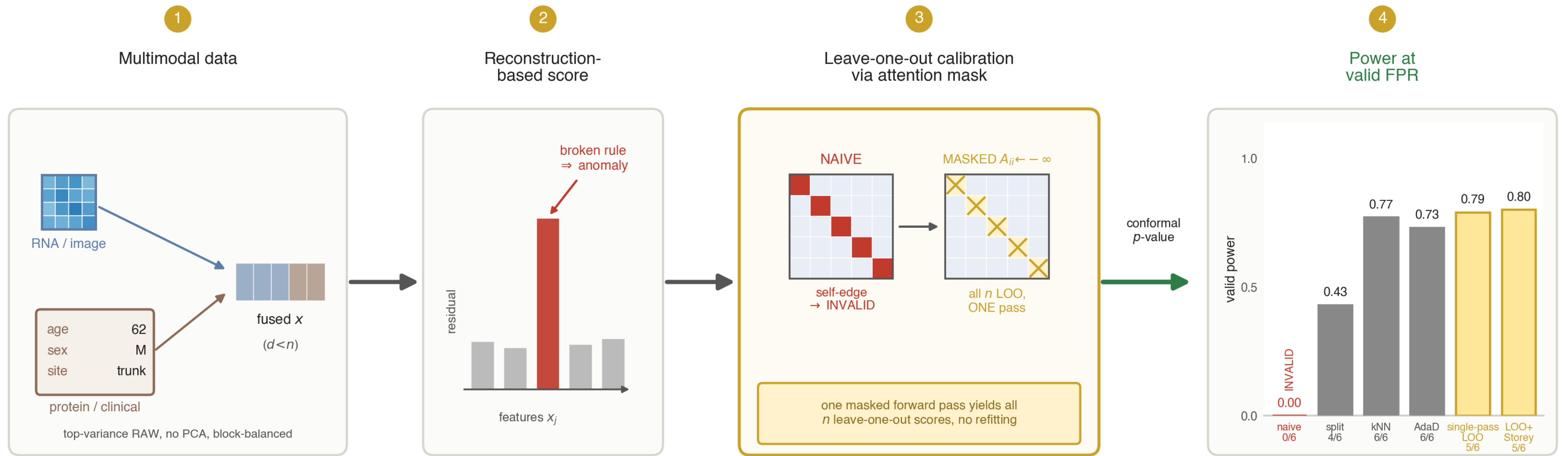
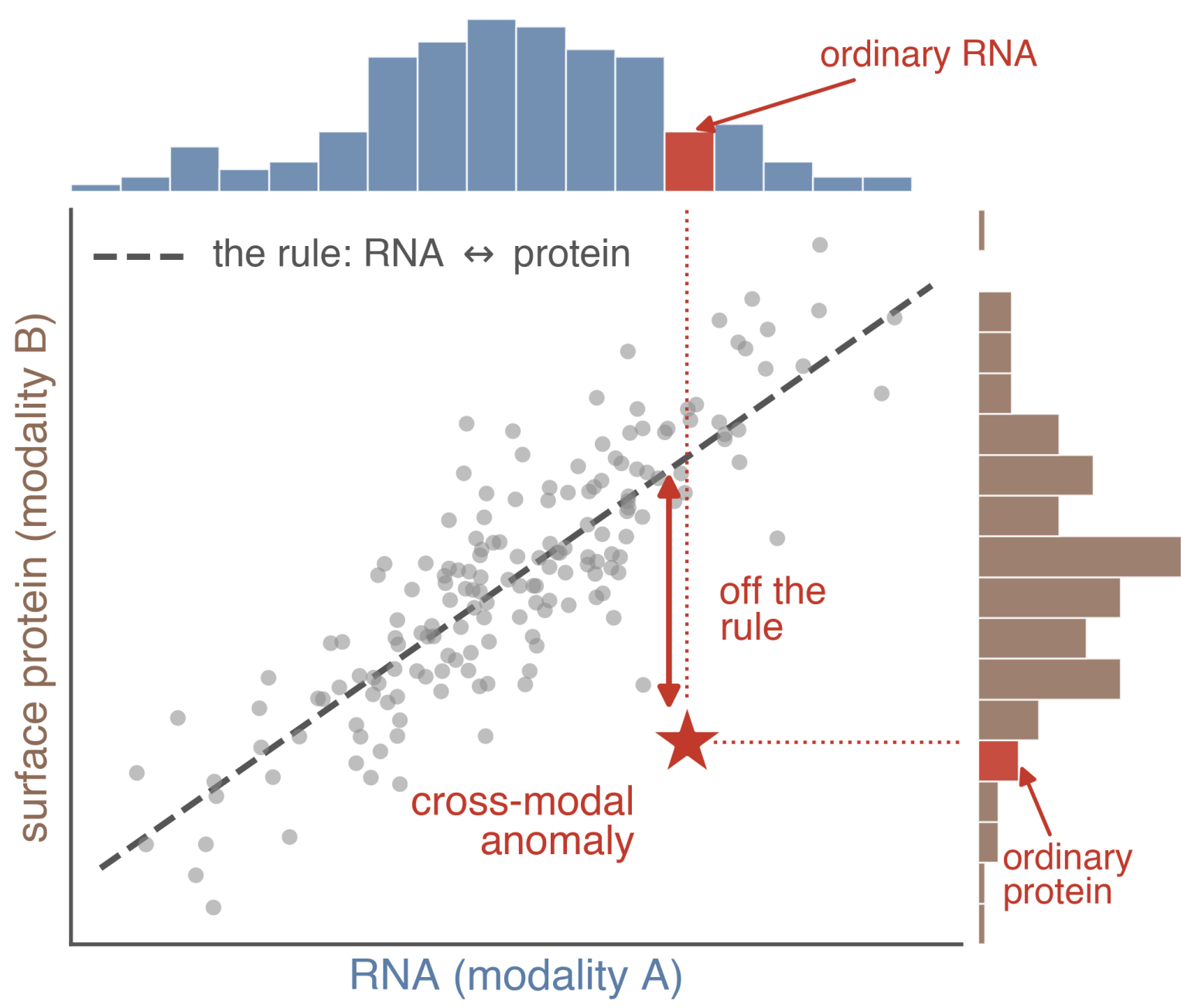




1 An anomaly no single modality sees



Two modalities measured on the same unit are tied by a rule. We flag records that are **ordinary in each modality yet off the joint rule**: a rare subpopulation that a unimodal screen can miss. (simulated illustration)

2 A training-free score

TabPFN-v2, in context on a pool R of assumed-clean inliers, predicts each feature from all the others. No training. (illustration below)



The score is the mean squared standardized residual:

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

Fusion keeps raw top-variance features per block, with **no PCA**, so the score reads the cross-modal rule.

3 The gate: conformal p -values

$$p(x_*) = \frac{1 + \#\{i : s_i \geq s(x_*)\}}{n_{\text{cal}} + 1}$$

Threshold at α for FPR control, exact *only if* calibration and query scores are exchangeable. Benjamini–Hochberg or Storey over a batch: FDR is reported, not certified.

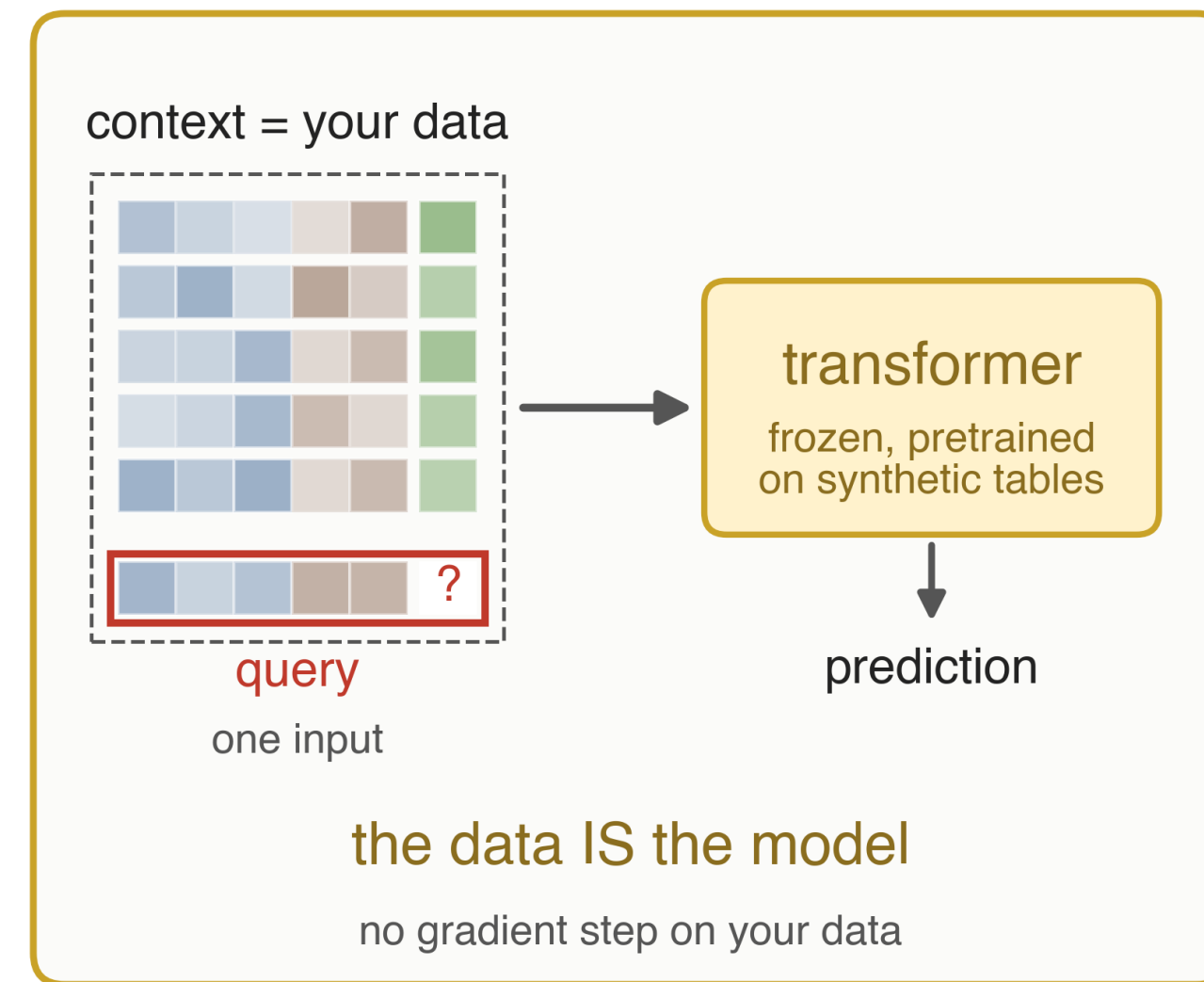
4 Evaluation suite

Six orthogonal pairs: three CITE-seq panels (RNA + protein), TCGA-BRCA (expression + clinical), ISIC (dermoscopy + clinical), 10x Visium (H&E + spatial expression).

Nine clinical tables test the calibration alone.

Valid power = mean power, with every dataset whose FPR exceeds 0.125 counted as 0. $\alpha = q = 0.10$.

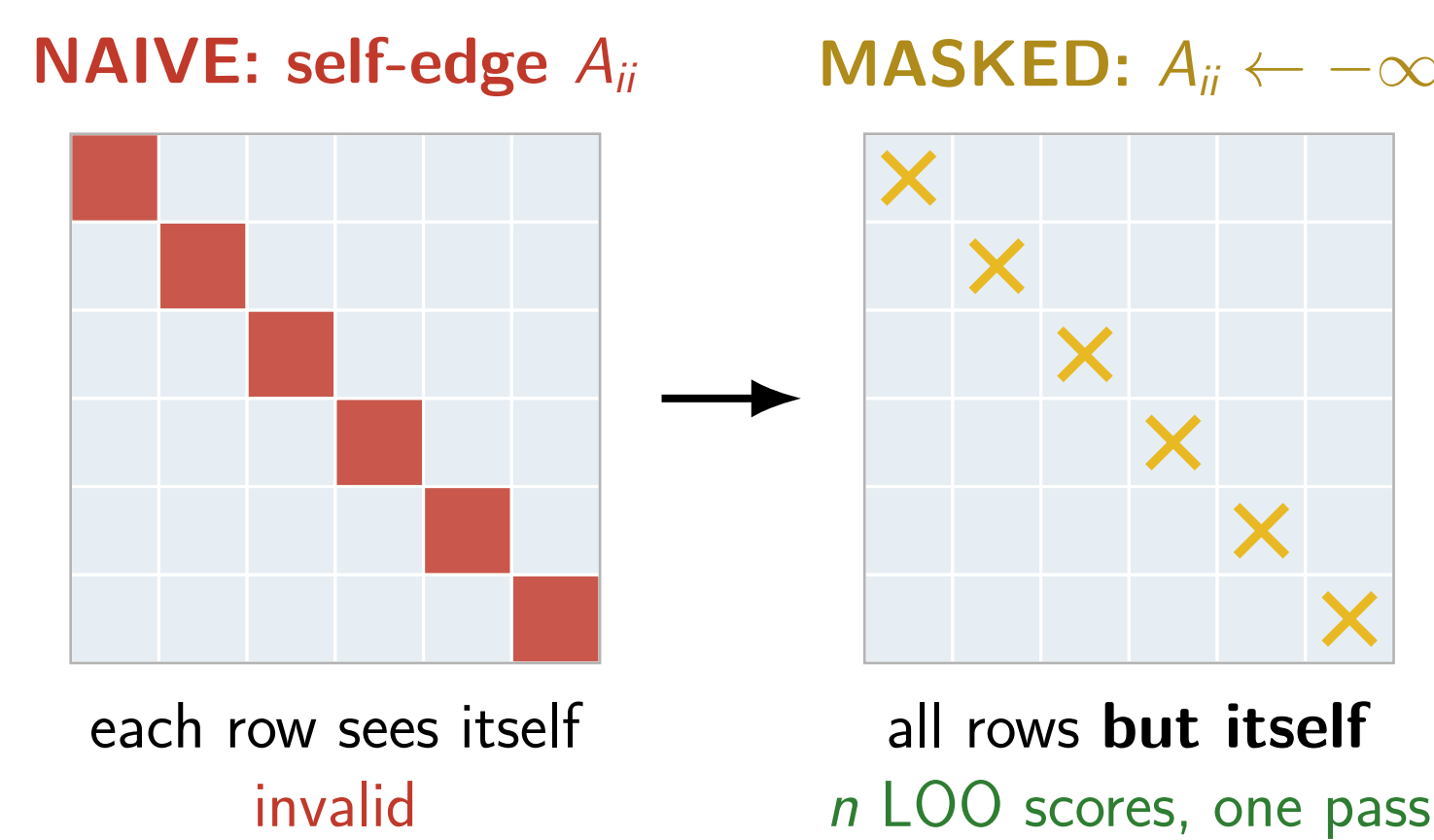
5 The obstacle: one pool, two jobs



TabPFN has no training step: the pool R is the context it reads at prediction time. Reusing R as the calibration set scores every calibration row with itself in the input, so its residual deflates and the gate over-flags.

Naive in-context conformal:
FPR **0.49–0.68** on the pairs, nominal 0.10
valid on **0/6** pairs, **0/9** tables

6 The fix: mask the diagonal



Feed R as context and as queries, and set the diagonal of the sample-attention matrix to $-\infty$ at every layer:

$$s_i \approx f(x_i | R \setminus \{x_i\}).$$

- ✓ All n approximate leave-one-out scores in **one masked pass** per target feature, instead of n refits; no held-out split.
- Needs maskable sample attention; tested on TabPFN-v2 only.

7 What the certificate is

- ✗ Correlated neighbours still leak a masked point: *anti-conservative*, grows with redundancy.
- Calibration sees $n - 1$ rows, a test point n : *conservative*.
- ✓ Masked vs. exact LOO on CBMC ($d=8$, $n=45$): Spearman **0.975**.

Empirical certificate: realized FPR ≤ 0.125 at $\alpha = 0.10$ on 5/6 pairs and 8/9 tables, i.e. at most 12–13 false flags per 100 inliers instead of 10. A triage tolerance, not an exact conformal guarantee.

8 Six cross-modal pairs

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

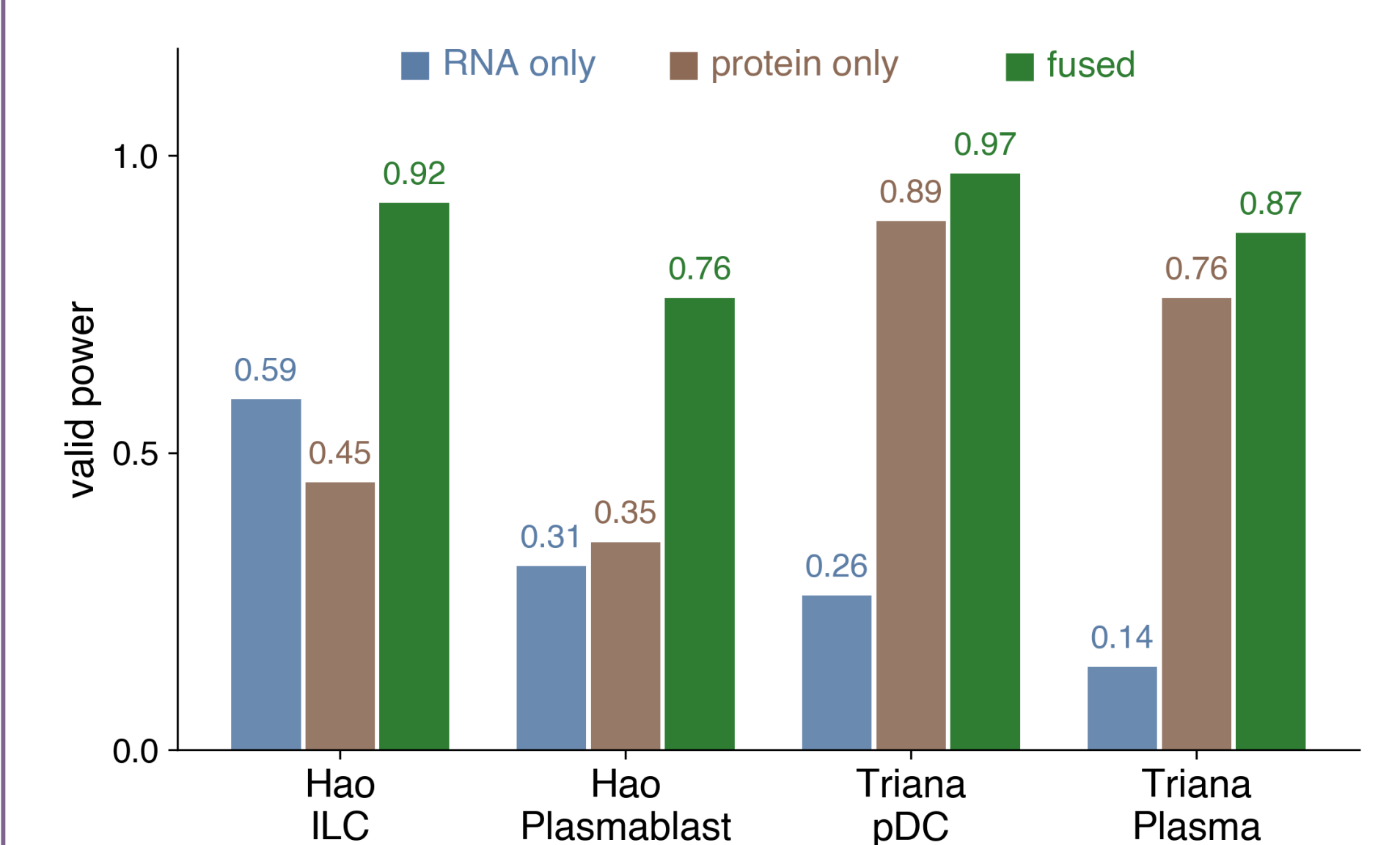
ISIC: power **0.93** (LOO+Storey) at FPR 0.10; k -NN 0.75. The one failing pair is Visium (FPR 0.13).

9 Nine clinical tables: validity alone

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 beats CV+ (0.45 vs 0.41) without K refits; CV+ is valid on 9/9.

10 The signal lives in the coupling



The fused screen beats **both** unimodal screens on four rare cell states. PCA fusion drops ISIC power to **0.18**.

11 Where it fails

Correlated blocks keep a subtype reconstructable (Multitime, glioma, CLL).

Redundant pair (Visium): leakage breaks the certificate, FPR 0.13.

Local outliers are not dependency breaks (CosMx: LOF beats reconstruction).

Take-home. One masked pass gives all n approximate leave-one-out scores. The certificate is empirical. Fusion must preserve correlation.