RFWALL: A STRUCTURAL DIAGNOSTIC OF GENERATIVE DATA VIA THE AUGMENTATION DIFFICULTY INDEX

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative data augmentation is widely used to reduce labeling cost, and RF fingerprint localization, where every label requires a site survey, has become a prominent application. Reported gains are hard to compare: evaluations rarely share data or splits, focus on dense Wi-Fi, and often fit the generator on data withheld from the low-label condition. We introduce RFWALL, a benchmark of 31 augmentation methods on 20 partitions spanning Wi-Fi, cellular, and non-RF controls, under a single leakage-controlled protocol. Fitting generators on the full labeled pools inflates the measured gain at a 1% label budget by $12.3\times$. Under leakage control, augmentation recovers 10.8% of the available accuracy headroom on Wi-Fi and none on cellular fingerprints, and its positive effects under a nearest-neighbor localizer reverse under an MLP. Grounded by the Data Processing Inequality (DPI) on information bounds, we conduct controlled sparsification experiments on high-density physical testbeds and identify representational feature vocabulary as the causal driver of this gap: shrinking the vocabulary from 520 columns to 17 elevates exact label collision from 0.004 to 0.853, collapsing manifold structure and reducing augmentation gains to zero. We distill these mechanisms into the Augmentation Difficulty Index (ADI), a data-inherent screen combining label collision, relational manifold sparsity, and noise level, which orders held-out dataset utility at Spearman $\rho = +0.739$ under leave-one-dataset-out validation. Beyond the RF domain, spatial voxel experiments on non-RF controls show identical label collision dynamics. Controlled noise injection confirms ADI's theoretical bounds, with structure-aware generators outperforming baseline models across all test cells. Our findings demonstrate that simple non-parametric interpolation frequently matches or exceeds deep generative models, and that standard distributional fidelity metrics fail to predict downstream utility.

1 INTRODUCTION

Synthesizing training data to substitute for expensive labels is standard practice across machine learning, yet its returns are uneven, and the field usually explains them method by method rather than through properties of the data. RF fingerprint localization makes the cost of labels concrete: a fingerprint database is built by physically surveying a site, and UJIIndoorLoc alone took more than 20 surveyors carrying 25 devices over several months (Torres-Sospedra et al., 2014). A large literature therefore synthesizes the missing fingerprints, typically reporting its largest gains in the low-label regime where survey cost is highest (Sec. 2).

These gains are difficult to compare. Most published evaluations use dense Wi-Fi RSSI and leave the sparse, quantized fingerprints of cellular indoor localization largely unstudied, and code releases are rare enough that no two papers share a building, split, or localizer. A less visible confound is the evaluation protocol itself: several evaluations fit the generator on data beyond the low-data training subset, an instance of the leakage documented across many fields of ML-based science (Kapoor & Narayanan, 2023). Without a shared protocol, it is unclear how much of the reported benefit reflects the generator and how much reflects what the generator was allowed to see.

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.

In this work, we introduce RFWALL, a benchmark of 31 augmentation methods on 20 partitions—dense Wi-Fi, outdoor and indoor cellular, and four non-RF controls—evaluated under one leakage-controlled protocol in 6,537 runs. The same localization task, labels, and methods are applied to representations whose feature vocabularies differ by an order of magnitude, and the non-RF controls test whether the resulting findings are specific to radio data. Using RFWALL, we find that augmentation helps little once leakage is removed, and not at all on cellular fingerprints. We trace this gap to a *sparsity wall*: a small feature vocabulary induces label aliasing that synthetic data cannot resolve. Per-scan missingness, the usual meaning of "sparse" in this literature, is not the operative variable. Our contributions are:

- **A leakage-controlled benchmark.** RFWALL reimplements or wraps 31 methods behind a common interface, fixes all splits in advance, and trains every generator only on the downstream model's training subset (Sec. 3).
- **A measurement of evaluation leakage.** Fitting the generator on the full pool before subsampling inflates the measured gain $12.3\times$ at a 1% label budget, enough to make a content-free noise control appear to add $+0.65$ accuracy (Sec. 4.1).
- **A map of where augmentation helps.** Under control, augmentation recovers 10.8% of the available accuracy headroom on Wi-Fi and -0.6% on aliased cellular partitions; the simplest interpolation methods rank first; and positive effects under a nearest-neighbor localizer reverse under an MLP (Secs. 4.2–4.4).
- **A structural account.** Shrinking one Wi-Fi building's vocabulary reproduces the cellular regime, identifying vocabulary as causal. Label collision, manifold sparsity, and noise level combine into the Augmentation Difficulty Index (ADI), which orders held-out datasets at Spearman $\rho = +0.739$ leave-one-dataset-out, and both mechanisms reappear on non-RF data (Sec. 5).

2 RELATED WORK

Reducing survey effort. Fingerprint localization pairs an offline survey phase with an online matching phase (He & Chan, 2016). Established ways to reduce the survey include crowdsourcing from inertial traces and floor-plan constraints (Rai et al., 2012; Wang et al., 2012; Yang et al., 2012) and transfer or semi-supervised learning across environments (Pan et al., 2007; Njima et al., 2022). Augmentation requires neither user trajectories nor a source environment, and has received the least systematic scrutiny of the three.

Fingerprint augmentation. Four families of methods have emerged (Appendix C). *Geostatistical interpolation* densifies radio maps from surveyed reference points (Jan et al., 2015; Wang et al., 2019; Tang et al., 2024). *Classical augmentation* perturbs, drops, or recombines fingerprints without a learned model (Sinha & Hwang, 2019; Hilal et al., 2021; Sinha & Hwang, 2020; Anagnostopoulos & Kalousis, 2021; Sugasaki & Shimosaka, 2021). *Deep generative models* sample labeled fingerprints from a small seed set using GANs (Njima et al., 2021; Yean et al., 2023; Quezada-Gaibor et al., 2022; Zhang et al., 2025), VAEs and graph VAEs (Liu et al., 2025; Kim et al., 2025; Wu et al., 2025), or conditional diffusion (Abdelmotlb et al., 2025). The related radio-map estimation literature is scored by pixel-wise reconstruction rather than downstream localization (Wang et al., 2025; Jiang et al., 2024; Jia et al., 2026) and is out of scope.

Benchmarks and evaluation leakage. Fixed data under a shared protocol has repeatedly turned incomparable claims into cumulative progress (Deng et al., 2009; Reddi et al., 2020). Fingerprint augmentation has no such benchmark, and fitting the generator on the full pool before subsampling is a form of the train–test leakage catalogued by Kapoor & Narayanan (2023). To our knowledge, no prior benchmark evaluates fingerprint augmentation across both dense Wi-Fi and sparse cellular fingerprints under a common protocol, reimplements a comparable number of published methods, or pre-registers structural predictors of benefit against held-out datasets.

Predicting when augmentation helps. Data-complexity measures characterize classification difficulty through class overlap, neighborhood structure, and dimensionality (Ho & Basu, 2002; Lorena et al., 2019), and information-theoretic measures estimate how difficult a dataset is for a given model

family (Ethayarajh et al., 2022). Theoretical analyses explain augmentation as regularization or invariance (Dao et al., 2019; Chen et al., 2020). These lines of work describe how hard a dataset is or why augmentation can help in principle; the ADI instead predicts whether synthetic data will improve accuracy on a specific dataset, and is validated on datasets withheld from its fit.

3 RFWALL

A RFWALL experiment draws a labeled subsample at a fixed budget, trains a generator on that subsample only, and compares a downstream localizer trained with and without the synthetic data. We distinguish released datasets, independently evaluated partitions, and task-specific labelings throughout; Appendix A gives the unit and count for every analysis.

3.1 DATA SUITE

The suite comprises twelve source datasets: sixteen RF partitions spanning dense Wi-Fi, outdoor cellular, and indoor cellular, plus four non-RF controls (Table 4). Cellular scans expose only one or two decodable cells among 12–37 possible features on UniCellular and 44–70 on IndoCell, quantized to integer dBm with a hard sensitivity floor. The non-RF controls—three dense time-series datasets and one sparse voxel-occupancy dataset—vary feature dimensionality and per-sample density independently of radio physics (Appendix B).

3.2 METHOD ZOO

All 31 methods share one training-and-sampling interface, so each is evaluated under identical data, budgets, and downstream models. Twenty-one come from the fingerprint literature, spanning geostatistical interpolation, classical perturbation and recombination, GANs, VAEs and graph generative models, and diffusion. Eight of these are reimplementations of published methods (Njima et al., 2021; Yean et al., 2023; Quezada-Gaibor et al., 2022; Zhang et al., 2025; Liu et al., 2025; Kim et al., 2025; Wu et al., 2025; Abdelmotlb et al., 2025), and three learned models are architectural controls of our own design. The remaining ten are general tabular synthesizers and time-series augmentations, included to test whether findings are specific to fingerprint methods. Tabular methods are run under both continuous and discrete column encodings and reported under their better arm (Appendices C, D).

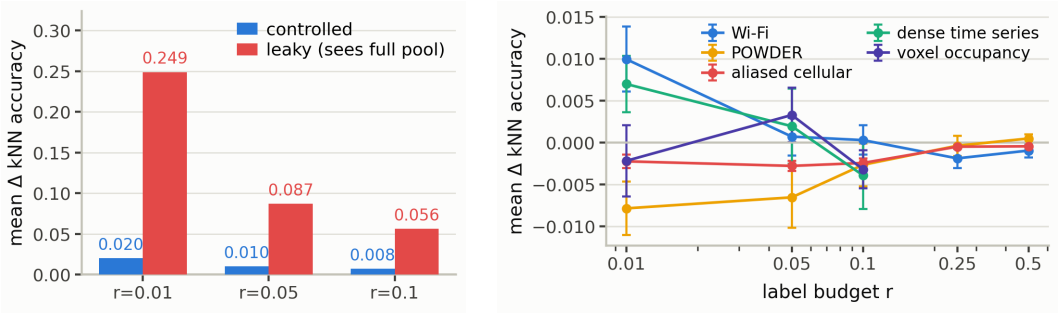
3.3 THE LEAKAGE-CONTROLLED PROTOCOL

Let $D = \{(x_i, \ell_i)\}$ be a labeled pool with $x_i \in \mathbb{R}^d$ a per-scan feature vector (per-AP RSSI or per-cell RSRP) and ℓ_i a location label, and let T be a fixed test set disjoint from D. A method G trained on $S \subseteq D$ produces synthetic pairs $\tilde{D}$, and we ask whether a localizer trained on $S \cup \tilde{D}$ outperforms one trained on S alone. At budget r, we draw a stratified subsample $S_r \subset D$, train G *only* on S_r, and evaluate the downstream model trained on S_r (real-only) and on $S_r \cup \tilde{D}$ (augmented). No stage observes $D \setminus S_r$. We sweep $r \in \{0.01, 0.05, 0.10, 0.25, 0.50, 1.00\}$ with three seeds per configuration. Coordinate-conditional methods and positioning error use only the coordinate-carrying fraction of each pool. The protocol tests densification at observed locations; Sec. 4.4 evaluates extrapolation with a spatial split.

What the protocol implies. Under this protocol the generator has access only to S_r and to label-independent randomness Z, with no pretrained weights, held-out data, or site side information. Hence $\tilde{D} = g(S_r, Z)$, and the data-processing inequality gives $I(\tilde{D}; Y) \leq I(S_r; Y)$ (Cover & Thomas, 2006). Augmentation can reshape the decision boundary using information already in S_r, but cannot add label information. A generator fit on all of D violates this premise, which is why its gains are not evidence about augmentation (Sec. 4.1).

3.4 METRICS AND CONTROLS

Every method is evaluated with two localizers—a one-nearest-neighbor (kNN) classifier and a four-layer MLP—and on two targets where labels allow: categorical accuracy and positioning error in



(a) What the generator is allowed to see

(b) How much real data it competes with

Figure 1: **Two ways a reported benefit disappears.** (a) Mean change in kNN accuracy for five methods on four datasets under the controlled and leaky protocols of Sec. 4.1; the gap closes as the subsample approaches the pool. (b) Mean change in kNN accuracy against label budget by structural regime, over all applicable methods, with standard errors over runs; the Wi-Fi benefit decays as real data accumulates, while the aliased regime has none to decay. Appendices E and G give the full sweeps.

native units. The four non-RF partitions carry no coordinates and are scored on accuracy only. Accuracy changes are absolute ($+0.08$ is eight percentage points). Fidelity is measured with maximum mean discrepancy (Gretton et al., 2012) and a classifier two-sample test (C2ST), the ROC AUC of a real-versus-synthetic discriminator (Lopez-Paz & Oquab, 2017). For statistical control we use a paired Wilcoxon test per method over all datasets, ratios, and seeds, Benjamini–Hochberg-corrected across the 31 methods at FDR 0.05 (Benjamini & Hochberg, 1995), and a Friedman test on within-dataset ranks with a Nemenyi post-hoc (Demšar, 2006). Every synthetic set is filled to the same total training-set size, so additive Gaussian noise serves as a content-free control that isolates training volume from the information contributed by synthesis.

4 Benchmark Results

4.1 Leakage inflates reported gains

We first measure how much of an apparent gain comes from what the generator is allowed to see. We rerun five methods on four datasets under two protocols that differ only in the generator's training data: S_r (controlled) or all of D before S_r is drawn (leaky); Appendix E gives the full sweep. At $r = 0.01$ the leaky protocol reports a mean gain of $+0.249$ against $+0.020$ under control, a $12.3\times$ inflation that falls to $8.4\times$ at $r = 0.05$ and $7.5\times$ at $r = 0.10$ as the subsample approaches the pool (Fig. 1a).

Additive Gaussian noise shows the mechanism most directly. It carries no information beyond the subsample and is worth -0.000 pooled under control, yet appears to add $+0.65$ accuracy on Tampere when fit on the full pool, because its per-class means and variances are then estimated from labels the low-data condition withholds. The inflation exceeds every pooled effect reported in the rest of this paper by more than an order of magnitude. Because the sweep covers only non-learned and rule-based generators, we treat the ratio as an estimate of scale rather than a benchmark-wide constant.

4.2 Where augmentation helps

Three structural regimes. We report results by structural regime rather than by sensor, because the outcome follows the aliasing of the representation more closely than its modality (Table 1). The *aliased* partitions—eight IndoCell and two UniCellular—have 12–70 column vocabularies and exact label collision of 0.12–0.85, and show no usable benefit at any budget (-0.002 at $r = 0.01$ under kNN). The *Wi-Fi* partitions keep collision below 0.004 and are where augmentation pays, averaging $+0.010$ with individual methods an order of magnitude higher. The dense time-series controls fall in between at $+0.007$. The ordering therefore follows aliasing first and per-scan density second.

Table 1: **Augmentation helps only a nearest-neighbor localizer, and only in unaliased regimes.** Mean accuracy change from augmentation at $r = 0.01$, pooled over all applicable methods, by structural regime. Active entries are the mean number of non-null features per sample, with the same quantity as a fraction of the feature vocabulary in parentheses; the Wi-Fi and aliased-cellular ranges overlap on that fraction, while their collision rates do not. Collision is the fraction of samples whose exact feature vector also occurs under another label. Per-method and per-partition results are in Fig. 2 and Appendix I; the fingerprint-methods-only breakdown is in Appendix H.

Regime	Partitions	Active entries	Collision	kNN Δ	MLP Δ
Wi-Fi	5	6.4–48.0 (3–92%)	≤ 0.004	$+0.0100$	-0.0570
Aliased cellular	10	1.4–4.5 (2–36%)	0.12–0.85	-0.0022	-0.0181
POWDER (outdoor cellular)	1	20.8 (83%)	0.000	-0.0078	-0.0550
Dense time series	3	94–658 (98–100%)	0.000–0.022	$+0.0070$	-0.0495
Voxel occupancy	1	109 (21%)	0.024	-0.0022	-0.0504

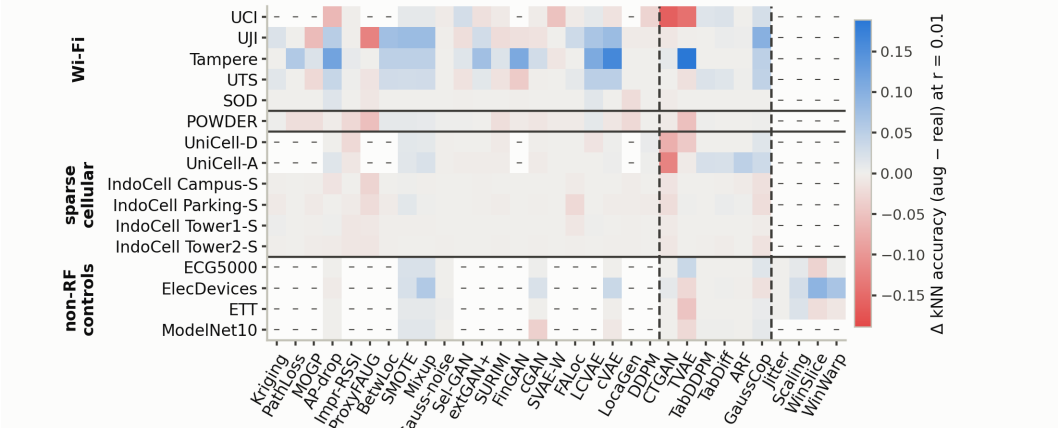
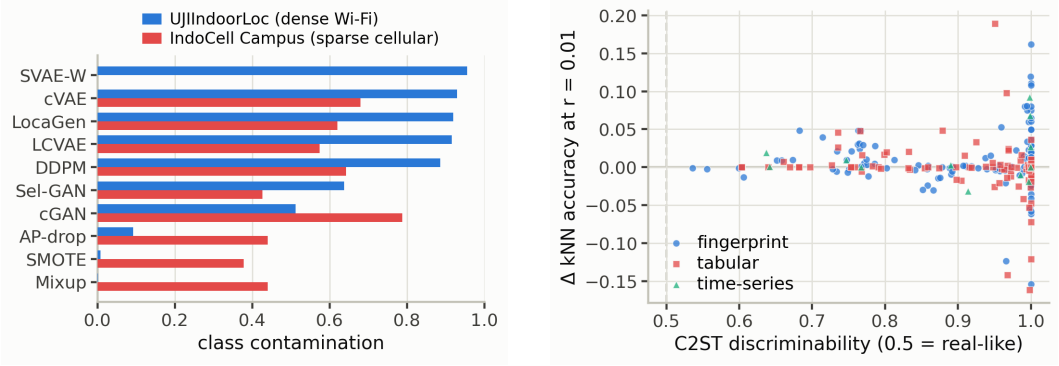


Figure 2: **The benefit is concentrated in a few cells.** Mean change in kNN accuracy from augmentation (augmented minus real-only) at $r = 0.01$, over all 31 methods and 16 of the 20 base partitions; every aggregate in this section is a mean over some slice of this grid. "–" marks cells where the method does not apply: a coordinate-conditional method on a dataset without coordinates, or a time-series augmentation off the three time-series partitions. The four mobile IndoCell partitions are omitted for legibility and behave like the stationary rows shown; dashed lines separate the 21 fingerprint methods, the 6 tabular synthesizers, and the 4 time-series augmentations. Appendix F reads the grid row by row and quotes the extreme cells.

POWDER. One partition does not fit this ordering. POWDER is dense (83% active) and unaliased (collision 0.000), yet augmentation is negative on it (-0.008). It shares a small vocabulary with the cellular partitions—25 receivers—so it is the one case where the vocabulary is small without aliasing, and the benefit is still absent.

Scarcity, not task difficulty. The Wi-Fi benefit decays toward zero as real data accumulates, while the aliased regime is flat at every budget, including the smallest (Fig. 1b, Appendix G); the cellular null is therefore not a consequence of fewer labels. Nor is it a consequence of less room to improve. Taking real-only accuracy at $r = 1.00$ as the ceiling, the aliased partitions leave headroom of 0.171 against 0.255 on Wi-Fi, of which augmentation captures -0.6% and 10.8% respectively.

Dependence on the downstream model. The positive effects above are kNN effects. With the MLP as the localizer, the mean change is negative in every regime (Table 1), and the reversal holds for the fingerprint-native methods alone (Wi-Fi -0.051; Appendix H). A nearest-neighbor rule is unaffected by a synthetic point placed beside an existing same-class neighbor, whereas an MLP fits a global boundary that systematic artifacts can move. Throughout the paper, "augmentation helps on Wi-Fi" should be read as "augmentation helps a nearest-neighbor localizer on Wi-Fi."



(a) Where synthetic points land

(b) Fidelity against utility

Figure 3: **Two diagnostics behind the ranking.** (a) Class contamination, the fraction of synthetic points whose nearest real neighbor carries a different label, for ten representative methods at $r = 0.01$, on dense Wi-Fi and on an aliased cellular partition. (b) C2ST discriminability against accuracy change, one point per method and partition at $r = 0.01$, over all 31 methods and 20 partitions; an AUC of 0.5 means real and synthetic data are indistinguishable, and the largest gains sit at the right edge. Appendices K and M give the per-family detail.

4.3 Interpolation versus deep synthesis

Method ranking. Ranked by mean within-dataset rank of accuracy change (Appendix I), SMOTE is first (3.53), mixup second (5.78), and CTGAN last (18.12). A Friedman test rejects equal ranks, but the Nemenyi critical difference is wide, so the ranking separates the extremes rather than the leading methods from each other. Eighteen of 31 methods have a significant pooled effect, but 13 of those are negative; only SMOTE, mixup, Between-Location, and two time-series augmentations are significantly positive. Per dataset, a non-parametric method wins 12 of 20 partitions and a deep generative method wins 8. The generic tabular synthesizers are high-variance: ARF ranks third overall, while TVAE produces the largest single gain in the benchmark ($+0.189$ on Tampere) and is significantly negative pooled (-0.012).

Where learned generators win. Tampere is the clearest case: TVAE reaches $+0.189$ and a conditional VAE $+0.162$, against $+0.048$ for SMOTE and mixup (Lohan et al., 2017). At $r = 0.01$ Tampere provides only seven real samples—too few for same-class interpolation to draw on—against a 992-dimensional fingerprint with enough shared structure for a jointly trained generator to exploit. This combination does not recur elsewhere in the suite. SVAE-WGAN is the opposite case: its adversarial objective produced non-finite output in 200 of 210 runs at these budgets, after which the guard emitted no synthetic data. Its fourth-place rank therefore mostly records abstention rather than a working generator, and the divergence rate is the result to carry forward.

Structural match. The non-RF controls show that a method extracts only the structure it assumes. On ModelNet10 voxels, methods that keep a synthetic grid on its class's occupancy topology gain $+0.0089$, while learned generators indifferent to that topology lose -0.0076 ($p = 7 \times 10^{-6}$); the structure-respecting family wins all 8 cells of the sparsification sweep ($p = 0.007$). On time series, window slicing, which assumes translation tolerance, is the best method on ElectricDevices ($+0.092$) and the worst of its family on ECG5000 and ETT, where class identity depends on the position of a beat complex or a quantile of the cropped window (Appendix J).

Where synthetic points land. For ten methods on one Wi-Fi and one cellular partition at $r = 0.01$, we measure the fraction of synthetic points whose nearest real neighbor has a different label (Fig. 3a, Appendix K). On Wi-Fi, interpolation contaminates almost nothing, while every deep generative method contaminates at least half its output. On the aliased partition every method degrades, and interpolation's contamination rises from 0.008 to 0.378, because the feature space leaves too little margin for a same-class convex combination to stay nearest its own class. The usual explanation for contamination in VAEs appeals to latent geometry; we instrumented four VAE-family models

and found no posterior collapse ($67/67$ runs), no preferential decoder stretching at class boundaries ($10/43$ runs), and no excess out-of-class decoding for different-class latent midpoints (0.44 against 0.48 for a same-class control). The mechanism therefore remains open (Appendix L).

Fidelity does not predict utility. The methods with the largest gains are among the easiest to distinguish from real data, because their points lie exactly on segments between same-class samples, which a discriminator detects and a downstream boundary benefits from. Several methods with more realistic output give no benefit (Fig. 3b, Appendix M), and switching tabular synthesizers to a discrete encoding improves fidelity while leaving utility flat or worse (Appendix D).

4.4 Beyond densification and accuracy

Spatial extrapolation. Withholding entire k-means clusters of reference coordinates from generator fitting tests whether coordinate-conditional methods extrapolate (Appendix N). Kriging degrades most: its reconstruction RMSE on UJIIndoorLoc worsens 73.7%, against 7.9% for MOGP, because its semivariogram constrains nothing outside the observed support. On IndoCell's Campus partition, every coordinate-conditional method is worse than predicting the training-set mean fingerprint.

Models and targets. Extending a six-method subset at $r = 0.01$ to random forest and logistic regression, only Tampere is positive under all four downstream models, and the positive kNN result on UJIIndoorLoc ($+0.044$) holds under none of the other three (Appendix H). The two targets also diverge: the largest accuracy gainers on a dataset often worsen its positioning error by several metres (Appendix O). Restricting stationary IndoCell partitions to reference-point rather than floor labels leaves every non-parametric method within ± 0.01.

5 The Sparsity Wall: A Structural Account

Sec. 3.3 implies that augmentation can only regularize with information already present in the subsample, so whether regularization is worthwhile must depend on the representation. We first identify the causal variable with a controlled experiment, then measure the mechanisms it acts through, and finally combine them into a screen.

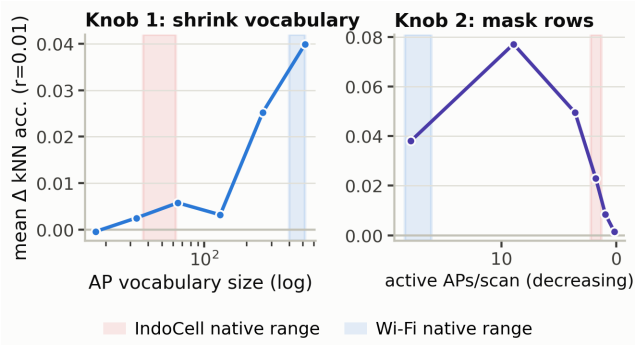
5.1 A controlled sparsification experiment

Sec. 4.2 compares different buildings, sensors, and modalities, so the gap could in principle reflect an unmeasured property of the cellular sites. We therefore move UJIIndoorLoc toward the cellular regime along two independent knobs while holding the building, labels, and every real RSSI measurement fixed: shrinking the retained AP vocabulary by dropping columns, and masking a growing fraction of each row's active APs at fixed vocabulary. Per-scan sparsity alone cannot be the operative variable, since natural UJIIndoorLoc scans are already 96.6% not-heard, inside IndoCell's range of 96.4–97.7%; the datasets differ instead in vocabulary, 520 columns against 44–70. Shrinking the vocabulary reduces the gain monotonically from $+0.040$ to zero, becoming indistinguishable from zero as the vocabulary enters IndoCell's native range (Fig. 4a). Both knobs eventually remove the benefit, but only the vocabulary knob does so along the path the cellular partitions follow.

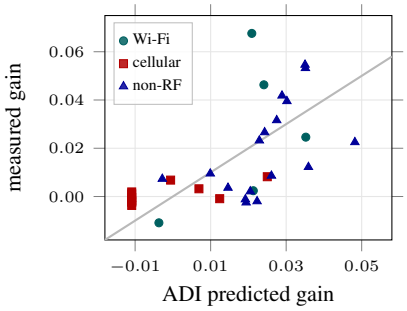
5.2 Label collision and manifold collapse

Label collision. Because columns are quantized, two scans can share an identical feature vector under different labels. The collision rate separates the modalities: every Wi-Fi dataset lies between 0.000 and 0.004, while the eight IndoCell partitions span 0.380–0.850, six of them above 0.71, implying Bayes error floors of 0.09–0.61 that no classifier can cross. The sparsification sweep reproduces this: reducing UJIIndoorLoc to 17 columns yields collision 0.853 and a floor of 0.580, close to the most aliased real cellular partition (0.850, 0.611) on identical data and labels.

Manifold collapse. Interpolation can only place synthetic points inside the convex hull of same-class real points, so we measure the fraction of test points reachable in this way. Sparse cellular test points are *more* reachable (0.596) than Wi-Fi ones (0.250), and on three of the eight IndoCell



(a) Controlled sparsification of UJIIndoorLoc



(b) ADI, leave-one-dataset-out

Figure 4: **Vocabulary size causes the wall, and the ADI orders held-out datasets.** (a) Two sparsification knobs on UJIIndoorLoc, holding the building, labels, and every real measurement fixed. Shrinking the AP vocabulary from 520 to 17 columns reduces the gain monotonically to zero; masking a growing fraction of each row's active APs at fixed vocabulary removes the benefit only eventually and non-monotonically. Bands mark each modality's native range. (b) Predicted against measured gain for the ADI (Sec. 5.3) over all 33 rows of the pooled suite, each predicted with its entire source dataset withheld; grey line is $y = x$. Spearman $\rho = +0.739$, MAE 0.0138 against a constant's 0.0149. Cellular points stack at the left edge because, with IndoCell withheld, no training fold contains a collision rate above 0.32.

Table 2: **The inherent-property ADI gives the best rank ordering of held-out datasets.** Leave-one-dataset-out performance on the pooled suite (33 rows, 12 sources, equally weighted). Sign accuracy is the fraction of withheld rows placed on the correct side of the 0.005 threshold.

Specification	LODO MAE	Spearman	Sign acc.
vocabulary + purity *(precursor, App. Q)*	0.0136	+0.508	0.55
vocabulary + purity + real-only accuracy	**0.0130**	+0.376	**0.76**
collision + manifold sparsity + noise (ADI)	0.0138	**+0.739**	0.70
collision + manifold sparsity	0.0219	+0.197	0.67
constant +0.005	0.0149	—	0.48

partitions the residual distance is numerically zero ($< 10^{-12}$). High reachability here indicates collapse: synthetic points land on existing real points and add nothing. Reachability predicts measured gain across partitions ($\rho = -0.598$, $p = 7.7 \times 10^{-4}$, $n = 28$), but it correlates with collision at $\rho = 0.71$, so we treat the two as views of one mechanism. Intrinsic dimension alone does not separate the modalities (TwoNN: Wi-Fi 0.08–5.9, IndoCell 0.10–1.7), so the index below uses it only as the denominator of a ratio.

Non-RF controls. Both mechanisms appear without radio physics. Coarsening the ModelNet10 voxel grid from 8^3 to 6^3 to 4^3 shrinks the vocabulary while leaving objects, labels, and task unchanged, raising collision from 0.024 to 0.072 to 0.338 while measured gain stays small at every grid. Relocating a fraction of each object's occupied voxels to random empty positions instead holds feature count, occupancy, and collision fixed while raising only the noise level; gain rises to +0.0395 at 50% noise before falling once objects become too corrupted to classify. Noise that leaves labels recoverable is where augmentation helps most, and aliasing that destroys them is where it cannot. Because each manipulation also degrades the real-only baseline, this evidence is directional rather than calibrated (Appendix Q).

5.3 The Augmentation Difficulty Index

Definition. The ADI is a linear model on three properties of the data, one per mechanism, excluding anything that depends on a trained model. *Collision rate* captures lost label information. *Relational manifold sparsity*, the effective rank of the covariance spectrum divided by the TwoNN intrinsic dimension, captures how many directions the data occupies relative to its manifold di-

mension. *Noise level*, within-class spread over between-class separation, captures what fills those directions. The last two enter in logs. Downstream accuracy is excluded because it is a property of the (data, model) pair rather than of the dataset. We arrived at this specification after a fingerprint-only precursor built from vocabulary size and neighbor purity failed pre-registered out-of-domain tests (Appendix Q).

Validation. We fit on 33 partitions and sweep cells from 12 source datasets, RF and non-RF together, and validate leave-one-*dataset*-out. Folds are defined by source because sweep variants of one dataset are near-duplicates, and rows are weighted so each source carries equal total weight, since ModelNet10 alone contributes 14 rows. The ADI reaches Spearman $\rho = +0.739$ ($p = 9 \times 10^{-7}$), against $+0.508$ and $+0.376$ for the model-dependent alternatives, at essentially the same absolute error, and beats a constant on both rank and sign accuracy (0.70 against 0.48; Table 2, Fig. 4b). Its coefficients are negative on collision and positive on log manifold sparsity and log noise: a representation is worth augmenting when its labels are not aliased, when it occupies many directions relative to its manifold dimension, and when noise rather than signal fills those directions. The ranking is stable when either large source is dropped.

Scope. The ADI orders the pooled RF and non-RF suite and correctly identifies the held-out cellular modality as a null regime. It is a screen for direction, not magnitude: its absolute error is no better than a constant's. Because collision above 0.32 occurs only in IndoCell, that coefficient is extrapolated whenever IndoCell is withheld.

6 Discussion and Conclusion

Within RFWall, the augmentation gap is more strongly associated with the structure of the feature space than with generator family. Augmentation cannot add label information, so it can only reshape a decision boundary with what the subsample already contains. A cellular fingerprint of the kind IndoCell releases is close to discrete, with roughly two decodable cell identities among 44–70 features; it collides, imposes a Bayes floor of up to 0.61, and occupies a manifold on which a synthetic point either duplicates a real point or falls off the manifold. A compatible reading of our results is that augmentation is broadly useful wherever a representation retains enough continuous local structure, so that the cellular null reflects a mismatch between the representation and the assumptions these methods encode rather than a limit of generative modeling.

Recommendations. Our results suggest four practices. Screen the representation before choosing a generator, since the expected return is near zero where labels are aliased. Try SMOTE or mixup before a deep generative model, and match the method's assumed structure to the data. Report at least two downstream models and, where coordinates exist, both accuracy and positioning error, since these are the main axes of disagreement in our results. Do not use fidelity as a substitute for utility.

Limitations. The positive findings are nearest-neighbor findings; whether this reflects augmentation itself or how these classifiers respond to duplicated local structure is the most important open question. The ADI is a directional screen, and both collision above 0.32 and the dense low-rank corner are extrapolated. Eight methods are reimplementations because many published systems report only on radio maps that were never released (Yoon et al., 2024; Yoo, 2024), so a single method's number is weaker evidence than the aggregate comparisons. The mechanism behind class contamination remains open. Appendix R lists all qualifications.

Conclusion. We introduced RFWall, a leakage-controlled benchmark for data augmentation across Wi-Fi, cellular, and non-RF representations. Bounded by the data-processing inequality to a regularization effect, augmentation pays only where the representation leaves something to regularize. Because the relevant properties can be measured from the data alone, this question can be answered before a generator is chosen, and the non-RF controls indicate that the answer concerns representations in general rather than radio data specifically.

Reproducibility Statement

All splits are fixed in advance and released with the benchmark code; every (dataset, method, ratio, seed) configuration runs independently and deterministically. The protocol is specified in Sec. 3.3, statistical procedures including the multiple-comparison correction and the content-free control in Sec. 3.4, the dataset suite and its per-partition processing in Sec. 3.1 and Appendix B, and every reimplemented method with each documented deviation from its published recipe in Appendix C. Eleven of the twelve source datasets are public and cited; the twelfth is released with this work. The four out-of-domain ADI predictions were registered before the corresponding sweeps were run; each registration file, carrying a UTC timestamp and a SHA-256 digest of the fitted coefficients, is released with the benchmark. Anonymized source code is provided as supplementary material.

Ethics Statement

Fingerprint localization can in principle support unwanted tracking, but the artifacts released here are radio maps and augmentation baselines, not a tracking system, and our central finding is that a class of proposed methods works less well than reported rather than a capability increase. We have read and adhere to the ICLR Code of Ethics.

AI Use Statement

In this work, we used generative AI tools to debug the benchmark code. We have not used generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for mathematical proofs, assist in writing proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, assist with translation, clean or reformat datasets, support qualitative or thematic data analysis, or interpret experimental results. The choices of methods, datasets, experimental design, and conclusions were made by the authors.

Additionally, we used generative AI tools to create and edit visualization code, improve paper readability, search for and identify relevant literature and information, format references, and develop the structure of the paper. In particular, AI tools were used to identify potentially relevant methods, datasets, and references, as well as to examine the structure of other ICLR papers; the authors independently evaluated these suggestions and made the final choices. The synthetic fingerprints evaluated throughout this paper are produced by the benchmarked augmentation methods themselves, not by any general-purpose generative AI tool.

We primarily used Claude for these purposes. We have reviewed all AI-assisted work: AI-assisted code was reviewed, debugged, and tested by the authors; experimental results were independently checked by running the relevant experiments; literature and references suggested by the AI were verified against the original sources; and AI-assisted text and visualizations were reviewed by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

References

Abdelrahman Abdelmotlb, Abdallah Taman, Sherif Mostafa, and Moustafa Youssef. LocaGen: Low-overhead indoor localization through spatial augmentation, 2025. Proc. 4th ACM SIGSPATIAL Int. Workshop on Spatial Big Data and AI for Industrial Applications (GeoIndustry '25).

Grigorios G. Anagnostopoulos and Alexandros Kalousis. ProxyFAUG: Proximity-based fingerprint augmentation, 2021. Proc. IPIN 2021.

Anonymous. IndoCell: A multi-site, multi-RAT cellular fingerprint dataset for indoor localization, 2026. Dataset details withheld for review.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57(1):289–300, 1995.

Jingxue Bi, Yunjia Wang, Baoguo Yu, Hongji Cao, Tongguang Shi, and Lu Huang. Supplementary open dataset for WiFi indoor localization based on received signal strength. *Satellite Navigation*, 3:25, 2022. doi: 10.1186/s43020-022-00086-y. SODIndoorLoc; 21,205 train / 2,720 test scans, 105 APs, 3 buildings, 1,630 RPs; data: `https://github.com/bijingxue/SODIndoorLoc`.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002. doi: 10.1613/jair.953.

Shuxiao Chen, Edgar Dobriban, and Jane H. Lee. A group-theoretic framework for data augmentation. *Journal of Machine Learning Research*, 21(245):1–71, 2020.

Yanping Chen, Yuan Hao, Thanawin Rakthanmanon, Jesin Zakaria, Bing Hu, and Eamonn Keogh. A general framework for never-ending learning from time series streams. *Data Mining and Knowledge Discovery*, 29(6):1622–1664, November 2015. ISSN 1573-756X. doi: 10.1007/s10618-014-0388-4.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006.

Tri Dao, Albert Gu, Alexander Ratner, Virginia Smith, Christopher De Sa, and Christopher Ré. A kernel theory of modern data augmentation. In *International Conference on Machine Learning (ICML)*, 2019.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *International Conference on Machine Learning (ICML)*, 2022.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Suining He and S.-H. Gary Chan. Wi-fi fingerprint-based indoor positioning: Recent advances and comparisons. *IEEE Communications Surveys & Tutorials*, 18(1):466–490, 2016. doi: 10.1109/COMST.2015.2464084.

Amr E. Hilal, Ismail Arai, and Samy El-Tawab. DataLoc+: A data augmentation technique for machine learning in room-level indoor localization, 2021. Proc. IEEE WCNC 2021.

Tin Kam Ho and Mitra Basu. Complexity measures of supervised classification problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):289–300, 2002.

Sau-Shian Jan, Shau-Jie Yeh, and Ya-Wen Liu. Received signal strength database interpolation by kriging for a Wi-Fi indoor positioning system. *Sensors*, 15(9):21377–21393, 2015. doi: 10.3390/s150921377.

Honggang Jia, Nan Cheng, Xiucheng Wang, Conghao Zhou, Ruijin Sun, and Xuemin Shen. RadioMamba: Breaking the accuracy-efficiency trade-off in radio map construction via a hybrid Mamba-UNet. *IEEE Trans. Netw. Sci. Eng.*, 13:2454–2468, 2026. arXiv:2508.09140.

Fan Jiang, Ting Li, Xin Lv, Hua Rui, and Depeng Jin. Physics-informed neural networks for path loss estimation by solving electromagnetic integral equations. *IEEE Trans. Wireless Commun.*, 23(10):15380–15393, 2024. doi: 10.1109/TWC.2024.3392330.

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.

Dongdeok Kim, Jae-Hyeon Park, and Young-Joo Suh. A Wi-Fi fingerprinting indoor localization framework using feature-level augmentation via variational graph auto-encoder. *Electronics*, 14 (14):2807, 2025. doi: 10.3390/electronics14142807.

Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. Tabddpm: Modelling tabular data with diffusion models. In *International Conference on Machine Learning (ICML)*, 2023.

Chunyuan Liu, Xiaomin Yu, Peilong Wu, and Haoyu Wang. A novel 3D indoor localization method integrating deep spatial feature augmentation and attention-based denoising. *Scientific Reports*, 15(1):33025, 2025. doi: 10.1038/s41598-025-18549-y.

Elena Simona Lohan, Joaquín Torres-Sospedra, Helena Leppäkoski, Philipp Richter, Zhe Peng, and Joaquín Huerta. Wi-Fi crowdsourced fingerprinting dataset for indoor positioning. *Data*, 2(4): 32, 2017. doi: 10.3390/data2040032. 4,648 fingerprints; 697 train / 3,951 test; data: `https://zenodo.org/records/889798`.

David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *Int. Conf. Learning Representations (ICLR)*, 2017.

Ana C. Lorena, Luís P. F. Garcia, Jens Lehmann, Marcilio C. P. Souto, and Tin Kam Ho. How complex is your classification problem? A survey on measuring classification complexity. *ACM Computing Surveys*, 52(5), 2019.

Frost Mitchell, Aniqua Baset, Sneha Kumar Kasera, and Aditya Bhaskara. A dataset of outdoor RSS measurements for localization. Zenodo, `https://zenodo.org/records/7259895`, 2022. 5,214 samples, 10–25 receivers; POWDER testbed, University of Utah.

Sherif Mostafa, Khaled A. Harras, and Moustafa Youssef. UniCellular: An accurate and ubiquitous floor identification system using single cell tower information. In *Proc. 31st ACM Int. Conf. Advances in Geographic Information Systems (SIGSPATIAL)*, 2023. doi: 10.1145/3589132.3625635.

Wafa Njima, Marwa Chafii, Arsenia Chorti, Raed M. Shubair, and H. Vincent Poor. Indoor localization using data augmentation via selective generative adversarial networks. *IEEE Access*, 9: 98337–98347, 2021. doi: 10.1109/ACCESS.2021.3095546.

Wafa Njima, Ahmad Bazzi, and Marwa Chafii. DNN-based indoor localization under limited dataset using GANs and semi-supervised learning. *IEEE Access*, 10:69896–69909, 2022. doi: 10.1109/ACCESS.2022.3187837.

Sinno Jialin Pan, James T. Kwok, Qiang Yang, and Jeffrey J. Pan. Adaptive localization in a dynamic WiFi environment through multi-view learning. In *Proc. 22nd AAAI Conf. Artificial Intelligence*, 2007.

Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. The synthetic data vault. In *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2016.

Darwin Quezada-Gaibor, Joaquín Torres-Sospedra, Jari Nurmi, Yevgeni Koucheryavy, and Joaquín Huerta. SURIMI: Supervised radio map augmentation with deep learning and a generative adversarial network for fingerprint-based indoor positioning, 2022. Proc. IPIN 2022; code: `https://github.com/darwinquezada/SURIMI`.

Anshul Rai, Krishna Kant Chintalapudi, Venkata N. Padmanabhan, and Rijurekha Sen. Zee: Zero-effort crowdsourcing for indoor localization. In *Proc. 18th Annual Int. Conf. Mobile Computing and Networking (MobiCom)*, pp. 293–304, 2012. doi: 10.1145/2348543.2348580.

Vijay Janapa Reddi, Christine Cheng, David Kanter, Peter Mattson, Guenther Schmuelling, Carole-Jean Wu, et al. MLPerf inference benchmark. In *2020 ACM/IEEE 47th Annual Int. Symp. Computer Architecture (ISCA)*, pp. 446–459, 2020. doi: 10.1109/ISCA45697.2020.00045.

Jayant G. Rohra et al. Wireless indoor localization data set. UCI Machine Learning Repository, `https://archive.ics.uci.edu/dataset/422`, 2017. 2,000 samples, 7 APs, 4 rooms (500/room).

Juntong Shi, Minkai Xu, Harper Hua, Hengrui Zhang, Stefano Ermon, and Jure Leskovec. Tabdiff: a mixed-type diffusion model for tabular data generation. In *International Conference on Learning Representations (ICLR)*, 2025.

Rashmi Sharan Sinha and Seung-Hoon Hwang. Comparison of CNN applications for RSSI-based fingerprint indoor localization. *Electronics*, 8(9):989, 2019. doi: 10.3390/electronics8090989.

Rashmi Sharan Sinha and Seung-Hoon Hwang. Improved RSSI-based data augmentation technique for fingerprint indoor localisation. *Electronics*, 9(5):851, 2020. doi: 10.3390/electronics9050851.

Xudong Song, Xiaochen Fan, Chaocan Xiang, Qianwen Ye, Leyu Liu, Zumin Wang, Xiangjian He, Ning Yang, and Gengfa Fang. A novel convolutional neural network based indoor localization framework with WiFi fingerprinting. *IEEE Access*, 7:110698–110709, 2019. doi: 10.1109/ACCESS.2019.2933921. UTSIndoorLoc dataset: `https://github.com/XudongSong/CNNLoc-Access`.

Masato Sugasaki and Masamichi Shimosaka. Robustifying Wi-Fi localization by between-location data augmentation. *IEEE Sensors Journal*, 2021. doi: 10.1109/JSEN.2021.3106765. Project page: `http://www.miubiq.cs.titech.ac.jp/indoorloc-bl-en/`.

Zhe Tang, Sihao Li, Kyeong Soo Kim, and Jeremy S. Smith. Multi-dimensional Wi-Fi received signal strength indicator data augmentation based on multi-output Gaussian process for large-scale indoor localization. *Sensors*, 24(3):1026, 2024. doi: 10.3390/s24031026.

Joaquín Torres-Sospedra, Raúl Montoliu, Adolfo Martínez-Usó, Joan P. Avariento, Tomás J. Arnau, Mauri Benedito-Bordonau, and Joaquín Huerta. UJIIndoorLoc: A new multi-building and multi-floor database for WLAN fingerprint-based indoor localization problems. In *Proc. Int. Conf. Indoor Positioning and Indoor Navigation (IPIN)*, pp. 261–270, 2014. doi: 10.1109/IPIN.2014.7275492.

TSML Archive Maintainers. ElectricDevices. Zenodo, October 2018.

Terry T. Um, Franz M. J. Pfister, Daniel Pichler, Satoshi Endo, Muriel Lang, Sandra Hirche, Urban Fietzek, and Dana Kulić. Data augmentation of wearable sensor data for Parkinson's disease monitoring using convolutional neural networks. In *Proc. 19th ACM Int. Conf. Multimodal Interaction (ICMI)*, pp. 216–220, 2017. doi: 10.1145/3136755.3136817.

Bairu Wang, Xiaodong Liu, Baoguo Yu, Ruicai Jia, and Xingli Gan. Fast radio map construction by using adaptive path loss model interpolation in large-scale building. *Sensors*, 19(3):712, 2019. doi: 10.3390/s19030712.

He Wang, Souvik Sen, Ahmed Elgohary, Moustafa Farid, Moustafa Youssef, and Romit Roy Choudhury. No need to war-drive: Unsupervised indoor localization. In *Proc. 10th Int. Conf. Mobile Systems, Applications, and Services (MobiSys)*, pp. 197–210, 2012.

Xiucheng Wang, Keda Tao, Nan Cheng, Zhisheng Yin, Zan Li, Yuan Zhang, and Xuemin Shen. RadioDiff: An effective generative diffusion model for sampling-free dynamic radio map construction. *IEEE Trans. Cogn. Commun. Netw.*, 11(2):738–750, 2025. doi: 10.1109/TCCN.2024.3504489. arXiv:2408.08593.

David S. Watson, Kristin Blesch, Jan Kapar, and Marvin N. Wright. Adversarial random forests for density estimation and generative modeling. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

Shixun Wu, Xinrui Zeng, Miao Zhang, Kanapathippillai Cumanan, Abdulhamed Waraiet, Zheng Chu, and Kai Xu. LCVAE-CNN: Indoor Wi-Fi fingerprinting CNN positioning method based on LCVAE. *IEEE Internet of Things Journal*, 12(16):33395–33410, 2025. IEEE Xplore document 11021455; add DOI when available.

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes, 2015. URL `https://arxiv.org/abs/1406.5670`.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Zheng Yang, Chenshu Wu, and Yunhao Liu. Locating in fingerprint space: Wireless indoor localization with little human intervention. In *Proc. 18th Annual Int. Conf. Mobile Computing and Networking (MobiCom)*, pp. 269–280, 2012.

Seanglidet Yean, Wayne Goh, Bu-Sung Lee, and Hong Lye Oh. extendGAN+: Transferable data augmentation framework using WGAN-GP for data-driven indoor localisation model. *Sensors*, 23(9):4402, 2023. doi: 10.3390/s23094402.

Jaehyun Yoo. Wi-fi fingerprint indoor localization by semi-supervised generative adversarial network. *Sensors*, 24(17):5698, 2024. doi: 10.3390/s24175698.

Namkyung Yoon, Wooyong Jung, and Hwangnam Kim. Deeprssi: Generative model for fingerprint-based localization. *IEEE Access*, 2024. doi: 10.1109/ACCESS.2024.3398970.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *Int. Conf. Learning Representations (ICLR)*, 2018.

Jiaming Zhang, Jiajun He, Jie Zhang, and Okan Yurduseven. FinGAN: An interpretable RSS generation network for scalable fingerprint localization, 2025. Accepted, IEEE ICC 2026.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Virtual Conference*, volume 35, pp. 11106–11115. AAAI Press, 2021.

A Unit of Analysis

Different analyses in this paper count different things, because they answer different questions. The benchmark is reported over *partitions*, the level at which a method is run. The structural model is fit over *representation variants*, since the sparsification sweep produces twelve structurally different versions of one building that must each contribute a point. Cross-validation folds at the level of the *source dataset*, since those twelve variants are near-duplicates and splitting them across folds would leak. Table 3 states the unit and count for every analysis.

Table 3: Unit of analysis for each study in the paper. A structural model is fit over representation variants, while cross-validation folds at the level of the source dataset to keep near-duplicates on one side of a split.

Analysis	**Unit, and how many**
Main benchmark	20 base partitions (16 RF, 4 non-RF); 24 partition–task combinations
Fingerprint ADI fit (Appendix Q)	28 representation variants: 16 real fingerprint partitions + 12 controlled UJIIndoorLoc sweep variants
Held-out modality test	8 IndoCell partitions, excluded from that fit
Pre-registered out-of-domain tests	4 datasets, each predicted before any method ran on it
ModelNet sparsification sweep (Appendix Q)	14 cells across two knobs, 441 runs
Pooled re-fit and cross-validation (Sec. 5.3)	33 dataset/sweep rows, 12 source-level folds

B Data Suite Detail

Coordinates are available only where a genuine floor plan exists. IndoCell's low-rise university building and two of three car-park levels carry real floor-plan pixel coordinates, while the two high-rise sites release only schematic, non-metric layouts that repeat across floors and are therefore excluded from positioning-error evaluation. UTSIndoorLoc contributes a single 16-floor building with 589 APs, evaluated on floor classification with the dataset's own released coordinates.

SODIndoorLoc contributes three buildings exposing one to three floors each. Because this yields only five distinct building-floor classes, its classification accuracy is near ceiling with only a handful of labeled samples, which we flag as a limitation even though its positioning-error results remain informative. Its three buildings also retain separate coordinate frames, so a coordinate-conditional generator sees three geometrically disjoint coordinate ranges that can numerically overlap without corresponding to the same physical location. In addition to the eight-partition, floor-level IndoCell task, its four stationary partitions with sufficient reference points support a reference-point-level task, evaluated in Sec. 4.4.

C Method Catalog

Table 5 catalogs all 31 methods. The **Re-impl.** column marks the eight methods we implemented from their papers because no usable release exists; their individual numbers are weaker evidence than the aggregates (Appendix R). The last two groups are the ten methods from outside the fingerprint literature, included so that fingerprint methods are not judged only against each other. This control proved informative: one of the ten (ARF) places third overall, while two others place last (Appendix I).

D Column Encoding for Tabular Baselines

The six general tabular synthesizers are designed for mixed continuous and discrete columns, so how a fingerprint column is declared is a substantive design choice. Fingerprint columns are quantized to integer dBm over a dominant not-heard floor, so a typical column takes only a handful of distinct values (median 5 on Tampere, 6 on IndoCell, 27 on UJIIndoorLoc). Declaring such a column continuous disables the mixed-type machinery these methods provide; declaring it discrete asks

Table 4: Datasets evaluated in RFWALL. #Samples are raw observations. †not stated by the source. *independent modem readings (1.08 M logged scans). ‡both UniCellular testbeds. For non-RF controls, d is the feature dimensionality.

Dataset (modality; scale)	#Samples	Labels
Public Wi-Fi RSSI		
UJIIndoorLoc (Torres-Sospedra et al., 2014); 520 APs, 3 bldg/933 RP	21,048	bldg/fl/coord
Tampere/TUT (Lohan et al., 2017); 992 APs, 1 bldg/4 fl	4,648	fl/coord
UTSIndoorLoc (Song et al., 2019); 589 APs, 1 bldg/16 fl	—†	fl/coord
SODIndoorLoc (Bi et al., 2022); 105 APs, 3 bldg/1,630 RP	23,925	fl/coord
UCI WSN-Indoor (Rohra et al., 2017); 7 APs, 1 fl/4 rooms	2,000	room
Cellular		
POWDER (Mitchell et al., 2022); 10–25 Rx, outdoor campus	5,214	coord
UniCellular (Mostafa et al., 2023); serving cell, 2 bldg/7–9 fl	∼29,000‡	floor
IndoCell (Anonymous, 2026); GSM+LTE, 4 sites/61 fl/549 RP	159,800*	floor/RP/coord
Non-RF controls		
ElectricDevices (TSML Archive Maintainers, 2018); 96 features, dense, high-rank	—	class
ECG5000 (Chen et al., 2015); 140 features, dense, low-rank	—	class
ETT (Zhou et al., 2021); 672 features (96×7), dense, low-rank	—	target quartile
ModelNet10 (Wu et al., 2015); $8^3 = 512$ voxel features, 21.3% active	—	shape class

them to handle roughly a thousand categorical columns on Wi-Fi. We therefore ran every tabular method in both encodings across all partitions, budgets, and seeds, and report each under its better arm elsewhere in the paper.

The discrete encoding costs CTGAN the most (-0.0186) and the two diffusion synthesizers about 0.004 each, while leaving TVAE and the copula unchanged. One plausible explanation, which we did not test, is that CTGAN's training-by-sampling conditions on one category drawn from one discrete column—a design suited to a handful of such columns rather than the ∼1000 a Wi-Fi fingerprint presents—and that an unordered categorical discards the ordinal structure of dBm that a truncated normal approximates.

For four of the five methods with a real switch, the discrete encoding makes synthetic data *harder* to distinguish from real—CTGAN $1.000 \to 0.971$, TabDDPM $0.980 \to 0.935$, TabDiff $0.980 \to 0.937$, ARF $0.881 \to 0.816$—while leaving downstream utility flat or worse in each case. This is a within-method instance of the fidelity–utility decoupling in Appendix M: a single configuration flag moves the two quantities in opposite directions on the same method and data. TVAE moves slightly up on utility and markedly down on fidelity, which is also consistent with the decoupling.

E LEAKAGE SWEEP

This appendix supports Sec. 4.1. The sweep runs five methods (Gaussian noise, SMOTE, mixup, AP-dropout, kriging) on four datasets (UJIIndoorLoc, Tampere, UCI, IndoCell Campus) at three label budgets and three seeds, under two protocols that differ only in the data the generator is fit on. Under the controlled protocol the generator sees only S_r; under the leaky protocol it sees all of D before S_r is subsampled. The localizer, test set, synthetic budget, and seeds are held fixed, so the difference between the protocols is attributable to the generator's information access.

Additive Gaussian noise perturbs each sample by its own class's estimated variance and carries no information about the task. Under control it is worth -0.000 pooled. Fit on the full pool, it reaches $+0.65$ on Tampere across three seeds, because the per-class means and variances it perturbs around are then estimated from labels the low-data condition should withhold. A generator that cannot add information thus appears to be the strongest method in the benchmark solely because of what it is allowed to see.

Two scope limits apply. The sweep is restricted to non-GPU methods, so the ratios are demonstrated on non-learned and rule-based generators; we do not claim that learned generators are inflated by the same factor, only that they depend on the same premise. The four datasets are a subset of the

Table 5: The 31 augmentation and generation methods benchmarked in this work, grouped by architecture family. The first 21 are drawn from the fingerprint literature; the last 10 are general tabular synthesizers and time-series augmentations. **Re-impl.** marks methods we implemented ourselves rather than using code released by their authors; the four methods marked "this work" are our own controls. Where a public implementation exists, the released code notes whether it was usable under our protocol.

Method	Venue/Yr	Type	Re-impl.	Core idea
Interpolation / geostatistical				
Kriging (Jan et al., 2015)	Sensors'15	Kriging (GP)	No	Per-AP semivariograms densify sparse RPs.
Adaptive path-loss (Wang et al., 2019)	Sensors'19	Path-loss interp.	No	Path-loss interpolation beats IDW/kriging.
MOGP (Tang et al., 2024)	Sensors'24	MOGP	No	Multi-output GP exploits cross-AP correlation.
Classical augmentation				
DataLoc+ (Hilal et al., 2021)	WCNC'21	AP-dropout	Yes	Dropout-inspired room-level AP dropping.
Improved RSSI (Sinha & Hwang, 2020)	Electron.'20	Temporal aug.	Yes	Single-RP mimicry of temporal fluctuation.
ProxyFAUG (Anagnostopoulos & Kalousis, 2021)	IPIN'21	Recomb. aug.	No	Proximity-based recombination of nearby scans.
Between-Location (Sugasaki & Shimosaka, 2021)	Sensors J.'21	Mixup + gen.	Yes	Mixup plus label-continuity correction.
SMOTE (Chawla et al., 2002)	JAIR'02	Interp. (1D)	No	Interpolation between same-class neighbors.
Mixup (Zhang et al., 2018)	ICLR'18	Interp. (1D)	No	Convex mix of two same-class samples.
Gaussian noise	this work	Noise aug.	—	Per-class perturbation; volume-matched control.
GAN-based				
Selective GAN (Njima et al., 2021)	Access'21	SS-GAN	Yes	Semi-supervised GAN with realism filter.
extendGAN+ (Yean et al., 2023)	Sensors'23	WGAN-GP	Yes	Transferred WGAN-GP, Dirichlet up-sampling.
SURIMI (Quezada-Gaibor et al., 2022)	IPIN'22	CNN-LSTM-GAN	Yes	Supervised fingerprint-database augmentation.
FinGAN (Zhang et al., 2025)	arXiv'25	InfoGAN	Yes	InfoGAN generation at unseen RPs.
Conditional GAN	this work	cGAN	—	Standard class-conditional generator pair.
VAE / graph-generative				
SVAE-WGAN (Liu et al., 2025)	Sci. Rep.'25	VAE + WGAN	Yes	Stacked VAE features plus attention denoising.
FALoc (Kim et al., 2025)	Electron.'25	VGAE	Yes	Graph auto-encoder over RP$\times$AP graph.
LCVAE (Wu et al., 2025)	IoT J.'25	cVAE	Yes	Dual-encoder cVAE with geospatial loss.
Conditional VAE	this work	cVAE	—	Standard class-conditional VAE.
Diffusion				
LocaGen (Abdelmotlb et al., 2025)	SIGSPAT.'25	Diffusion	Yes	Conditional diffusion at unseen locations.
Conditional DDPM	this work	Diffusion	—	Standard class-conditional denoising diffusion.
General tabular synthesizers (outside the fingerprint literature)				
CTGAN (Xu et al., 2019)	NeurIPS'19	Cond. GAN	No	Mode-specific normalization, training-by-sampling.
TVAE (Xu et al., 2019)	NeurIPS'19	VAE	No	Same transformer as CTGAN, VAE objective.
TabDDPM (Kotelnikov et al., 2023)	ICML'23	Diffusion	No	Gaussian + multinomial diffusion on mixed types.
TabDiff (Shi et al., 2025)	ICLR'25	Diffusion	No	Joint continuous-discrete diffusion, learned schedule.
ARF (Watson et al., 2023)	AISTATS'23	Random forest	No	Adversarial random forests with density estimation.
Gaussian copula (Patki et al., 2016)	DSAA'16	Copula	No	Marginals via empirical quantiles, Gaussian dependence.
Time-series augmentation (time-series partitions only)				
Jitter (Um et al., 2017)	ICMI'17	Noise aug.	No	Additive per-timestep Gaussian noise.
Scaling (Um et al., 2017)	ICMI'17	Amplitude aug.	No	Random global amplitude rescaling.
Window slicing (Um et al., 2017)	ICMI'17	Crop aug.	No	Random crop, resampled to full length.
Window warping (Um et al., 2017)	ICMI'17	Time warp	No	Random local speed-up/slow-down of a window.

twenty partitions, chosen to span both the paying and the null regime, so the pooled ratio estimates scale rather than a benchmark-wide constant.

F TRANSFER HEATMAP

Figure 2 is the per-cell view behind Sec. 4.2. Every aggregate in the main text is a mean over some slice of that grid, which shows how much of the benefit is concentrated in a few cells.

Table 6: Discrete minus continuous encoding, paired within each (dataset, budget, seed) cell; 153 paired runs per method. Δ is the change in mean kNN accuracy gain; the C2ST columns give mean real-versus-synthetic discriminability under each arm. The Gaussian copula has no such switch—it inverts through the empirical quantile function and is on-grid by construction—so its two arms agreeing to 0.0000 is a consistency check on the harness.

Method	**Δ gain (disc. $-$ cont.)**	**C2ST cont.**	**C2ST disc.**
TVAE	$+0.0005$	0.897	0.975
Gaussian copula	$+0.0000$	0.834	0.834
ARF	-0.0013	0.881	0.816
TabDiff	-0.0036	0.980	0.937
TabDDPM	-0.0038	0.980	0.935
CTGAN	-0.0186	1.000	0.971

Table 7: Mean accuracy gain under the two protocols, pooled over five methods, four datasets, and three seeds. The inflation shrinks as the subsample approaches the pool; at $r = 1.00$ the protocols coincide by construction.

Label budget	**Controlled**	**Leaky**	**Inflation**
$r = 0.01$	$+0.0202$	$+0.2489$	$12.3\times$
$r = 0.05$	$+0.0104$	$+0.0870$	$8.4\times$
$r = 0.10$	$+0.0075$	$+0.0563$	$7.5\times$

Three patterns are visible in Fig. 2. First, the benefit is concentrated: the Tampere and UJIIndoorLoc rows hold nearly every cell above $+0.05$, and the sparse-cellular rows are almost uniformly at the display threshold. Second, the negative cells are systematic: ProxyFAUG is negative across nearly the entire matrix, and Improved-RSSI and MOGP are negative on the outdoor cellular row, which the pooled Wilcoxon detects as significant harm. Third, the tabular synthesizers are volatile: CTGAN and TVAE produce both the largest positive cell (TVAE, $+0.19$ on Tampere) and some of the largest negative ones (CTGAN, -0.16 and -0.14 on UCI), a range no fingerprint-native method matches.

G Benefit Decay with Label Budget

This appendix supports the scarcity claim in Sec. 4.2. If augmentation substituted for survey effort in general, its benefit would persist as real data accumulates; if it only regularizes a boundary that few samples leave underdetermined, the benefit should vanish once enough real samples are available. We observe the second, as the information bound of Sec. 3.3 predicts. Figure 1b plots the curves; Table 8 pools the two non-RF halves, which that figure keeps apart because they follow different paths to similarly small values.

Table 8: Mean change in kNN accuracy by label budget, over every method applicable to each family. The Wi-Fi effect is concentrated at the smallest budget: by $r = 0.05$ it is an order of magnitude smaller, and from $r = 0.25$ it is negative. The aliased partitions show no benefit at any budget.

Family	$r = 0.01$	$r = 0.05$	$r = 0.10$	$r = 0.25$	$r = 0.50$
Wi-Fi	$+0.0100$	$+0.0007$	$+0.0003$	-0.0019	-0.0009
Aliased cellular	-0.0022	-0.0028	-0.0024	-0.0005	-0.0004
POWDER	-0.0078	-0.0065	-0.0026	-0.0004	$+0.0005$
Non-RF controls	$+0.0052$	$+0.0022$	-0.0038	—	—

On Wi-Fi, the mean gain falls by an order of magnitude between $r = 0.01$ and $r = 0.05$ and is at or below zero at every larger budget, so the benefit is a small-sample effect rather than a substitute for survey effort. On the aliased cellular partitions the curve is flat and slightly negative at every budget, including the smallest. This rules out the explanation that the cellular partitions would behave like Wi-Fi if scarcity were matched: they show no effect at the budget where Wi-Fi shows its largest.

POWDER converges to the same floor from below, starting as the worst curve at $r = 0.01$ and meeting the aliased partitions by $r = 0.10$. The non-RF controls were run only at the three budgets shared by the later method additions, so their curves stop at $r = 0.10$; over that range dense time series and voxel occupancy cross rather than move together, which is why the figure separates them.

H Downstream-Model Diversity

This appendix supports the downstream-model result in Sec. 4.2: the sign of the augmentation effect depends on the classifier, and on the Wi-Fi partitions it reverses.

Table 9: Mean accuracy change from augmentation at $r = 0.01$ under the two localizers, by structural regime. The left pair pools every method applicable to the regime; the right pair restricts to the 21 fingerprint-literature methods, showing that the reversal is not driven by the generic baselines. Every MLP cell is negative.

	All methods		**Fingerprint methods**	
Regime	**kNN**	**MLP**	**kNN**	**MLP**
Wi-Fi (5 partitions)	$+0.0100$	-0.0570	$+0.0114$	-0.0511
Aliased cellular (10)	-0.0022	-0.0181	-0.0013	-0.0159
POWDER (1)	-0.0078	-0.0550	-0.0076	-0.0558
Dense time series (3)	$+0.0070$	-0.0495	$+0.0112$	-0.0214
Voxel occupancy (1)	-0.0022	-0.0504	-0.0025	-0.0177

A 1-NN classifier is essentially unaffected by a synthetic point beside an existing same-class real point, since it changes no decision outside a region it already agreed with. An MLP fits a global decision boundary, so the same synthetic mass reweights the training distribution, and individually harmless points can move the boundary collectively. This is consistent with the observation that the methods with the largest kNN gains produce convex combinations of same-class real points.

Two explanations remain open: augmentation may help only nearest-neighbor localizers on these representations, or the MLP may be under-regularized at these training sizes, with synthetic data interacting with that rather than with the task. Distinguishing them requires a downstream-model sweep that varies capacity and regularization, which we did not run.

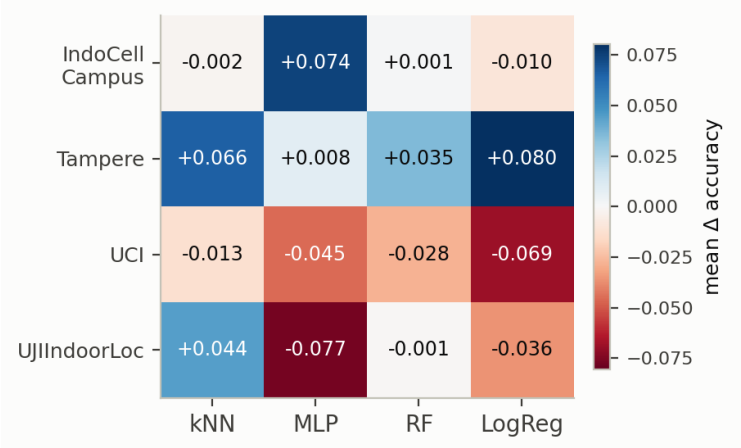


Figure 5: Mean accuracy change from augmentation at $r = 0.01$ for a fixed six-method subset on four datasets, by supervised downstream model. Each classifier is trained on the synthetic set's generator-assigned labels, as elsewhere in the benchmark.

Extending to random forest and logistic regression does not resolve the instability. UCI is negative under every model, and IndoCell's Campus partition is within noise of zero under three of four, with a single positive MLP result ($+0.074$) that does not replicate under RF or LogReg. Only Tampere is positive under all four.

We also ran a semi-supervised sweep, not plotted. Label propagation and label spreading build a k-nearest-neighbor graph over real-labeled and synthetic-unlabeled points and infer labels for the synthetic points from the graph; self-training accepts only pseudo-labels that a logistic-regression base classifier assigns with confidence above 0.75. Inferring labels is substantially worse than using the generator's labels on UCI and UJIIndoorLoc (label spreading -0.280 and -0.161; label propagation -0.208 and -0.166), better on IndoCell Campus and Tampere ($+0.015$ to $+0.030$ for all three models on both), and self-training is least damaging overall (-0.017 mean against -0.081 and -0.100). The conditional VAE and conditional GAN are worst under every semi-supervised model (-0.10 mean), consistent with the contamination diagnostic: because their points already lie near another class's real neighbor, inferring labels from local density lets contamination overwrite the correct label. Self-training's confidence threshold partially guards against this.

I Full Method Ranking

Table 10 is the full version of the ranking in Sec. 4.3, over all 31 methods and 20 base partitions. Its three kinds of columns answer different questions: per-family deltas show where a method helps, the pooled Wilcoxon shows whether its overall effect differs from zero, and the mean rank shows how it compares with other methods on shared partitions.

Eighteen of the 31 methods have a pooled effect that survives Benjamini–Hochberg correction at FDR 0.05, but 13 of those 18 are negative, so the benchmark detects harm more often than benefit. Only SMOTE, mixup, Between-Location, and two of the four time-series augmentations are significantly positive. The learned fingerprint generators cluster around zero or below, and three of them (Selective-GAN, DDPM, SURIMI) are significantly negative despite looking harmless on per-family averages.

Three observations follow. First, the ordering is driven by the extremes: SMOTE and mixup lead, CTGAN trails, and the methods between them are separated by less than the Nemenyi critical difference. Second, family means and pooled means can disagree. The cVAE has the second-largest Wi-Fi delta but is not significantly different from zero pooled, because its Wi-Fi gains are concentrated on Tampere; a paper reporting only its best family would present it as a leading method. Third, provenance does not predict performance: five of the top ten methods were not designed for fingerprints, and neither were the bottom two.

Two coverage caveats apply to the rank column. The ten methods from outside the fingerprint literature were run only at $r \in \{0.01, 0.05, 0.10\}$, and the time-series augmentations apply only to the three time-series partitions, so ranking all 31 together would compare methods on different grids. We therefore rank only the 20 with complete coverage. The pooled column has no such restriction, since the Wilcoxon is paired within each (dataset, ratio, seed) cell.

J Time-Series Augmentation on a Held-Out Domain

On the three time-series partitions we additionally ran the four standard augmentations of Um et al. (2017), which assume the feature axis is time. On ElectricDevices, window slicing gains $+0.0919$ at $r = 0.01$ and $+0.0747$ averaged over budgets—the largest gain of any method on that dataset, and one of the few that survives the MLP ($+0.0573$)—and the within-family ordering tracks how much temporal structure each method uses: jitter ($+0.0004$), scaling ($+0.0200$), warping ($+0.0646$), slicing ($+0.0747$).

This ordering does not hold on the other two time-series datasets. At $r = 0.01$, window slicing is the *worst* of the four on both ECG5000 (-0.0319) and ETT (-0.0192), and window warping is negative on ETT (-0.0100). Only scaling is positive on all three ($+0.0268$, $+0.0094$, $+0.0192$ at $r = 0.01$). A likely explanation is that slicing and warping assume the discriminative signal tolerates translation and dilation, which holds for appliance-usage profiles but not for ECG morphology, where class depends on the position and width of a beat complex, nor for our ETT construction, where the label is a quantile of the window's target variable and cropping changes that quantile. Structure-aware augmentation helps only when its assumed structure matches the data, and a mismatched assumption is worse than none.

Table 10: All 31 methods. Δ columns are the mean change in kNN accuracy at $r = 0.01$ within each family (Wi-Fi: 5 partitions; POWDER: 1; aliased cellular: 10; non-RF: 4); "–" marks a family the method does not apply to. **Pooled** is the mean over every dataset, ratio, and seed the method ran on, with * marking a Wilcoxon effect significant at FDR 0.05 across the 31 methods. **Rank** is the mean within-dataset rank over the 20 methods with complete coverage of the shared partitions and budgets (1 = best); the 11 methods without complete coverage are unranked. • marks learned generative methods (GAN, VAE, or diffusion); † marks methods not designed for fingerprints.

Method	**Wi-Fi**	**POWDER**	**Cellular**	**non-RF**	**Pooled**	**Rank**
SMOTE †	$+0.033$	$+0.009$	$+0.004$	$+0.017$	$+0.0116^*$	3.53
Mixup †	$+0.034$	$+0.009$	$+0.003$	$+0.026$	$+0.0103^*$	5.78
ARF †	$+0.001$	$+0.000$	$+0.003$	-0.001	-0.0014	8.62
SVAE-WGAN •	-0.013	-0.005	-0.000	–	-0.0018^*	8.81
Gaussian noise	-0.001	$+0.002$	-0.000	$+0.001$	-0.0004^*	9.25
cVAE •	$+0.055$	-0.010	$+0.000$	$+0.007$	$+0.0053$	9.62
TabDiff †	$+0.008$	$+0.001$	$+0.002$	$+0.001$	-0.0011	9.62
LCVAE •	$+0.049$	$+0.012$	-0.003	–	$+0.0042$	9.94
extendGAN+ •	$+0.020$	$+0.001$	-0.001	–	$+0.0025$	10.03
TabDDPM †	$+0.008$	$+0.002$	$+0.002$	$+0.001$	-0.0012	10.28
Conditional GAN •	$+0.001$	-0.011	-0.001	-0.003	-0.0031^*	10.78
Selective GAN •	$+0.002$	$+0.002$	-0.001	–	-0.0011^*	10.81
AP-dropout	$+0.034$	-0.004	$+0.001$	-0.001	$+0.0017$	10.88
DDPM •	-0.007	-0.001	$+0.000$	–	-0.0015^*	11.72
Gaussian copula †	$+0.043$	$+0.003$	-0.004	$+0.004$	$+0.0047$	11.78
SURIMI •	-0.005	-0.019	-0.002	–	-0.0021^*	12.16
FALoc •	$+0.007$	-0.006	-0.004	–	-0.0007	12.56
Improved-RSSI	$+0.001$	-0.026	-0.007	–	-0.0036^*	12.59
TVAE † •	$+0.006$	-0.053	-0.005	-0.016	-0.0122^*	13.09
CTGAN † •	-0.033	-0.006	-0.028	$+0.002$	-0.0195^*	18.12
Incomplete coverage: unranked						
Window warping †	–	–	–	$+0.020$	$+0.0214^*$	—
Scaling †	–	–	–	$+0.019$	$+0.0143^*$	—
Window slicing †	–	–	–	$+0.014$	$+0.0162$	—
Jitter †	–	–	–	$+0.001$	$+0.0002$	—
Between-Location	$+0.038$	$+0.009$	-0.002	–	$+0.0103^*$	—
Adaptive path-loss	$+0.015$	-0.019	-0.000	–	$+0.0007$	—
Kriging	$+0.008$	$+0.003$	-0.002	–	$+0.0016$	—
FinGAN •	$+0.013$	-0.007	-0.001	–	-0.0052^*	—
LocaGen •	-0.006	-0.024	-0.003	–	-0.0033^*	—
MOGP	-0.016	-0.019	-0.004	–	-0.0078^*	—
ProxyFAUG	-0.036	-0.055	-0.017	–	-0.0188^*	—

This bears on the ADI's calibration. Generic tabular methods hurt ElectricDevices (-0.0053 at $r = 0.01$, better-arm basis) while a well-matched augmentation gains nearly 0.1, so the achievable benefit depends on the method as well as the representation. A screen that reads only feature-space statistics cannot capture this, which is one reason we claim the ADI for the direction of the effect rather than its size (Sec. 5.3).

K Failure-Mode Diagnostics

Sec. 4.3 and Fig. 3a report where synthetic points land relative to the real manifold. This appendix gives a descriptive, family-by-family account: each generator family fails in a characteristic way, and averaging over all "deep generative methods" would obscure the differences.

GANs impose no class-boundary constraint through the discriminator, and their generators can partially mode-collapse in ways that blur adjacent classes (diversity ratio 0.15 for the conditional GAN and 0.13 for Selective-GAN, both far below 1.0). The VAE family contaminates 0.90 to 0.96 of synthetic points despite conditioning on the correct label. The diffusion methods (DDPM and LocaGen) over-disperse, generating points far from any real sample (novelty 79.1 and 79.0, roughly an order

of magnitude above every other method) with a diversity ratio near 10, consistent with structure the few-shot training set cannot constrain.

L Latent-Geometry Instrumentation

Sec. 4.3 reports class contamination as measured and its mechanism as open. The standard account for the VAE family appeals to latent geometry: a collapsed posterior, or a decoder that stretches sharply across class boundaries, would place decoded samples in the wrong class. We measured each link in that account.

We instrumented four VAE-family models (our conditional VAE, LCVAE, SVAE-WGAN, FALoc) with a uniform probe returning the posterior mean and log-variance and a decoder closure, across seven partitions and three seeds. SVAE-WGAN returns a non-finite posterior in 17 of 21 runs, consistent with its divergence (Sec. 4.3); these runs are recorded as diverged rather than dropped, leaving 67 valid runs.

None of the links holds in the 67 valid runs. *Posterior collapse does not occur*: every latent dimension is active in all 67 runs. *The decoder is not preferentially stretched at class boundaries*: its Jacobian norm exceeds the interior value in only 10 of the 43 runs where both are measurable, with per-model mean ratios between 0.26 and 0.68. *The encoder does not destroy class structure*: 5-NN purity in latent space is slightly higher than in input space (0.346 against 0.334). Finally, decoding the midpoint between two different-class latent codes lands outside both endpoint classes 0.441 of the time, against 0.475 for a same-class control, a non-significant difference in the opposite direction to the hypothesis.

The contamination measured in Sec. 4.3 is real, but these measurements do not support the latent-geometry mechanism commonly offered for it, so we leave the mechanism open. The probe covers only the VAE family, and a negative result over 67 runs bounds the size of an effect rather than excluding one.

M Fidelity versus Utility

Distributional fidelity is the metric this literature most often reports when downstream accuracy is unavailable. On our data it does not predict utility (Fig. 3b).

The methods with the largest accuracy gains are among the most easily distinguished from real data. SMOTE and mixup reach C2ST 0.77 and 0.84 pooled across budgets while carrying two of the best pooled effects, and their synthetic points lie exactly on segments between same-class samples—a signature a discriminator detects and the placement that reinforces the correct boundary downstream. Conversely, several methods with more realistic output deliver no benefit: FALoc reaches C2ST 0.72 pooled, the lowest in the benchmark, with a pooled effect of -0.001.

The same pattern holds at the family level. At $r = 0.01$, averaging one point per method, the tabular group is more realistic than the fingerprint-native group (mean C2ST 0.929 against 0.966), while its mean accuracy effect is negative where the fingerprint group's is not (-0.002 against $+0.001$).

A discriminator is rewarded for finding any systematic difference between real and synthetic points, including differences that make the synthetic set useful; interpolation deliberately places mass where no real sample lies, which is both easy to detect and the source of its benefit. Fidelity measures whether a generator has reproduced the data distribution, while utility measures whether it has reshaped a decision boundary, and our results show these objectives can conflict. Appendix D shows the same decoupling within single methods.

N Spatial-Holdout Results

The label-budget protocol of Sec. 3.3 withholds samples at locations the generator has seen, which is a densification task. This appendix tests extrapolation: producing a fingerprint at a coordinate no training sample came from. We withhold entire k-means clusters of reference coordinates from generator fitting, query each method at the withheld coordinates, and score the reconstruction against

the real fingerprints there, in standardized feature space. We compare three conditions—*known* (the cluster was available during fitting), *heldout* (it was not), and *fewshot* (five real samples per withheld cluster are returned)—to separate degradation from partial recovery.

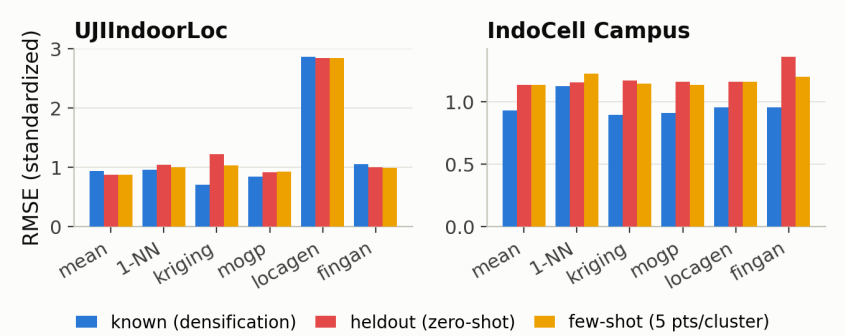


Figure 6: Fingerprint-reconstruction RMSE (standardized feature space) at externally specified query coordinates, under the three holdout conditions. Lower is better.

Kriging has the lowest RMSE on UJIIndoorLoc under the known condition (0.704) but degrades 73.7% under holdout, the most of any method, against 7.9% for MOGP; on IndoCell's Campus partition every coordinate-conditional method degrades 21–42%. The few-shot condition recovers 35.3% of kriging's degradation on UJIIndoorLoc but only 8.9% on IndoCell. On UJIIndoorLoc, the few-shot condition leaves MOGP's degradation unchanged (−13.5%, a mild worsening), because its single joint Gaussian process absorbs a handful of new points as noise rather than as a local correction. On IndoCell's Campus partition FinGAN recovers 39.4%, the largest fraction on either dataset, but starts from a degradation nearly three times kriging's and remains worse than every non-parametric baseline in absolute terms.

The comparison against a trivial baseline is the most informative. On IndoCell's Campus partition under holdout, the mean training fingerprint—a constant vector—reaches an RMSE of 1.138, and every coordinate-conditional method is worse: kriging 1.168, MOGP 1.161, LocaGen 1.161, FinGAN 1.359. Conditioning on the query coordinate carries no usable information in that representation, consistent with the sparsity wall, and a densification gain would not reveal this. On UJIIndoorLoc the constant reaches 0.882 and kriging under holdout 1.223, so the ordering holds there as well; only MOGP (0.916) approaches the constant.

O Positioning-Error Results

This appendix supports the claim in Sec. 4.4 that accuracy and positioning error diverge. Positioning error is evaluated only on the coordinate-carrying fraction of each pool, in metres, so the datasets here are a subset of the twenty partitions.

On UJIIndoorLoc, non-parametric methods reduce mean positioning error by 3 to nearly 14 metres relative to real-only training at $r = 0.01$. Tampere shows the sharpest divergence between targets: none of the largest accuracy gainers improves positioning error, AP-dropout adds 1.7 m, and mixup and SMOTE each gain $+0.05$ accuracy while adding about 5.7 m of error. MOGP is the only method on Tampere that improves both targets, by a slight $+0.02$. UTSIndoorLoc mirrors UJIIndoorLoc at smaller scale: non-parametric and geostatistical methods trim 0.3 to 1.9 m from a 13.2 m baseline, while FinGAN and LCVAE increase error. On SODIndoorLoc the same family nearly halves the 14.1 m baseline, by 6 to almost 8 metres, while LocaGen adds 26.9 m and FinGAN 13.2 m, the two worst positioning results in the benchmark, plausibly compounded by SODIndoorLoc's disjoint coordinate frames (Appendix B).

A result on one target is therefore not a weaker version of a result on the other; on Tampere the two targets disagree on the sign of the effect for every leading method. We report both wherever coordinates exist.

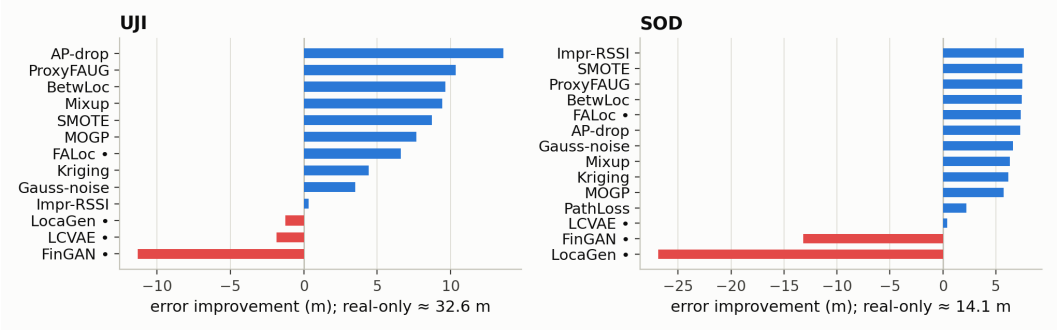


Figure 7: Positioning-error improvement over real-only training at $r = 0.01$ (metres), on UJIIndoorLoc and SODIndoorLoc. Positive is better. Coordinate-conditional learned methods are marked with •.

P Structural Correlations

This is the univariate screen behind the predictor choices in Sec. 5 and Appendix Q: each structural property, correlated alone against measured augmentation gain, over the 16 real fingerprint partitions and the 12 controlled UJIIndoorLoc sweep variants.

Table 11: Pearson correlation of eight structural properties with measured augmentation gain, over 28 real and controlled-sweep dataset variants.

Property	r	p
Feature vocabulary size	$+0.784$	5.0×10^{-7}
Effective rank	$+0.675$	6.0×10^{-5}
Intrinsic dim. (PCA)	$+0.521$	3.8×10^{-3}
Active APs per scan	$+0.364$	5.2×10^{-2}
Neighbor purity (5-NN)	$+0.320$	9.0×10^{-2}
Active fraction	-0.195	3.1×10^{-1}
Mean mutual information	-0.177	3.6×10^{-1}
Active-value entropy	-0.149	4.4×10^{-1}

Feature vocabulary size is the strongest univariate predictor, which motivated making it the manipulated variable in Sec. 5.1. Active fraction—the per-scan density that "sparse" usually denotes in this literature—has the wrong sign and is not significant, as is active-value entropy. This quantifies the distinction drawn in Sec. 1: the two modalities differ in vocabulary rather than per-scan density, and only vocabulary tracks the outcome.

These are univariate correlations over 28 points, several of which are variants of one building, so they identify candidates only. A linear model on all eight predictors reaches $R^2 = 0.870$, but with 28 points and 8 predictors it is likely overfit; the index we adopt uses three predictors chosen for mechanism (Sec. 5.3).

Q The Fingerprint-Only ADI Precursor, and Why It Does Not Transfer

The main text (Sec. 5.3) presents the inherent-property index. We arrived at it after fitting and testing a fingerprint-only precursor, reported here because its failure motivated the re-derivation.

Q.1 Fitting on fingerprints alone

We compute eight structural properties on the 16 real fingerprint partitions and the 12 controlled UJIIndoorLoc sweep variants of Sec. 5.1, and correlate each with the mean gain of a fixed method

subset (Appendix P). Rather than use the two best correlates, we fit a small model whose predictors correspond to mechanisms. Vocabulary size is the variable the sparsification experiment identified as causal, and it sets the collision rate: a small vocabulary forces distinct locations onto identical quantized vectors. Neighbor purity is a local form of reachability, measuring whether a same-class convex combination lands nearer its own class than another's. The precursor, a linear model on these two, reaches $R^2 = 0.664$ ($n = 28$) with both coefficients positive. Refit with the eight IndoCell partitions held out, it predicts $+0.0004$ against a measured -0.0001 (Fig. 8), a correct qualitative verdict; a near-zero prediction of a near-zero effect is not evidence of fine-grained accuracy.

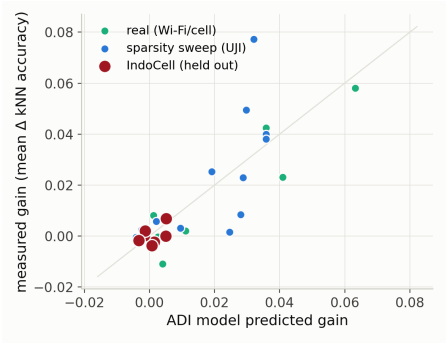


Figure 8: Predicted versus measured augmentation gain for the *precursor* index of this appendix, not the ADI adopted in Sec. 5.3 (whose leave-one-dataset-out behavior is Fig. 4b). The eight IndoCell partitions (red) were held out of fitting.

Q.2 Four pre-registered out-of-domain tests

We then tested the precursor outside the fingerprint domain four times, each time computing the dataset's structural properties and recording the prediction with a timestamp and a SHA-256 digest of the fitted coefficients *before any method was run on it* (Table 12). Three are dense time series chosen to vary column count and effective rank. The fourth, ModelNet10 voxel occupancy, was constructed to match the fingerprint regime on the quantities the index reads, so that it tests the index inside its intended regime.

Table 12: Four pre-registered out-of-domain tests of the precursor. Every prediction was recorded before any method ran on that dataset. "Meas." is the mean gain of the fixed method subset at $r = 0.01$ with a 95% CI over its 18 paired runs; "call" is the pre-registered binary verdict against the 0.005 threshold. "In regime" marks whether the dataset satisfies the sparsity condition the index is built for; the Wi-Fi partitions it was fit on are shown for reference. "PCA dim" is the participation-ratio dimension of Appendix P, not the TwoNN estimate used in Sec. 5.3.

Dataset	d	act.	PCA dim	pred.	meas. [95% CI]	sign	call	in regime
Dense: outside the regime the index is built for								
ElectricDevices (UCR)	96	0.98	54.9	$+0.0050$	$+0.0225$ [$+0.002$, $+0.044$]	✓	match	no
ECG5000 (UCR)	140	1.00	5.4	$+0.0119$	$+0.0073$ [-0.003, $+0.018$]	✓	match	no
ETT (load forecasting)	672	0.98	8.8	$+0.0427$	$+0.0036$ [-0.005, $+0.012$]	✓	**miss**	no
Sparse: inside the regime								
ModelNet10 (voxels)	512	0.21	23.5	$+0.0345$	-0.0025 [-0.015, $+0.010$]	×	**miss**	**yes**
(Wi-Fi partitions, fitted)	*444*	*0.24*	*33.0*	—	—	—	—	*yes*

Three of the four datasets are dense (98–100% of entries active) and fall outside the regime the index is built for. On dense data the vocabulary term reads raw column count, which is why it over-predicts ETT by a factor of twelve (672 columns that are 96 timesteps of 7 correlated channels, effective rank 19.5) while under-predicting the high-rank ElectricDevices. The precursor should not be applied to low-dimensional dense data.

ModelNet10 is the decisive test, because it is the one out-of-domain dataset inside the regime. At 21% active over 512 columns, with intrinsic dimension 23.5 and neighbor purity 0.87, it is structurally close to the Wi-Fi partitions the index was fit on, and it is sparse for the same reason a

fingerprint is: an empty voxel is unobserved occupancy, as an unheard AP is unobserved signal. The precursor predicts $+0.0345$; the measured gain is -0.0025, the wrong sign. This is not extrapolation, since at $d = 512$ the dataset lies 1.0 standard deviations above the fitted mean column count.

The subset mean conflates what a representation admits with what a given method extracts. Over the three shared budgets, the structure-respecting methods—SMOTE, mixup, and AP-dropout, each a convex combination or masking of same-class real points—gain $+0.0089$ on ModelNet10 and are significantly positive (Wilcoxon $p = 0.004$), while the learned generative methods lose -0.0076 ($p = 7 \times 10^{-6}$ for the difference). The pre-registered statistic averages both and understates what the representation admits. The generic tabular synthesizers are weak on both domains—CTGAN and TVAE average -0.0150 on Wi-Fi against -0.0072 here ($p = 0.89$)—so the domain-specific reversal is carried by the conditional VAE and GAN, which gain $+0.0124$ on Wi-Fi and lose -0.0132 on ModelNet10 ($p < 10^{-4}$).

Q.3 A CONTROLLED MODELNET10 SPARSIFICATION SWEEP

A single dataset cannot distinguish a mis-scaled index from one that reads the wrong variable. We therefore moved ModelNet10 along two knobs that the fingerprint suite cannot separate, holding objects, labels, and task fixed, and pre-registered a prediction for every cell (Table 13). The *density* knob subsamples the surface point cloud before voxelization, so occupancy falls from 0.213 to 0.052 while the feature count stays at 512. The *resolution* knob changes the grid from 8^3 to 4^3, so the feature count falls while the object does not.

Table 13: ModelNet10 sparsification sweep, 441 runs. Every prediction was registered before its cell was run. "Real" is real-only kNN accuracy at $r = 0.01$, shown because the density knob degrades the task as well as the representation; "best" is the strongest structure-respecting method. Predicted and measured values order the cells in opposite directions on both knobs.

Cell	d	occ.	purity	real	predicted	measured	best
Density knob: point cloud subsampled at source, d fixed at 512							
4096 points	512	0.213	0.87	0.557	$+0.0345$	-0.0025	$+0.0143$
512 points	512	0.195	0.87	0.557	$+0.0345$	$+0.0005$	$+0.0048$
256 points	512	0.169	0.85	0.539	$+0.0342$	$+0.0021$	$+0.0121$
128 points	512	0.129	0.76	0.431	$+0.0330$	$+0.0266$	$+0.0452$
64 points	512	0.086	0.60	0.330	$+0.0307$	$+0.0418$	$+0.0745$
32 points	512	0.052	0.43	0.238	$+0.0285$	$\mathbf{+0.0532}$	$+0.0863$
Resolution knob: full point budget							
4^3 grid	64	0.262	0.75	0.462	$+0.0047$	$+0.0095$	$+0.0224$
6^3 grid	216	0.238	0.85	0.535	$+0.0156$	$+0.0086$	$+0.0206$
8^3 grid	512	0.213	0.87	0.557	$+0.0345$	-0.0025	$+0.0143$

Density knob. With the feature count fixed at 512, only the neighbor-purity term can move the prediction, and its coefficient is the smaller of the two ($+0.0037$ against $+0.0165$). Predictions therefore span only 0.006 ($+0.0345$ to $+0.0285$), while measured gain rises by 0.056 (-0.0025 to $+0.0532$), 9.2 times more. The predicted spread is also below the standard error of a single cell mean (≈ 0.008 over 18 paired runs), so we do not report a rank correlation over these near-tied predictions; its sign flipped across three refits on different fingerprint subsets. Two manipulations rule out candidate explanations. First, no label information is destroyed: collision is 0.000 in every subsampled cell against 0.024 at full density, so the loss of purity here is not the cellular mechanism. Second, purity is not the operative variable: replacing uniform subsampling with farthest-point sampling recovers 0.16 of neighbor purity at matched occupancy (0.679 against 0.520) and leaves measured gain unchanged ($+0.0546$ against $+0.0532$).

Task difficulty. Every manipulation that degrades a representation also degrades the real-only baseline, which falls from 0.557 to 0.117 across the 14 cells and is the strongest single correlate of gain ($r = -0.70$, $p = 0.005$). A third sweep relocates a fraction of each object's occupied voxels to random empty positions, fixing feature count and occupancy; it raises gain to $+0.0395$ at 50% noise

with an interval excluding zero, but the effect is non-monotone, only one of five cells is individually significant, and the baseline falls there too. Controlling for the baseline reverses the apparent role of both noise and purity, and three reasonable adjustments disagree, so we do not attribute the inversion to any single variable. The result that does not depend on the adjustment is that the precursor's raw ordering of these representations is wrong, and neither quantity it reads survives direct manipulation.

Two causes of low purity. Purity lowered by *label collision* marks a representation where augmentation cannot help because the information is gone, as in the cellular regime (collision 0.380–0.850, Bayes floors up to 0.61). Purity lowered by *measurement noise* marks a representation where augmentation helps most. Every partition in the precursor's fit exhibits only the first cause, so it learned a single positive coefficient; the density knob supplies the second cause, and the coefficient points the wrong way. Refitting cannot repair this confound, which requires a separate collision term.

Resolution knob. The resolution knob does vary collision—0.024 at 8^3, 0.072 at 6^3, 0.338 at 4^3—and measured gain stays small throughout ($+0.0095$, $+0.0086$, -0.0025), with only the 4^3 interval excluding zero. We read this as consistent with the wall: once collision is present, the available benefit is small at every grid, and differences among the three cells are within the noise of an 18-run estimate. Across the whole sweep, structure-respecting augmentation beats the topology-indifferent generators in 8 of 8 cells ($+0.0311$ against $+0.0132$ pooled, $p = 0.007$). Structure-awareness governs what a method extracts from a representation; it does not correct the precursor's estimate of what is available.

Q.4 Rejected repairs

Before re-deriving the index, we tried two repairs of the precursor. Substituting effective rank for raw column count places all four pre-registered predictions inside their confidence intervals, but so does predicting a constant $+0.005$, at lower error, so passing this test shows only that the intervals are wide. Adding a task-difficulty term, motivated by the baseline correlation above, did not improve out-of-sample prediction over a constant in any of four specifications. Across the four pre-registered datasets the raw precursor is beaten by predicting no effect at all (mean absolute error 0.069 against 0.009).

These results, together with two conceptual problems, motivated re-deriving the index from properties of the data alone. Vocabulary size is not comparable across modalities (512 voxels and 512 access points are different kinds of axes), and neighbor purity summarizes only a single 5-NN neighborhood and depends on a model choice.

R Extended Limitations

Sec. 6 states the limitations that bear most on the main claims. This appendix gives the complete list, ordered from those that most constrain interpretation to those that constrain coverage.

1. The positive findings are nearest-neighbor findings. Under the MLP localizer the mean effect at $r = 0.01$ is negative in every regime, including Wi-Fi (-0.057 over all methods, -0.051 over fingerprint methods alone; Appendix H). We lead with kNN because it is the convention in this literature, but "augmentation helps on dense Wi-Fi" is conditional on the localizer. We have not separated a genuine interaction with nearest-neighbor decision rules from an under-regularized MLP at these training sizes.

2. Eight methods are reimplementations, and none was validated against its original paper. None of the eight reimplemented methods in Table 5 was checked against the accuracy reported by its authors on their own data, so for any single method, an underperforming reimplementation cannot be fully distinguished from an underperforming method. This does not extend to the aggregate cross-method and cross-modality comparisons, which would require all eight to be wrong in the same direction.

3. The ADI is a directional screen, and two regions of its input space are extrapolated. The adopted index uses three predictors—collision rate, log relational manifold sparsity, and log noise

level—and we claim it only for ordering (Sec. 5.3). Collision above 0.32 occurs only in IndoCell and the dense, low-rank corner is sparsely covered, so both are extrapolated. Leave-one-dataset-out Spearman is $+0.739$ with sign accuracy 0.70 against a constant's 0.48, at essentially the same absolute error as the constant. The binary calls are also underpowered: with 18 paired runs per dataset, no interval excludes the 0.005 threshold. A larger and more diverse set of held-out datasets is needed before treating the ADI as a general planning tool. The fingerprint-only precursor, its pre-registered tests, and two rejected repairs are reported in Appendix Q.

4. Several experiments are narrower than the full sweep. The leakage quantification and the downstream-model-diversity experiment each cover five or six methods on four datasets at one or three label budgets, rather than the full 6,537-run grid. The leakage subset is further restricted to non-GPU methods, so its ratios are demonstrated on non-learned and rule-based generators and make no claim about learned ones. The spatial-holdout evaluation covers the two datasets with enough reference points for a k-means cluster holdout, and the coordinate-carrying subset spans only nine datasets, limiting the power of the restricted Friedman–Nemenyi analysis.

5. Two sources carry little signal. SODIndoorLoc's building-floor task has only five classes and a real-only baseline near ceiling from $r = 0.05$ onward, so its classification results have little power, although its positioning-error results remain informative. SVAE-WGAN produced non-finite output in 200 of 210 runs and emitted no synthetic data in those runs, so its rows mostly measure abstention rather than a working generator (Sec. 4.3).

6. The cellular evidence comes from one operator in one city. IndoCell is collected from a single operator in a single city, and generalization to other operators, regions, and 5G-dominant deployments is untested. The sparsification experiment of Sec. 5.1 reproduces the cellular collision structure on a Wi-Fi building, so the mechanism does not rest on IndoCell alone, although the natural cellular measurements do.

7. The 31 methods are not a census. Recent systems we do not evaluate include DeepRSSI (Yoon et al., 2024) and semi-supervised GAN augmentation (Yoo, 2024), among several GAN and VAE methods. The main reason is data availability: much of this work reports results only on radio maps that were never released, so neither the training data nor the published numbers can be reproduced. We restricted the benchmark to methods specified precisely enough to reimplement and testable on public fingerprint data. The excluded methods are concentrated in the GAN and VAE families, so those families may be under-represented by their strongest recent entries.