An Unidentifiability Oracle for H&E → Spatial Transcriptomics

ML Lab Research Proposal — trustworthy selective prediction for virtual spatial transcriptomics

Team: Aayan Alwani

1. Problem Statement

Predicting spatial gene expression from routine H&E histology promises cheap, slide-wide molecular maps, but current models give point predictions with no trustworthy notion of when to believe them. A weak prediction–truth correlation is ambiguous: the gene may be *biologically unpredictable* from morphology, or the measurement may simply be *noisy / dropout-corrupted* — and existing uncertainty and abstention methods conflate the two. We build an **unidentifiability oracle** that separates intrinsic biological unpredictability from technical measurement noise, so a selective-prediction layer abstains on genuinely unidentifiable genes rather than merely noisy ones.

2. Proposed Approach

We use a **cleaner-reference** design. Frozen pathology foundation-model embeddings (UNI / Virchow2) feed light predictive heads; per-gene unidentifiability U is the deep-ensemble out-of-fold aleatoric residual **minus** an externally measured noise floor (a true Xenium technical replicate) **minus** registration-induced variance. A post-hoc spatial-Mondrian conformal layer turns the surviving signal into coverage-controlled selective risk. We chose this over Bayesian UQ (uncalibrated here) and naive variance-based abstention (which tracks dropout, not biology) because the aleatoric/epistemic split is provably not recoverable from a predictor's own outputs — it requires the external clean reference and replicate that Xenium uniquely provides. Ensembling across distinct architectures isolates recoverable signal, so U is a principled upper bound on intrinsic unidentifiability, not one model's under-fitting.

3. Dataset

All data are public. The primary source is GEO **GSE243280** (Janesick 2023): paired breast **Xenium** (two serial technical replicates, GSM7780153/4), **Visium**, and post-stain **H&E**, on a 313-gene panel. We extend to five co-registrable organ triads (breast, colon, lung, liver/HCC, ovary), including a Xenium-5K panel for a scaling test — on the order of hundreds of thousands of cells per section. Preprocessing: niche-level spatial binning, serial-section rigid registration into a shared frame, panel-gene intersection, and one-time cached UNI/Virchow2 tile embeddings.

4. Evaluation Metrics

- **Selective-risk–coverage curve** (selective risk vs. retained fraction) — headline.
- **Separation gap:** AUROC(deferral↔high-U) − AUROC(deferral↔dropout), ≥ 0.10, holding under two independent predictor architectures.
- **Marginal coverage** within each Mondrian (gene×region) group, within ±2% of nominal $1-\alpha$.
- **Noise-floor clearing:** fraction of panel genes with reproducible signal above the replicate floor. **Retained-fraction × utility** ≥ 15%. **Cross-panel scaling gap** (v1 → Xenium-5K).

5. Compute Requirements

Frozen encoders + pre-cached embeddings + light heads + CPU-side conformal → comfortably over-provisioned: training fits in 24–48 GB GPU memory (192 GB available). The only meaningful cost is a one-time WSI tiling + embedding-extraction pass; downstream training and conformal run in hours, not days. Dependencies: PyTorch, scanpy/anndata, `timm` (UNI/Virchow2), MAPIE/crepes, numpy/scipy/scikit-learn. No distributed training or special hardware.

6. Expected Deliverables

- Working, documented oracle implementation with a green gate/test suite (public repository already live).
- A reusable **benchmark** for biological-vs-technical separation on H&E→ST, released with the code.
- Selective-risk–coverage results across the organ panel + a panel-scaling ablation.
- A short paper targeting **WACV 2026**, pre-registered and fully reproducible (per-run git SHA + config hash).

7. Risk & Mitigation

- **Signal-vs-noise floor.** Addressed by a **solved pre-flight**: a cheap, no-training control on the real breast data already shows the technical-replicate noise floor clears at niche resolution, with biologically coherent markers (ESR1, keratins, SFRP1) leading. The staged design gates all downstream work behind this control.
- **Serial-section registration.** Handled at niche (not single-cell) resolution, using 10x-provided H&E↔Xenium homographies plus a built-in perturbation-sensitivity check; residual error stays in the conservative, signal-shrinking direction.
- **Scope & timeline.** Staged with a guaranteed fallback — a feasibility-vehicle result stands alone as a complete contribution, with the full conformal oracle layered on top, so a shippable deliverable always exists within two weeks.
- **Engineering quality.** Version-controlled, test-gated, and already passed an independent multi-model code review — engineering risk to the timeline is low.

Public data (GEO GSE243280). Code & pre-registration: github.com/aayanAW/unidentifiability-oracle-he-st