

Single-Pass Conformal Cross-Modal Anomaly Screening with Tabular Foundation Models

José Lucas De Melo Costa Seong Woo Ahn Fabrice Popineau Arpad Rimmel
Bich-Liên Doan

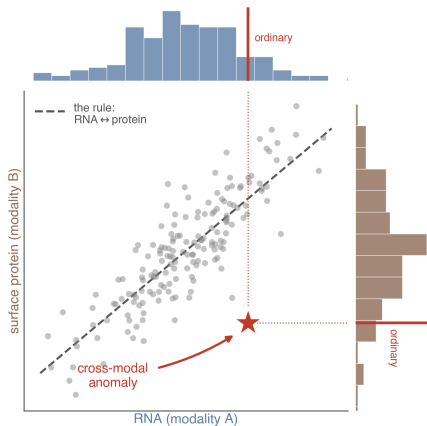
LISN, Université Paris-Saclay, CNRS, CentraleSupélec
Gif-sur-Yvette, France

MultiTab Workshop @ MICCAI 2026

Strasbourg, 27 September 2026

Paper #24 · Oral · github.com/jose-melo/single-pass-conformal-calibration

An anomaly that no single modality can see



Co-measured modalities are tied by a rule.

Transcriptome $\leftrightarrow$ surface protein, image $\leftrightarrow$ clinical record.

The record we want to flag:

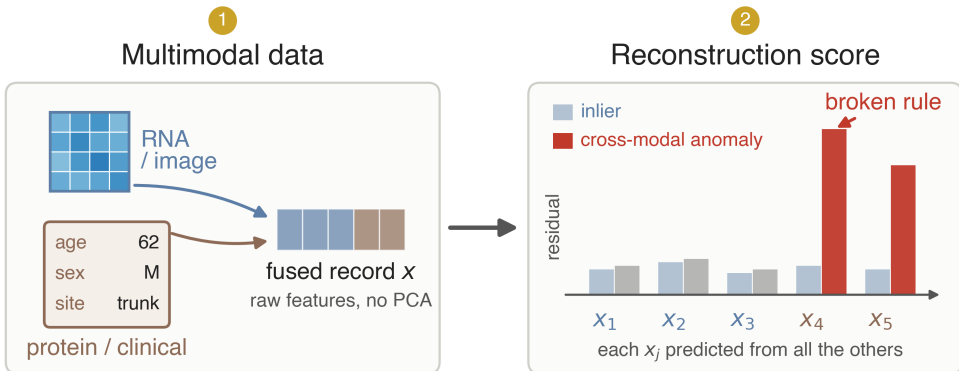
- ordinary in **modality A**,
- ordinary in **modality B**,
- ✗ off the joint rule.

A unimodal screen misses it *by construction*.

Target

A rare **biological subpopulation** (aberrant cell state, clinical stage), not a corrupted record: a *dependency* anomaly, not a second mode.

A training-free screen from a tabular foundation model

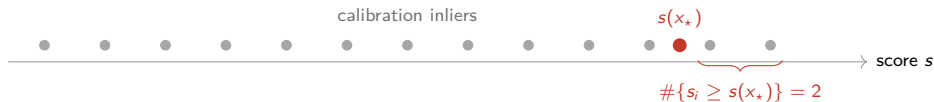


$$s(x) = \frac{1}{d} \sum_{j=1}^d \frac{(x_j - \hat{x}_j)^2}{\hat{\sigma}_j^2}, \quad \hat{x}_j = \mathbb{E}[x_j \mid x_{-j}, R]$$

TabPFN-v2 in context on a pool R of inliers: no training, no fine-tuning. [Hollmann+25]

The gate: conformal p -values and Benjamini–Hochberg

$$p(x_*) = \frac{1 + \#\{i : s_i \geq s(x_*)\}}{n_{\text{cal}} + 1}, \quad \{s_i\}_{i=1}^{n_{\text{cal}}} \text{ calibration inlier scores}$$



$$p(x_*) = \frac{1 + 2}{13 + 1} \approx 0.21 \Rightarrow \text{not flagged at } \alpha = 0.10$$

- Exchangeability of calibration and inlier query scores $\Rightarrow p$ is super-uniform.
- Threshold at $\alpha \Rightarrow$ false-positive rate $\leq \alpha$. [Vovk+05, Bates+23]
- BH or Storey over a batch $\Rightarrow$ FDR at level q . [BH95, Storey02, AdaDetect]

The one premise

Everything above is exact *only if* the calibration point and the query are scored the same way.

Where do the calibration scores come from?

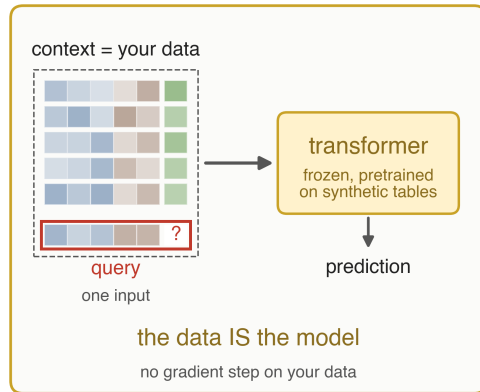
A TFM has no training step: fit only stores a context. So the natural recipe uses the one clean pool for everything.

What anyone would write

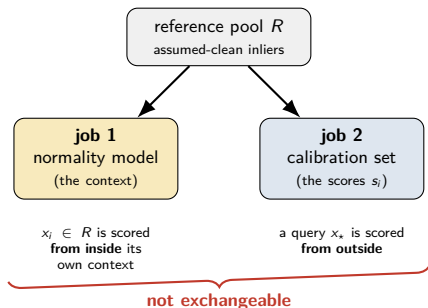
```
# one clean pool R
m = TabPFN().fit(R)                                # context
s_cal = score(m, R)                                # calibration
s_test = score(m, X_test)
thr = quantile(s_cal, 0.90)
flag = s_test > thr
```

`score(m, R)` is in-sample evaluation in disguise: every calibration row is part of the input it is scored with.

TabPFN: in-context learning



The obstacle: the reference pool does two jobs



What happens. A point that attends to itself reconstructs itself: its residual **deflates**, calibration scores sit too low, the gate **over-flags**.

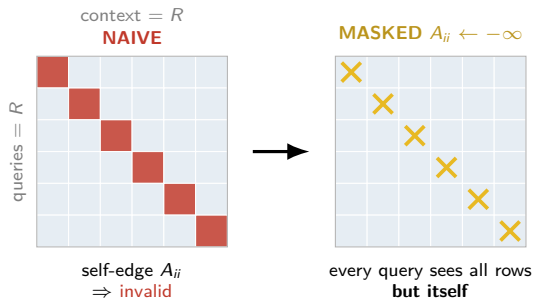
Measured on the suite (naive in-context conformal)

realized FPR **0.49 – 0.68** at nominal **0.10**
controls FPR on **0/6** pairs and **0/9** clinical tables

Textbook exits, and their price

- ✗ held-out split: exact, but it *spends* the context.
- ✗ jackknife / CV+: n (or K) refits of a foundation model.

The fix: mask the diagonal of the sample-attention matrix



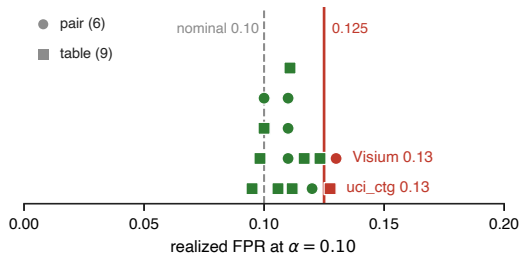
Place R as *both* context and query block. At every layer the model forms $A \in \mathbb{R}^{n \times n}$; the diagonal is query i 's edge to its own token.

$$s_i \approx f(x_i \mid R \setminus \{x_i\}) \quad \text{after} \quad A_{ii} \leftarrow -\infty$$

- ✓ **All n leave-one-out scores in one forward pass.** No refit, no held-out split.
- ✓ Test points keep the full pool; both sets enter the p -value.
- ➔ Needs maskable quadratic sample attention, in-context conditioning, a per-feature mean and scale. Run on **TabPFN-v2** only.

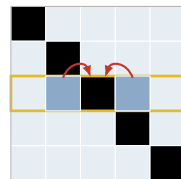
What the certificate is, and what it is not

What it is



Realized per-inlier FPR of single-pass LOO+Storey at $\alpha = 0.10$; 15 seeds (pairs), 10 (tables). Valid iff $\text{FPR} \leq 0.125$ (nominal + one binomial SE at $n \approx 10^2$).

What it is not



masked point **still reconstructable** through correlated neighbours $\Rightarrow$ **anti-conservative**

context of $n - 1$ vs. $n \Rightarrow$ **conservative** offset (scale ratio ≈ 1.02)

Not an exact conformal guarantee. FDR is characterized, not certified.

Sanity check vs. exact n -refit LOO on CBMC ($d=8$, $n=45$): Spearman **0.975**

In clinical terms: ≈ 12 – 13 false flags per 100 screened inliers instead of 10: a **triage tolerance**.

One suite, one conformal scaffold, two regimes

One record, two modalities



image block
(modality A)

field	value
age_approx	30
sex	female
anatom_site_general	torso
melanocytic	true

clinical block
(modality B)



one fused row



image features

clinical

ISIC image CC-0 · table schematic

Six orthogonal instrument pairs

pair	modalities
CBMC CITE-seq	RNA + protein
10x PBMC CITE-seq	RNA + 14 proteins
BMMC CITE-seq	RNA + 134 proteins
TCGA-BRCA	expression + clinical
ISIC	dermoscopy + clinical
10x Visium	H&E + spatial expression

Reference pool: majority phenotype, $n_{\text{ref}} \sim 10^2$.

Target: a rare phenotype, independently annotated.

Valid power: mean power, every dataset
with FPR > 0.125 counted as 0.

Six cross-modal pairs: the only single-pass route that stays valid

construction	valid power $\uparrow$	valid pairs
naive (dual-role)	0.00	0/6
split conformal	0.43	4/6
CV+ (K refits)	–	1/6 tractable
AdaDetect	0.73	6/6
best classical (k -NN)	0.77	6/6
single-pass LOO	0.79	5/6
single-pass LOO + Storey	0.80	5/6

Naive tops *raw* power on every pair, at FPR 0.49–0.68: raw power without validity is not a result.

An ordinary mole in the wrong context (ISIC)

Lesion benign on its own, age and site unremarkable on their own; the *combination* is rare.

Fused screen: power **0.93** at FPR 0.10
 k -NN: 0.75 naive: FPR **0.56**

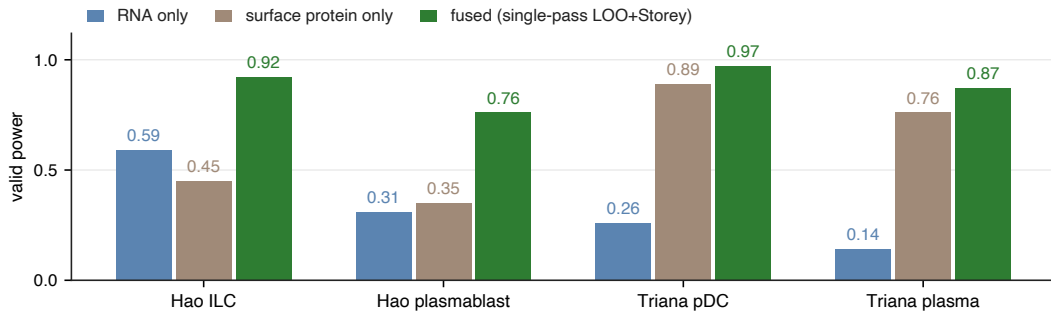
The one failure is **Visium**: FPR 0.13, just past the line. Its raw power 0.56 tops every baseline (≤ 0.17) but is *uncertified*.

Nine clinical tables: the calibration holds without any cross-modal help

construction	raw	valid $\uparrow$	valid tables
naive	0.63	0.00	0/9
split conformal	0.27	0.25	7/9
AdaDetect	0.18	0.18	9/9
best classical	0.29	0.29	8/9
CV+ (K refits)	0.41	0.41	9/9
single-pass LOO	0.41	0.37	8/9
LOO + Storey	0.49	0.45	8/9

- ✓ LOO+Storey leads valid power at **one** forward pass; CV+ pays K refits for 0.41.
- ✗ Naive has the best raw mean and controls FPR nowhere.
- Pooled over the suite, 8,400 inlier p -values stay within **0.027** of the uniform diagonal, at every threshold, not only at α .
- Realized FDR 0.05 / 0.12 / 0.23 at $q = 0.05$ / 0.10 / 0.20.

The signal lives in the coupling, not in either margin



Valid power (LOO+Storey) on four rare cell states of the deep CITE-seq panels Hao (228 proteins) and Triana (197).

The fused screen clears both unimodal screens on all four rare cell states; per-modality PCA fusion collapses ISIC power to 0.18.

Missing modality: recalibrate per observed pattern. Valid on all 21 ISIC dropout regimes; power $0.81 \rightarrow 0.79$ without the image, $0.81 \rightarrow 0.45$ without the clinical block.

Take-home

- 1 **One masked forward pass = all n leave-one-out calibration scores.**

Setting $A_{ii} \leftarrow -\infty$ removes the dual-role bias that drives naive FPR to 0.49–0.68; no refit, no held-out split.

- 2 **The guarantee is an empirical certificate, not an exact one.**

$\text{FPR} \leq 0.125$ on 5/6 pairs and 8/9 clinical tables, at power on par with the strongest training-free baselines (0.80 vs 0.77).

- 3 **Fusion must preserve correlation; blocks must be orthogonal; the anomaly must break a dependency.**

The fused score flags rare cell states neither modality resolves alone; PCA, redundant pairs and local outliers are the named failure modes.

Next: masked-vs-exact LOO at higher dimension, the contamination sweep, a finite-sample guarantee.



code and eight baselines
[jose-melo/single-pass-conformal-calibration](#)



[paper \(open access\)](#)

Thank you.

Questions?

Open to collaborations

- NeurIPS 2026** Drive vs. Decay: the training dynamics of JEPAs
- CIKM 2026 (oral)** Convergence collapse in fixed-target detectors (KCM)
- CIKM 2026** Knowledge-informed causal discovery (with S. W. Ahn)
- ESANN 2026** High performance, low reliability: tabular FM uncertainty
- ICLR 2025** T-JEPA: augmentation-free SSL for tables (joint first)

PhD ending autumn 2027 · research visits, postdocs, joint work

Poster in Hall Rhin at 17:20 · reception at 18:00



paper · code · slides
poster · contact

josecosta.fr/multitab



**José Lucas
De Melo Costa**

jose-lucas.de-melo-costa@centralesupelec.fr

Backup: Table 1, six pairs and the backbone summary

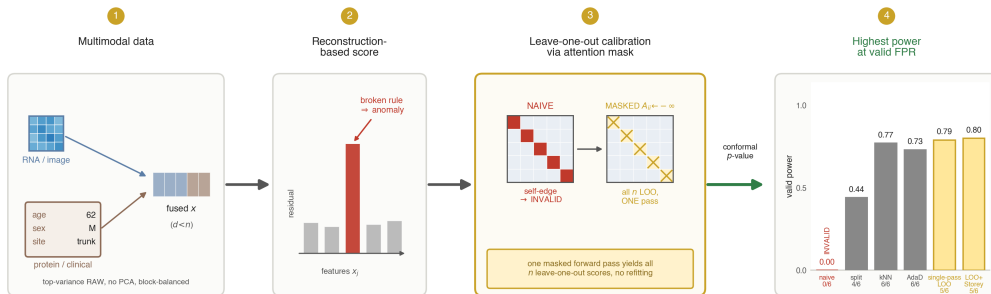
dataset	modalities	conformal constructions				CV+	AD baselines	
		naive	split	single-pass LOO	LOO+Storey		AdaDetect	best classical
Orthogonal instrument pairs (cross-modal)								
CBMC CITE-seq	RNA+prot	1.00 [97,1.00]	1.00 [97,1.00]	0.99 [95,1.00]	1.00 [97,1.00]	—	0.96 [91,98]	1.00 [97,1.00]
		0.49 [40,58] ✗	0.07 [04,13]	0.11 [09,13]	0.11 [09,13]		0.08 [04,14]	0.08 [04,14]
10x PBMC CITE-seq	RNA+14 prot	1.00 [97,1.00]	1.00 [97,1.00]	1.00 [97,1.00]	1.00 [97,1.00]	—	0.99 [95,1.00]	1.00 [97,1.00]
		0.53 [44,62] ✗	0.13 [08,20] ✗	0.10 [10,14]	0.12 [10,14]		0.05 [02,10]	0.11 [07,18]
BMMC CITE-seq	RNA+134 prot	1.00 [97,1.00]	0.93 [87,96]	0.99 [95,1.00]	0.99 [95,1.00]	—	0.98 [94,99]	1.00 [97,1.00]
		0.57 [48,66] ✗	0.08 [04,14]	0.12 [09,13]	0.11 [09,13]		0.07 [04,13]	0.12 [07,19]
TCGA-BRCA	expr+clin	0.94 [88,97]	0.68 [59,76]	0.83 [75,89]	0.88 [81,93]	—	0.77 [69,84]	0.85 [78,90]
		0.61 [52,69] ✗	0.09 [05,15]	0.11 [09,13]	0.11 [09,13]		0.11 [07,18]	0.05 [02,10]
ISIC	image+clin	0.99 [95,1.00]	0.00 [00,03]	0.90 [83,94]	0.93 [87,96]	0.92 [86,96]	0.50 [41,59]	0.75 [67,82]
		0.56 [47,65] ✗	0.13 [08,20] ✗	0.10 [08,11]	0.10 [08,11]	0.09 [05,15]	0.10 [06,17]	0.08 [04,14]
10x Visium (spatial)	H&E+expr	0.94 [88,97]	0.00 [00,03]	0.14 [09,21]	0.56 [47,65]	—	0.17 [11,25]	0.00 [00,03]
		0.68 [59,76] ✗	0.08 [04,14]	0.13 [10,13] ✗	0.13 [10,13] ✗		0.08 [04,14]	0.09 [05,15]
valid power ↑		0.00	0.43	0.79	0.80	—	0.73	0.77
controls FPR		0/6	4/6	5/6	5/6	—	6/6	6/6
Validity backbone (unimodal, 9 clinical+omics tables)								
valid power ↑		0.00	0.25	0.37	0.45	0.41	0.18	0.29
controls FPR		0/9	7/9	8/9	8/9	9/9	9/9	8/9

Each cell: raw power over realized per-inlier FPR, 95% CIs; FPR > 0.125 marked X for every method including ours.

Backup: Table 2, per-table backbone

table	conformal constructions				AD baselines		
	naive	split	single-pass LOO	LOO+Storey	CV+	AdaDetect	best classical
annthyroid	0.52 [.23,.80]	0.26 [.00,.53]	0.44 [.20,.70]	0.55 [.27,.83]	0.47 [.26,.68]	0.23 [.00,.46]	0.29 [.11,.48]
	0.17 [.14,.20] ✗	0.11 [.08,.13]	0.10 [.08,.12]	0.10 [.08,.12]	0.10 [.07,.12]	0.07 [.04,.09]	0.10 [.07,.14]
thyroid	0.82 [.74,.89]	0.45 [.18,.72]	0.77 [.66,.86]	0.81 [.70,.90]	0.75 [.66,.84]	0.41 [.16,.66]	0.38 [.09,.67]
	0.17 [.14,.19] ✗	0.10 [.08,.12]	0.10 [.08,.12]	0.10 [.08,.12]	0.10 [.07,.12]	0.08 [.05,.11]	0.09 [.07,.10]
cardio	0.82 [.75,.87]	0.35 [.13,.57]	0.48 [.29,.66]	0.68 [.53,.80]	0.46 [.33,.57]	0.24 [.06,.44]	0.37 [.13,.60]
	0.37 [.34,.40] ✗	0.10 [.07,.13]	0.11 [.09,.13]	0.11 [.09,.13]	0.08 [.05,.11]	0.10 [.06,.14]	0.12 [.09,.16]
Cardiotocography	0.35 [.23,.48]	0.05 [.00,.14]	0.09 [.02,.16]	0.20 [.14,.26]	0.11 [.04,.19]	0.14 [.04,.28]	0.09 [.00,.23]
	0.32 [.26,.39] ✗	0.08 [.06,.10]	0.11 [.08,.13]	0.11 [.08,.13]	0.09 [.08,.10]	0.08 [.06,.11]	0.11 [.08,.14]
mammography	0.60 [.44,.73]	0.45 [.29,.58]	0.40 [.24,.55]	0.51 [.38,.62]	0.36 [.22,.48]	0.09 [.00,.21]	0.17 [.00,.35]
	0.23 [.18,.29] ✗	0.09 [.07,.11]	0.12 [.09,.14]	0.12 [.09,.14]	0.10 [.07,.13]	0.08 [.06,.10]	0.11 [.07,.15]
breastw	0.98 [.97,.99]	0.68 [.38,.95]	0.93 [.84,.98]	0.99 [.99,1.00]	0.97 [.93,.98]	0.14 [.00,.35]	0.93 [.87,.98]
	0.16 [.12,.19] ✗	0.10 [.07,.13]	0.11 [.09,.14]	0.11 [.09,.14]	0.10 [.08,.13]	0.04 [.01,.08]	0.10 [.06,.13]
Pima	0.07 [.00,.14]	0.00 [.00,.00]	0.00 [.00,.00]	0.01 [.00,.03]	0.00 [.00,.00]	0.00 [.00,.00]	0.04 [.00,.11]
	0.17 [.14,.21] ✗	0.08 [.06,.10]	0.10 [.08,.13]	0.10 [.08,.13]	0.10 [.08,.11]	0.07 [.05,.10]	0.14 [.11,.17] ✗
uci_ctg	0.50 [.36,.60]	0.11 [.00,.28]	0.35 [.20,.49]	0.41 [.28,.53]	0.44 [.28,.56]	0.28 [.11,.45]	0.28 [.06,.51]
	0.33 [.30,.38] ✗	0.13 [.09,.19] ✗	0.13 [.10,.16] ✗	0.13 [.10,.16] ✗	0.11 [.09,.14]	0.09 [.06,.11]	0.08 [.04,.12]
uci_wdbc	0.97 [.93,1.00]	0.08 [.00,.24]	0.25 [.10,.41]	0.26 [.10,.43]	0.17 [.00,.34]	0.07 [.00,.22]	0.09 [.00,.26]
	0.70 [.65,.76] ✗	0.14 [.11,.17] ✗	0.12 [.10,.14]	0.12 [.10,.14]	0.08 [.05,.10]	0.09 [.07,.12]	0.09 [.07,.11]
mean power (raw) ↑	0.63	0.27	0.41	0.49	0.41	0.18	0.29
valid power ↑	0.00	0.25	0.37	0.45	0.41	0.18	0.29
controls FPR	0/9	7/9	8/9	8/9	9/9	9/9	8/9

Backup: the pipeline in one figure



Backup: baselines and the eight classical detectors

Same score, same scaffold, different calibration

- **naive**: pool scored inside its own context.
- **split**: 60% of the pool as context, 40% held out as calibration.
- **single-pass LOO**: diagonal mask; + **Storey**: $\hat{\pi}_0$ at $\lambda = 0.5$ sharpens BH.
- **CV+**: K -fold jackknife+ (Barber+21); tractable on the backbone only.
- **AdaDetect** (Marandon+24): learned score under BH.

Best classical = argmax over eight PyOD detectors on the backbone (k -NN, LOF, isolation forest, OCSVM, PCA, HBOS, COPOD, ECOD); on the pairs the classical column is k -NN.

Trained fusers (totalVI, Seurat-WNN, TIP, HEALNet) return embeddings, not p -values, and need per-dataset training: outside the training-free regime of the table.

Seeds and CIs: 15 seeds on the pairs, 10 on the backbone; seed-bootstrap or Wilson intervals per cell.

Backup: which model properties the argument needs

- ① Full quadratic attention over the context samples, exposed and **maskable at every layer**.
- ② In-context conditioning with **no per-query refit**.
- ③ A per-feature predictive **mean and residual scale** to feed the score.

Whether other in-context tabular models with these properties behave the same way is **untested**: every number in the paper is TabPFN-v2.

Related routes. LOO-CAD (Hennhöfer & Preisach 24) pays n folds with classical detectors; LOO-StabCP (Lee & Zhang 25) bounds LOO via algorithmic stability of a refit learner; ICL-for-full-CP (Deng+25) trains a CP-aware transformer and needs one pass per candidate. Ours: frozen model, one pass total, leave-one-out over *context* examples.

Backup: numbers a reviewer asked about

- **ISIC valid arm** runs at a reduced d/n : $n_{\text{ref}} = 300$, $d = 8$ (image embedding + clinical, top-variance).
- **Visium**: plain LOO raw power 0.14, LOO+Storey 0.56, both at FPR 0.13; split 0.00 at 0.08; AdaDetect 0.17 at 0.08.
- **Null control**: a within-cohort record swap perturbs no dependency and stays at chance (AUROC ≈ 0.49).
- **Backbone scope**: nine tables where every construction including CV+ is defined; the two $d > n$ omics tables (METABRIC, PBMC scRNA) sit outside that regime and are not averaged in.
- **Storey** is used only when power-limited; the default is plain single-pass LOO.
- **Reproducibility**: pair splits are seeded (`default_rng`); the nine backbone splits used an unpinned Python hash, so cells reproduce up to seed noise, not bit-for-bit.