PATCH: Panel-Aware Hierarchical Conformal Cell Typing for Spatial Proteomics under Marker-Panel Shift

Tianhao Luo [1] **Bao Li** [1] **Kun-Hsing Yu** [1]

Abstract

Cell phenotyping in spatial proteomics converts multiplexed tissue images into cell-type maps used to study tissue microenvironments, discover spatial biomarkers, and support patient stratification. It is hindered by cross-cohort shift in platforms and marker panels, and by the labor and subjectivity of manual gating. We present PATCH, a deployment workflow with three components: a panel-aware hierarchical conformal layer, a label-free cohort-level shift diagnostic, and Vision Language Model (VLM) quality control. The conformal layer removes classes unsupported by the deployment panel and backs predictions off to the finest ontology ancestor identifiable from the observed markers; under exchangeability on a fixed panel, it inherits the standard split-conformal guarantee. Treating cell typing as a monitored decision pipeline, PATCH turns mean per-marker Kolmogorov–Smirnov distance, computed on unlabeled deployment cells, into a label-free pre-deployment triage signal: it is negatively correlated with cross-cohort coverage and flags cohorts likely to need in-platform recalibration before any labels arrive. The VLM is applied only to ambiguous cells and may nominate one supported label for review, abstain, or flag artifacts; the original conformal set is retained and no unsupported label is added. On a competitive backbone (an XGBoost head at 69.9% flat accuracy on the matched-input CRC30–40 benchmark), PATCH attains 0.98 empirical ontology coverage at target 0.90 with mean set size $|\mathcal{S}| = 2.44$ on $200{,}000$ held-out colorectal cells. Across held-out public cohorts, coverage decreases with mean per-marker KS distance; in a seven-cohort analysis, the correlation is $r = -0.85$ ($p = 0.015$). For flagged cohorts, recalibration is two-tier: appending $5{,}000$ pooled in-platform labeled cells restores target coverage on part of the flagged set, while at a fixed 500-cell budget, conditioning the conformal threshold on cohort identity rather than pooling shrinks the mean coverage gap from 0.120 to 0.009. Aggregated to the patient level, the calibrated cell-type proportions remain predictive in an exploratory single-cohort downstream analysis, closing the loop from per-cell uncertainty to patient-level summaries.

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. Correspondence to: Kun-Hsing Yu <Kun-Hsing_Yu@hms.harvard.edu>.

Accepted to the ICML 2026 Workshop on Statistical Frameworks for Uncertainty in Agentic Systems, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

1. Introduction

Cell phenotyping is a foundational step in spatial proteomics: it converts multiplexed tissue images into cell-type maps that link tissue architecture to patient outcomes (Schürch et al., 2020; Lin et al., 2023). In practice, it faces two distinct bottlenecks. The first is *distributional and technical shift* across samples, cohorts, platforms, and marker panels: marker intensities are affected by batch effects, acquisition technology, and staining scale, and different datasets often measure different subsets of proteins, so a model trained on one cohort may be deployed on cells whose defining markers were never measured (Shaban et al., 2024; Sun et al., 2025). The second bottleneck is the labor and subjectivity of *manual annotation*. In the standard workflow, experts threshold each marker at the sample level, convert markers into present-or-absent calls for each cell, and assign cell types from the resulting marker-positivity patterns (Zhang et al., 2022; Xiong et al., 2024). In such settings, the appropriate prediction is often not a fine cell subtype, but a calibrated set or lineage-level label that reflects what the observed panel can support.

Existing approaches address parts of this problem but not the full deployment setting: manual gating, clustering, and supervised, deep-learning, or foundation-model cell typers (Shaban et al., 2024; Brbić et al., 2022; Shaban et al., 2025) return hard labels or uncalibrated scores and can overcommit when distinguishing markers are absent at deployment, while off-the-shelf conformal prediction (Vovk et al., 2005; Romano et al., 2020) guarantees coverage under exchangeability but does not account for panel-dependent

identifiability (we expand this comparison in Section 2). None directly addresses a common practical question in spatial proteomics: when a new cohort arrives with a partially mismatched panel and no labeled cells yet, should an existing cell-typing pipeline be trusted, or should the cohort first be re-annotated and recalibrated?

PATCH is a proof-of-concept workflow for this setting. At the cell-label level, it adds a panel-aware hierarchical conformal layer on top of a statistical cell-typing pipeline. PATCH treats cell typing as a staged decision pipeline with explicit, uncertainty-gated decision points at the panel, cohort, and cell levels. Classes whose required markers are absent are removed from the candidate set, and unresolved leaf predictions are backed off to the finest ontology ancestor supported by the observed markers. At the cohort level, PATCH computes mean per-marker Kolmogorov–Smirnov (KS) distance on unlabeled deployment cells as a cohort-level empirical triage statistic for whether the existing calibration is likely to transfer or whether in-platform recalibration is needed. At the cell level, an optional vision–language model (VLM) is applied only to routed ambiguous cases as a quality-control (QC) step. The VLM operates inside the conformal support and may nominate one supported label for review, abstain, or flag artifacts; the original conformal set is retained and no unsupported label is introduced.

This paper makes three contributions. First, we formalize panel-aware hierarchical conformal prediction (Section 4) for deployment under marker-panel shift. Second, we evaluate a label-free shift diagnostic and recalibration protocol across an in-house colorectal cohort and multiple public datasets (Section 5.2), showing that mean per-marker KS distance is a useful empirical predictor of cross-cohort coverage degradation and that modest in-platform recalibration can recover target coverage on some flagged cohorts. Third, we study a constrained VLM-based QC step (Section 5.3) that reviews ambiguous cells after conformalization while preserving the underlying statistical support set. Together, these components instantiate a monitored, uncertainty-aware decision pipeline in which a label-free cohort-level diagnostic flags whether the existing calibration is likely to transfer or needs refreshing, the panel-aware hierarchical conformal layer controls per-cell resolution, and a constrained VLM audits the residual ambiguous cases, reducing silent deployment failure under panel shift without repeated cohort-specific manual re-annotation.

2. Related Work

Cell typing for multiplexed imaging. Cell typing in multiplexed imaging has traditionally relied on manual gating or clustering (PhenoGraph, Louvain, FlowSOM (Levine et al., 2015; Traag et al., 2019; Van Gassen et al., 2015)), which need post-hoc annotation and give no calibrated uncertainty. Recent supervised and foundation-model approaches (MAPS, STELLAR, foundation probes (Shaban et al., 2024; Brbić et al., 2022; Shaban et al., 2025); spatial-prior methods such as CELESTA (Zhang et al., 2022)) improve point accuracy but return hard labels or uncalibrated scores and can overcommit when distinguishing markers are absent. None combines set-valued uncertainty, panel-aware backoff, cohort-level deployment triage, and a constrained QC step for ambiguous cells.

Distribution-free uncertainty quantification. Split-conformal prediction (Vovk et al., 2005; Lei et al., 2018) gives marginal coverage under exchangeability, while adaptive prediction sets (Romano et al., 2020) target approximate conditional coverage and RAPS (Angelopoulos et al., 2021) regularize set size. Mondrian conformal (Vovk et al., 2003) conditions on observed strata, including the class label for class-conditional coverage, and broader frameworks such as learn-then-test (Angelopoulos et al., 2025) and conformal risk control (Angelopoulos et al., 2024) extend calibration ideas beyond set prediction. Related work has studied conformal confidence sets for structured multiclass and multilabel outputs (Cauchois et al., 2021). Our setting differs in two ways: the deployment panel is itself an observed conditioning variable, and the set of admissible fine labels changes with that panel.

Uncertainty under spatial or structured shift. Conformal ideas have also been applied to structured biological data: TISSUE (Sun et al., 2024) stratifies spatial-transcriptomic imputation by local neighborhood, and UTOPIA (Jin et al., 2026) adds multiscale uncertainty for virtual spatial transcriptomics. These are regression rather than hierarchical-classification settings, but share the theme that calibration may need to respect structured covariates; here that covariate is the deployment panel, which determines which labels are identifiable.

Selective prediction and second-stage adjudication. A second model can abstain, defer, or refine uncertain predictions, as in selective classification (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2019) and LLM-as-judge evaluation (Saad-Falcon et al., 2024). We use this conservatively: the VLM is a post-conformalization QC step restricted to the conformal support—it may only return a label within that support, abstain, or flag an artifact, and the original set is retained.

3. The PATCH Workflow

Overview. PATCH is a deployment workflow rather than a single classifier. The statistical pipeline converts raw marker measurements into panel-aware features and ontology probabilities. The hierarchical conformal layer then turns those

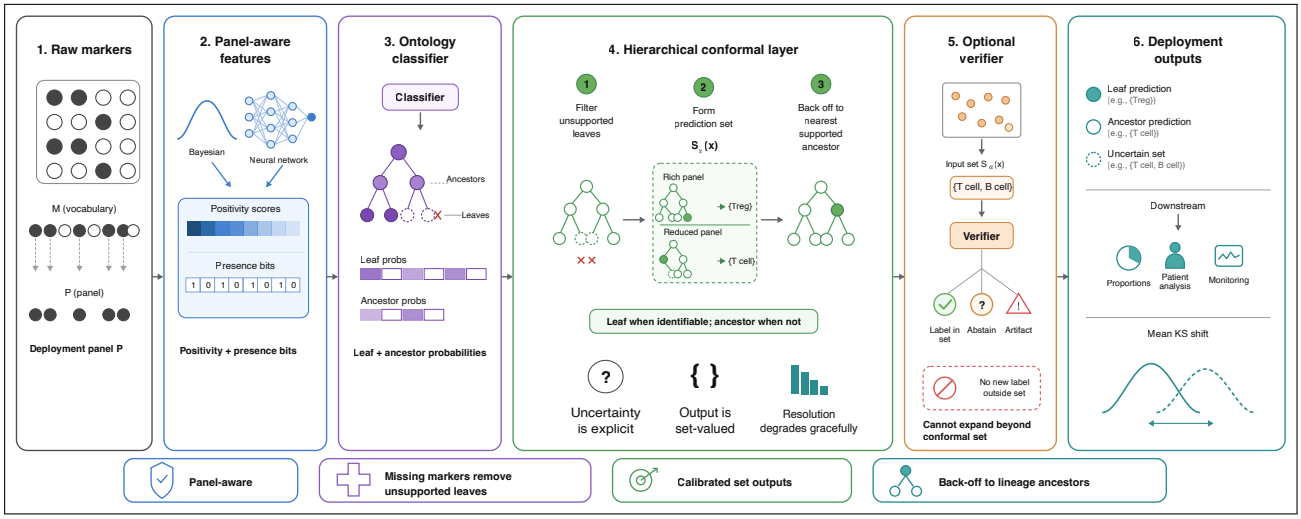

Figure 1. **PATCH workflow.** Raw marker intensities are mapped to a universal marker vocabulary and converted into positivity scores with explicit presence bits. An ontology classifier predicts probabilities over leaves and internal nodes. A panel-aware hierarchical conformal layer removes unsupported leaves and backs unresolved cells off to the finest ancestor supported by the deployment panel. A cohort-level KS diagnostic flags shifted cohorts for recalibration, and an optional VLM QC step reviews ambiguous cells without introducing labels outside the conformal support.

probabilities into calibrated set-valued predictions whose resolution depends on the observed deployment panel. In parallel, a cohort-level KS diagnostic is computed on unlabeled deployment cells to assess whether the calibration cohort is close enough for zero-shot deployment or whether in-platform recalibration is likely to be needed. Finally, an optional VLM is used as a quality-control step on cells whose conformal set remains ambiguous. Figure 1 summarizes these three levels of uncertainty handling.

Notation. Let $\mathcal{M}$ denote a fixed 40-marker universal vocabulary, defined as the union of marker panels across the training datasets. A deployment panel $P \subseteq \mathcal{M}$ is specified per cohort, and every cell in that cohort shares the same panel. A cell i carries raw marker intensities $x_i \in \mathbb{R}^{|P|}$. The label y_i takes values in a 26-node immunology hierarchy $\mathcal{T}$ with 19 leaves and 7 internal lineage nodes. For each class $c \in \mathcal{T}$ we specify a set of required markers $R(c) \subseteq \mathcal{M}$, curated from the published gating definitions of the constituent datasets. A class is identifiable under panel P if $R(c) \subseteq P$.

Stages 1–2: raw markers to panel-aware features. For each observed marker, a per-region Gaussian mixture model (GMM) converts raw intensities into positivity scores. When the fitted mixture separates background and signal components, we use the posterior probability of the signal component:

$$p_{i,m}^{\mathrm{gmm}} = \frac{\pi_s \mathcal{N}(x_{i,m}; \mu_s, \sigma_s^2)}{\pi_b \mathcal{N}(x_{i,m}; \mu_b, \sigma_b^2) + \pi_s \mathcal{N}(x_{i,m}; \mu_s, \sigma_s^2)}. \quad (1)$$

When the marker distribution is effectively unimodal, we fall back to a monotone bounded score that preserves within-region intensity rank but is not interpreted as a calibrated posterior. We also include a complementary positivity estimate from Nimbus, a multi-platform U-Net trained on the Pan-Multiplex (Pan-M) dataset (Rumberger et al., 2025). Nimbus operates on raw multiplexed images together with cell segmentation masks, not on the GMM scores. In this study, Nimbus output is available for the two in-house cohorts, Orion high-plex IF and Pheno CODEX; for these cohorts, $p_{i,m}^{\mathrm{nim}}$ is the per-pixel positivity score aggregated over each cell's segmentation mask. For the four external cohorts (Hoch IMC, IMMUcan IMC, Schürch CODEX, and Liu MIBI), Nimbus output is not available, so we set the corresponding Nimbus score to 0.5 and the presence bit to 0.

Each cell is represented by

$$\begin{aligned}\phi_i &= \left[p_i^{\mathrm{gmm}},\ s_i^{\mathrm{gmm}},\ p_i^{\mathrm{nim}},\ s_i^{\mathrm{nim}}\right] \\ &\in [0,1]^{40} \times \{0,1\}^{40} \times [0,1]^{40} \times \{0,1\}^{40}.\end{aligned} \quad (2)$$

where the binary presence bits s distinguish missing markers from observed markers whose estimated positivity is genuinely uncertain. For markers or channels not available under the deployment setting, we set the corresponding score to 0.5 and the presence bit to 0. This distinction is important under panel shift because a missing marker or missing auxiliary channel should route differently from an observed feature with borderline value.

Stage 3: ontology classifier. An XGBoost classifier (Chen & Guestrin, 2016) maps ϕ_i to probabilities over all 26 ontology nodes, including both leaves and internal lineage nodes. Internal nodes are first-class outputs rather

than post-hoc aggregates because the deployment setting often requires a biologically meaningful coarse prediction rather than a forced leaf decision. To expose the classifier to missing-marker configurations seen at deployment, training augments each cell with randomly masked copies in which selected markers are reset to the missing-marker state and the label is rolled up to the finest ancestor still identifiable under the masked panel.

Stage 4: panel-aware hierarchical conformal layer. Given ontology probabilities, the conformal layer performs three steps that correspond to the center panel of Figure 1. First, leaf classes unsupported by the deployment panel are removed from consideration. Second, a class-conditional conformal set is formed. Third, unresolved fine labels are mapped to the nearest ancestor supported by the observed panel. The result is a calibrated set-valued output in which leaf labels are reported only when they are identifiable; otherwise, the prediction degrades to a lineage-level node. The formal construction and fixed-panel coverage result are given in Section 4.

Stage 5: optional VLM quality control. The VLM is used only after conformalization and only for cells whose prediction set remains ambiguous, typically when $|\mathcal{S}| \geq 2$. It receives the candidate set together with marker tiles, lineage composites, and local tissue context, and emits a verdict in $\mathcal{S} \cup \{\texttt{abstain}, \texttt{artifact}\}$, that is, either an element of the conformal set or one of two refusal tokens. Because the VLM is not allowed to output a new label outside $\mathcal{S}$, it functions as a quality-control and adjudication step rather than a competing classifier. In deployment and evaluation, we retain the original conformal set alongside any VLM verdict, so the statistical guarantee remains attached to the conformal set.

Stage 6: deployment outputs and cohort-level triage. The statistical output of PATCH is a set-valued cell-level prediction that may contain leaves and/or internal nodes. These outputs can be aggregated downstream into per-patient or per-region summaries such as cell-type proportions. Separately, mean per-marker KS distance is computed once per deployment cohort on unlabeled cells by comparing shared-marker post-GMM posterior distributions between the calibration cohort and the new cohort. For calibration cohort C and deployment cohort D, with shared markers $P_C \cap P_D$, let $\hat{F}_{C,m}$ and $\hat{F}_{D,m}$ denote the empirical CDFs of $p^{\text{gmm}}_{i,m}$ for marker m on unlabeled cells. We define

$$\mathrm{KS}_{\text{mean}}(C, D) = \frac{1}{|P_C \cap P_D|} \sum_{m \in P_C \cap P_D} \sup_t \left| \hat{F}_{C,m}(t) - \hat{F}_{D,m}(t) \right|. \tag{3}$$

This statistic is not part of the per-cell conformal score. Instead, it is used before deployment as a cohort-level empirical triage signal for whether the existing calibration is likely to transfer or whether in-platform labeled cells are likely to be needed.

4. Panel-Aware Hierarchical Conformal

We describe PATCH at two levels. First, we give a clean fixed-panel formulation that captures the intended panel-aware back-off mechanism and admits a simple coverage statement. We then describe the multi-level thresholding procedure used in the experiments, which follows the same intuition but is implemented with separate calibrated thresholds at multiple ontology levels.

Notation for the hierarchy. Let $\mathcal{T}$ denote the 26-node ontology used throughout the paper, including both leaves and internal lineage nodes. Let $\mathcal{L} \subset \mathcal{T}$ denote the leaf labels. The ground-truth cell type y is assumed to be a leaf, so $y \in \mathcal{L}$. For any node $c \in \mathcal{T}$, let $\mathrm{Anc}(c) \subseteq \mathcal{T}$ denote the set of ancestors of c, optionally including c itself when convenient. The classifier outputs a score $\hat{p}(c \mid \phi)$ for every ontology node $c \in \mathcal{T}$, where ϕ is the feature vector of the cell. For a deployment panel $P \subseteq \mathcal{M}$, recall that $R(c) \subseteq \mathcal{M}$ denotes the set of markers required to identify class c, and that c is identifiable under P if $R(c) \subseteq P$.

Idealized fixed-panel formulation. Let $\mathcal{D}_{\text{cal}} = \{(\phi_i, y_i)\}_{i=1}^n$ be a held-out calibration set and define nonconformity scores

$$s_i = 1 - \hat{p}(y_i \mid \phi_i).$$

For each leaf class $c \in \mathcal{L}$, let $n_c = |\{i : y_i = c\}|$ and let $\hat{q}_c^{(\alpha)}$ be the $\lceil (1-\alpha)(n_c+1) \rceil / n_c$ empirical quantile of $\{s_i : y_i = c\}$. The class-conditional split-conformal set is

$$\mathcal{S}_\alpha(\phi) \;=\; \{c \in \mathcal{L} : 1 - \hat{p}(c \mid \phi) \leq \hat{q}_c^{(\alpha)}\}. \tag{4}$$

Thus, the baseline conformal set is formed over leaves, since calibration labels are leaf labels. We assume each calibrated leaf has at least one calibration example, i.e. $n_c \geq 1$ for all calibrated c; in practice, leaves without calibration support are excluded from class-conditional calibration or merged to a coarser calibrated node.

Panel filter. When the deployment panel P omits markers needed to identify some leaves, those leaves should not remain as candidates. We therefore restrict the candidate set to the panel-identifiable subset

$$\mathcal{Y}_P \;=\; \{c \in \mathcal{L} : R(c) \subseteq P\}, \qquad \mathcal{S}_\alpha^P(\phi) \;=\; \mathcal{S}_\alpha(\phi) \cap \mathcal{Y}_P. \tag{5}$$

A naive filtered set $\mathcal{S}_\alpha(\phi) \cap \mathcal{Y}_P$ can discard an unsupported true leaf without replacing it. The guarantee-preserving construction in Equation (6) therefore maps every selected leaf to its finest supported ancestor via π_P.

Hierarchical back-off. For any leaf class $c \in \mathcal{L}$, let $\pi_P(c)$ denote the finest node in $\mathcal{T}$ on the path from c to the root whose required markers are contained in P. In other words, $\pi_P(c)$ is the most specific ontology node supported by the deployment panel for leaf c. The idealized panel-aware hierarchical prediction set is

$$\mathcal{S}_\alpha^{P,\mathcal{T}}(\phi) \;=\; \big\{\pi_P(c) : c \in \mathcal{L},\; 1 - \hat{p}(c \mid \phi) \le \hat{q}_c^{(\alpha)}\big\}. \quad (6)$$

Operationally, this means that leaves are reported only when the deployment panel supports them; otherwise the prediction backs off to the finest supported lineage node.

Proposition 4.1 (Idealized fixed-panel ontology coverage)**.** *Assume calibration and deployment cells are exchangeable on a common panel P. Then for each leaf class $c \in \mathcal{L}$,*

$$\Pr\big[\pi_P(y) \in \mathcal{S}_\alpha^{P,\mathcal{T}}(\phi) \;\big|\; y = c\big] \;\ge\; 1 - \alpha.$$

Proof. Let $(\phi_i, y_i)_{i=1}^{n+1}$ be exchangeable on a common panel P, and let $\hat{q}_c^{(\alpha)}$ denote the class-conditional split-conformal threshold computed on $(\phi_i, y_i)_{i=1}^{n}$. Let $\mathcal{S}_\alpha$ denote the unfiltered split-conformal set of Equation (4). By the standard class-conditional split-conformal argument for Mondrian stratification on the class label (Vovk et al., 2003; Lei et al., 2018),

$$\Pr\big[y_{n+1} \in \mathcal{S}_\alpha(\phi_{n+1}) \mid y_{n+1} = c\big] \ge 1 - \alpha.$$

The hierarchical set $\mathcal{S}_\alpha^{P,\mathcal{T}}$ of Equation (6) is the image of $\mathcal{S}_\alpha$ under the deterministic mapping π_P. Therefore, whenever $y_{n+1} \in \mathcal{S}_\alpha(\phi_{n+1})$, we also have $\pi_P(y_{n+1}) \in \mathcal{S}_\alpha^{P,\mathcal{T}}(\phi_{n+1})$. This gives

$$\Pr\big[\pi_P(y_{n+1}) \in \mathcal{S}_\alpha^{P,\mathcal{T}}(\phi_{n+1}) \mid y_{n+1} = c\big] \ge 1 - \alpha.$$

□

Implementation used in experiments. The experimental pipeline uses a closely related but not identical *set-valued* procedure. Instead of computing a single leaf-level conformal set and then taking its image under π_P, the code aggregates leaf probabilities into ontology-level probabilities, calibrates separate class-conditional thresholds at multiple ontology levels (lineage, subtype, and fine), and returns the set of deepest supported nodes whose level-specific thresholds are satisfied. Ancestor nodes are pruned whenever a more specific returned descendant is already present. This preserves the panel filter $\mathcal{Y}_P$ and the finest-supported-ancestor logic π_P, while allowing the implementation to operate directly on multiple ontology levels.

Ontology coverage in the experiments. In the experiments, the returned output is a set $\mathcal{S}_{\mathrm{impl}}(\phi) \subseteq \mathcal{T}$, and ontology-accurate coverage is defined operationally as

$$\mathcal{S}_{\mathrm{impl}}(\phi) \cap \mathrm{Anc}(y) \neq \varnothing,$$

that is, at least one returned node is the true leaf or one of its ancestors. This is exactly the quantity evaluated by the implemented system. Theorem 4.1 applies to the idealized fixed-panel construction above and formalizes the intended panel-aware back-off mechanism. The ontology-coverage curves reported in Section 5.2 for the multi-level thresholding implementation are therefore empirical rather than a direct corollary of the proposition.

Scope of the guarantee. Even for the idealized construction, Theorem 4.1 is a fixed-panel statement. When calibration and deployment differ by platform, cohort composition, or label definition, exchangeability no longer holds, so cross-cohort coverage is empirical rather than guaranteed. The experiments therefore ask two separate questions: whether panel-aware back-off behaves sensibly at the cell level, and whether a cohort-level diagnostic can warn when recalibration is needed (Table 3 maps which objects are guaranteed versus empirical).

This construction can be viewed as conditioning on an observable panel variable, in the spirit of Mondrian conformal (Vovk et al., 2003), but without recomputing panel-specific quantiles at deployment.

5. Experiments

Datasets and setup. In-domain experiments use the Orion high-plex IF colorectal cancer cohort of Lin et al. (2023) (CRC01–40; train CRC01–15, calibration CRC16–29, test CRC30–40). Cross-dataset evaluation uses four held-out public datasets: Schürch CRC CODEX (Schürch et al., 2020), Hoch melanoma IMC (Hoch et al., 2022), IMMUcan multi-cancer IMC (Schulz et al., 2024), and Liu multi-tissue MIBI (Liu et al., 2023). In each leave-one-dataset-out (LODO) run, training uses CRC01–15 together with all external datasets except the held-out one; calibration always uses CRC16–29 only. Panel sizes and overlaps with the 40-marker vocabulary are reported in Section A. All experiments target $\alpha = 0.10$ with fixed hyperparameters across LODO variants. For context, the underlying XGBoost head reaches 69.9% flat accuracy and 57.5% macro-F1 on the matched-input CRC30–40 benchmark; these point-prediction metrics are computed on the 11-class harmonized leaf task—the coarse leaf label set shared by all point-prediction baselines, used so that the methods are compared on matched inputs—whereas set-valued and ontology coverage are evaluated on the full 26-node hierarchy of Section 3 (19 leaves and 7 internal nodes). Comparative

point-prediction baselines are in Section K. The four LODO held-outs are our primary held-out design; we extend to n=7 cohorts only to probe the robustness of the shift diagnostic, while the eight-cohort fixed-model transfer (Section G) and the Sorin LUAD cross-disease (negative) control (Section I) are stress-tests rather than new claims (Table 1).

Table 1. Cohort roster. The four LODO held-outs are the primary design; the n=7 KS set and n=8 transfer are robustness stress-tests, and Sorin LUAD an independent cross-disease (negative) control.

Role	Cohorts
In-domain	Orion CRC (train/cal/test)
LODO held-out	Hoch, Schürch, IMMUcan, Liu
KS set (n=7)	+ Tietscher, Hartmann, MAPS
Cross-cohort (n=8)	+ Cords
Downstream	Orion; Sorin LUAD

5.1. In-domain coverage

On CRC30–40 (∼200K cells), the full in-domain pipeline attains empirical ontology coverage $\mathbf{0.98}$ at target 0.90 with mean set size $|\mathcal{S}| = \mathbf{2.44}$ and $\mathbf{52.0}\%$ leaf-resolved predictions. On the CRC 16-marker panel, the panel filter removes 9 of 26 ontology nodes that are not identifiable (unsupported leaves together with the internal nodes left without a supported descendant), so unresolved cells back off to lineage-level nodes rather than unsupported leaves. Against two recently published hierarchical-conformal methods (HCC and scConform), PATCH is the only one simultaneously calibrated and resolved—committing many cells to leaves rather than ballooning to whole lineages—on the matched-panel benchmark (Section F). Under fixed-model cross-platform transfer this advantage is *conditional*: PATCH does not dominate on mean coverage across the eight transfer cohorts (0.68 versus HCC's 0.72) and under-covers on the high-shift cohorts the KS diagnostic flags, but on cohorts where the transferred base model stays well-ordered it converts scConform's whole-lineage sets into covered sets roughly half as large (Section G). Figure 1 summarizes the end-to-end deployment workflow and the three points at which uncertainty is handled: panel-aware back-off, cohort-level shift triage, and optional cell-level QC. A qualitative spatial cell-type map for one held-out region is shown in Figure 5.

5.2. Cross-cohort deployment, shift triage, and recalibration

To stress-test deployment, we retrain four times, each time removing one external dataset D entirely from the training pool. Each LODO model is calibrated on CRC16–29 at $\alpha = 0.10$ and evaluated on the held-out dataset. Figure 2 and Table 2 (top block) shows that ontology coverage varies sharply across the four principled LODO held-outs, from 0.80 on Liu MIBI to 0.26 on Hoch IMC. This variation is not explained by platform label alone. Instead, it tracks a label-free covariate computed before any labels arrive: mean per-marker KS distance between the calibration cohort and the new deployment cohort, measured on shared-marker post-GMM posterior distributions.

Table 2. Ontology coverage and mean set size at $\alpha = 0.10$. The first row is the in-domain CRC reference, the next four rows are the LODO held-outs evaluated under their own holdout models, and the last three rows are additional external cohorts used to extend the cohort-level KS diagnostic to $n = 7$.

Dataset	Platform	Cov.	$\lvert\mathcal{S}\rvert$	Leaf-res %
(in-domain)	Orion	0.98	2.44	52.0
Liu	Multi-tiss. MIBI	0.80	2.70	17.8
Schürch	CRC CODEX	0.49	2.47	16.0
IMMUcan	Multi-canc. IMC	0.39	1.36	76.2
Hoch	Melanoma IMC	0.26	2.01	18.7
Tietscher	Breast IMC	0.79	1.20	81.3
Hartmann	CRC MIBI	0.59	1.96	26.5
MAPS	Hodgkin CODEX	0.58	1.97	34.8

To stress-test the diagnostic itself, we additionally evaluate on three independent public cohorts not part of the LODO protocol: Tietscher breast IMC cohort (Tietscher et al., 2023), Hartmann colorectal MIBI cohort (Hartmann et al., 2021), and MAPS Hodgkin lymphoma CODEX (Shaban et al., 2024). These three cohorts are evaluated under the Hoch-holdout model; none was present in any training pool, so they serve as additional zero-shot external cohorts for the KS diagnostic. Across all $n = 7$ **external cohorts** spanning three platforms (CODEX (Goltsev et al., 2018), IMC (Giesen et al., 2014), MIBI (Angelo et al., 2014)) and a diverse range of tissues—colorectal (Schürch, Hartmann), melanoma (Hoch), Hodgkin lymphoma (MAPS), breast (Tietscher), multi-tissue (Liu), and multi-cancer (IMMUcan)—mean KS is negatively correlated with ontology coverage, with **Pearson** $r = -0.85$ ($p = 0.015$) and **Spearman** $\rho = -0.89$ ($p = 0.007$) (Figure 3). The trend is stable to dropping any single cohort (leave-one-out $r \in [-0.90, -0.82]$, all $p < 0.05$). Restricted to the four LODO held-outs, the same relationship remains visible, giving $r = -0.95$ ($p = 0.048$) under the per-cohort shared-marker normalization used in the original LODO analysis. The four-cohort LODO analysis provides the strictest held-out design, while the seven-cohort extension offers supportive evidence that the same negative trend persists across a broader set of tissues and panels; a complementary fixed-model cross-cohort conformal benchmark is reported in Section G.

Operationally, mean KS is used only at the cohort level, not inside the per-cell conformal predictor. Before deployment on a new cohort, we compute it on unlabeled cells as an empirical triage signal for whether in-platform recalibration is likely to be needed. We then test this protocol by appending

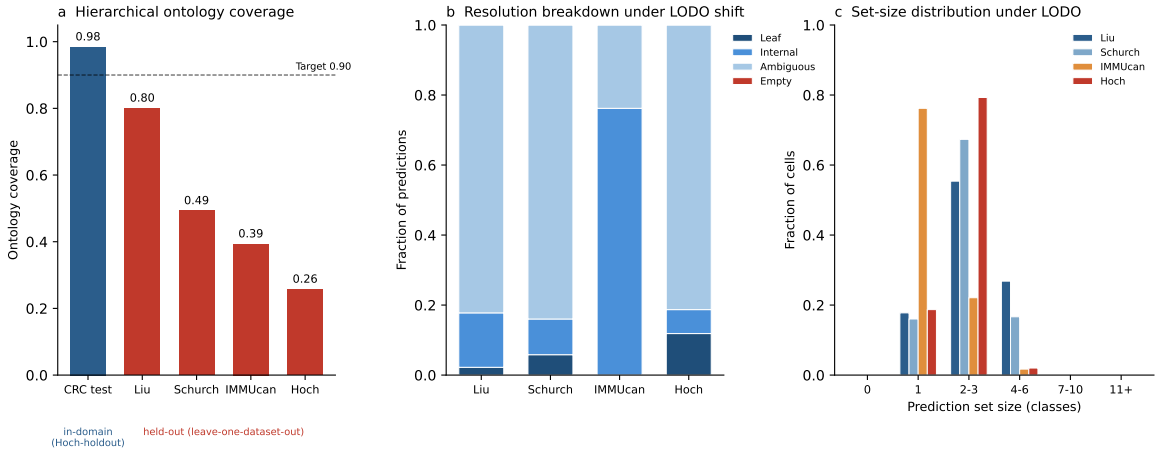

Figure 2. **Cross-cohort coverage and back-off.** (a) Empirical ontology coverage at target 0.90 for in-domain CRC and four held-out public cohorts. (b) Under cross-dataset shift, predictions back off rather than forcing unsupported leaves: most cells are returned as multi-element *ambiguous* sets, and on IMMUcan they resolve predominantly to internal lineage nodes. (c) Under cross-dataset shift, panel-aware prediction-set sizes stay concentrated at small values (mostly 1–3 classes, with a modest 4–6 tail on Liu) rather than inflating to large flat sets.

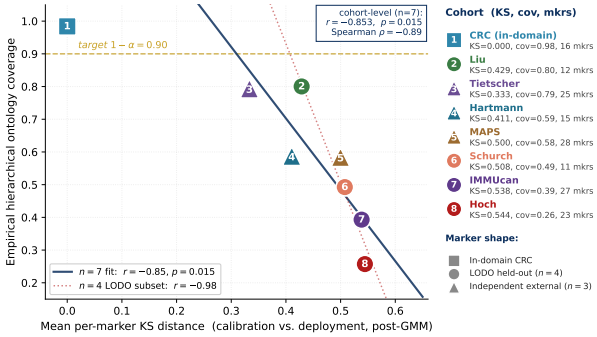

Figure 3. **Label-free pre-deployment shift diagnostic.** Mean per-marker KS distance between the calibration cohort and each deployment cohort versus empirical ontology coverage at target $1 - \alpha = 0.90$. The seven-cohort fit (solid line) gives Pearson $r = -0.85$ ($p = 0.015$) and Spearman $\rho = -0.89$ ($p = 0.007$); the dashed line is the fit restricted to the four LODO held-out cohorts and corresponds to $r = -0.98$ under the unified 40-marker normalization used in this scatter, consistent with the $r = -0.95$ ($p = 0.048$) reported in the main text under the per-cohort shared-marker normalization. The in-domain CRC anchor is shown separately from the seven external cohorts. KS is computed on unlabeled cells and used only at the cohort level.

N_{recal} labeled cells from the deployment cohort to the Orion calibration pool. At $N_{\text{recal}} = 5000$, Hoch coverage recovers from 0.257 to **0.900**, Liu from 0.800 to **0.907**, IMMUcan from 0.393 to 0.688, and Schürch from 0.493 to 0.809. Extended sweeps are reported in Section C. The pattern suggests two regimes under this single pooled threshold appended to the Orion pool: calibration-addressable shift, for which a modest labeled pool restores target coverage, and a residual shortfall on IMMUcan and Schürch that this single Orion-appended threshold does not close. This residual gap is largely a pooled-calibration artifact rather than deep structural mismatch: a per-cohort conditional-conformal threshold restores per-cohort coverage from a comparably small labeled slice (under a separate fixed-transfer setup; Section H), with the largest gain on Schürch. Any remaining mismatch is secondary and may stem from label-definition drift or panel/ontology differences. Extended cohort-level KS results are reported in Section J.

Removing panel-awareness confirms that the filter is not cosmetic. On CRC30–40 with 2 of 16 markers held out at random, ontology coverage drops from 0.89 to 0.82 without the filter; with 4 held-out markers, it drops from 0.87 to 0.68. Extended ablations are reported in Section B.

5.3. Cell-level QC with a constrained VLM

In deployment the VLM is invoked as a QC step only on cells whose conformal set remains ambiguous ($|\mathcal{S}| \geq 2$); it sees marker tiles, lineage composites, a context crop, and the candidate set, and its output is constrained to $\mathcal{S} \cup \{\text{abstain}, \text{artifact}\}$, so it can flag or shrink a set but never add an unsupported label. On a deliberately stratified 100-cell stress-test, four multimodal verifiers span very different

operating points—from Claude Opus 4.7 (artifact recall 0.867, but over-flagging 48% of cells) and the abstaining GPT-5.5, to Qwen3-VL-32B (recall 0.133) and the newer Qwen3.6-35B-A3B, the most aggressive flagger ($59/100$)—confirming the VLM as a conservative auditing layer rather than a replacement classifier. Full per-tier results and the four-verifier comparison are in Section E.

6. Discussion

PATCH decomposes deployment uncertainty in spatial-proteomics cell typing into three cases and addresses each: marker-panel mismatch creates *structural non-identifiability* when defining markers are absent (handled by panel-aware hierarchical conformal prediction); cross-cohort shift can break calibration transfer (flagged by a cohort-level KS diagnostic); and a few cells remain visually or biologically suspicious after statistical prediction (reviewed by a constrained VLM).

The mechanisms are complementary: in-domain, panel-aware back-off preserves high ontology coverage without collapsing cells to coarse labels; under cross-cohort shift, predictions move toward internal ontology nodes rather than forcing unsupported leaves; the cohort-level mean per-marker KS distance is a useful (if empirical, not formally guaranteed) pre-deployment covariate; and a modest labeled pool restores coverage on the calibration-addressable cohorts while others remain limited by deeper panel or label-semantics mismatch. Aggregated to the patient level, the calibrated cell-type proportions remain predictive in an exploratory single-cohort survival analysis that does not replicate in an independent lung cohort (Section I).

The VLM is similarly conservative: it neither provides the guarantee nor replaces the classifier, but audits routed ambiguous cells—surfacing likely artifacts and optionally nominating one supported label—while leaving the conformal set intact, so review stays within the statistically supported label space.

Limitations and future work. The formal guarantee holds for an idealized fixed-panel construction and does not automatically extend to cross-platform deployment, so the implemented multi-level system should be read empirically; the KS diagnostic is likewise empirical, though the seven-cohort and leave-one-cohort-out analyses support it as a useful covariate. The VLM audit is a small proof of concept; a local domain-specific VLM is a natural next step. The present system is *monitored* rather than autonomous (fixed gates, a single stateless VLM call); the agentic extension we see as most promising is a learned routing policy that decides *when* to back off, recalibrate, or invoke review from the uncertainty signals quantified here. More broadly, the study covers limited panels, cohorts, and tissues, and ablations of the Nimbus and GMM branches under shift are left to future work.

Reproducibility. The PATCH code is released at `https://github.com/JSluo888/PATCH`. The LODO artifacts, calibration tables, held-out predictions, and VLM QC example bundle are available on request; the private in-house datasets cannot be shared (IRB and data-governance restrictions).

Impact Statement

This work targets cell phenotyping in spatial proteomics, a step often performed with hand-tuned marker thresholds that can fail silently under panel or platform shift. A panel-aware, set-valued workflow can improve robustness by making uncertainty explicit, backing off to lineage-level predictions when fine labels are not identifiable, and flagging shifted cohorts for recalibration rather than forcing unsupported labels.

The method is not a replacement for expert review: the fixed-panel conformal guarantee is population-level under exchangeability, cross-cohort performance remains empirical and can degrade under shift, and the VLM module is only an auxiliary auditor for ambiguous or artifactual cells. The intended use is research and translational analysis, not direct clinical decision-making; we hope this style of workflow reduces silent deployment failure and makes uncertainty handling in spatial proteomics more explicit and auditable.

Generative AI Statement

Coding agents (Codex, Claude Code) assisted with code and data-processing scripts, and large language models helped revise wording; all scientific ideas, experimental design, analyses, and reported results are the authors' own.

Acknowledgments

T.L. was supported by an NHGRI training grant (T32HG002295) and an NSF Graduate Research Fellowship. K.-H.Y. is supported in part by NIGMS R35GM142879, NHLBI R01HL174679, the DoD Peer Reviewed Cancer Research Program Career Development Award HT9425-23-1-0523, the American Cancer Society Research Scholar Grant RSG-24-1253761-01-ESED, and the Harvard Medical School Dean's Innovation Award.

References

Angelo, M., Bendall, S. C., Finck, R., Hale, M. B., Hitzman, C., Borowsky, A. D., Levenson, R. M., Lowe, J. B., Liu, S. D., Zhao, S., et al. Multiplexed ion beam imaging of

human breast tumors. *Nature Medicine*, 20(4):436–442, 2014.

Angelopoulos, A. N., Bates, S., Malik, J., and Jordan, M. I. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations (ICLR)*, 2021.

Angelopoulos, A. N., Bates, S., Fisch, A., Lei, L., and Schuster, T. Conformal risk control. In *International Conference on Learning Representations (ICLR)*, 2024.

Angelopoulos, A. N., Bates, S., Candès, E. J., Jordan, M. I., and Lei, L. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998.

Brbić, M., Cao, K., Hickey, J. W., Tan, Y., Snyder, M. P., Nolan, G. P., and Leskovec, J. Annotation of spatially resolved single-cell data with STELLAR. *Nature Methods*, 19:1411–1418, 2022. doi: 10.1038/s41592-022-01651-8.

Cauchois, M., Gupta, S., and Duchi, J. C. Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction. *Journal of Machine Learning Research*, 22(81):1–42, 2021.

Chen, C.-J., Chen, Y., Yun, S., Stanley, N., and Chen, T. Spatial coordinates as a cell language: A multi-sentence framework for imaging mass cytometry analysis. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 13241–13252, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.685.

Chen, T. and Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. ACM, 2016.

Corbetta, D., Finos, L., Geistlinger, L., and Risso, D. Conformal inference for cell type annotation with graph-structured constraints. *arXiv preprint arXiv:2410.23786*, 2024.

Cords, L., Tietscher, S., Anzeneder, T., et al. Cancer-associated fibroblast classification in single-cell and spatial proteomics data. *Nature Communications*, 14:4294, 2023. doi: 10.1038/s41467-023-39762-1.

den Hengst, F., Blin, I., Mohammadi, M., Shah, S. I. H., and Younesian, T. Hierarchical conformal classification. *arXiv preprint arXiv:2508.13288*, 2025.

El-Yaniv, R. and Wiener, Y. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

Geifman, Y. and El-Yaniv, R. SelectiveNet: A deep neural network with an integrated reject option. In *International Conference on Machine Learning (ICML)*, 2019.

Gibbs, I., Cherian, J. J., and Candès, E. J. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B*, 87(4):1100–1126, 2025. doi: 10.1093/jrsssb/qkaf008.

Giesen, C., Wang, H. A., Schapiro, D., Zivanovic, N., Bodenmiller, B., et al. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature Methods*, 11(4):417–422, 2014.

Goltsev, Y., Samusik, N., Kennedy-Darling, J., Bhate, S., Hale, M., Gonzalez, G., and Nolan, G. P. Deep profiling of mouse splenic architecture with CODEX multiplexed imaging. *Cell*, 174(4):968–981, 2018.

Hartmann, F. J., Mrdjen, D., McCaffrey, E., Glass, D. R., Greenwald, N. F., et al. Single-cell metabolic profiling of human cytotoxic T cells. *Nature Biotechnology*, 39: 186–197, 2021.

Hoch, T., Schulz, D., Eling, N., Gómez, J. M., Levesque, M. P., and Bodenmiller, B. Multiplexed imaging mass cytometry of the chemokine milieus in melanoma characterizes features of the response to immunotherapy. *Science Immunology*, 7(70):eabk1692, 2022. doi: 10.1126/sciimmunol.abk1692.

Jin, K., Chen, Z., Yu, X., Yuan, M., Schroeder, A., Dumoulin, B., Liu, Y., Wang, L., Park, J. H., Hwang, T. H., Susztak, K., Ren, Z., Zhang, N. R., and Li, M. Multiscale confidence quantification for virtual spatial transcriptomics with UTOPIA. *bioRxiv*, 2026. doi: 10.64898/2026.03.01.708850. Preprint.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Levine, D., Rizvi, S. A., Lévy, S., Pallikkavaliyaveetil, N., Zhang, D., Chen, X., Ghadermarzi, S., Wu, R., Zheng, Z., Vrkic, I., Zhong, A., Raskin, D., Han, I., De Oliveira Fonseca, A. H., Ortega Caro, J., Karbasi, A., Dhodapkar, R. M., and Dijk, D. V. Cell2sentence: Teaching large language models the language of biology. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 27299–27325. PMLR, 2024.

Levine, J. H., Simonds, E. F., Bendall, S. C., Davis, K. L., Amir, E.-a. D., Tadmor, M. D., Litvin, O., Fienberg, H. G., Jager, A., Zunder, E. R., et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell*, 162(1):184–197, 2015.

Lin, J.-R., Chen, Y.-A., Campton, D., Cooper, J., Coy, S., Yapp, C., Tefft, J. B., McCarty, E., Ligon, K. L., Rodig, S. J., et al. High-plex immunofluorescence imaging and traditional histology of the same tissue section for discovering image-based biomarkers. *Nature Cancer*, 4(7): 1036–1052, 2023.

Liu, C. C., Greenwald, N. F., Kong, A., et al. Robust phenotyping of highly multiplexed tissue imaging data using pixel-level clustering. *Nature Communications*, 14: 4618, 2023. doi: 10.1038/s41467-023-40068-5.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Romano, Y., Sesia, M., and Candès, E. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020.

Rumberger, J. L., Greenwald, N. F., Ranek, J. S., Boonrat, P., et al. Automated classification of cellular expression in multiplexed imaging data with Nimbus. *Nature Methods*, 22:2161–2170, 2025. doi: 10.1038/s41592-025-02826-9.

Saad-Falcon, J., Khattab, O., Potts, C., and Zaharia, M. ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 338–354, 2024. arXiv:2311.09476.

Schulz, D., Eling, N., Palau Fernandez, B., Daniel, M., and Bodenmiller, B. Immucan panel 1 cell type classification, 2024.

Schürch, C. M., Bhate, S. S., Barlow, G. L., Phillips, D. J., Noti, L., Zlobec, I., Chu, P., Black, S., Demeter, J., McIlwain, D. R., et al. Coordinated cellular neighborhoods orchestrate antitumoral immunity at the colorectal cancer invasive front. *Cell*, 182(5):1341–1359.e19, 2020. doi: 10.1016/j.cell.2020.07.005.

Shaban, M., Bai, Y., Qiu, H., Mao, S., Yeung, J., Yeo, Y. Y., Shanmugam, V., Chen, H., Zhu, B., Weirather, J. L., Mahmood, F., et al. MAPS: pathologist-level cell type annotation from tissue images through machine learning. *Nature Communications*, 15:28, 2024. doi: 10.1038/s41467-023-44188-w.

Shaban, M., Chang, Y., Qiu, H., Yeo, Y. Y., Song, A. H., Jaume, G., et al. A foundation model for spatial proteomics. *arXiv preprint arXiv:2506.03373*, 2025.

Sorin, M., Karimi, E., Fiset, B., Desharnais, L., Perus, L. J. M., Milette, S., Yu, M. W., Maritan, S. M., Doré, S., Wei, Y., Rayes, R., Siegel, P. M., Spicer, J. D., Quail, D. F., and Walsh, L. A. Single-cell spatial landscapes of the lung tumour immune microenvironment. *Nature*, 614: 548–554, 2023. doi: 10.1038/s41586-022-05672-3.

Sun, E. D., Ma, R., Navarro Negredo, P., Brunet, A., and Zou, J. TISSUE: uncertainty-calibrated prediction of single-cell spatial transcriptomics improves downstream analyses. *Nature Methods*, 21(3):444–454, 2024. doi: 10.1038/s41592-024-02184-y.

Sun, H., Yu, S., Martinez Casals, A., Bäckström, A., Lu, Y., Lindskog, C., Ruffalo, M., Lundberg, E., and Murphy, R. F. Flexible and robust cell-type annotation for highly multiplexed tissue images. *Cell Systems*, 16(9):101374, 2025. doi: 10.1016/j.cels.2025.101374.

Tietscher, S., Wagner, J., Anzeneder, T., Langwieder, C., Rees, M., Sobottka, B., de Souza, N., and Bodenmiller, B. A comprehensive single-cell map of T cell exhaustion-associated immune environments in human breast cancer. *Nature Communications*, 14:98, 2023. doi: 10.1038/s41467-022-35238-w.

Traag, V. A., Waltman, L., and van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9:5233, 2019.

Van Gassen, S., Callebaut, B., Van Helden, M. J., Lambrecht, B. N., Demeester, P., Dhaene, T., and Saeys, Y. FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry Part A*, 87(7):636–645, 2015.

Vovk, V., Lindsay, D., Nouretdinov, I., and Gammerman, A. Mondrian confidence machine. Technical report, Royal Holloway, University of London, 2003.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic Learning in a Random World*. Springer, 2005.

Xiong, J., Kaur, H., Heiser, C. N., McKinley, E. T., Roland, J. T., Coffey, R. J., Shrubsole, M. J., Wrobel, J., Ma, S., Lau, K. S., and Vandekar, S. Gammagater: semi-automated marker gating for single-cell multiplexed imaging. *Bioinformatics*, 40(6):btae356, June 2024. doi: 10.1093/bioinformatics/btae356.

Zhang, W., Li, I., Reticker-Flynn, N. E., et al. Identification of cell types in multiplexed in situ images by combining protein expression and spatial information using CELESTA. *Nature Methods*, 19:759–769, 2022. doi: 10.1038/s41592-022-01498-z.

Table 3. What is guaranteed versus empirical in PATCH. The fixed-panel coverage guarantee (Theorem 4.1) attaches only to the unpruned leaf-image set of Equation (6); the implemented multi-level display set and all cross-cohort results are empirical.

Object	Role	Guarantee
Unpruned leaf-image set (Equation (6))	statistical support	fixed-panel split-conformal
Multi-level pruned display set	compact output	empirical
Mean per-marker KS (Equation (3))	cohort triage	none (label-free)
Pooled recalibration	labeled repair	empirical
Cohort-conditional conformal	labeled per-cohort repair	holds over $\mathrm{span}(\Phi)$
VLM verdict	QC suggestion	none (constrained to $\mathcal{S}$)

A. Dataset and Panel Details

Panel sizes and shared markers per dataset are summarized in Table 4.

Table 4. Panel overlap with the 40-marker universal vocabulary.

Dataset	Markers in panel	Overlap
CRC Orion	16	16/40
Hoch Melanoma IMC	45	23/40
IMMUcan IMC	39	27/40
Liu MIBI	15	12/40
Schürch CODEX	49	11/40
Hartmann CRC MIBI	36	15/40
MAPS Hodgkin CODEX	49	28/40
Tietscher Breast IMC	41	25/40

B. Panel-Awareness Ablation (extended)

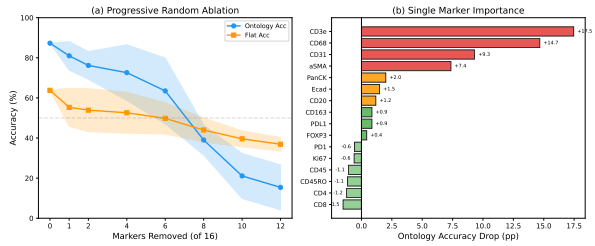

Figure 4. **Extended panel-awareness ablation.** (a) Progressive random marker removal on CRC30–40: fine-grained ontology accuracy degrades faster than coarse flat (leaf) accuracy as distinguishing markers are removed, motivating back-off to coarser, panel-supported predictions when distinguishing markers are missing. The coverage benefit of the panel filter itself is quantified in Section 5.2. (b) Single-marker importance, measured by the drop in ontology accuracy after removing one marker at a time, concentrates on lineage-defining markers (CD3e, CD68, CD31, αSMA) rather than within-lineage discriminators (CD4, CD8, FOXP3, which contribute little or even slightly help): losing the marker that anchors a lineage is what forces back-off to a coarser node, which is exactly the regime the panel filter is built for.

Extended panel-awareness ablation. To test whether the panel-aware design is functionally important rather than cosmetic, we perform two additional ablations. First, we progressively remove markers from the CRC panel at random and track fine-grained ontology accuracy against coarse flat accuracy. As shown in Figure 4a, ontology accuracy degrades faster than flat accuracy as distinguishing markers are removed, motivating back-off to coarser predictions once the panel can no longer support a leaf; the coverage benefit of the panel filter itself—panel-aware versus naive filtering—is quantified in the main text (Section 5.2). Second, we test whether the explicit presence channels are necessary by zeroing them at inference time without retraining. Together, these experiments ask whether PATCH is genuinely using panel availability information to control resolution under shift.

Presence-channel inference ablation. To verify that the explicit presence bits are not merely a re-encoding of the scalar fill value, we zero the presence channels at inference time on each LODO model without retraining. This forces the classifier to treat observed markers as if they were structurally absent. The resulting drop in ontology coverage, reported in Table 5, measures how strongly the classifier relies on the presence channel under panel and platform shift.

Table 5. Presence-channel inference ablation: ontology coverage drop when the presence bits are zeroed at test time without retraining.

Holdout	$\mathrm{cov}_{\mathrm{onto}}^{\mathrm{full}}$	$\mathrm{cov}_{\mathrm{onto}}^{\mathrm{zeroed}}$	Δ
Hoch	0.257	0.259	$+0.002$
Liu	0.800	0.726	-0.074
IMMUcan	0.393	0.204	-0.189
Schürch	0.493	0.456	-0.037

As summarized in Table 5, the largest degradation appears on IMMUcan (-18.9 pp) and Liu (-7.4 pp), showing that the presence channel is a load-bearing feature under deployment shift. The small change on Hoch is consistent with that cohort already operating in a severely degraded regime, where removing additional panel information has little further effect.

C. Cross-Platform Recalibration

Section 5.2 reports the headline $N_{\mathrm{recal}} = 5000$ results. Here we show the full recalibration sweep over $N_{\mathrm{recal}} \in \{500, 1000, 2000, 5000\}$, where labeled cells from the deployment cohort are appended to the Orion calibration pool and ontology coverage is re-evaluated at target 0.90. The full trajectory for each cohort is given in Table 6.

As shown in Table 6, Hoch and Liu represent calibration-addressable shift: Hoch reaches the 0.90 target at $N_{\mathrm{recal}} = 5000$, and Liu exceeds target with as few as 1000 labeled cells. IMMUcan and Schürch also improve substantially, but remain below target at $N_{\mathrm{recal}} = 5000$ under this pooled single-threshold recalibration. This is consistent with the

Table 6. Recalibration sweep (3-seed mean ontology coverage at target 0.90) after appending N_{recal} labeled in-platform cells to the Orion calibration pool.

Holdout	Baseline	500	1000	2000	5000
Hoch	0.257	0.826	0.864	0.892	**0.900**
Liu	0.800	0.899	0.901	0.905	**0.907**
IMMUcan	0.393	0.585	0.647	0.669	0.688
Schürch	0.493	0.606	0.709	0.785	0.809

scoped claim in the main text that cross-cohort deployment failures are heterogeneous: some are largely repaired by modest in-platform recalibration, whereas others are not fully addressed by appending labeled cells to a single shared threshold. The latter case is recovered once cohort identity is conditioned on: a per-cohort conditional-conformal threshold (Section H) brings each of the flagged cohorts to within $\sim$0.015 of the 0.90 target under cohort-specific calibration (gaps Hoch 0.002, Schürch 0.011, IMMUcan 0.015), with the largest recovery on Schürch ($0.589 \rightarrow 0.911$). That conditional-conformal benchmark uses a separate fixed CRC-transfer model with its own $50/50$ split (Section H), so its pooled coverages are not directly comparable to the LODO-holdout pooled@500 values in this table. Much of the residual gap is therefore a shared-threshold limitation rather than an irreducible mismatch; any remaining shortfall may stem from deeper label-definition drift or panel/ontology mismatch.

D. Qualitative phenotyping example

E. Verifier details and full per-tier breakdown

PATCH curation. Audit cells are drawn from the held-out CRC test samples CRC30–35, so they have not been seen by either the XGBoost head or its calibration pool (CRC16–29). Each candidate is run through the full PATCH pipeline at $\alpha = 0.10$, and within each tier cells are sorted by the top-1 minus top-2 probability margin (smallest margin first) so that the audit probes the tightest cases the conformal layer would surface in production. Cells within 80 px of the region edge are excluded to avoid truncated patches, and source samples are interleaved in round-robin order within each tier to preserve cohort diversity.

Each selected cell is rendered as a 64×64 px native crop ($\sim$20 µm focal field, matching the DeepCell/KRONOS receptive field), using per-region per-channel $p_{0.1}$–$p_{99.9}$ linear percentile rescaling with no gamma compression, so the verifier reads the same raw fluorescence intensities used by the GMM positivity model rather than a visually remapped version. Twelve display channels are tiled into a 4×3 grid and bilinearly upscaled $4\times$ to 256 px for legibility. Two RGB lineage composites are also rendered at higher resolution: a general-lineage composite (R=CD45, G=PanCK, B=DAPI) and a T-cell composite (R=CD3e, G=CD4, B=CD8). On every focal patch, the DeepCell segmentation mask is overlaid with the target-cell boundary in solid white and neighboring cells in faint gray, so the verifier can identify the subject cell while still reading local morphology. A separate 256×256 px native context crop ($\sim$82 µm) is rendered as an additional RGB composite at 768×768 with the same segmentation overlay, allowing the verifier to judge whether the cell sits in an epithelial nest, an immune aggregate, a stromal region, or a mixed microenvironment. The four PNGs (channel grid, lineage composite, T-cell composite, context view), together with a structured text panel listing the GMM marker posteriors and the candidate ontology labels, constitute the input bundle (Figure 6). The verifier never sees the model's top prediction, the tier label, or the gating ground truth; it does see the conformal candidate set, because the output is constrained to $\mathcal{S} \cup \{\mathrm{abstain}, \mathrm{artifact}\}$.

Audit composition. The 100-cell audit is stratified into four tiers defined by the conformal output on each cell: *easy* ($n = 20$, leaf-resolved and ontology-correct), *medium* ($n = 30$, leaf-resolved but wrong, or resolved only to a subtype node), *hard* ($n = 35$, the conformal layer backed off to the `Immune` lineage internal node or the root), and *artifact* ($n = 15$, flagged by the cross-lineage conflict detector as a doublet or segmentation failure). This is a deliberately stratified stress-test sample and is not intended to estimate the prevalence of routed cases in deployment, where the VLM is invoked only on ambiguous cells. The tiers are intended to probe VLM behavior as post-conformal ambiguity increases, with the artifact tier specifically testing whether the verifier can identify segmentation failures, debris, or biologically implausible cells. The verifier never sees the tier label. We evaluate four multimodal models under identical inputs (temperature 0, single image-text turn): Claude Opus 4.7 with max reasoning, GPT-5.5 with xHigh reasoning, and two open-weights models served via vLLM on a single A100 each—Qwen3-VL-32B-Instruct (FP8) and the newer Qwen3.6-35B-A3B (FP8, a mixture-of-experts model). The constraint $\mathrm{verdict} \in \mathcal{S} \cup \{\mathrm{abstain}, \mathrm{artifact}\}$ is specified in the prompt and enforced post hoc by a YAML-schema validator; outputs that do not parse to one of the allowed labels are recorded as *abstain*.

Per-tier behavior. Table 7 reports the per-tier verdict breakdown for the four verifiers used in Section 5.3. *Hit*: verdict is a leaf or internal node that is the gating ground truth or one of its ontology ancestors. *Wrong*: verdict is a label inside the allowed set but neither the ground truth nor any of its ancestors. *Art.*: verdict was *artifact*. *Abs.*: verdict was *abstain*. The four columns sum to n within each tier.

The four verifiers occupy visibly different QC operating points in Table 7; Figure 8 shows representative cells at

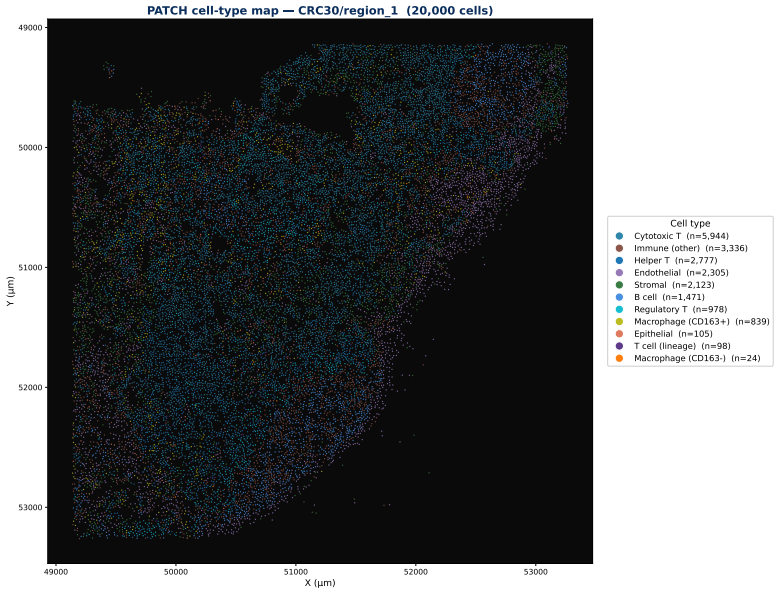

Figure 5. **Qualitative PATCH phenotyping on a held-out colorectal region (CRC30, region 1; ∼20,000 cells).** Segmented cells are plotted at their tissue coordinates and colored by the top-1 PATCH cell-type label (legend gives per-type counts). The spatial organization—epithelial sheets, stromal bands, vessel-tracing endothelium, and immune aggregates—is recovered from the per-cell marker features alone, with no spatial input, illustrating what a multiplexed spatial-proteomics cohort and the resulting cell-type map look like in practice. Calibrated set-size and coverage behavior for the pipeline are reported separately in Figure 2 and Section 5.1.

Table 7. Per-tier verdict breakdown for the four verifiers on the 100-cell audit.

		Claude Opus 4.7				GPT-5.5				Qwen3-VL-32B				Qwen3.6-35B-A3B			
Tier	n	Hit	Wrong	Art.	Abs.	Hit	Wrong	Art.	Abs.	Hit	Wrong	Art.	Abs.	Hit	Wrong	Art.	Abs.
easy	20	8	6	5	1	9	4	4	3	7	13	0	0	6	4	10	0
medium	30	4	15	11	0	3	15	7	5	4	22	4	0	4	9	17	0
hard	35	14	2	19	0	19	3	8	5	16	12	6	1	9	3	23	0
artifact	15	2	0	13	0	4	3	5	3	6	7	2	0	2	4	9	0
Overall	**100**	28	23	48	1	35	25	24	16	33	54	12	1	21	20	59	0

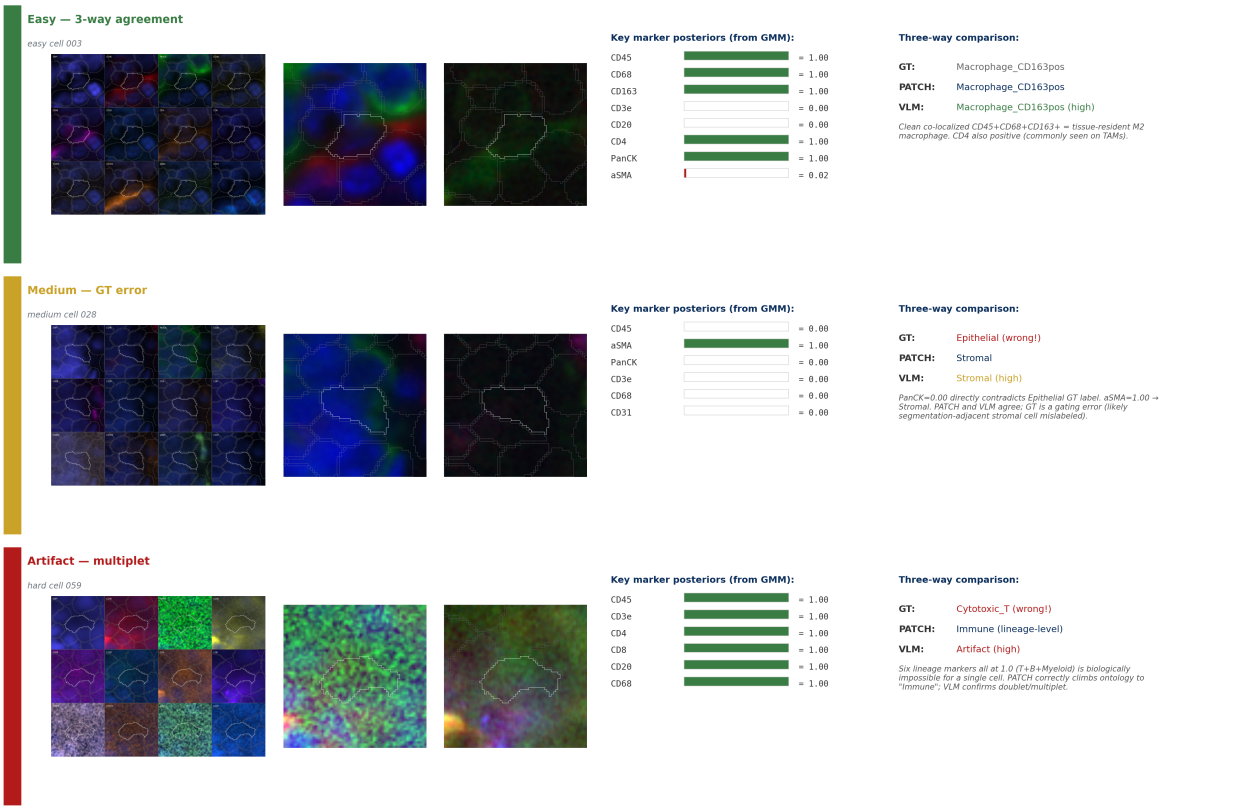

Figure 6. **Example VLM QC input format.** The verifier receives only the left block: a marker grid, lineage composites, a context view, and a text panel of GMM marker posteriors and candidate labels; its output is constrained to $\mathcal{S} \cup \{\text{abstain}, \text{artifact}\}$. The tier label and the *GT* and PATCH rows of the three-way comparison are shown for the reader only and are *withheld* from the verifier.

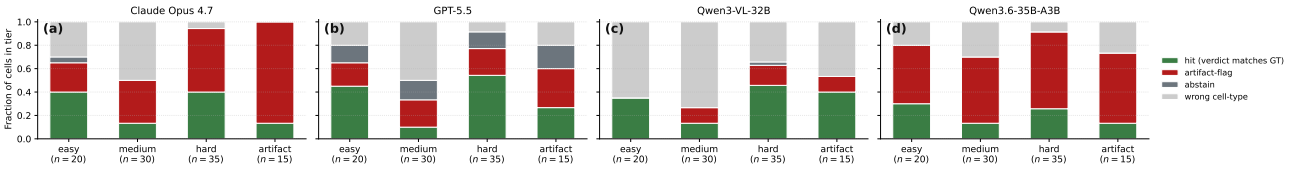

Figure 7. **VLM quality-control trade-offs on audited cells.** Different VLMs occupy different points on the abstention versus artifact-detection trade-off. In deployment, the VLM is invoked only after conformalization on routed ambiguous cells; the audit shown here is a deliberately stratified stress-test sample spanning ambiguity and artifact tiers. Panels show the four audited verifiers—Claude Opus 4.7 (a), GPT-5.5 (b), and the open-weights Qwen3-VL-32B (c) and Qwen3.6-35B-A3B (d); the bar heights are the per-tier verdict fractions of Table 7.

each routing tier together with the real per-cell verdicts of two open-weights verifiers (Qwen3-VL-32B and Qwen3.6-35B-A3B). Claude Opus 4.7 is the most aggressive artifact flagger, especially on the artifact tier, but also labels many non-artifact cells as artifact, leaving only $23/100$ wrong-label calls. GPT-5.5 uses abstention more often and is therefore more conservative ($25/100$ wrong-label, $16/100$ abstain). Qwen3-VL-32B essentially never abstains (a single unparsed output) and detects fewer artifacts overall, which translates into the highest wrong-label rate ($54/100$); operationally this makes it the least suitable as a stand-alone adjudicator. The newer Qwen3.6-35B-A3B sits at the opposite extreme: it is the most aggressive artifact-flagger of all four ($59/100$ artifact calls and the lowest hit rate, $21/100$), flagging the majority of cells—including many true negatives—as artifacts. This pattern, spanning a $5\times$ range in artifact-flag rate across the four verifiers, is consistent with the main-text interpretation that the VLM is best treated as an auxiliary QC layer rather than as a replacement classifier, and that the operating point is verifier-dependent.

On the artifact tier, artifact recall is $13/15 = 0.867$ for Claude Opus 4.7, $9/15 = 0.600$ for Qwen3.6-35B-A3B, $5/15 = 0.333$ for GPT-5.5, and $2/15 = 0.133$ for Qwen3-VL-32B.

Pairwise exact-string verdict agreement is 0.55 (Claude

Table 8. Conformal variants evaluated at ontology coverage on the CRC test split at target $\alpha = 0.10$, calibrated on CRC16–29. Variants (a)–(e) use the Hoch-holdout model: (a)–(c) are standard flat conformal baselines and (d)–(e) are evaluated at ontology coverage, with and without the panel filter. Variants (f)–(g) port two recent hierarchical-conformal methods onto the same CRC ontology task on the matched-panel CRC test split; leaf-res. is the fraction of cells resolved to a leaf.

Variant	Coverage	$\lvert\mathcal{S}\rvert$	Leaf-res %
(a) Global split	0.981	3.64	—
(b) Class-cond. split	0.895	17.71	—
(c) RAPS (k=1, λ=0.01)	0.968	3.51	—
(d) Class-cond., ontology cov. (no filter)	0.984	17.71	—
(e) Panel-aware hier. (ours)	**0.984**	**8.71**	—
(f) HCC (den Hengst et al., 2025)	0.645	1.07	99.4
(g) scConform (Corbetta et al., 2024)	0.990	15.0	0.0

Opus 4.7, GPT-5.5), 0.42 (Claude Opus 4.7, Qwen3-VL-32B), and 0.46 (GPT-5.5, Qwen3-VL-32B). This modest agreement indicates substantial residual ambiguity even under matched prompts and inputs, and further supports the use of the VLM as a constrained auditing tool rather than a source of autonomous labels.

Coverage-preservation. By construction, the verifier's verdict either lies in $\mathcal{S}$, equals *abstain*, or equals *artifact*. The verifier therefore cannot add a label outside the conformal support. In deployment and evaluation, the original conformal set is retained alongside the VLM verdict, so the statistical guarantee remains attached to the pre-VLM conformal set. The body's claim of "adding resolution without degrading coverage" refers to this retained-set construction: the audit numbers above measure the verifier's QC-side behavior (artifact recall, abstention rate, wrong-label rate within $\mathcal{S}$), not a new post-VLM coverage guarantee on the verdict alone.

F. Benchmark of conformal variants

To situate panel-aware hierarchical conformal among standard alternatives, we benchmark five conformal variants on CRC30–40 at target $\alpha = 0.10$, using the Hoch-holdout classifier and CRC16–29 for calibration. The variants are: global split-conformal, class-conditional split-conformal (Romano et al., 2020), RAPS (Angelopoulos et al., 2021), class-conditional conformal evaluated at ontology coverage without the panel filter, and panel-aware hierarchical conformal. We additionally port two recent hierarchical-conformal methods onto the same CRC ontology task, calibrated and evaluated on the matched-panel CRC test split: HCC (den Hengst et al., 2025), which forms label sets along the ontology, and scConform (Corbetta et al., 2024), a hierarchical conformal method developed for single-cell annotation. The results are summarized in Table 8.

As shown in Table 8, the main gain of the panel-aware hierarchy is not higher ontology coverage than the unfiltered ontology-aware baseline, but substantially tighter sets at the same coverage level. Relative to variant (d), the panel filter reduces mean set size from 17.71 to 8.71, a 51% reduction, by removing leaves that are not identifiable on the deployment panel before forming the final hierarchical prediction. These removed leaves are not incidental: on the CRC test set every class-conditional no-filter set contains at least one panel-unidentifiable leaf (dendritic cell, NK, neutrophil, or plasma cell, whose required markers are absent from the CRC panel), so the panel filter takes the unsupported-label rate from 100% to 0% at the same coverage. Flat global-split and RAPS avoid unsupported labels only by emitting small non-hierarchical sets that never back off to a lineage when a leaf is unidentifiable; the panel-aware hierarchy is the only variant here that is simultaneously biologically valid (no unsupported leaves) and ontology-resolved.

The two ported hierarchical-conformal methods (f)–(g) illustrate the same calibration-versus-resolution tension from opposite ends. HCC keeps sets very small ($\lvert\mathcal{S}\rvert = 1.07$, 99.4% leaf-resolved) but under-covers badly, reaching only 0.645 ontology coverage against the 0.90 target, so its compact sets are not calibrated. scConform reaches the target coverage (0.990) but only by inflating sets to $\lvert\mathcal{S}\rvert = 15.0$ and never committing to a leaf (0.0% leaf-resolved), so its calibrated sets are not informative. Panel-aware hierarchical conformal (e) is the only point in this comparison that is simultaneously calibrated (coverage $\geq$ target) and resolved: it holds ontology coverage at 0.984 while resolving a substantial fraction of cells to leaves, whereas on the in-domain matched-panel setting it attains the same coverage at mean set size $\lvert\mathcal{S}\rvert = 2.44$ with 52.0% leaf-resolved predictions (Section 5.1).

G. Cross-cohort compact-display benchmark under cross-platform transfer

The benchmark above is matched-panel (CRC test). To stress-test the three hierarchical-conformal methods under genuine cross-platform shift, we transfer a single fixed CRC-trained classifier unchanged to eight external cohorts (CODEX, IMC, MIBI)—the seven cohorts of Table 13 plus the Cords pan-cancer CAF (cancer-associated fibroblast) cohort (Cords et al., 2023)—split each cohort 50/50 (stratified, seed 0) into calibration and test, and run all three variants at $\alpha = 0.10$ (Table 9). No method is re-trained, so every method is evaluated under heavy base-model miscalibration.

scConform reaches target coverage on all eight cohorts but only by ballooning to entire lineage blocks ($\lvert\mathcal{S}\rvert \approx 16.6$, never leaf-resolved), reproducing its matched-panel behavior and confirming it is uninformative for resolution. HCC keeps near-leaf sets but under-covers, clearing the target

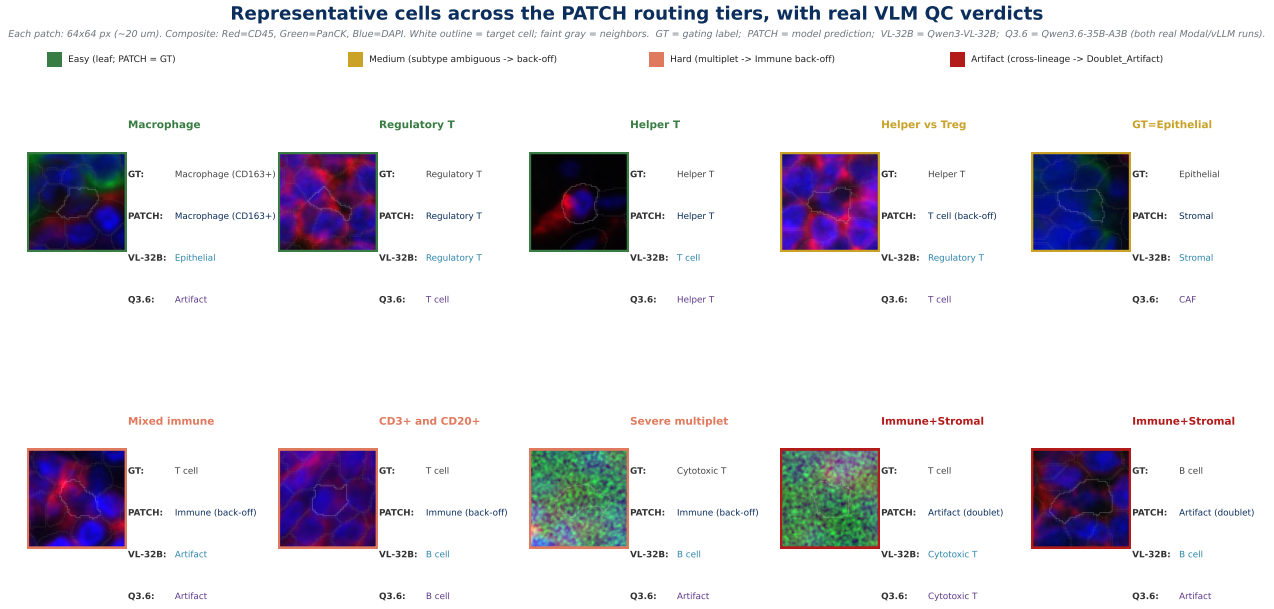

Figure 8. **Representative cells across the PATCH routing tiers, with real VLM QC verdicts.** For each cell, the composite image patch (Red=CD45, Green=PanCK, Blue=DAPI; solid white outline = the target cell's segmentation boundary, faint gray = neighbors) is shown with its gating ground truth (GT), the PATCH prediction, and the constrained QC verdict of two open-weights vision-language models—Qwen3-VL-32B (VL-32B) and Qwen3.6-35B-A3B (Q3.6)—each served with vLLM on the identical image bundle and prompt. Every label is real: GT and PATCH from the per-cell records, the two verdicts from the model runs. *Easy* cells (green) resolve to a leaf matching GT; *medium* cells (gold) are where the subtype is unidentifiable (e.g. helper vs. regulatory T without FOXP3) so PATCH backs off to a coarser node, or where the markers contradict the gating GT; *hard* cells (orange) are lineage-level multiplets that back off to the Immune node; the *artifact* tier (red) collects cross-lineage conflicts on a single segmentation mask that PATCH flags as `Doublet_Artifact`. The two verifiers act as independent auditors: they agree with PATCH on many cells but also commit some PATCH-flagged cells to a leaf (e.g. cell 086, which PATCH flags as a doublet, both verifiers call Cytotoxic T) and flag others as *artifact*, with Q3.6 the more aggressive artifact-flagger. Their aggregate per-tier rates over the full 100-cell audit are in Table 7.

Table 9. Cross-cohort compact-display benchmark: a single fixed CRC-trained model transferred *unchanged* to eight external cohorts (no recalibration), so every method runs under heavy base-model miscalibration. OntCov = empirical ontology coverage (target ≥ 0.90), not a transfer guarantee; $|\mathcal{S}|$ = mean set size. PATCH applies the panel filter; HCC and scConform run panel-blind as published.

Cohort (#mk)	PATCH cov/$|\mathcal{S}|$	HCC cov/$|\mathcal{S}|$	scConf. cov/$|\mathcal{S}|$
Schürch (11)	**0.96**/6.1	0.53/1.0	0.98/16.3
Liu (12)	**0.94**/6.7	0.69/1.9	0.99/15.3
Hartmann (15)	**0.97**/7.7	0.78/12.3	0.99/16.7
MAPS (28)	**0.99**/10.2	0.79/12.0	1.00/17.0
Tietscher (25)	0.68/1.2	0.67/1.2	0.97/16.6
Cords (14)	0.48/8.9	0.51/1.7	1.00/16.9
Hoch (23)	0.35/3.7	0.83/1.2	1.00/17.0
IMMUcan (27)	0.11/0.1	0.97/2.1	1.00/17.0
mean	0.68/5.6	0.72/4.2	0.99/16.6

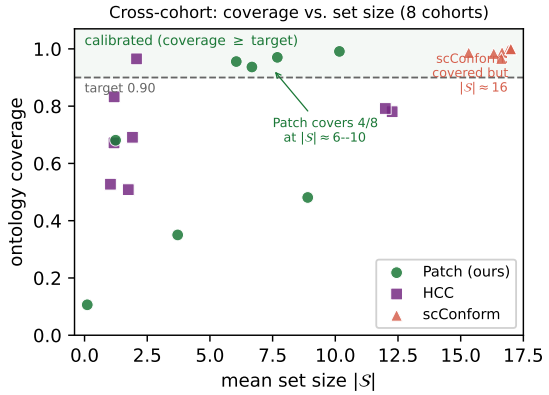

Figure 9. Cross-cohort coverage versus mean set size (eight external cohorts; one point per method per cohort, so eight points per method). The shaded band marks the calibrated regime (coverage ≥ 0.90). scConform always clears the band but only with whole-lineage sets ($|\mathcal{S}| \approx 16$, right wall); HCC keeps near-leaf sets on most cohorts (inflating to $|\mathcal{S}| \approx 12$ on Hartmann and MAPS) but under-covers on seven of eight cohorts. PATCH is the only method that clears the band while keeping sets well below the lineage-block size—reaching coverage at $|\mathcal{S}| \approx 6$–10 on the four narrow-panel CODEX/MIBI cohorts (Schürch, Liu, Hartmann, MAPS) where the transferred base model stays well-ordered.

on only one of eight cohorts: thresholds calibrated on a miscalibrated transferred model are too tight to generalize. PATCH is the only method that combines real coverage with sets well below scConform's lineage blocks (Figure 9), clearing 0.90 on four of eight cohorts—all the narrow-panel CODEX/MIBI cohorts (Schürch, Liu, Hartmann, MAPS)—there converting scConform-sized balloons into covered sets roughly half as large ($|\mathcal{S}| \approx 6$–10 versus ≈ 16). The cohorts where PATCH drops (IMMUcan 0.11, Hoch 0.35) are exactly the high-KS cohorts the label-free diagnostic flags (Section 5.2): the panel filter is a resolution win *conditional on a reasonably-calibrated base model*, and the diagnostic correctly identifies when that condition fails. The next section addresses this directly.

H. Conditional conformal recovers per-cohort coverage

The label-free diagnostic identifies *when* a cohort needs recalibration; conditional conformal prediction (Gibbs et al., 2025) provides a principled *how*. We pool the CRC calibration set with a small in-platform labeled slice (500 cells) from each deployment cohort and fit a conditional-conformal threshold $\tau(x) = \Phi(x)^\top \beta$ with $\Phi(x) = [\text{intercept, one-hot(cohort)}]$, against a pooled split-conformal threshold on the same data (Table 10). Conditioning on cohort identity fits a separate threshold per cohort and shrinks the mean coverage gap to target—averaged over the three flagged cohorts of Table 10 (Hoch, Schürch, IMMUcan; Liu already meets target under the pooled threshold)—from 0.120 (pooled) to 0.009 (conditional), most dramatically on Schürch, where a single pooled threshold collapses to 0.589 coverage while the conditional threshold recovers 0.911.

Table 10. Conditional conformal (Gibbs et al., 2025) with $\Phi = [\text{intercept, one-hot(cohort)}]$ vs. a pooled split-conformal threshold, $\alpha = 0.10$, 500-cell in-platform slice per cohort. Gap $= |\text{coverage} - 0.90|$; means are over these three flagged cohorts (Liu is omitted as its pooled threshold already meets target). This benchmark uses the fixed CRC-transfer model and 50/50 split of Section G, so the pooled coverages here differ from the LODO-holdout pooled@500 values of Table 6.

Cohort	Cond. cov	Pooled cov	Cond. gap	Pooled gap
Hoch	0.898	0.927	0.002	0.027
Schürch	0.911	0.589	0.011	**0.311**
IMMUcan	0.915	0.921	0.015	0.021
mean			**0.009**	0.120

The exact finite-sample guarantee holds over the shifts encoded in $\text{span}(\Phi)$—the cohorts that appear in Φ and contribute calibration cells—so this is precisely the small-labeled-slice recalibration regime of Section 5.2, not zero-label deployment: a brand-new cohort maps to the intercept column and inherits the calibration-cohort threshold. Within

that regime, conditional conformal turns the label-free KS flag into an actionable, coverage-restoring recalibration step.

I. Exploratory patient-level aggregation

To assess whether calibrated cell-level predictions remain useful after aggregation, we convert per-cell outputs on the CRC Orion cohort (Lin et al., 2023) ($n = 40$ patients) into per-patient cell-type composition features and use these as inputs for downstream patient-level analyses. Concretely, each cell contributes one count to the bin of its top-1 ontology label, which may be a lineage, subtype, or leaf depending on where the conformal layer resolves it. Per-patient counts are then normalized by the total number of cells for that patient to obtain a 26-dimensional vector of cell-type proportions. This construction preserves the hierarchy-aware behavior of the predictor: when the conformal layer backs off to an internal node, the cell contributes to that broader lineage rather than forcing a potentially unsupported leaf label.

We evaluate four binary clinical endpoints by leave-one-out logistic regression and compare against a baseline that uses the same underlying data modality without cell typing. The baseline is a raw mean-intensity vector with one feature per measured marker on the Orion panel (16 dimensions). Logistic regression is L2-regularized, class-weight balanced, and fitted after a per-fold `StandardScaler`. We report LOOCV AUC together with 95% bootstrap confidence intervals from 1,000 bootstrap replicates on the leave-one-out predictions. The four endpoints are defined directly from the Orion clinical metadata: *any recurrence* (Recurrence flag), *recurrence <2y* (Recurrence with PFSDays ≤ 730), *MSI deficient* (`MMRIHC` reports absence), and *any-cause death* (Death flag). In addition, we test whether selected cell-type proportions stratify progression-free survival by a median split followed by a `lifelines` log-rank test. The results are summarized in Table 11.

Table 11. Patient-level LOOCV AUC (95% bootstrap CI, $n = 40$ CRC Orion patients) of cell-type proportions vs. raw mean-intensity, and log-rank p-values for progression-free-survival stratification.

Endpoint / stratifier	Cell typing	Raw intensity
Classification AUC		
Any recurrence	**0.80** [0.65, 0.93]	0.63 [0.44, 0.79]
Recurrence < 2 y	**0.71** [0.55, 0.86]	0.56 [0.36, 0.74]
MSI deficient	**0.70** [0.38, 0.97]	0.64 [0.22, 0.97]
Any-cause death	0.60 [0.42, 0.78]	**0.66** [0.47, 0.85]
PFS log-rank p		
Cytotoxic-T proportion	**0.008**	—
Regulatory T	**0.044**	—
Immune lineage total	**0.006**	—

As shown in Table 11, cell-type proportion features outper-

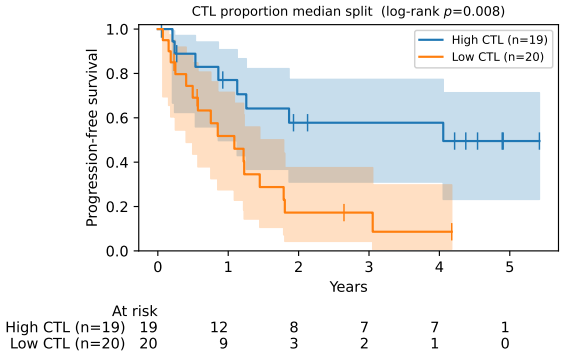

Figure 10. Progression-free survival in the Orion CRC cohort (median split on the n=39 patients with progression-free-survival follow-up; high-CTL n=19 vs. low-CTL n=20) by the calibrated cytotoxic-T (CTL) proportion. High-CTL patients have markedly longer PFS (median 1481 vs. 397 days; log-rank p=0.008). Tick marks denote censoring; the table beneath the axis reports numbers at risk.

Table 12. Survival analysis (Orion CRC, PFS; median-split log-rank on the n=39 patients with PFS follow-up, multivariable Cox on n=40). The Cox model is fit on the continuous proportion (hazard ratio per +1 standard deviation, adjusted for age and stage; Harrell's concordance c); Cox p-values are CTL 0.038, regulatory-T 0.020, Immune total 0.009. The CTL and regulatory-T effects survive adjustment; the aggregated Immune total runs the opposite direction.

PFS stratifier	log-rank p	Cox HR/+1SD [95% CI]	c
Cytotoxic T (CTL)	0.008	0.47 [0.23, 0.96]	0.70
Regulatory T	0.044	0.50 [0.28, 0.90]	0.74
Immune total	0.006	1.66 [1.13, 2.44]	0.69

form raw-intensity baselines on recurrence- and immune-associated endpoints, while performing worse on any-cause death, which is plausibly more influenced by bulk tumor burden than by immune composition alone. The significant progression-free-survival stratifications further suggest that the hierarchy-aware cell-type summaries retain biologically meaningful signal after aggregation. A multivariable Cox model adjusting for age and stage (Table 12, Figure 10) shows the CTL association is not merely a confound of clinical staging: each one-standard-deviation increase in the calibrated CTL proportion carries hazard ratio 0.47 (95% CI $[0.23, 0.96]$, p=0.038) with concordance c=0.70, and the regulatory-T effect is comparable (HR=0.50, p=0.020, c=0.74). The aggregated Immune total runs in the *opposite* direction (high $\rightarrow$ shorter PFS, HR=1.66, p=0.009): an internal-node count is only as interpretable as the populations it pools, and the Immune bin mixes cytotoxic with suppressive and myeloid lineages—a caution for hierarchy-aware features, not a contradiction of the leaf-level result. At the same time, these downstream analyses are exploratory: the cohort is small ($n = 40$; $n = 39$ with progression-free-survival follow-up), the endpoints are partially overlapping, and we apply no multiple-comparison correction across the reported survival tests—so the borderline associations (e.g. the regulatory-T log-rank p=0.044) would not survive a Bonferroni adjustment, whereas the headline CTL effect (log-rank p=0.008, Cox p=0.038) is the most robust. The results should be interpreted as evidence that the calibrated cell-typing outputs remain informative at the patient level rather than as a definitive clinical model. We further note that this CTL$\rightarrow$PFS association is colorectal-specific: in the independent lung-adenocarcinoma (LUAD) cohort of Sorin et al. (2023) (n=416) the same median-split test does not reach significance (log-rank p=0.97), consistent with the cross-platform, cross-disease shift this paper otherwise quantifies.

J. External validation of the KS diagnostic

To extend the four-cohort LODO result from the main text, we re-ran the Hoch-holdout model on three additional public cohorts: Hartmann colorectal MIBI cohort (Hartmann et al., 2021), MAPS Hodgkin lymphoma CODEX (Shaban et al., 2024), and Tietscher breast IMC cohort (Tietscher et al., 2023). Together with Liu, Schürch, IMMUcan, and Hoch, these form the seven-cohort external set used for the extended cohort-level KS diagnostic. The cohort-level values are summarized in Table 13.

As shown in Table 13, the same monotone trend seen in the four-cohort LODO analysis persists in the larger seven-cohort set: cohorts with larger mean KS tend to have lower ontology coverage. Across these seven external cohorts, Pearson correlation between mean KS and ontology coverage is $r = -0.853$ ($p = 0.015$), and Spearman rank correlation is $\rho = -0.893$ ($p = 0.007$). This extended analysis does not replace the stricter LODO design in the main text, but it strengthens the empirical case that mean per-marker KS distance is a useful label-free deployment covariate across a broader range of tissues and platforms.

Table 13. Cohort-level KS shift versus ontology coverage at $\alpha = 0.10$ across the seven external cohorts used in the extended KS diagnostic.

Cohort	Mean KS	Onto. cov	$\lvert\mathcal{S}\rvert$	n
Liu	0.429	0.800	2.70	100K
Tietscher	0.333	0.794	1.20	297K
Hartmann	0.411	0.587	1.96	64K
MAPS	0.500	0.584	1.97	123K
Schürch	0.508	0.493	2.47	100K
IMMUcan	0.538	0.393	1.36	74K
Hoch	0.544	0.257	2.01	100K

K. Point-prediction baselines

For reference, we evaluated several primary cell-typing baselines on the matched-input CRC30–40 benchmark. A zero-shot open-weights LLM (Qwen Team, 2025) reaches 34.3% flat accuracy on the 11-class task, below the rule-based baseline at 44.4%. Cell2Sentence (Levine et al., 2024) fine-tuned on CRC01–15 reaches 60.8% accuracy and 46.1% macro-F1. Our reimplementation of Spatial2Sentence (Chen et al., 2025) reaches 68.2% accuracy and 51.5% macro-F1 on a 44K-cell subsample. Our XGBoost head reaches 69.9% accuracy and 57.5% macro-F1 on the same input information; MAPS (Shaban et al., 2024) reaches 71.7% accuracy under matched inputs. These results show that the XGBoost head is competitive with the strongest supervised baseline (MAPS, 71.7%) and far above the zero-shot LLM (34.3%), motivating its use as PATCH's statistical backbone, with the multimodal VLM reserved for constrained QC rather than primary classification.