Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Augmentation-induced sibling leakage in small-data surface imaging: measurement and a reporting protocol

Shaoxuan Zhang

2023113033@stu.hit.edu.cn

Harbin Institute of Technology

Research Article

Keywords: data leakage, evaluation protocol, data augmentation, surface imaging, scanning tunnelling microscopy, object detection

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11679740/v1

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

Additional Declarations: No competing interests reported.

Augmentation-induced sibling leakage in small-data surface imaging: measurement and a reporting protocol

Shaoxuan Zhang[1*]

[1]Faculty of Computing, Harbin Institute of Technology, Harbin, Heilongjiang, China
ORCID 0009-0001-4103-6099.

Corresponding author(s). E-mail(s): 2023113033@stu.hit.edu.cn;

Abstract

Deep learning papers on surface and microscopy imaging routinely build their training sets by augmenting a handful of acquisitions into hundreds or thousands of images, and then split the augmented pool at random. Every test image then has a near-copy in the training set, and the reported metric measures recognition rather than generalisation. We quantify the cost. Holding the annotation budget fixed and varying only the provenance of the training images, we compare a random split (A) with a split by acquisition (B) on two public benchmarks and two tasks: surface-defect detection with Faster R-CNN, and point-defect segmentation in scanning tunnelling microscopy with a U-Net. Random splitting overstates mAP@0.5:0.95 by 0.546 at two source images per class and by 0.303 at ten; on the STM data it overstates Dice by 0.278–0.448. At the smallest budgets the random split drives mAP@0.5 to 1.000 ± 0.000, a number that describes the test set rather than the model. The gap narrows as acquisitions increase; what declines is the benefit of leaking, not the quality of grouped evaluation. We propose a four-item reporting protocol and a sibling-leakage statistic that costs seconds to compute and whose error behaviour we characterise.

Keywords: data leakage, evaluation protocol, data augmentation, surface imaging, scanning tunnelling microscopy, object detection

1 Introduction

Consider a recent study of molecular recognition in scanning tunnelling microscopy (STM). Its training set was built by manually annotating "only a dozen of molecules" and expanding them with rotation, flipping, blur and colour jitter into "thousands" of images; the resulting detector reports mAP@0.5 of 0.996 [1]. As a description of the data the model was fitted to, the number is unremarkable. As an estimate of what the detector will do on the next scan, it cannot be an unbiased one, because its test set was drawn from the same augmented pool as its training set.

The mechanism is simple. When a dataset is built by augmenting a small number of acquisitions, and the augmented pool is then split at random, every test image has a sibling, another transformation of the same source, in the training set. The test set is no longer a sample of the distribution the model will meet; it is a sample of the distribution it has already met. A detector under these conditions is asked to recognise, not to generalise, and the metric records that.

The practice is widespread. In the thirty papers we coded from this area (STM, AFM, SEM, TEM and optical surface inspection) none describes a split at the level of acquisitions, none reports an evaluation on acquisitions independent of its training data, and twenty-three do not state whether augmentation preceded or followed the split. The constraint behind this silence is real: with three scans and two classes, holding out a scan leaves one scan per class. The sample also contains the counter-case. One study works with 25 images, splits before augmenting, and states its protocol in a single sentence.

What has been missing is a measurement. Leakage is documented as a threat to reproducibility in machine learning generally [2–5], and the audits that established this are convincing. They concern classification, and mostly duplicate records or preprocessing fitted on the full dataset. The channel we study (near-duplicates manufactured into the training set by augmentation) is a different mechanism with a different signature, and it dominates small-data imaging, where augmentation is not optional. The field also lacks a cheap way to check for it: the check costs one pass over the dataset.

We ask a narrow question and answer it experimentally. Holding the annotation budget fixed and varying only the provenance of the training images, how much does a random split over an augmented pool overstate performance, and does the overstatement shrink as more independent acquisitions become available?

Our answer, from two public benchmarks and two tasks:

- **The effect is large.** On a surface-defect detection benchmark the gap between a random split and a split by acquisition is 0.546 mAP@0.5:0.95 at two source images per class and still 0.303 at ten. On a public STM dataset of point defects the Dice gap is 0.278–0.448 for two to five scans.
- **It saturates the metric.** Under the random split, mAP@0.5 reaches 1.000 ± 0.000 on every partition at $k \leq 3$, and 0.998 ± 0.003 at $k = 5$. Perfect detection on a real defect dataset is a property of the test set, not of the model, and we propose treating it as a diagnostic signal.

- **It is not uniform.** Leakage inflates the classes that depend on the particular acquisition and leaves the texture-separable ones roughly intact, so a leaky protocol can make a benchmark look solved while part of it remains unsolved.
- **It does not close within the range that matters.** The gap narrows as acquisitions increase, but what narrows is the benefit of leaking; at ten acquisitions per class the random split is still worth thirty points.

These findings lead to a reporting protocol rather than an algorithm: report the number of independent acquisitions, the unit of the split, the order of augmentation and splitting, and a near-duplicate audit of the test set. Each of the four costs a sentence or a short script, and together they make the constraint visible instead of hiding it behind a random seed.

The contribution is fourfold: a definition and a cheap estimator of sibling leakage, including its error behaviour; a controlled measurement of the effect on two public benchmarks and two tasks with the annotation budget held fixed; evidence that the effect is class-dependent and that its attenuation comes from the collapse of the leaky protocol rather than from any improvement in grouped evaluation; and a minimum reporting standard motivated by a survey of current practice.

As these pipelines move from offline analysis towards closed-loop control of experiments, the reported metric stops being a summary and becomes a signal that decides what to do next. A metric that measures recognition of the training set is a poor input to that decision.

2 Related work

Leakage in ML-based science. That a model evaluated on data related to its training set will appear better than it is has been known since the data-mining literature, where Kaufman et al. formalised leakage and gave a taxonomy of its sources [3]. The issue returned to prominence with the COVID-19 imaging literature, where Roberts et al. showed that essentially every published chest-radiograph model suffered from some form of leakage or bias, including duplicate images shared between public datasets [4], and with Kapoor and Narayanan's audit of seventeen fields, which identifies leakage between train and test as the single most common threat to reproducibility in ML-based science [2]. Varoquaux and Cheplygina reach the same conclusion from the medical-imaging side, emphasising that methodological failures, not model capacity, dominate reported performance [5].

These audits concentrate on classification. Detection and segmentation introduce a leakage channel that classification lacks: a model is rewarded for localising an object it has already seen in a near-identical form, and the metric responds strongly to that. The leakage they describe arises mostly from duplicate records or from preprocessing fitted on the full dataset. The channel we study here (augmentation applied before splitting, so that near-duplicates of test images are manufactured into the training set) is a distinct mechanism, and it is the dominant one in small-data imaging, where augmentation is not optional.

Deep learning on surface and microscopy images. Automatic analysis of scanning probe and electron micrographs has moved in a few years from hand-crafted

features to off-the-shelf detectors. These have been applied to keypoint localisation in STM images [1, 6], defect classification in MoS_2 [7] and layer segmentation in SEM [8], and AFM pipelines have been trained on simulated images and evaluated on real ones [9]. What these studies share is a training set built from a handful of acquisitions and expanded by rotation, flipping and photometric jitter into hundreds or thousands of images. Reported metrics in this literature are correspondingly high (mAP@0.5 above 0.99 appears in more than one study), and we found no measurement of how far such numbers are inflated by the split.

Our own survey of thirty papers in this area (Section 5.2) finds that none reports a split at the level of acquisitions, and twenty-three do not state whether augmentation preceded or followed the split. This is the practice our experiments are designed to evaluate.

Grouped evaluation. Splitting by group is long established outside computer vision: blocked, hierarchical and leave-one-site-out cross-validation are standard in ecology, geostatistics and clinical prediction, precisely because observations sharing a site, a patient or a session are not independent [10]. Within imaging, site-wise splits are routine in multi-centre medical studies. That convention has not reached small-data materials imaging, where the "site" is the acquisition and the number of sites is often in the single digits. Our contribution is to make that transfer explicit, to measure what its absence costs on two public benchmarks, and to give a statistic cheap enough to apply routinely.

Relationship to a companion preprint. A preprint by the same author applies leave-one-group-out evaluation to magnetic-tile defect segmentation, as a robustness check for parameter-efficient fine-tuning, and frames grouped splits as one response to leakage [11]. That work takes the grouped protocol as given and asks where low-rank capacity should be placed. The present study takes the protocol itself as the object of study: it varies the number of acquisitions, compares the two protocols at an equal annotation budget, and measures what the random split costs. The two share a dataset and a motivation but address different questions, and no result reported here is taken from that preprint.

3 Method

3.1 Source-level partitioning and the two evaluation protocols

We write S for the set of source images (the original micrographs as acquired) and treat every augmented copy as a descendant of exactly one source. This distinction is the object of study: a split that is random at the level of images is not random at the level of sources.

Data. We use NEU-Seg, a public surface-defect dataset of 840 images of 200×200 px annotated with pixel-level masks for three defect classes [12]. Masks were converted to axis-aligned boxes by connected-component analysis, discarding components smaller than 25 px; each image contributes 3.2 boxes on average (range 1–19). We use a public dataset and a detector that fits in 8 GB of GPU memory because the phenomenon is a property of the data-construction protocol, independent of any particular model or instrument.

Variants. Each source image yields 80 deterministic variants: eight dihedral transforms (four rotations composed with a mirror) combined with ten photometric transforms (five gain factors between 0.85 and 1.15 and five brightness offsets between -12 and $+12$ grey levels). Boxes are transformed together with the image. The photometric stage leaves geometry unchanged, so the dihedral composition is the only source of geometric variation. Variants are regenerated on the fly from the source image; no augmented copy is stored.

Fixed annotation budget. Every configuration trains on 240 images for 30 epochs and is evaluated on the same 120 images. Only the provenance of the training images changes; the number of images, the optimiser, the schedule and the evaluation set are held fixed. This budget is a small fraction of what the NEU-Seg training split could supply: the regime of interest is the one in which only a few acquisitions are available.

Protocol A (pool-level split). Variants from all k sources per class are pooled, and 240 training and 120 test images are drawn uniformly at random, ignoring which source each variant descends from. A test image may therefore have a sibling, another variant of the same source image, in the training set. This is the protocol used by most small-data microscopy papers we surveyed.

Protocol B (source-level split). One source per class is reserved before any sampling takes place. Training images are drawn only from the remaining $k - 1$ sources; the reserved sources supply the test images and are never seen during training. Protocol B measures what protocol A cannot: performance on an acquisition the model has not encountered. Fig. 1 shows the two constructions side by side.

Three independent partitions. The whole construction, including which sources are reserved, is repeated with three random partitions, and reported values are means over partitions. This tests whether the effect depends on a particular choice of held-out images rather than on the protocol itself.

Fixed test set across *k*. Within a partition the test set is identical at every k. Differences across k are attributable to the number of distinct sources alone, not to a changing evaluation set.

The official NEU-Seg split is not our object of study: we use the dataset as a substrate and impose the k-source construction above, so that the two protocols can be compared at an equal annotation budget.

Training. We use Faster R-CNN with a ResNet-50-FPN-v2 backbone initialised from COCO weights [13, 14], an input resolution of 256 px, SGD with learning rate 0.005, momentum 0.9 and weight decay 5×10^{-4}, batch size 4, 30 epochs, step decay at 60% and 85% of the schedule, and gradient clipping at norm 5. mAP follows COCO's all-point interpolation; we report mAP@0.5 and mAP@0.5:0.95, together with per-class AP. Every run completed on a single 8 GB consumer GPU in 15–21 minutes.

Scope of the comparison. At $k = 2$, protocol B trains on one source per class whereas protocol A trains on two, so the smallest budget is not matched in training-source count. At $k \geq 5$ the difference is at most one source in five, and we base our conclusions on $k = 5$ and $k = 10$.

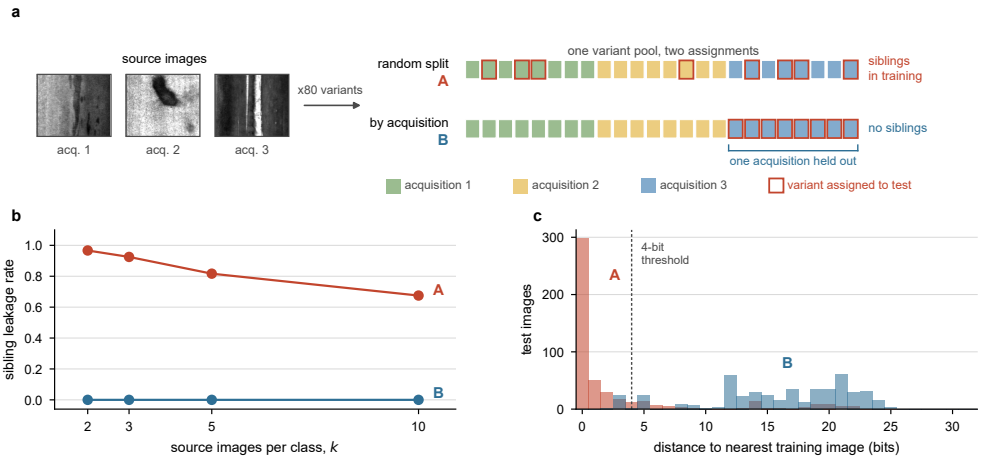


Fig. 1 The construction, and a statistic that measures it without training. **a**, Three acquisitions (NEU-Seg images) are expanded into 80 deterministic variants each and pooled. Red outlines mark the variants assigned to test: under a random split (A) they are scattered across acquisitions, so every test variant has siblings in training; under a split by acquisition (B) one acquisition is held out whole. **b**, Sibling leakage rate, the fraction of test images whose nearest training neighbour shares an acquisition, for both protocols as a function of the number of sources per class k. **c**, Distribution of the 64-bit dHash distance from each test image to its nearest training image, pooled over k; the dashed line marks the 4-bit threshold.

3.2 A sibling-leakage statistic and a minimum reporting protocol

Protocols A and B differ in one respect that can be measured directly, without training a model. We call two images *siblings* if they descend from the same acquisition: the same scan, the same specimen, or the same tiling operation applied to one parent image. A test set is compromised to the extent that its members have siblings in the training set, and this can be estimated from the images alone.

Definition. Let T and S be the test and training sets and let $d(\cdot,\cdot)$ be a distance between image descriptors. For each $t \in T$ let $s\backslash\{\}(t) = arg\ min_\{s \in S\}\ d(t, s^*)$ be its nearest training neighbour. The sibling leakage rate is

$SLR(T, S) = (1/—T—)\ \Sigma_\{t \in T\}\ 1[sib(t, s\backslash\{\}(t^*))]$,

the fraction of test images whose closest training image shares an acquisition with them. SLR is 1 when every test image has a near-copy in training and 0 when no test image does; a random split over an augmented pool drives it towards 1, and a grouped split drives it to 0. The informative case is a split that is meant to be grouped and is not (crops from one scan assigned to both sides), and that is the case SLR detects.

Estimator. We use a 64-bit difference hash (dHash) computed on the greyscale image resized to 9×8, with Hamming distance as d. The estimator is approximate; its error behaviour is characterised below. On NEU-Seg, a 4-bit threshold gives an SLR between 0.68 and 0.97 for the leaky construction, depending on k, and at most 0.01 for every grouped split. The estimate is stable from 0 to 4 bits: the leaky construction stays above 0.42 and the grouped splits below 0.01, so the separation does not depend on where in that range the threshold sits. Above 6 bits, distinct acquisitions that

Table 1 Detection performance on NEU-Seg (mean ± s.d. over three partitions; 240 training images, 120 test images).

k	A mAP@0.5	B mAP@0.5	gap	A mAP@0.5:0.95	B mAP@0.5:0.95	gap
2	1.000 ± 0.000	0.811 ± 0.044	+0.189	0.981 ± 0.006	0.435 ± 0.074	+0.546
3	1.000 ± 0.000	0.817 ± 0.074	+0.183	0.955 ± 0.013	0.411 ± 0.104	+0.544
5	0.998 ± 0.003	0.842 ± 0.054	+0.156	0.910 ± 0.019	0.455 ± 0.072	+0.455
10	0.962 ± 0.015	0.844 ± 0.047	+0.118	0.800 ± 0.047	0.497 ± 0.090	+0.303

happen to look alike fall inside the threshold and the grouped estimate rises to 0.11 at 6 bits and 0.42 at 12 bits for $k = 10$; a non-nested pilot showed the same residue, up to 17%, which is why a nonzero SLR computed on an undocumented split is an upper bound. We report the median nearest-neighbour distance alongside the rate: 0–1 bits for the leaky construction, 14–19 bits for grouped splits. The statistic costs one pass over the dataset, seconds for a few hundred images.

Protocol. When the number of acquisitions permits, the split is drawn at the level of acquisitions, and the number of acquisitions per class is reported in place of the number of augmented images. When it does not permit (three scans and two classes, say), both numbers are reported: performance under a grouped split, however unstable, and performance under a random split, labelled as an upper bound. We take the second route in Section 4.2, where 14 scans do not support a grouped split at $k > 5$ and the grouped estimate is reported with its variance.

Near-duplicate auditing is used to clean datasets before training: a recent SEM classification study removed 332 exact and 95 near-duplicate images before freezing its split [15]. What remains unmeasured is the cost of leaving the leak in place: how far a random split over an augmented pool inflates the reported metric, and how that inflation scales with the number of acquisitions. Section 4 measures exactly that, and the statistic above is what predicts it.

4 Results

4.1 A surface-defect benchmark under a fixed annotation budget

We first isolate the effect of provenance on a single benchmark, holding everything else constant. NEU-Seg supplies three defect classes with pixel-level annotations; masks are converted to boxes, giving 3.2 instances per image on average. Each configuration trains a Faster R-CNN on 240 images for 30 epochs and is evaluated on the same 120 images. Protocol A draws its 240 training images uniformly from the pooled variants of all k sources per class, so a test image may have a sibling in the training set. Protocol B reserves one source per class before sampling and trains only on the remaining $k - 1$. Three independent partitions are drawn, each re-selecting the reserved sources; all values are means over partitions.

Fig. 2 plots the same gaps as a function of k.

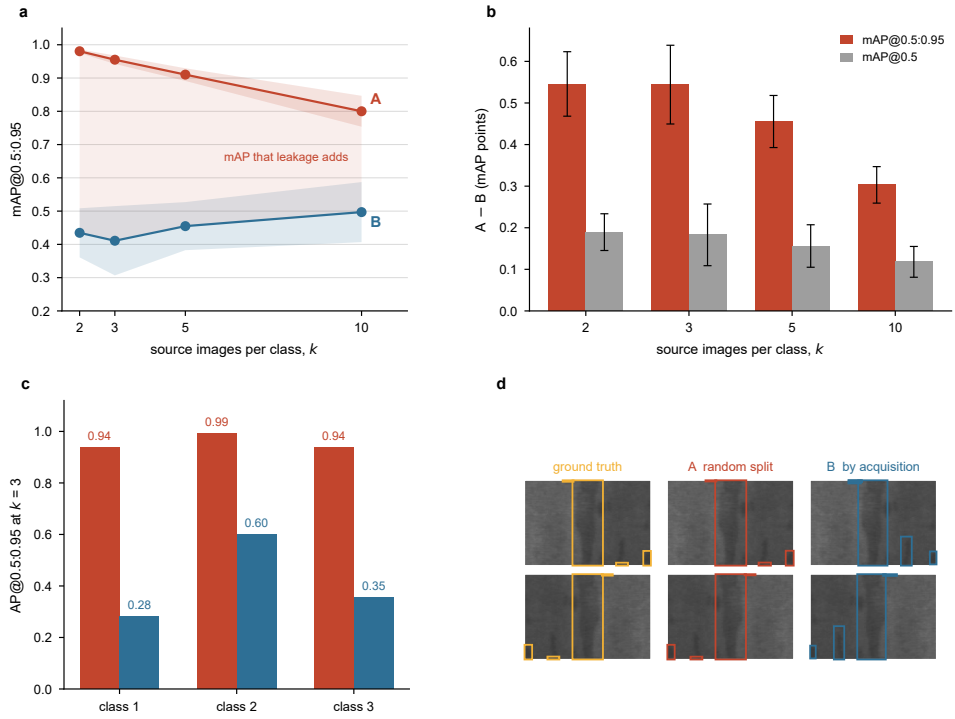


Fig. 2 Random splitting overstates detection on a surface-defect benchmark. **a**, mAP@0.5:0.95 on the fixed 120-image test set as a function of the number of source images per class, for a random split (A) and a split by acquisition (B); shading is s.d. over three independent partitions and the shaded band between the curves is the performance that leakage adds. **b**, The same difference at both IoU thresholds. **c**, Per-class AP@0.5:0.95 at $k = 3$: leakage inflates the acquisition-dependent classes and leaves the texture-separable one intact. **d**, Detection examples, cropped around the largest annotated defect; boxes are ground truth (yellow), protocol A (red) and protocol B (blue). The crops come from the reserved test sources, one of which also appears in Fig. 1a.

Protocol A reaches a perfect score. At $k \leq 3$, mAP@0.5 is 1.000 ± 0.000 on every partition, and 0.998 ± 0.003 at $k = 5$: essentially every ground-truth box is detected at IoU 0.5 with no false positive ranked above it. A perfect score on a real defect dataset describes the test set. Saturation at 1.000 is a diagnostic symptom: a model that reaches IoU 0.5 saturation on a visually heterogeneous task should prompt an audit of the split before it prompts a claim about the model.

The gap narrows with source diversity and does not close. At the strict threshold it falls from 0.546 to 0.303 as k grows from 2 to 10; at IoU 0.5 from 0.189 to 0.118. Both sequences are monotone. At $k = 10$, which is ten independent source images per class and more than this literature typically reports, a random split is still worth 30 mAP points at IoU 0.5:0.95. No threshold below which a random split becomes safe appears in this range.

Protocol A carries the narrowing; protocol B stays within its standard deviation. Protocol A drops from 0.981 to 0.800 as k increases, a change far outside the error bars. Protocol B moves from 0.435 to 0.497 over the same range, a change of the same magnitude as its standard deviation (0.07–0.10). The slope of the gap against $\log_2 k$ is -0.110 ± 0.036 per doubling at IoU 0.5:0.95 and -0.032 ± 0.043 at IoU 0.5, while protocol B's own slope is $+0.032 \pm 0.044$ and $+0.016 \pm 0.043$ (mean $\pm$ s.d. over the three partitions). At three partitions the grouped slope is not distinguishable from zero. More source diversity does not make grouped evaluation better; it removes the advantage the random split was exploiting.

The effect is not uniform across classes. One of the three defect classes retains AP@0.5:0.95 of 0.60–0.68 under protocol B, while the other two fall to 0.28 and 0.35 at $k = 3$. The class that transfers is also the one with the highest AP under protocol A (0.99), i.e. the easiest to recognise from texture alone. Leakage does not scale a model's performance uniformly: it inflates the classes that depend on the specific acquisition and leaves the texture-driven classes roughly intact.

The smallest budget is not matched in training-source count: at $k = 2$, protocol B trains on one source per class and protocol A on two; at $k = 5$ and $k = 10$ the difference is one source in five and one in ten, and our conclusions rest on the larger budgets. The three partitions vary the training data and the reserved sources together, so the reported standard deviations include partition-to-partition variation in task difficulty as well as training noise.

4.2 A scanning-tunnelling dataset with real scan-level structure

The surface-defect benchmark in Section 4.1 is a proxy: its "sources" are images, so the grouping used by protocol B is a convention we impose rather than a property of the data. To test whether the effect survives when the grouping is real, we repeat the experiment on a public STM dataset of point defects on Si(001)-H-AsH_3, Ge(001)-AsH_3 and TiO_2(110) [16]. It contains 14 scans acquired on different days from different samples and three substrates, each with a binary defect mask. Each scan is a natural source: its acquisition parameters, tip state and surface preparation are held constant within the scan and vary across scans.

We tile every scan into 64×64 patches with a stride of 32, giving 225 patches per 512×512 scan and 961 per 1024×1024 scan. Patches are the variants; scans are the sources. Three scans, one per substrate, are reserved for testing and contribute 400 test patches; the remaining 400 training patches are drawn from k scans. Protocol A samples its training patches from the k training scans *and* the three test scans; protocol B samples only from the k training scans. Everything else is held fixed: a four-level U-Net [17], BCE plus Dice loss, 30 epochs, 400 patches. We repeat the whole construction, including which scans are reserved, with three random partitions.

IoU follows the same ordering (A: 0.645, 0.623, 0.571; B: 0.224, 0.191, 0.291). Fig. 3 shows the Dice curves and the per-scan predictions.

The effect survives the change of task (detection $\rightarrow$ segmentation), of model (Faster R-CNN $\rightarrow$ U-Net) and of imaging modality (optical surface inspection $\rightarrow$ scanning tunnelling microscopy), with a gap of 28–45 Dice points, the same order of magnitude as the 30–55 mAP points measured there. The grouping is now physically defined, so

Table 2 Dice and IoU on the 400 held-out patches (mean ± s.d. over three partitions).

k	A: random patch split	B: held-out scans	gap
2	0.783 ± 0.032	0.360 ± 0.106	+0.424
3	0.767 ± 0.010	0.320 ± 0.039	+0.448
5	0.725 ± 0.050	0.447 ± 0.070	+0.278

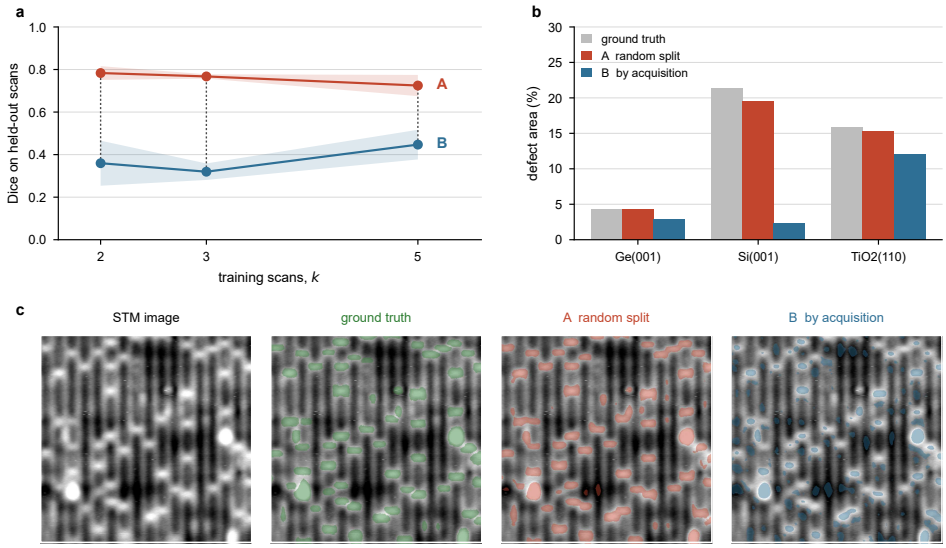


Fig. 3 The same effect on real scanning-tunnelling data. **a**, Dice on the 400 held-out patches as a function of the number of training scans, means over three independent partitions with s.d. shading. **b**, Defect area predicted by each protocol on the three held-out scans, against the annotated area. **c**, A held-out TiO2(110) scan with the annotation and the two predictions overlaid.

protocol B cannot be read as an artefact of how we defined a source. The two protocols also respond differently: in Section 4.1 the gap narrowed almost entirely because A fell, whereas here B rises from 0.320 to 0.447 between $k = 3$ and $k = 5$. Fitting the Dice gap against $\log_2 k$ gives -0.116 ± 0.086 per doubling here, and $+0.071 \pm 0.082$ for protocol B; at three partitions neither estimate is resolved. On real scans, additional sources do improve cross-scan generalisation; they do not make the random split trustworthy.

The gap in Table 2 is an upper bound on the part attributable to leakage alone: the reserved scans span all three substrates and none of them is guaranteed to appear in the training set, so protocol B absorbs both the cross-scan and the cross-substrate shift. With 14 scans available, k cannot exceed 5. Fourteen scans is the scale at which this literature operates, and it is the regime in which the random split fails.

A single U-Net run takes 21 s on one 8 GB consumer GPU, and the entire sweep (18 runs) takes 7 min.

5 Discussion

5.1 One mechanism, two benchmarks

The two experiments differ in task, model, imaging modality and source granularity, yet they agree on the quantity that matters: under a fixed annotation budget, changing only the provenance of the training images moves the reported metric by 28–55 points. On NEU-Seg the gap at IoU 0.5:0.95 is 0.546 at $k = 2$ and 0.303 at $k = 10$; on the STM data the Dice gap is 0.278–0.448 for $k \leq 5$. Neither benchmark finds a source count at which a random split becomes safe within the range this literature operates in.

The two experiments disagree. On NEU-Seg the gap narrows because protocol A falls; protocol B barely moves. On the STM data both protocols move, with B rising from 0.320 to 0.447 between $k = 3$ and $k = 5$. The difference follows from what the two datasets are. NEU-Seg images are crops of a small number of steel surfaces, so additional "sources" add little new appearance; the STM scans come from different samples, days and substrates, so additional scans extend the model's experience. In both benchmarks the size of the gap tracks how much new appearance each additional source brings. The leak does not disappear as data grow; its benefit does, and grouped performance improves only when the additional sources are genuinely new acquisitions.

5.2 Why the leak persists

In the thirty papers we coded, none describes a split that separates sources, and none reports an evaluation on acquisitions independent of its training data; three augment the pool before dividing it, four divide before augmenting, and twenty-three do not state the order. Six do not report the number of acquisitions even approximately, and one releases code.

A study built on 25 images states its split in one sentence and augments only the training portion. A paper that says "we trained on 240 patches from 3 scans and evaluated on a held-out scan" is making a weaker claim than one that says "we trained on 240 images and evaluated on 120", and the difference should be legible in the text rather than hidden in a random seed. Following the convention costs an author one sentence about the data.

5.3 A minimum reporting standard

We propose four items, each cheap to report and each sufficient to expose the problem:

1. **The number of independent acquisitions**, not the number of augmented images. If the dataset is built by tiling, report the number of parent images.
2. **The unit of the split**: image, source, session, or sample. If the split is random at the image level, say so.
3. **Whether augmentation precedes or follows the split.** Augmenting before splitting changes what the test set measures.
4. **A near-duplicate audit of the test set**: for example the fraction of test images whose nearest training neighbour is a sibling of the same source. In our data the

leaky construction gives a sibling rate of 0.68–0.97 depending on the source count, versus 0 for every grouped split, at a cost of seconds per dataset.

Items 1–3 cost a sentence each; item 4 costs one script. None requires additional computation.

5.4 Scope

The claim concerns small-data imaging, where the number of independent acquisitions is in the single or low double digits; with thousands of independent images the effect must shrink, and our own $k = 10$ point shows the direction. It concerns evaluation: models trained on augmented data are not the subject, and the subject is what a random split can measure. The STM benchmark also reserves scans across three substrates, so protocol B absorbs a cross-substrate shift in addition to the cross-scan shift; the Dice gap in Section 4.2 is an upper bound on the leakage component alone.

Both benchmarks reward localisation accuracy, and both can be saturated by a model that has seen near-identical images. A metric that measured something else (calibration, or performance on the classes that are not texture-separable) might be less vulnerable. We have not tested this, and it is the question we leave open.

6 Conclusion

Small-data imaging papers are constrained by their own data. With a few acquisitions per class, a split by acquisition leaves a test set too small to support any claim, so the augmented pool is split at random, and the resulting number is read as generalisation. The number these papers report describes the training set. On both benchmarks it is dominated by how many acquisitions the training set contains, and the dependence is strong enough that no source count in the range studied removes it. The study in our sample with the fewest acquisitions (25 images) documents a split before augmenting.

The remedy we propose is procedural: report the number of acquisitions, the unit of the split, the order of augmentation and splitting, and a near-duplicate audit of the test set. None of the four requires additional hardware. As these pipelines move from offline analysis towards closed-loop control, a metric that measures recognition of the training set stops being a harmless overstatement and becomes a wrong signal in a control loop.

Declarations

Funding. No funding was received for conducting this study.

Competing interests. The author has no relevant financial or non-financial interests to disclose.

Ethics approval. Not applicable. The study analyses public datasets and involves no human participants, human tissue or animals.

Consent to participate. Not applicable.

Consent for publication. Not applicable.

Data availability. NEU-Seg is publicly available [12]; the STM dataset is available from Zenodo under CC-BY 4.0 [16]. Every number in this manuscript is produced from these public data; no private data were used.

Code availability. All split definitions, training scripts and the leakage-audit script are available at https://github.com/ZhangShaoXuan-HIT/sibling-leakage and archived at https://doi.org/10.5281/zenodo.24082524 (v1.0.1).

Author contributions. S.Z. conceived the study, implemented the experiments, analysed the data and wrote the manuscript.

Use of AI tools. DeepSeek V4.1 Flash (version 10.17) was used to verify reference metadata, to assist with the LaTeX typesetting, and to translate and edit the manuscript text. The author is responsible for all scientific content, analysis and conclusions, and checked every AI-assisted output against the primary sources.

Acknowledgements. Not applicable.

References

[1] Yuan, S., Zhu, Z., Lu, J., Zheng, F., Jiang, H., Sun, Q.: Applying a deep-learning-based keypoint detection in analyzing surface nanostructures. Molecules **28**(14), 5387 (2023) https://doi.org/10.3390/molecules28145387

[2] Kapoor, S., Narayanan, A.: Leakage and the reproducibility crisis in machine-learning-based science. Patterns **4**(9), 100804 (2023) https://doi.org/10.1016/j.patter.2023.100804

[3] Kaufman, S., Rosset, S., Perlich, C., Stitelman, O.: Leakage in data mining. ACM Transactions on Knowledge Discovery from Data **6**(4), 1–21 (2012) https://doi.org/10.1145/2382577.2382579

[4] Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A.I., Etmann, C., McCague, C., Beer, L., Weir-McCall, J.R., Teng, Z., Gkrania-Klotsas, E., AIX-COVNET, Ruggiero, A., Korhonen, A., Jefferson, E., Ako, E., Langs, G., Gozaliasl, G., Yang, G., Prosch, H., Preller, J., Stanczuk, J., Tang, J., Hofmanninger, J., Babar, J., Sánchez, L.E., Thillai, M., Gonzalez, P.M., Teare, P., Zhu, X., Patel, M., Cafolla, C., Azadbakht, H., Jacob, J., Lowe, J., Zhang, K., Bradley, K., Wassin, M., Holzer, M., Ji, K., Ortet, M.D., Ai, T., Walton, N., Lio, P., Stranks, S., Shadbahr, T., Lin, W., Zha, Y., Niu, Z., Rudd, J.H.F., Sala, E., Schönlieb, C.-B.: Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans. Nature Machine Intelligence **3**(3), 199–217 (2021) https://doi.org/10.1038/s42256-021-00307-0

[5] Varoquaux, G., Cheplygina, V.: Machine learning for medical imaging: methodological failures and recommendations for the future. npj Digital Medicine **5**(1), 48 (2022) https://doi.org/10.1038/s41746-022-00592-y

[6] Zhu, Z., Huang, Q., Yang, T., Jiang, H., Yuan, S., Xiang, J., Cai, L., Sun, Q.: Deep learning drives autonomous molecular reactions with single-bond selectivity in tetra-brominated porphyrins on au(111). Nature Communications **17**(1), 2348 (2026) https://doi.org/10.1038/s41467-026-69080-1

[7] Wu, S., Chen, G., Shen, S., Yan, J.: Point defect detection and classification in mos2 scanning tunneling microscopy images: A deep learning approach. Molecules **30**(12), 2644 (2025) https://doi.org/10.3390/molecules30122644

[8] Zhao, L., Wang, Y., Xu, F.: Accurate segmentation of localized fuel cladding chemical interaction layers in sem micrographs with deep learning method. Scientific Reports **15**(1), 28878 (2025) https://doi.org/10.1038/s41598-025-14927-8

[9] Millan-Solsona, R., Checa, M., Brown, S.R., Bible, A.N., Srijanto, B., Wiggins, L., Madugula, S.S., Pyne, A.L.B., Morrell-Falvey, J.L., Retterer, S., Vasudevan, R.K., Collins, L.: Synthetic data-driven deep learning for label-free autonomous atomic force microscopy. Nature Communications **17**(1), 3886 (2026) https://doi.org/10.1038/s41467-026-70421-3

[10] Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Hartig, F., Dormann, C.F.: Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography **40**(8), 913–929 (2017) https://doi.org/10.1111/ecog.02881

[11] Zhang, S.: Where to Put Low-Rank Capacity? A Grouped-Split Study of LoRA for Industrial Defect Segmentation. Research Square, PREPRINT (Version 1). 01 November 2026 (2026). https://doi.org/10.21203/rs.3.rs-11517758/v1

[12] Yan, K., et al.: NEU_Seg, MT, KolektorSDD2, DAGM2007. Mendeley Data, V1, dataset (2025). https://doi.org/10.17632/z8g47vzt4p.1

[13] Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(6), 1137–1149 (2017) https://doi.org/10.1109/tpami.2016.2577031

[14] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Lecture Notes in Computer Science, pp. 740–755 (2014). https://doi.org/10.1007/978-3-319-10602-1_48

[15] Orji, D.O.: Intelligent Microscopy for Materials: Leakage-Controlled SEM Morphology Classification with Calibration, Explainability, and Artifact Perturbation. ChemRxiv, preprint (2026). https://doi.org/10.26434/chemrxiv.15008876/v1

[16] Kolev, N.L., Stock, T.: Overcoming Data Scarcity in STM Image Segmentation. Zenodo, dataset (2025). https://doi.org/10.5281/zenodo.15525935

[17] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Lecture Notes in Computer Science, pp. 234–241 (2015). https://doi.org/10.1007/978-3-319-24574-4_28