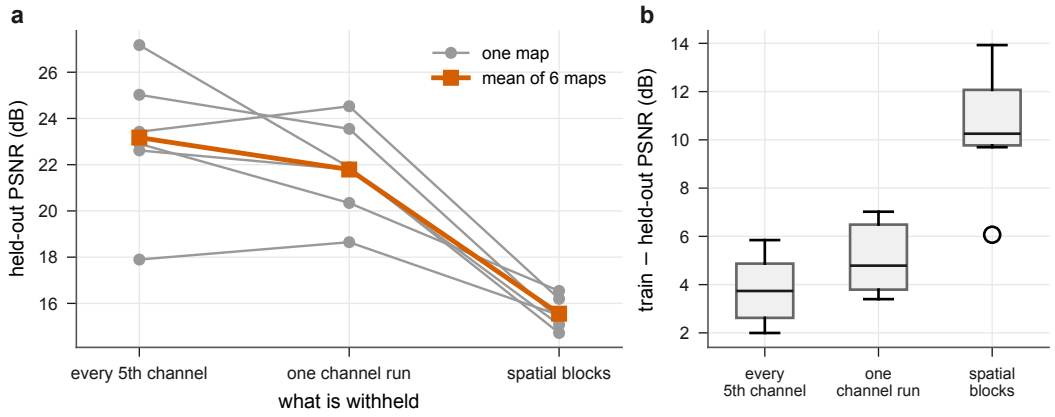


Figure 17: The same fields, scored three ways. The 6 maps of the corpus study, each fitted with one configuration under three hold-out geometries: every fifth channel of the binned axis (the protocol the corpus study reported), one contiguous channel run, and spatial blocks. (a) Held-out PSNR falls from 23.2 to 21.8 to 15.6 dB on average, a spread larger than the differences between methods that papers usually argue over, while nothing about the model or the data changes. (b) The gap between training and held-out PSNR grows in the same order. These fits used the smoothed corpus and so carry the contamination of §4, which is largest for the interleaved geometry; the benchmark itself never smooths.

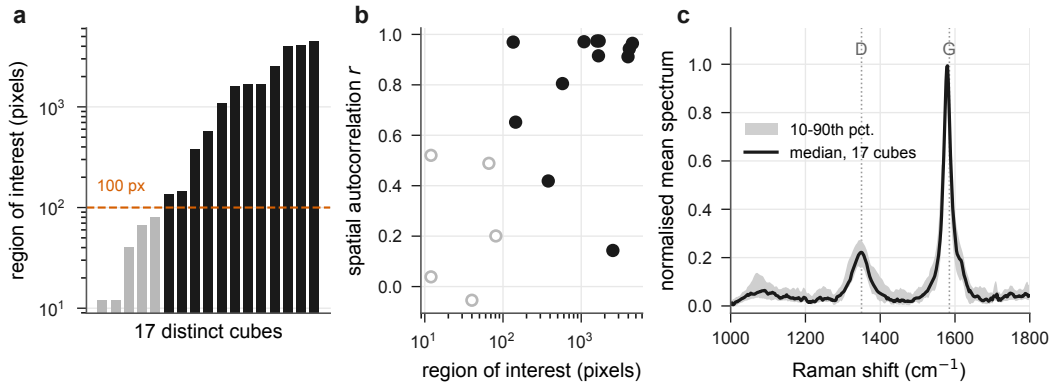


Figure 18: The cubes the benchmark runs on. (a) Region-of-interest size of the 17 distinct cubes the sweep ran on (the corpus's one larger cube exceeds the sweep's size cap); cubes below the 100-pixel cut (grey) are reported but excluded from the headline statistics. (b) Spatial autocorrelation against size: the covariate of §5.3 spans almost its whole range within the retained cohort. (c) The corpus's normalised mean spectra, with the D and G bands marked.

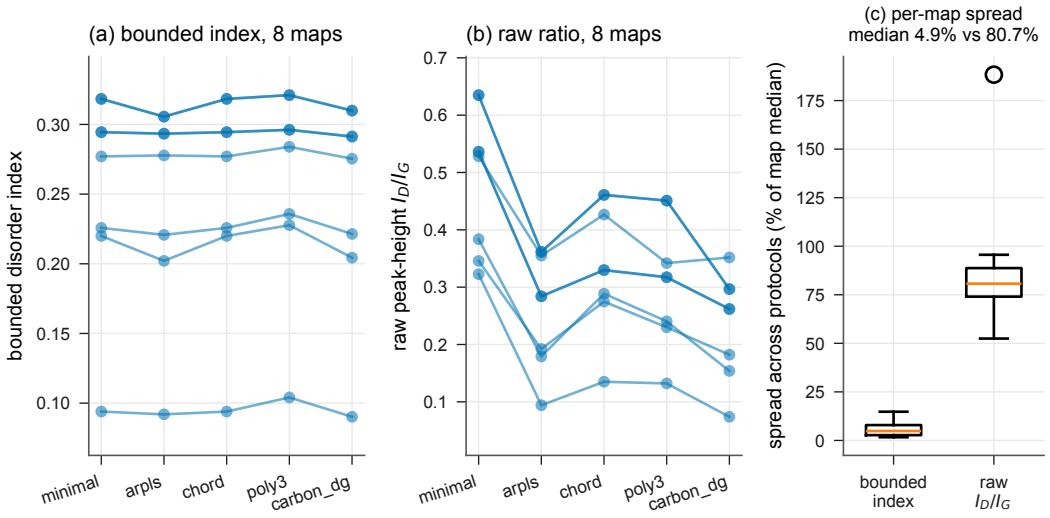


Figure 19: Preprocessing changes the numbers a corpus reports. The same 8 maps summarised under 5 preprocessing protocols. (a) A bounded disorder index moves little between protocols; (b) the raw D-to-G peak-height ratio moves a lot; (c) the per-map spread across protocols has median 4.9 % for the bounded index against 80.7 % for the raw ratio. The protocol is part of the measurement, which is the premise of §4.