AUTOMARKER: OUTCOME-GUIDED DISCOVERY OF MULTISCALE BIOMARKER CANDIDATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Clinical outcomes are observed at the patient level, whereas associated biomarker candidates emerge across cells, local microenvironments, and tissue organization. AutoMarker introduces an outcome-guided framework for discovering multiscale biomarker candidates directly from pathology representations. Protein-aligned morphology representations organize tissue into four coordinated views capturing cellular phenotypes, neighborhood composition, tissue organization, and context-dependent cellular states. Adaptive local divergence search identifies representation-space regions enriched in different outcome groups, while sign-consistent graph refinement consolidates local responses into coherent candidate components. WSI-level abundance comparisons, pathological characterization, and spatial back-tracing connect candidates to tissue context. Across four cohorts, we evaluate multiscale candidate discovery, including independent external validation in TCGA-COAD of candidates discovered and frozen in Orion-CRC. Comparative and ablation experiments demonstrate the contribution of adaptive search and graph refinement, while scalability experiments support mining at the million-entity scale.

1 INTRODUCTION

Clinical outcomes are defined at the patient level, whereas their associated biomarker candidates emerge across cells, local microenvironments, and larger tissue structures. Spatial profiling has shown that cellular organization is closely linked to antitumoral immunity (Schürch et al., 2020). This motivates an outcome-guided discovery paradigm: *without local phenotype annotations, how can differences between outcome groups be traced back to the multiscale biomarker candidates from which they arise?* We therefore start from observed clinical outcomes and work backward to localize, characterize, and spatially map the biomarker candidates associated with those differences.

Weakly supervised learning supports slide-level prediction without exhaustive local annotations (Ghaffari Laleh et al., 2022). Multiple-instance learning (MIL) aggregates patch representations under slide-level supervision (Ilse et al., 2018; Campanella et al., 2019; Lu et al., 2021), and graph models incorporate spatial interactions (Chen et al., 2021). These approaches provide representations and localization cues. A prediction or attention map, however, does not by itself specify a candidate's membership, scale, abundance, or context. Here, we seek explicit candidate regions whose membership, scale, group abundance, and tissue context can be examined together.

Local phenotype annotations are unavailable, and morphology does not directly specify molecular identity. Paired hematoxylin and eosin (H&E) and multiplex immunofluorescence (mIF) images offer molecularly informed representations (Lin et al., 2023). Outcome-associated variation may also involve cellular identity, neighborhood composition, or tissue organization, each requiring an explicit view (Hu et al., 2024). Within a view, fixed-radius neighborhoods span unequal amounts of data, while fixed-k neighborhoods have variable geometric extent.

AutoMarker integrates these stages into an outcome-guided representation-space localization framework for WSI-level candidate discovery and screening. Four protein-aligned pathology views capture cellular phenotypes, local microenvironments, tissue organization, and context-dependent cellular states. Within each view, adaptive neighborhood expansion tracks local log-fold changes between outcome groups, while a curvature-based stopping rule determines candidate-specific search boundaries. Sign-consistent graph refinement then consolidates neighboring responses into coherent

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

candidate components, whose WSI-level abundance and source coordinates enable statistical characterization and tissue back-tracing. We make three contributions: (1) a formulation of outcome-guided biomarker candidate discovery that jointly defines candidate membership, scale, enrichment direction, and spatial support; (2) adaptive representation-space search that selects candidate-specific neighborhood scales from outcome-conditioned divergence trajectories; and (3) a traceable refinement and validation workflow that consolidates local responses, links candidates back to tissue, and evaluates frozen discoveries in an external cohort.

2 Related Work

Weak supervision and spatial representations. To reduce dependence on detailed annotations, MIL models learn from slide-level labels by aggregating instance features, enabling prediction and localization of outcome-associated regions (Ilse et al., 2018; Campanella et al., 2019; Lu et al., 2021). CLAM further combines attention pooling with instance-level clustering to refine discriminative features and localize relevant tissue without spatial annotations (Lu et al., 2021). Graph-based models represent relationships among tissue patches to incorporate context beyond individual patch appearance (Chen et al., 2021). CTransPath and UNI provide transferable morphology features, while HIPT models tissue organization hierarchically across scales (Wang et al., 2022; Chen et al., 2024; 2022). Together, these approaches reduce annotation requirements and support heterogeneous tissue analysis. However, predictive importance and morphological similarity do not by themselves specify discrete phenotype candidates with explicit membership, candidate-specific scale, and measurable abundance. Relating predictive regions to reproducible phenotype definitions remains a separate analytical task.

Differential abundance and spatial phenotype analysis. Differential abundance methods identify cell populations whose prevalence varies across biological conditions. Citrus combines hierarchical clustering with predictive modeling to identify stratifying cellular subpopulations (Bruggner et al., 2014). diffcyt couples high-resolution clustering with differential testing, while scCODA accounts for compositional structure of cell-type abundances (Weber et al., 2019; Büttner et al., 2021). DA-seq and Milo move beyond fixed cell-type partitions by assessing local abundance changes in multiscale or graph-based neighborhoods (Zhao et al., 2021; Dann et al., 2022). Squidpy characterizes spatial cellular organization, while CytoCommunity uses cell phenotypes and spatial distributions to discover tissue cellular neighborhoods (Palla et al., 2022; Hu et al., 2024). UTAG combines morphology, protein expression, and spatial proximity to identify microanatomical domains without manual annotation (Kim et al., 2022). These efforts extend analysis from predefined populations to local cellular states and tissue communities. Nevertheless, candidate granularity remains sensitive to clustering, graph construction, or neighborhood-scale choices, and overlapping signals can complicate interpretation. Jointly defining adaptive candidate boundaries, consolidating redundant responses, and retaining enrichment direction and tissue traceability remains an open challenge.

3 Method

3.1 Problem formulation and overview

Let $\mathcal{D} = \{(X_s, Y_s)\}_{s=1}^{S}$ be WSIs and binary outcome-group labels. Each slide provides entities with morphology and spatial coordinates. A representation view ℓ contains records

$$\mathcal{S}^{(\ell)} = \{(z_i^{(\ell)}, Y_{s(i)}, \pi_i)\}_{i=1}^{N_\ell}, \qquad z_i^{(\ell)} \in \mathbb{R}^{D_\ell}, \tag{1}$$

where $s(i)$ identifies the slide and π_i stores spatial metadata. Outcomes are defined at the patient level and propagated to all corresponding slides and entities. Abundance is quantified per WSI, and each WSI contributes one observation to the cohort-level comparison. Slides from the same patient share the patient outcome label. Each candidate is defined by its representation view, query-center component, expanded coverage, enrichment direction, WSI-level abundance, and source coordinates. Query centers form candidate components in representation space, their adaptive neighborhoods determine candidate coverage, and back-traced instances locate the candidate in tissue. Outcome labels directly guide candidate discovery without requiring cell- or region-level annotations. The resulting biomarker candidates capture outcome-associated spatial patterns for subsequent pathological and prognostic analysis.

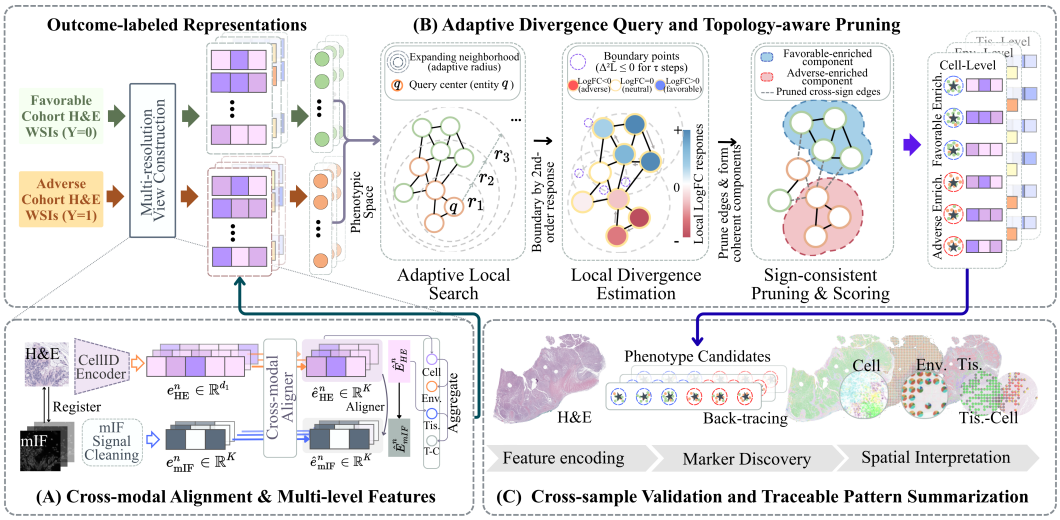

Figure 1: **Overview of AutoMarker.** Paired H&E–mIF data train a protein-aligned morphology encoder, yielding four hierarchical views of cellular and tissue organization. Adaptive divergence search localizes outcome-associated candidates, sign-consistent refinement consolidates representation structure, and spatial back-tracing maps them to source WSIs.

Figure 1 summarizes the workflow. Outcome-associated candidates are discovered in representation space, with physical adjacency providing tissue context and stored coordinates enabling direct localization back to the source WSI. Components are defined by phenotypic similarity in representation space and may correspond to multiple spatially separated tissue regions.

3.2 Protein-aligned hierarchical representations

Registered H&E–mIF patches provide matched cells through morphology- and DAPI-based segmentation. CTransPath encodes each masked H&E cell crop into a d_1-dimensional morphology descriptor, paired with a measured K-channel mIF vector e_i^{mIF}. The H&E descriptor is projected to K dimensions. Two attention branches contextualize cellular inputs and aggregate them into patch tokens. Their outputs are protein-aligned H&E cell vectors $\hat{e}_i$ and patch tokens $\hat{E}_P^{\mathrm{HE}}$ and $\hat{E}_P^{\mathrm{mIF}}$.

For matched cell i, the branch inputs are $\tilde{e}_i^{\mathrm{HE}} = \mathcal{F}_{\mathrm{fc}}(\mathcal{F}_{\mathrm{enc}}(m_i \otimes x_{\mathrm{HE}}))$ and $e_i^{\mathrm{mIF}} = [v_{i,1}, \ldots, v_{i,K}]$. Here m_i is the cell mask and $v_{i,k}$ aggregates channel-k intensities within the radius-constrained, thresholded cell region. DAPI supplies segmentation information; the other channels define the target features. Attention contextualizes cells within each patch, while an aggregation token summarizes their regional representation.

Table 1: **Hierarchical views.** A separate index is built for each view. Env.: microenvironment; Tis.: tissue; T–C: tissue-conditioned cell.

View	Dim.	Unit	Captures
Cell	K	Cell	Cellular phenotype
Env.	K	Local region	Neighborhood composition
Tis.	K	Tissue region	Regional representation
T–C	K	Cell in tissue	Contextual cell variation

The paired-training objective combines patch-level contrastive alignment with cell-level regression. Contrastive learning aligns paired views (Chen et al., 2020); here, positives are registered H&E–mIF patches and negatives are mIF patches from other locations:

$$\mathcal{L}_{\mathrm{patch}} = \mathrm{CTR}\big((\hat{E}_P^{\mathrm{HE}}, \hat{E}_P^{\mathrm{mIF}}), \{\hat{E}_t^{\mathrm{neg}}\}_{t=1}^{T}\big), \tag{2}$$

$$\mathcal{L}_{\mathrm{cell}} = \mathrm{MSE}\big(\{\hat{e}_i\}_{i\in P}, \{e_i^{\mathrm{mIF}}\}_{i\in P}\big). \tag{3}$$

The patch objective aligns regional context; cell regression anchors morphology-derived vectors to measured protein profiles. At inference, only H&E is required: cell vectors $\hat{e}_i$ and patch tokens

$\hat{E}_P = \hat{E}_P^{\mathrm{HE}}$ supply the four views below. *Appendix A.1* details feature extraction and token construction; Figure 5 provides paired spatial examples.

Four views. Four views represent cellular phenotypes, local composition, tissue organization, and cellular variation within tissue contexts. They specify different questions about tissue structure (Table 1). The cell view uses the protein-aligned embedding directly, $z_i^{\mathrm{cell}} = \hat{e}_i$.

For a local region P, let A_P be the cell-to-cell self-attention matrix and $d_i = \sum_{j \in P} A_P[i, j]$ its row-sum aggregation degree. The microenvironment pooling weights are $w_i = d_i / \sum_{j \in P} d_j$, giving

$$z_P^{\mathrm{env}} = \frac{\sum_{i \in P} w_i \hat{e}_i}{\sum_{i \in P} w_i}, \qquad z_P^{\mathrm{tis}} = \hat{E}_P, \qquad z_{i|\mathcal{P}}^{\text{T-C}} = \hat{e}_i,\ i \in \Omega(\mathcal{P}). \tag{4}$$

Here $\hat{E}_P$ is the aligned patch representation. For a tissue-level candidate $\mathcal{P}$, tissue conditioning uses

$$\Omega(\mathcal{P}) = \bigcup_{p \in \mathcal{B}(\mathcal{P})} \mathcal{C}(p),$$

where $\mathcal{B}(\mathcal{P})$ is the set of constituent H&E patches and $\mathcal{C}(p)$ is the set of cell instances spatially contained in patch p. Thus the tissue-conditioned view contains cells from the patches constituting the tissue candidate, with membership determined by spatial containment. The tissue-conditioned view asks which cellular patterns occur within the selected tissue context. The cell, microenvironment, and tissue views describe individual phenotype, pooled local composition, and regional representation, respectively. All four views have K dimensions. Each view is indexed separately, with cell coordinates or region identifiers retained. A grid-balanced sampler and per-WSI cap limit the contribution of densely sampled slides.

3.3 Adaptive Local Divergence Search

For a query center q in one view, a KD-tree (Bentley, 1975) retrieves the neighborhood $\mathcal{N}(q, r) = \{i : \|z_i - z_q\|_2 \le r\}$. Let $n_g(q, r)$ count its entities from group $g \in \{0, 1\}$. The local response is computed as

$$L_q(r) = \mathrm{LogFC}(q, r) = \log_2 \frac{\gamma\, n_0(q, r) + \psi(r)}{n_1(q, r) + \psi(r)}, \qquad \psi(r) = \psi_{\mathrm{base}} \frac{r_{\min}}{r + \epsilon}, \tag{5}$$

where $N_g = |\{i \in \mathcal{S}^{(\ell)} : Y_{s(i)} = g\}|$ is the total entity count for outcome group g in the entire indexed cohort view, not a query-neighborhood count. The factor $\gamma = N_1/N_0$ compensates for these unequal group totals, and $\psi(r)$ regularizes low-count neighborhoods, where ψ_{base} is the pseudo-count base and ϵ is a numerical stabilizer. Labels are assigned as $Y = 0$ for favorable and $Y = 1$ for adverse prognosis, following the figure convention; positive LogFC therefore indicates favorable-group enrichment. The same response is used within each view. Cellular queries compare cell vectors, microenvironment queries compare pooled region vectors, and tissue queries compare patch tokens. Tissue-conditioned queries operate on the cells selected by $\Omega(\mathcal{P})$, with group totals defined for the indexed population. The radius measures distance in representation space, not physical distance on the slide; tissue coordinates are retained for subsequent back-tracing.

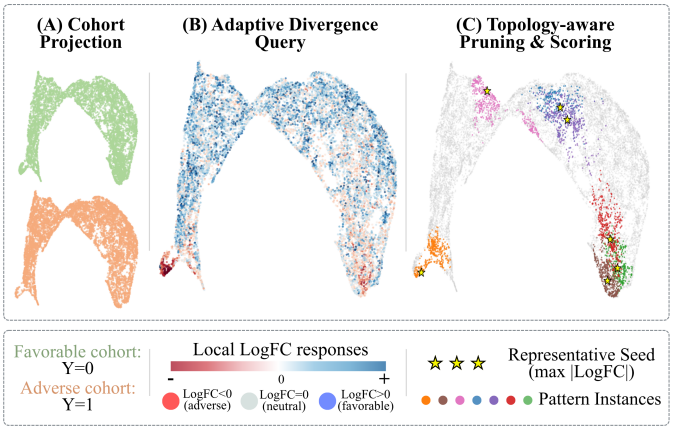

Figure 2: **Visualization after discovery.** UMAP shows the CellID representation, adaptive responses, and selected candidate components. Search is performed in the original per-view representation space.

Data-adaptive radius initialization. The radius range is estimated separately from the feature distribution of each indexed view. Let $Z^{(\ell)}$ be its feature matrix and $\sigma_{\text{global}} = \text{std}(Z^{(\ell)})$ the scalar standard deviation over all matrix elements. We sample up to n_{init} entities and compute each sampled entity's mean distance to its k_{init} nearest neighbors, $\bar{d}_i = \frac{1}{k_{\text{init}}} \sum_{j=1}^{k_{\text{init}}} d_{i,j}$. The bounds are

$$r_{\min} = \max\{P_\eta(\{\bar{d}_i\}),\, \alpha_{\min}\sigma_{\text{global}}\}, \qquad r_{\max} = \alpha_{\max}\sigma_{\text{global}}, \tag{6}$$

where P_η denotes the ηth percentile and $\alpha_{\min}$ and $\alpha_{\max}$ scale the radius bounds. If $r_{\min} \geq r_{\max}$, we set $r_{\max} = \beta r_{\min}$, where β is the fallback multiplier. The search uses J linearly spaced radii,

$$r_t = r_{\min} + \frac{t}{J-1}(r_{\max} - r_{\min}), \qquad t = 0, \ldots, J-1. \tag{7}$$

Choosing the neighborhood scale. The search evaluates radii in increasing order. The first valid nonzero response fixes the query's enrichment direction $s_q = \text{sign}(L_{q,\text{first}})$. When a later response reverses this sign, the previous retained response is reused. This suppresses sign reversals without imposing monotonic growth of response magnitude. A moving average of width w gives $\widetilde{L}_{q,t}$. Boundary selection uses the direction-normalized trajectory

$$U_{q,t} = s_q \widetilde{L}_{q,t}, \qquad s_q = \text{sign}(L_{q,\text{first}}),$$

whose second difference is

$$\Delta^2 U_{q,t} = U_{q,t} - 2U_{q,t-1} + U_{q,t-2}. \tag{8}$$

The stopping counter c_t increments when $\Delta^2 U_{q,t} \leq 0$ and resets to zero otherwise. This tests progression in the query's established enrichment direction rather than the arbitrary favorable/adverse sign convention. The original signed LogFC is retained for enrichment interpretation, component scoring, and sign-consistent graph refinement. Boundary acceptance requires stopping persistence τ and the neighborhood support defined below.

Minimum neighborhood support. A radius is accepted as the adaptive boundary of query q when the stopping condition persists for at least τ consecutive evaluations and the current query neighborhood contains at least $n_{\min}$ entities:

$$c_t \geq \tau, \qquad |\mathcal{N}(q, r_t)| \geq n_{\min}. \tag{9}$$

At the first radius satisfying both conditions, the search sets $r_q^* = r_t$ and $L_q^* = \widetilde{L}_{q,t}$. If the curvature condition is met before the neighborhood reaches this entity count, expansion continues. Neighborhood support counts the entities around an individual query center, not the size of a subsequent graph component or its WSI coverage.

Fallback boundary. If no radius satisfies both conditions over all J steps, the search evaluates a fallback at $r_{\max} = r_{J-1}$. The query is retained with $r_q^* = r_{\max}$ and $L_q^* = \widetilde{L}_{q,J-1}$ only if

$$|\widetilde{L}_{q,J-1}| > \delta_{\text{fb}}, \qquad |\mathcal{N}(q, r_{\max})| \geq n_{\text{fb}}. \tag{10}$$

Here δ_{fb} and n_{fb} are the fallback response and neighborhood-support thresholds. Otherwise, the query is discarded before graph construction. Each retained query stores its view, center identity, selected radius, signed response, and source coordinates. *Appendix B* gives the search procedure.

3.4 Component refinement and spatial back-tracing

Sign-consistent graph refinement. Nearby query centers generate redundant neighborhoods. A k-nearest-neighbor graph is built on retained centers, and edges joining opposite signs of L_q^* are removed. Connected components of pruned graph group phenotypically adjacent responses with consistent enrichment direction. For each center component $\mathcal{R}$, its retrieved entity coverage is

$$\mathcal{U}(\mathcal{R}) = \bigcup_{q \in \mathcal{R}} \mathcal{N}(q, r_q^*). \tag{11}$$

The graph partitions query centers, whereas $\mathcal{U}(\mathcal{R})$ collects their neighborhood members. Coverage sets may overlap even when center components are disjoint. Neither representation-space connectivity nor shared component membership requires physical adjacency in tissue.

WSI-level abundance screening. For a query-center component $\mathcal{R}$, its per-slide hit ratio counts component members, rather than the union of expanded neighborhoods $\mathcal{U}(\mathcal{R})$:

$$H_s(\mathcal{R}) = 100\frac{|\mathcal{R} \cap \mathcal{S}_s^{(\ell)}|}{|\mathcal{S}_s^{(\ell)}|}, \tag{12}$$

where $\mathcal{S}_s^{(\ell)}$ is the corresponding WSI's entity set used in the abundance calculation. Thus H_s measures the frequency of component-member entities relative to that set, whereas $\mathcal{U}(\mathcal{R})$ describes expanded retrieval coverage. The two quantities distinguish component membership from neighboring entities and need not yield the same abundance. Each WSI contributes one abundance observation to the cohort-level comparison; repeated WSIs from the same patient are not collapsed in the primary analysis. A two-sided Mann–Whitney U test (Mann & Whitney, 1947) compares WSI-level abundances between outcome groups. The screening threshold is $p \leq \theta_p$. Among retained components, the ranking score is

$$\text{Score}(\mathcal{R}) = \left| \frac{1}{|\mathcal{R}|} \sum_{q \in \mathcal{R}} L_q^* \right| \sqrt{|\mathcal{R}|}. \tag{13}$$

This combines average response magnitude with sublinear weighting of component support. The center with largest $|L_q^*|$ serves as a representative seed. Back-tracing associates each candidate with its view, seed, member identities, WSI-level abundances, and source coordinates. Cell views return cellular locations; regional views return patch locations and extents. *Appendix B.2* gives refinement and screening procedure.

Cross-cohort candidate transfer. For external evaluation, each discovered candidate is represented by its view, preprocessing transform, query centers, adaptive radii, and Euclidean metric, all fixed from the discovery cohort. A target entity is assigned to a candidate when its transformed representation falls within at least one corresponding frozen center–radius neighborhood. Candidate geometry and assignment thresholds therefore remain unchanged across cohorts; *Appendix C* gives the formal membership and abundance definitions.

4 Experiments

4.1 Data and evaluation protocol

Orion-CRC (40 slides) provides colorectal discovery candidates frozen for independent external validation in TCGA-COAD (462 slides). Private bile-duct carcinoma (80 slides) and breast cancer (100 slides) cohorts provide additional within-cohort discovery analyses (*Appendix C*). CellID is trained on paired H&E–mIF Orion-CRC data (Lin et al., 2023); its 19-channel colorectal representation transfers to TCGA-COAD H&E images (Liu et al., 2018) as morphology-derived projections, not measured mIF profiles. Bile-duct and breast representations have 22 and 10 channels, respectively (Table 6, *Appendix C*). Cohort sizes count WSIs, each contributing one abundance observation without patient aggregation. The separate Orion-CRC cell-level sensitivity workload includes 24 WSIs from 23 patients, including two from CRC33.

Adverse prognosis denotes recurrence within 24 months or death within 60 months; favorable prognosis requires no recurrence within 24 months and survival beyond 60 months, including later censoring. TCGA-COAD labels are used only after frozen-candidate assignment and abundance aggregation, without repeating discovery (*Appendix C*).

Comparators and metrics. The marker-discovery comparison uses common CellID features and outcome labels. UMAP-based clustering, diffcyt-style clustering and differential testing, scCODA-style compositional modeling, and Squidpy-style spatial cluster-pair analysis generate biomarker candidates. Candidates are compared by WSI-level abundance differences. These adaptations use partition-first or predefined-neighborhood strategies under a common representation; they do not reproduce each method's original inferential model. Separate patient-level evaluations examine Milo neighborhood outputs and HIPT representations. Their distinct tasks and outputs are reported separately from Table 3 (*Appendix E.3*). Metrics include median $-\log_{10}(p)$, minimum p-value, number of regions passing the raw threshold, mean spot entropy, and peak memory. *Appendix C* details baseline adaptations and metrics, including entropy calculation and two-stage averaging.

Table 2: **Representative multiscale candidates and external evaluation.** Colorectal candidates are discovered in Orion-CRC (Lin et al., 2023) and frozen before TCGA-COAD (Liu et al., 2018) evaluation. Separate columns report raw discovery and validation p-values. Displayed signs are from Orion-CRC and were concordant in TCGA-COAD. Breast and bile-duct entries give within-cohort candidate ID, sign, and raw p-value. Negative/positive signs indicate adverse/favorable enrichment.

Layer	Colorectal			Breast (private)	Bile duct (private)
	Candidate (sign)	Orion-CRC discovery p	TCGA-COAD validation p		
Cell	$M_1^{cell}(-)$	3.64×10^{-5}	9.84×10^{-7}	$M_1^{cell}(-, 4.44 \times 10^{-11})$	$M_1^{cell}(-, 1.59 \times 10^{-2})$
	$M_3^{cell}(+)$	5.98×10^{-5}	6.04×10^{-11}	$M_2^{cell}(-, 1.36 \times 10^{-8})$	$M_2^{cell}(+, 1.59 \times 10^{-2})$
	$M_5^{cell}(+)$	7.42×10^{-4}	3.03×10^{-2}	$M_9^{cell}(+, 9.27 \times 10^{-3})$	$M_7^{cell}(+, 4.26 \times 10^{-2})$
Env.	$M_1^{env}(-)$	3.66×10^{-5}	2.61×10^{-10}	$M_1^{env}(-, 3.99 \times 10^{-10})$	$M_1^{env}(+, 3.65 \times 10^{-2})$
	$M_3^{env}(+)$	4.69×10^{-5}	2.29×10^{-2}	$M_2^{env}(+, 5.07 \times 10^{-9})$	$M_2^{env}(-, 1.59 \times 10^{-2})$
	$M_5^{env}(+)$	4.78×10^{-4}	9.16×10^{-7}	$M_4^{env}(-, 3.72 \times 10^{-2})$	$M_3^{env}(+, 1.71 \times 10^{-2})$
Tis.	$M_2^{tis}(+)$	2.09×10^{-2}	1.14×10^{-6}	$M_1^{tis}(-, 4.63 \times 10^{-9})$	$M_1^{tis}(+, 3.17 \times 10^{-2})$
	$M_3^{tis}(+)$	4.58×10^{-2}	2.32×10^{-2}	$M_2^{tis}(+, 4.27 \times 10^{-7})$	$M_2^{tis}(+, 3.65 \times 10^{-2})$
	$M_1^{tis}(-)$	4.83×10^{-2}	3.13×10^{-2}	$M_3^{tis}(-, 1.56 \times 10^{-2})$	$M_3^{tis}(-, 1.59 \times 10^{-2})$
T–C	$M_{1\text{-}2}^{\text{T-C}}(+)$	5.94×10^{-9}	2.65×10^{-5}	$M_{1\text{-}4}^{\text{T-C}}(-, 2.46 \times 10^{-16})$	$M_{1\text{-}1}^{\text{T-C}}(+, 1.51 \times 10^{-2})$
	$M_{1\text{-}1}^{\text{T-C}}(-)$	5.30×10^{-10}	2.85×10^{-6}	$M_{1\text{-}2}^{\text{T-C}}(+, 1.15 \times 10^{-27})$	$M_{1\text{-}2}^{\text{T-C}}(+, 3.49 \times 10^{-2})$
	$M_{2\text{-}1}^{\text{T-C}}(-)$	8.33×10^{-10}	6.96×10^{-6}	$M_{2\text{-}1}^{\text{T-C}}(-, 2.49 \times 10^{-21})$	$M_{1\text{-}3}^{\text{T-C}}(+, 3.14 \times 10^{-2})$

Implementation. Multiscale sampling uses 1,500 uniformly sampled patches per WSI. Morphology descriptors have $d_1 = 1000$ dimensions. Radius initialization uses $n_{\text{init}} = 10{,}000$, $k_{\text{init}} = 5$, percentile $\eta = 10$, scale factors $\alpha_{\min} = 0.05$ and $\alpha_{\max} = 0.5$, and fallback multiplier $\beta = 5$. Default search uses $J = 70$ radius steps, smoothing width $w = 5$, stopping persistence $\tau = 1$, pseudo-count base $\psi_{\text{base}} = 5.0$, and numerical stabilizer $\epsilon = 10^{-6}$. Boundary acceptance requires $n_{\min} = 200$ entities; fallback thresholds are $\delta_{\text{fb}} = 0.1$ and $n_{\text{fb}} = 20$. Graph refinement uses $k = 3$ neighbors, and abundance screening uses $\theta_p = 0.05$. Efficiency is measured on a Linux server with two Intel Xeon Platinum 8358 CPUs and 128 logical threads; adaptive querying and graph refinement run on CPU. *Appendix D* summarizes the representation and search parameters. Search-only, component-ablation, and full-pipeline runtimes are reported separately for their distinct workloads.

4.2 Multiscale candidates and tissue interpretation

Figure 2 descriptively projects the representations, adaptive responses, and selected components. AutoMarker identified candidates across four views in discovery cohorts. Table 2 presents 36 representative candidates: 12 from Orion-CRC, frozen for TCGA-COAD evaluation, and 12 each from private breast and bile-duct cohorts, extending within-cohort discovery to other cancer types. Colorectal IDs and signs refer to Orion-CRC; private-cohort IDs are specific to each cohort and view.

Orion-CRC candidates span adverse- and favorable-enriched patterns at cellular, microenvironment, tissue, and tissue-conditioned cell scales. Externally evaluated candidates retained their enrichment directions, with TCGA-COAD validation p-values reported separately in Table 2. Discovery-cohort abundance distributions show both between-group differences and within-group heterogeneity (Figure 3), linking outcome associations to complementary biological scales.

The broader component inventory in *Appendix C.1*, Table 7, complements these screened candidates by showing the candidate landscape across cohorts and representation views.

Tissue interpretation and spatial inspection. Pathological inspection linked representative cell, microenvironment, and tissue candidates to recognizable glandular, stromal, lymphocytic, and desmoplastic patterns (Figure 4). Patient-level Kaplan–Meier plots (Kaplan & Meier, 1958) and WSI back-tracing illustrate their prognostic associations and spatial localization (Figure 4; *Appendix C.2*, Figures 6–7).

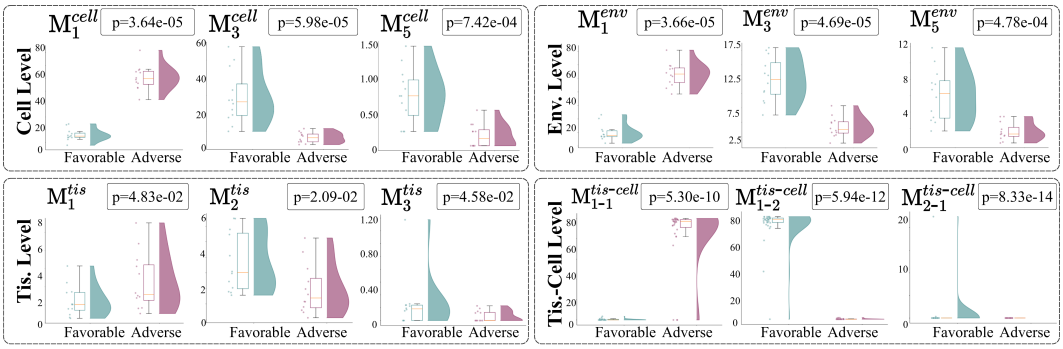

Figure 3: **Colorectal marker abundance.** The top three representative candidates in each view are compared between favorable- and adverse-outcome groups.

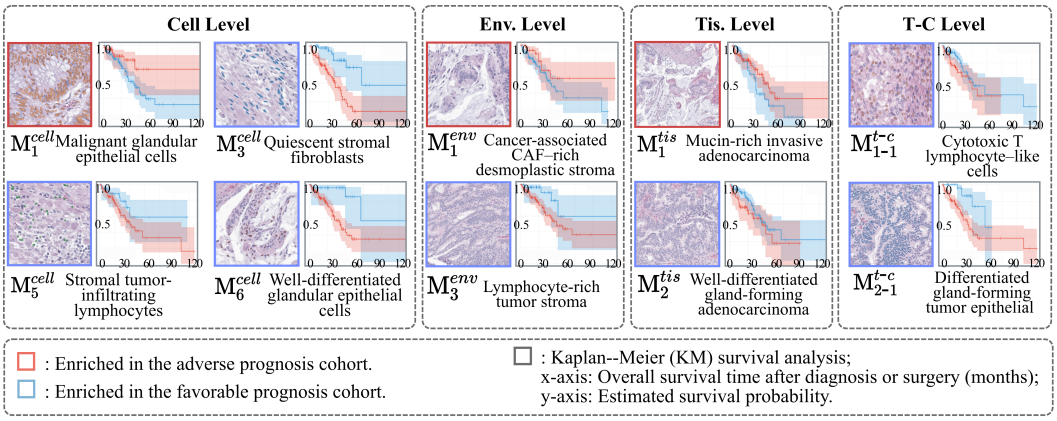

Figure 4: **Candidate and survival examples.** Expert-assigned descriptions accompany Kaplan–Meier plots. Frames indicate candidate enrichment direction, whereas Kaplan–Meier curve colors distinguish high- and low-abundance patient groups defined by the cohort median. Candidate abundance is averaged within patient before stratification (*Appendix C*).

4.3 Method Assessment and Computational Efficiency

Candidate outputs under a shared protocol. Under the shared screening protocol (Table 3), AutoMarker produced 11 regions meeting the raw threshold, with a median $-\log_{10}(p)$ of 4.437, mean entropy of 0.548, and peak memory usage of 426.6 MB. The association statistics describe discovery-set outputs: AutoMarker uses outcome labels during candidate localization, whereas the adaptations differ in when labels enter candidate generation. The comparison therefore characterizes the resulting candidates and computational requirements, rather than establishing out-of-sample statistical superiority. Counts below the raw threshold report absolute candidate yield, not the fraction of all tested candidates.

Boundary selection and component ablations. Search and refinement ablations use the same embeddings, labels, and LogFC scoring rule (Table 4). Fixed-radius and single-step search reduced runtime to about 100 seconds but yielded five regions passing the raw threshold versus 11 for the full method, with lower median $-\log_{10}(p)$.

Curvature stopping reduced runtime from 4157.7 to 558.3 seconds (7.4×), with little change in the median association statistic (4.437 versus 4.431). Removing graph pruning increased coverage overlap from 0.052 to 0.241 and reduced regions passing the threshold from 11 to two. Graph refinement thus reduced redundancy while retaining more candidates under the screening threshold. *Appendix E* reports radius-step sensitivity and its normalized signal measure.

Table 3: **Marker-discovery comparison.** Baselines are task-specific adaptations; statistics use raw p-values.

Method	Median $-\log_{10}(p)$	Best p	Entropy	Memory (MB)	#Regions ($p < 0.05$)
UMAP+clustering	0.713	5.197×10^{-3}	0.596	1223.6	2
diffcyt-style	0.605	2.623×10^{-2}	0.565	1267.1	3
scCODA-style	2.027	7.259×10^{-3}	0.563	1197.1	2
Squidpy-style	2.308	3.142×10^{-3}	0.592	1013.2	9
AutoMarker	**4.437**	$\mathbf{3.648 \times 10^{-5}}$	**0.548**	**426.6**	**11**

Table 4: **Search and refinement ablations.** The full method balances observed association strength, coverage redundancy, and runtime. All significance counts use the raw threshold.

Variant	Median $-\log_{10}(p)$	**#** $p < 0.05$	**Overlap**[a]	Mean entropy	**Time (s)**[b]	**#Regions**
Fixed-radius	1.027	5	0.118	0.562	100.3	8
Single-step	1.161	5	0.093	0.559	100.0	8
w/o curvature stopping	4.431	10	0.067	0.553	4157.7	13
w/o graph pruning	0.851	2	0.241	0.571	426.6	5
AutoMarker	**4.437**	**11**	**0.052**	**0.548**	558.3	**13**

[a]Overlap is the mean Jaccard similarity between retained coverage sets; lower is less redundant.
[b]Time includes querying, component construction, pruning, WSI-level abundance screening, and summarization.

A separate cell-level analysis compared adaptive search with fixed-radius, k-distance, and Kneedle alternatives. Adaptive search retained 13 outputs with 42.79 mean radius steps; removing sign pruning produced one low-contrast component, while omitting screening retained 95 components instead of 13 (*Appendix E.1*, Table 14).

Patient-level and robustness analyses. Milo identified 922 of 21,941 neighborhoods at $\mathrm{SpatialFDR} < 0.05$, with median support of five patients; significant neighborhoods also occurred in all ten patient-label permutations. HIPT achieved a higher leave-one-patient-out AUROC point estimate (0.758 versus 0.697), but the paired difference interval crossed zero. Under outcome-label reversal, the direction-normalized search reproduced all stopping radii and final candidates, while patient aggregation preserved the enrichment direction of all 13 candidates in the separate cell-level workload (*Appendices E.1–E.3*).

Computational efficiency. AutoMarker was $5.8\times$ and $2.0\times$ faster than fixed-window and graph-based search on 200,000 entities and scaled to two million entities with approximately 3.6 GB peak memory (*Appendix E*).

5 Conclusion

AutoMarker connects clinical outcome contrasts to multiscale biomarker candidates with explicit membership, adaptive scale, enrichment direction, spatial support, and tissue location. Protein-aligned views, direction-normalized search, and sign-consistent refinement enable traceable discovery across complementary cellular, microenvironmental, and tissue contexts. By linking representation-space candidates back to their supporting WSIs, the framework couples outcome association with inspectable pathological structure rather than predictive importance alone. Evaluation combines frozen-candidate testing in TCGA-COAD, pathological interpretation, spatial backtracing, and million-entity mining, while complementary Milo and HIPT analyses further clarify the contribution of consolidated candidate discovery. Retrospective sampling, WSI-level screening, outcome-guided selection, and feature-metric dependence remain limitations. Patient-level and label-reversal analyses support robustness within the tested workloads, while causal interpretation, clinical utility, and broader generalization require further evidence through selection-aware, independent, and prospective evaluation (*Appendices E–F*).

AI USE STATEMENT

ChatGPT was used for language editing, manuscript organization, and LaTeX assistance. The scientific ideas, methodology, experiments, code, figures, results, and conclusions were developed, verified, and finalized by the authors.

ETHICS STATEMENT

This study used de-identified retrospective data and was approved by the relevant institutional ethics committee, with informed consent waived under applicable institutional procedures. The protocol identifier is withheld for double-blind review.

REPRODUCIBILITY STATEMENT

Section 3 defines the representations, divergence response, and component screening, and Section 4.1 specifies the experimental setup and main parameters. *Appendices A–D* provide the full equations, pseudocode, cohort configurations, and implementation details; *Appendix E* reports computational measurements. An anonymous repository providing the implementation and supporting materials is available at `https://anonymous.4open.science/r/automarker-code-E0AC/`. Orion-CRC and TCGA-COAD provide public data for colorectal discovery and external validation, respectively; private-cohort data are subject to institutional authorization.

REFERENCES

Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975. doi: 10.1145/361002.361007.

Robert V. Bruggner, Bernd Bodenmiller, David L. Dill, Robert J. Tibshirani, and Garry P. Nolan. Automated identification of stratifying signatures in cellular subpopulations. *Proceedings of the National Academy of Sciences*, 111(26):E2770–E2777, 2014. doi: 10.1073/pnas.1408792111.

M. Büttner, J. Ostner, C. L. Müller, F. J. Theis, and B. Schubert. scCODA is a Bayesian model for compositional single-cell data analysis. *Nature Communications*, 12(1):6876, 2021. doi: 10.1038/s41467-021-27150-6.

Gabriele Campanella, Matthew G. Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J. Busam, Edi Brogi, Victor E. Reuter, David S. Klimstra, and Thomas J. Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8):1301–1309, 2019. doi: 10.1038/s41591-019-0508-1.

Richard J. Chen, Ming Y. Lu, Muhammad Shaban, Chengkuan Chen, Tiffany Y. Chen, Drew F. K. Williamson, and Faisal Mahmood. Whole Slide Images are 2D Point Clouds: Context-Aware Survival Prediction Using Patch-Based Graph Convolutional Networks. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pp. 339–349, 2021. URL `https://miccai2021.org/openaccess/paperlinks/2021/09/01/530-Paper2410.html`.

Richard J. Chen, Chengkuan Chen, Yicong Li, Tiffany Y. Chen, Andrew D. Trister, Rahul G. Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16144–16155, 2022. URL `https://openaccess.thecvf.com/content/CVPR2022/html/Chen_Scaling_Vision_Transformers_to_Gigapixel_Images_via_Hierarchical_Self-Supervised_Learning_CVPR_2022_paper.html`.

Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas

Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024. doi: 10.1038/s41591-024-02857-3.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607, 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Emma Dann, Neil C. Henderson, Sarah A. Teichmann, Michael D. Morgan, and John C. Marioni. Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nature Biotechnology*, 40(2):245–253, 2022. doi: 10.1038/s41587-021-01033-z.

Narmin Ghaffari Laleh, Hannah Sophie Muti, Chiara Maria Lavinia Loeffler, Amelie Echle, Oliver Lester Saldanha, Faisal Mahmood, Ming Y. Lu, Christian Trautwein, Rupert Langer, Bastian Dislich, Roman D. Buelow, Heike Irmgard Grabsch, Hermann Brenner, Jenny Chang-Claude, Elizabeth Alwers, Titus J. Brinker, Firas Khader, Daniel Truhn, Nadine T. Gaisa, Peter Boor, Michael Hoffmeister, Volkmar Schulz, and Jakob Nikolas Kather. Benchmarking weakly-supervised deep learning pipelines for whole slide classification in computational pathology. *Medical Image Analysis*, 79:102474, 2022. doi: 10.1016/j.media.2022.102474.

Yuxuan Hu, Jiazhen Rong, Yafei Xu, Runzhi Xie, Jacqueline Peng, Lin Gao, and Kai Tan. Unsupervised and supervised discovery of tissue cellular neighborhoods from cell phenotypes. *Nature Methods*, 21(2):267–278, 2024. doi: 10.1038/s41592-023-02124-2.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2127–2136, 2018. URL https://proceedings.mlr.press/v80/ilse18a.html.

E. L. Kaplan and Paul Meier. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958. doi: 10.1080/01621459.1958.10501452.

Junbum Kim, Samir Rustam, Juan Miguel Mosquera, Scott H. Randell, Renat Shaykhiev, André F. Rendeiro, and Olivier Elemento. Unsupervised discovery of tissue architecture in multiplexed imaging. *Nature Methods*, 19(12):1653–1661, 2022. doi: 10.1038/s41592-022-01657-2.

Jia-Ren Lin, Yu-An Chen, Daniel Campton, Jeremy Cooper, Shannon Coy, Clarence Yapp, Juliann B. Tefft, Erin McCarty, Keith L. Ligon, Scott J. Rodig, Steven Reese, Tad George, Sandro Santagata, and Peter K. Sorger. High-plex immunofluorescence imaging and traditional histology of the same tissue section for discovering image-based biomarkers. *Nature Cancer*, 4(7):1036–1052, 2023. doi: 10.1038/s43018-023-00576-1.

Jianfang Liu, Tara Lichtenberg, Katherine A. Hoadley, Laila M. Poisson, Alexander J. Lazar, Andrew D. Cherniack, Albert J. Kovatich, Christopher C. Benz, Douglas A. Levine, Adrian V. Lee, Larsson Omberg, Denise M. Wolf, Craig D. Shriver, Vesteinn Thorsson, The Cancer Genome Atlas Research Network, and Hai Hu. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell*, 173:400–416.e11, 2018. doi: 10.1016/j.cell.2018.02.052.

Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, 2021. doi: 10.1038/s41551-020-00682-w.

H. B. Mann and D. R. Whitney. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947. doi: 10.1214/aoms/1177730491.

Giovanni Palla, Hannah Spitzer, Michal Klein, David Fischer, Anna Christina Schaar, Louis Benedikt Kuemmerle, Sergei Rybakov, Ignacio L. Ibarra, Olle Holmberg, Isaac Virshup, Mohammad Lotfollahi, Sabrina Richter, and Fabian J. Theis. Squidpy: a scalable

framework for spatial omics analysis. *Nature Methods*, 19(2):171–178, 2022. doi: 10.1038/s41592-021-01358-2.

Christian M. Schürch, Salil S. Bhate, Graham L. Barlow, Darci J. Phillips, Luca Noti, Inti Zlobec, Pauline Chu, Sarah Black, Janos Demeter, David R. McIlwain, Shigemi Kinoshita, Nikolay Samusik, Yury Goltsev, and Garry P. Nolan. 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.

Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis*, 81:102559, 2022. doi: 10.1016/j.media.2022.102559.

Lukas M. Weber, Malgorzata Nowicka, Charlotte Soneson, and Mark D. Robinson. diffcyt: Differential discovery in high-dimensional cytometry via high-resolution clustering. *Communications Biology*, 2(1):183, 2019. doi: 10.1038/s42003-019-0415-5.

Jun Zhao, Ariel Jaffe, Henry Li, Ofir Lindenbaum, Esen Sefik, Ruaidhrí Jackson, Xiuyuan Cheng, Richard A. Flavell, and Yuval Kluger. Detection of differentially abundant cell subpopulations in scRNA-seq data. *Proceedings of the National Academy of Sciences*, 118(22):e2100293118, 2021. doi: 10.1073/pnas.2100293118.

APPENDIX

TABLE OF CONTENTS

A Representation and View Construction **13**

A.1 Cell feature extraction and cross-modal alignment 13

B Adaptive Search and Component Algorithms **14**

B.1 Per-view query semantics and retained outputs 15

B.2 Graph refinement 16

C Cohorts, Baselines, and Reporting Scope **16**

C.1 Candidate-component inventory 19

C.2 Spatial back-tracing across hierarchical views 19

D Implementation Parameters **21**

E Extended Efficiency and Sensitivity Results **21**

E.1 Supplementary component measurements 25

E.2 Outcome-label reversal and direction-normalized boundary selection 26

E.3 Complementary patient-level baseline evaluations 27

F Limitations and Discussion **28**

A REPRESENTATION AND VIEW CONSTRUCTION

Entity correspondence. Let x_{HE} and $x_{\mathrm{mIF}} = \{c_0, \ldots, c_K\}$ be a registered patch pair, with a DAPI channel for nuclear segmentation. Morphological and DAPI-based segmentations identify the matched entity set $\mathcal{M} = \{m_n\}_{n=1}^{N_{\mathrm{cell}}}$. The intersection notation used to describe this set denotes correspondence after registration; it does not specify an instance-matching algorithm. Masked H&E crops yield morphology features $e^n_{\mathrm{HE}} \in \mathbb{R}^{d_1}$, while radius-constrained aggregation within each mask yields $e^n_{\mathrm{mIF}} \in \mathbb{R}^K$. The two attention branches provide cell embeddings $\hat{e}^n_{\mathrm{HE}}$ and global patch tokens $\hat{E}_{\mathrm{HE}}$ and $\hat{E}_{\mathrm{mIF}}$. Patch-level contrastive alignment and cell-level regression train CellID. The patch objective aligns regional representations, and the cell objective anchors morphology-derived vectors to the measured protein profiles.

View records and metadata. The cell and tissue-conditioned cell views store (z^n, Y, π^n), where π^n includes cell coordinates, patch offsets, and WSI identifiers. Microenvironment and tissue records store (z_P, Y, ϕ_P) with region-level spatial identifiers. Tissue conditioning collects cells from the H&E patches constituting tissue candidate $\mathcal{P}$: $\Omega(\mathcal{P}) = \bigcup_{p \in \mathcal{B}(\mathcal{P})} \mathcal{C}(p)$. Here $\mathcal{B}(\mathcal{P})$ is the set of constituent H&E patches and $\mathcal{C}(p)$ is the set of cell instances spatially contained in patch p. Each cellular record contains the K aligned features. For microenvironment pooling, d_i and w_i are aggregation quantities, not extra indexed feature dimensions. Tissue conditioning retains the aligned cell vectors of the selected tissue context; it does not append a tissue token to each cell vector.

A.1 CELL FEATURE EXTRACTION AND CROSS-MODAL ALIGNMENT

The following equations detail the paired-training formulation introduced in Section 3.2, including cellular inputs, token construction, and alignment objectives.

Matched cellular inputs. For a registered patch pair, modality-specific segmentations provide corresponding cell masks m_n. Each matched H&E crop is encoded as

$$e^n_{\mathrm{HE}} = \mathcal{F}_{\mathrm{enc}}(m_n \otimes x_{\mathrm{HE}}) \in \mathbb{R}^{d_1}. \tag{14}$$

The morphology encoder $\mathcal{F}_{\mathrm{enc}}$ is CTransPath. Its $d_1 = 1000$ descriptor is projected by $\mathcal{F}_{\mathrm{fc}}$ to the K-dimensional branch input; the descriptor dimension is distinct from the aligned representation dimension. For mIF, the DAPI channel is used for segmentation, and the remaining K channels define the target vector

$$e^n_{\mathrm{mIF}} = [v_{n,1}, \ldots, v_{n,K}], \qquad v_{n,k} = \mathrm{Avg}(c_k, m_n, r_{\mathrm{cell}}, v_{\mathrm{th}}). \tag{15}$$

Here Avg averages channel-k pixels in the masked cellular region subject to a radius constraint and an intensity threshold. The settings are $r_{\mathrm{cell}} = 8.5$ pixels at $20\times$ magnification and $v_{\mathrm{th}} = 25$.

Token construction. Let N_P be the number of matched cells in patch P. A fully connected map projects each H&E descriptor before the Transformer branch:

$$\widetilde{e}^n_{\mathrm{HE}} = \mathcal{F}_{\mathrm{fc}}(e^n_{\mathrm{HE}}) \in \mathbb{R}^K, \tag{16}$$

$$(\{\widehat{e}^n_{\mathrm{HE}}\}_{n=1}^{N_P}, \widehat{E}_{\mathrm{HE}}) = \mathcal{T}_{\mathrm{HE}}(\{\widetilde{e}^n_{\mathrm{HE}}\}_{n=1}^{N_P}, E_{\mathrm{HE}}), \tag{17}$$

$$(\{\widehat{e}^n_{\mathrm{mIF}}\}_{n=1}^{N_P}, \widehat{E}_{\mathrm{mIF}}) = \mathcal{T}_{\mathrm{mIF}}(\{e^n_{\mathrm{mIF}}\}_{n=1}^{N_P}, E_{\mathrm{mIF}}). \tag{18}$$

E_{HE} and E_{mIF} denote aggregation tokens supplied in the corresponding branch's input space. Their output tokens and contextual cell vectors are K-dimensional. Self-attention contextualizes the cellular inputs, and attention-based aggregation forms the patch token. Tokens from corresponding registered patches form positive pairs. Negative mIF patch tokens are drawn from other spatial locations, with $T = 2048$ per positive pair.

Alignment objectives. The patch contrastive and cell regression objectives are defined in Eqs. 2–3. Here n indexes cells within the current patch P, whereas i denotes the corresponding entity in the main text. For the same cell, $\widehat{e}^n_{\mathrm{HE}} = \hat{e}_i$ and $e^n_{\mathrm{mIF}} = e_i^{\mathrm{mIF}}$. The patch tokens $\widehat{E}_{\mathrm{HE}}$ and $\widehat{E}_{\mathrm{mIF}}$ correspond to $\hat{E}_P^{\mathrm{HE}}$ and $\hat{E}_P^{\mathrm{mIF}}$, respectively; $\hat{E}_P$ denotes the H&E output at inference. The encoder $\mathcal{F}_{\mathrm{enc}}$ and projection $\mathcal{F}_{\mathrm{fc}}$ use the same notation in both sections. Cell regression targets the extracted mIF vectors e^n_{mIF}, not the transformed cell outputs $\widehat{e}^n_{\mathrm{mIF}}$. The paired-training formulation jointly updates the two branches and the H&E feature projection. At target-slide inference, the H&E branch supplies $\widehat{e}^n_{\mathrm{HE}}$ for cellular views and $\widehat{E}_{\mathrm{HE}}$ for tissue representations without requiring target mIF measurements.

View pooling. For cells in a region P, the row sums of the cell self-attention matrix provide aggregation degrees $d_i = \sum_{j \in P} A_P[i, j]$. Normalizing these degrees gives the weights used in Eq. 4. The weights derive from cell attention rather than a prescribed function of spatial distance. This pooling produces a regional composition vector, whereas the tissue view uses the H&E branch's output patch token. If A_P is row-stochastic, all row sums equal one and this pooling reduces to an arithmetic mean; nonuniform pooling requires nonuniform row sums.

Paired spatial examples. Figure 5 complements the representation specification with matched morphology and channel-level spatial views. Examining the same field across channels connects cellular location to the displayed feature profile. The paired fields provide spatial examples of the morphology and channel profiles represented by the encoder.

B Adaptive Search and Component Algorithms

The bounds and grid in Eqs. 6–7 are initialized once per indexed view and shared by its queries. Algorithm 1 summarizes direction-normalized adaptive search. The first valid nonzero response establishes the query direction, sign reversals reuse the preceding response, and curvature is computed on $U_{q,t} = s_q \widetilde{L}_{q,t}$. The selected response L^*_q remains signed. A search boundary is retained when the curvature-based stopping condition persists for τ evaluations and the current neighborhood contains at least 200 entities. The current radius and its smoothed response are accepted together. If no such boundary is found, the maximum tested radius is evaluated under the fallback response and neighborhood-support criteria in Eq. 10. Zero-only trajectories yield no informative candidate.

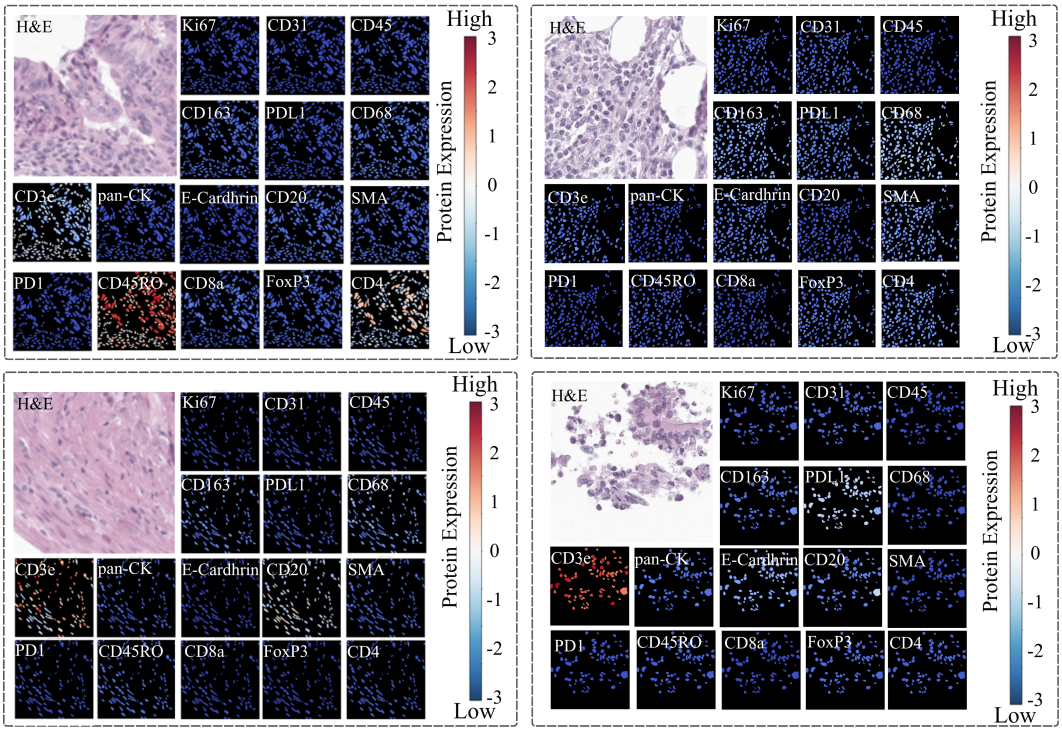

Figure 5: **Paired morphology and channel-level spatial profiles.** Four representative fields show H&E morphology alongside cell-resolved overlays for the indicated channels. The displayed normalized values use the accompanying color scales. The panels illustrate the spatial correspondence underlying protein-aligned representations.

B.1 Per-view query semantics and retained outputs

For view ℓ, the indexed neighborhood and its group subsets are

$$\mathcal{N}_g^{(\ell)}(q,r) = \{i : \|z_i^{(\ell)} - z_q^{(\ell)}\|_2 \le r,\ Y_{s(i)} = g\}, \qquad n_g^{(\ell)}(q,r) = |\mathcal{N}_g^{(\ell)}(q,r)|. \tag{19}$$

Equation 5 is applied to these per-view counts. A cell query compares cellular feature neighborhoods; a microenvironment query compares pooled region vectors; a tissue query compares patch tokens. A tissue-conditioned query is restricted to the cells selected by its tissue context $\Omega(\mathcal{P})$. Each resulting record retains the view, center identity, selected radius, response, slide identity, and spatial metadata. Radius is measured in representation space, while physical tissue locations are represented separately by the stored spatial coordinates.

Discrete expansion and sign handling. The search evaluates 70 radii and smooths the retained signed responses with a five-step moving average. The first nonzero response fixes the direction; a later sign reversal reuses the preceding response. The second difference of the direction-normalized response $U_{q,t} = s_q \widetilde{L}_{q,t}$ defines the discrete curvature used by the stopping rule; the stored LogFC remains signed. Algorithm 1 specifies the finite-step boundary selection and fallback convention.

Centers, coverage, and tissue coordinates. Equation 11 defines the union of retained query neighborhoods. The center set $\mathcal{R}$ defines graph membership and the score in Eq. 13; $\mathcal{U}(\mathcal{R})$ defines the expanded neighborhood coverage. Equation 12 expresses WSI-level abundance using the center-component set; each WSI contributes one abundance observation to screening. Stored coordinates map component members and neighborhood instances to their source tissue locations.

Algorithm 1 Direction-normalized adaptive search with neighborhood-support filtering

Require: View index $\mathcal{I}^{(\ell)}$, query q, radii $r_0 < \cdots < r_{J-1}$ ($J = 70$), smoothing width $w = 5$, persistence $\tau = 1$, minimum neighborhood support $n_{\min} = 200$, fallback neighborhood support $n_{\text{fb}} = 20$, fallback response threshold $\delta_{\text{fb}} = 0.1$, γ, $\psi_{\text{base}} = 5.0$, $\epsilon = 10^{-6}$

Ensure: Selected radius r_q^*, neighborhood, and smoothed response L_q^*, or no candidate

$s_q \leftarrow 0$; $c \leftarrow 0$; $\mathit{valid} \leftarrow$ false; initialize signed and oriented response arrays
for $t = 0, \ldots, J-1$ **do**
  Retrieve $\mathcal{N}(q, r_t)$; compute L_t using Eq. 5
  if $s_q = 0$ and $L_t \neq 0$ **then**
    $s_q \leftarrow \text{sign}(L_t)$
  end if
  if $s_q \neq 0$ and $\text{sign}(L_t) \neq s_q$ **then**
    $L_t \leftarrow L_{t-1}$
  end if
  $\widetilde{L}_t \leftarrow \text{Mean}(L_{\max(0,t-w+1)}, \ldots, L_t)$
  $U_t \leftarrow s_q \widetilde{L}_t$
  if $s_q \neq 0$ and $t \geq 2$ **then**
    $a_t \leftarrow U_t - 2U_{t-1} + U_{t-2}$
    if $a_t \leq 0$ **then**
      $c \leftarrow c + 1$
    else
      $c \leftarrow 0$
    end if
    if $c \geq \tau$ and $|\mathcal{N}(q, r_t)| \geq n_{\min}$ **then**
      $r_q^* \leftarrow r_t$; $L_q^* \leftarrow \widetilde{L}_t$; $\mathit{valid} \leftarrow$ true
      break
    end if
  end if
end for
if not valid **then**
  if $|\widetilde{L}_{J-1}| > \delta_{\text{fb}}$ and $|\mathcal{N}(q, r_{J-1})| \geq n_{\text{fb}}$ **then**
    $r_q^* \leftarrow r_{J-1}$; $L_q^* \leftarrow \widetilde{L}_{J-1}$; $\mathit{valid} \leftarrow$ true
  else
    return no informative candidate
  end if
end if
return $r_q^*, \mathcal{N}(q, r_q^*), L_q^*$

B.2 Graph refinement

Algorithm 2 summarizes the graph stage using only valid query centers returned by Algorithm 1. An intermediate component is a connected component before WSI-level abundance screening; a retained candidate passes screening, and a representative marker is a selected candidate used for interpretation. Table 7 lists component-level statistics before the final filter, whereas Algorithm 2 returns ranked candidates after filtering. The graph partitions query centers, whereas expanded neighborhood coverage can overlap. Component membership and ranking use the retained centers; coverage overlap is evaluated separately on their neighborhood unions. The procedure returns all components that pass filtering, ranked by score; it imposes no fixed maximum output count. Selecting three representative components for display does not truncate the algorithm output.

C Cohorts, Baselines, and Reporting Scope

Table 5 summarizes the cohort sizes and analysis settings.

Algorithm 2 Sign-consistent component refinement and WSI-level abundance screening

Require: Valid query centers V retained by Algorithm 1, responses $\{L_q^*\}$, graph parameter $k = 3$, threshold $\theta_p = 0.05$, WSI identifiers, outcome labels assigned to each WSI
Ensure: All retained components ranked by score, with spatial metadata
Build a k-nearest-neighbor graph $G = (V, E)$
Remove $(u, v) \in E$ whenever $\mathrm{sign}(L_u^*) \neq \mathrm{sign}(L_v^*)$
$\mathcal{R}_{\mathrm{out}} \leftarrow \varnothing$
for all connected components $\mathcal{R}$ of G **do**
  Compute one abundance value per WSI using Eq. 12
  Compute a two-sided Mann–Whitney U test between outcome groups using WSIs as observations
  if $p(\mathcal{R}) \leq \theta_p$ **then**
    Compute $\mathrm{Score}(\mathcal{R})$ using Eq. 13
    Select seed $q^* \in \arg\max_{q \in \mathcal{R}} |L_q^*|$
    Add component, score, seed, and spatial metadata to $\mathcal{R}_{\mathrm{out}}$
  end if
end for
return all entries in $\mathcal{R}_{\mathrm{out}}$, sorted by decreasing $\mathrm{Score}(\mathcal{R})$

Table 6: **Dataset-specific imaging-channel configurations.** Channel labels follow the supplied dataset configurations and include nuclear stains. TCGA-COAD uses the 19-channel representation projected from Orion-CRC; it has no measured mIF panel in this study.

Dataset	**Cancer type**	**Channel configuration**	K
Orion-CRC	Colorectal	Hoechst, AF1, CD31, CD45, CD68, Argo550, CD4, FOXP3, CD8a, CD45RO, CD20, PD-L1, CD3e, CD163, E-Cadherin, PD1, Ki67, panCK, SMA	19
TCGA-COAD	Colorectal	Projected features aligned to the Orion-CRC channel configuration above; H&E-only evaluation.	19
Private cohort	Breast	DAPI, P53, KI67, PanCK, GATA3, ER, PR, AR, SYN, HER2	10
Private cohort	Bile duct	DAPI, CD20, CD3, CD56, CD117, SMA, CK19, CD8, CD103, PD1, GZMB, PD-L1, CD4, FoxP3, Mum-1, CD206, CD68, CD66b, CD15, CD11c, CD83, HBsAg	22

Table 5: **Reported cohort configuration.** Counts refer to slides. The TCGA channel count denotes the projected representation, not a measured mIF panel.

Cohort	Slides	Channels	Analysis setting
Orion-CRC	40	19	Paired encoder training and discovery
TCGA-COAD	462	19	Independent external validation
Private bile-duct carcinoma	80	22	Within-cohort analysis
Private breast cancer	100	10	Within-cohort analysis

Channel configurations. Table 6 lists the channel labels for Orion-CRC and the private cohorts and identifies the corresponding projected representation used for TCGA-COAD. The listed configurations include nuclear-stain channels, so their entries are not all protein markers.

Discovery and external validation. Orion-CRC is the colorectal discovery cohort, and TCGA-COAD is the independent external validation cohort. Candidate center–radius geometry, view, and representation/preprocessing parameters are established in Orion-CRC and held fixed during TCGA-COAD assignment. After all target assignments and abundance aggregation are complete, TCGA-COAD outcome labels are used to test the associations of these fixed candidates with clinical outcomes. Each WSI contributes one abundance observation to the two-sided Mann–Whitney U

comparison; repeated WSIs from the same patient are not collapsed. The private breast and bile-duct cohorts are analyzed separately within cohort.

Frozen geometry and external membership. For candidate j, the frozen discovery specification consists of its view ℓ_j, the Orion-fitted representation and preprocessing transform $T_{\text{disc}}^{(\ell_j)}$, the Euclidean distance metric in that transformed space, and all center–radius pairs $\{(z_q^{\text{disc}}, r_q^*) : q \in \mathcal{R}_j\}$. Here z_q^{disc} is the stored representation of Orion query center q. The complete set of pairs defines the candidate; it is not replaced by a centroid or a common radius. All fitted normalization/scaling parameters and any fitted projection parameters are carried over from Orion. They are not re-estimated on TCGA, and the local membership thresholds remain the Orion radii r_q^*.

Let z denote a target entity's input to the frozen view transform. Membership is determined by

$$\mathbf{1}_j(z) = \mathbb{I}\Big[\exists q \in \mathcal{R}_j : \Big\| T_{\text{disc}}^{(\ell_j)}(z) - z_q^{\text{disc}} \Big\|_2 \le r_q^*\Big]. \tag{20}$$

Thus $\mathcal{M}_j^{\text{ext}} = \{z : \mathbf{1}_j(z) = 1\}$ is the union of the frozen Orion balls evaluated on target entities, corresponding to the coverage construction in Eq. 11. TCGA supplies new H&E-derived entities; it supplies neither new candidate centers nor a new radius-selection procedure.

Candidate assignments are evaluated independently. A target entity may belong to multiple candidates when their frozen coverage regions overlap; no nearest-candidate competition or exclusive assignment is imposed. Within one candidate, an entity is counted once even if it lies in several constituent balls. Consequently, abundances across different candidates need not sum to 100%.

For the target entities $\mathcal{S}_{s,\text{ext}}^{(\ell_j)}$ in WSI s, external abundance is

$$H_s^{\text{ext}}(j) = 100\,\frac{\sum_{z \in \mathcal{S}_{s,\text{ext}}^{(\ell_j)}} \mathbf{1}_j(z)}{|\mathcal{S}_{s,\text{ext}}^{(\ell_j)}|}. \tag{21}$$

The factor 100 retains the percentage convention used in Eq. 12. This target coverage statistic counts assigned target entities, not the Orion query centers. Discovery center membership and external geometric coverage are distinct quantities and are defined separately.

TCGA outcome labels, association p-values, and survival data do not enter the representation transform, candidate assignment, or membership-threshold selection. Assignment and abundance aggregation precede their use in the WSI-level Mann–Whitney U comparison or patient-level survival analysis. For survival analysis, target WSI abundances are averaged within patient before cohort-median stratification, as described below. The outcome-independent assignment does not change the primary WSI-level observation-unit convention.

Observation unit. Statistical abundance screening uses WSIs as the observation units. Each WSI contributes one abundance observation, and repeated WSIs from the same patient are not collapsed in the primary analysis. The 24-WSI Orion-CRC cell-level benchmark contains 23 patients; CRC33 contributes two WSI observations. These repeated observations share a patient outcome label and may be correlated.

Survival stratification. For each fixed biomarker candidate, abundance is first calculated separately for every WSI, using Eq. 12 for discovery-cohort center membership and Eq. 21 for external frozen-neighborhood coverage. For this survival analysis, let $\mathcal{S}_p$ denote the WSIs from patient p. The applicable WSI abundances, denoted generically by $H_s(\mathcal{R})$ below, are averaged to obtain one value per patient:

$$H_p(\mathcal{R}) = \frac{1}{|\mathcal{S}_p|} \sum_{s \in \mathcal{S}_p} H_s(\mathcal{R}). \tag{22}$$

The cohort median patient-level abundance defines the cutoff between high- and low-abundance groups. Patients are the independent units for Kaplan–Meier estimation and the log-rank test. The cutoff is not optimized using survival outcomes.

Baseline adaptations. The UMAP (Becht et al., 2019) baseline projects the cellular features, then clusters and tests abundance. The diffcyt-style (Weber et al., 2019) baseline clusters in the original feature space and tests cluster abundance differences. The scCODA-style (Büttner et al., 2021) baseline applies compositional analysis to the common cluster-count matrix. The Squidpy-style (Palla et al., 2022) baseline creates spatially adjacent cluster-pair markers and evaluates their WSI-level abundance differences. The scCODA-style candidate markers are evaluated through WSI-level abundance comparisons using the same downstream marker-evaluation protocol as the other methods (Table 3).

Evaluation metrics. Association statistics are calculated from WSI-level candidate abundances, with one observation per WSI and no within-patient aggregation. Mean spot entropy instead measures local spatial heterogeneity of reference phenotype-cluster composition. We standardize the instance features dimension by dimension across the full instance set and fit MiniBatch K-means with $C = 12$ clusters and random seed 42. Let $c_j \in \{1, \ldots, C\}$ be the reference cluster label of instance j. For each instance i, $\mathcal{V}_i$ contains eight spatial neighbors from the same WSI, ranked second through ninth by distance among non-self neighbors. The implementation uses scikit-learn 1.7.2, requests nine neighbors without an explicit query matrix, and then discards the first returned neighbor. The local cluster proportions and normalized entropy are

$$p_{i,c} = \frac{|\{j \in \mathcal{V}_i : c_j = c\}|}{|\mathcal{V}_i|}, \tag{23}$$

$$h_i = -\frac{\sum_{c:p_{i,c}>0} p_{i,c} \ln(p_{i,c} + \varepsilon_H)}{\ln(C + \varepsilon_H)}, \qquad \varepsilon_H = 10^{-8}. \tag{24}$$

The logarithms are natural logarithms. The normalization uses all 12 reference clusters, although a local neighborhood contains only eight observations; the maximum is approximately $\ln 8 / \ln 12 \approx 0.837$.

For a benchmark region $\mathcal{R}$, its mean spot entropy is the arithmetic mean over its member instances,

$$h_{\mathcal{R}} = \frac{1}{|\mathcal{R}|} \sum_{i \in \mathcal{R}} h_i. \tag{25}$$

Let $\mathcal{B}_{\text{valid}}$ denote a method's retained benchmark regions with non-NaN region-level entropy. The method-level summary is

$$\overline{h} = \frac{1}{|\mathcal{B}_{\text{valid}}|} \sum_{\mathcal{R} \in \mathcal{B}_{\text{valid}}} h_{\mathcal{R}}. \tag{26}$$

This two-stage mean weights instances equally within each region and regions equally within each method. The benchmark region family includes regions above and below the abundance-screening threshold; it is distinct from the final screened candidate set. NaN region-level values are omitted using `nanmean`. Thus, the mean entropy in Tables 3 and 4 is a method-level average of spatial composition measurements, not an entropy of outcome labels or WSI-level abundances.

C.1 Candidate-component inventory

Table 7 provides the inventory of candidate components before final WSI-level abundance screening, together with their signed LogFC values and group-abundance p-values. This broader inventory complements the screened representative candidates in Table 2 and shows the candidate landscape across cohorts and representation views.

C.2 Spatial back-tracing across hierarchical views

Figure 6 shows representative spatial maps of biomarker candidates across the four hierarchical views. Representation-space components are mapped to their supporting entities using the stored spatial coordinates; a component need not form a contiguous region in tissue.

Representation-to-morphology correspondence. Figure 7 complements the whole-slide maps in Figure 6 by linking the cell-level embedding to local H&E instances. The paired views support inspection of cellular shape and surrounding tissue context at the displayed locations. Component

Table 7: **Candidate-component landscape across cohorts and representation views.** Entries report component ID, signed LogFC, and raw abundance-test p-values. For colon, parenthesized p-values give TCGA-COAD validation results for frozen Orion-CRC candidates; LogFC values refer to Orion-CRC, with concordant enrichment directions in TCGA-COAD. Negative/positive LogFC indicates adverse/favorable enrichment. Candidates on both sides of the screening threshold are shown. Blank cells denote no entry; T–C denotes tissue-conditioned cell.

	Colon (Orion-CRC & TCGA-COAD)	**Breast** (private)	**Bile Duct** (private)
Cell	$M_1^{cell}(-0.69, 3.64\times10^{-5}(9.84\times10^{-7}))$	$M_1^{cell}(-0.67, 4.44\times10^{-11})$	$M_1^{cell}(-1.87, 1.59\times10^{-2})$
	$M_2^{cell}(+0.50, 1.35\times10^{-3}(7.21\times10^{-2}))$	$M_2^{cell}(-0.32, 1.36\times10^{-8})$	$M_2^{cell}(+0.89, 1.59\times10^{-2})$
	$M_3^{cell}(+0.36, 5.98\times10^{-5}(6.04\times10^{-11}))$	$M_3^{cell}(+3.20, 8.14\times10^{-2})$	$M_3^{cell}(+0.94, 5.50\times10^{-2})$
	$M_4^{cell}(-0.61, 1.18\times10^{-1}\ (1.91\times10^{-1}))$	$M_4^{cell}(-0.97, 3.97\times10^{-3})$	$M_4^{cell}(+3.22, 1.31\times10^{-1})$
	$M_5^{cell}(+0.24, 7.42\times10^{-4}\ (3.03\times10^{-2}))$	$M_5^{cell}(-0.76, 9.16\times10^{-2})$	$M_5^{cell}(+2.09, 1.31\times10^{-1})$
	$M_6^{cell}(-0.50, 1.60\times10^{-2}\ (9.39\times10^{-2}))$	$M_6^{cell}(+0.43, 9.34\times10^{-2})$	$M_6^{cell}(+1.33, 1.31\times10^{-1})$
	$M_7^{cell}(+0.34, 8.89\times10^{-3}\ (1.10\times10^{-2}))$	$M_7^{cell}(-0.44, 8.37\times10^{-2})$	$M_7^{cell}(+1.26, 4.26\times10^{-2})$
	$M_8^{cell}(-0.34, 9.60\times10^{-2}\ (1.08\times10^{-1}))$	$M_8^{cell}(-0.37, 1.16\times10^{-1})$	$M_8^{cell}(+1.26, 4.16\times10^{-2})$
	$M_9^{cell}(+0.21, 8.31\times10^{-2}\ (1.98\times10^{-1}))$	$M_9^{cell}(+0.30, 9.27\times10^{-3})$	$M_9^{cell}(+0.76, 4.81\times10^{-2})$
	$M_{10}^{cell}(+0.24, 1.36\times10^{-1}\ (2.00\times10^{-2}))$	$M_{10}^{cell}(-0.23, 1.17\times10^{-3})$	$M_{10}^{cell}(+1.58, 1.28\times10^{-1})$
	$M_{11}^{cell}(+0.23, 4.52\times10^{-2}\ (1.41\times10^{-1}))$	$M_{11}^{cell}(+0.30, 2.25\times10^{-2})$	$M_{11}^{cell}(+0.80, 1.06\times10^{-1})$
	$M_{12}^{cell}(+0.24, 1.16\times10^{-1}\ (1.89\times10^{-1}))$	$M_{12}^{cell}(-6.65, 4.43\times10^{-11})$	$M_{12}^{cell}(+0.78, 1.25\times10^{-1})$
	$M_{13}^{cell}(+0.15, 1.21\times10^{-1}\ (3.37\times10^{-1}))$	$M_{13}^{cell}(+0.24, 4.29\times10^{-3})$	$M_{13}^{cell}(+0.71, 6.06\times10^{-2})$
Env.	$M_1^{env}(-0.77, 3.66\times10^{-5}\ (2.61\times10^{-10}))$	$M_1^{env}(-0.85, 3.99\times10^{-10})$	
	$M_2^{env}(+0.45, 5.11\times10^{-3}\ (9.12\times10^{-7}))$	$M_2^{env}(+0.46, 5.07\times10^{-9}\ )$	$M_1^{env}(+0.75, 3.65\times10^{-2})$
	$M_3^{env}(+0.36, 4.69\times10^{-5}\ (2.29\times10^{-2}))$	$M_3^{env}(-1.81, 5.74\times10^{-2})$	$M_2^{env}(-0.42, 1.59\times10^{-2})$
	$M_4^{env}(+0.43, 6.09\times10^{-3}\ (2.78\times10^{-3}))$	$M_4^{env}(-1.47, 3.72\times10^{-2})$	$M_3^{env}(+0.60, 1.71\times10^{-2})$
	$M_5^{env}(+0.32, 4.78\times10^{-4}\ (9.16\times10^{-7}))$	$M_5^{env}(-1.82, 2.41\times10^{-2})$	$M_4^{env}(+0.60, 9.63\times10^{-2})$
	$M_6^{env}(-1.40, 4.81\times10^{-2}\ (2.62\times10^{-2}))$	$M_6^{env}(-0.55, 6.81\times10^{-2})$	$M_5^{env}(+0.88, 1.31\times10^{-1})$
	$M_7^{env}(-0.68, 1.72\times10^{-1}\ (8.15\times10^{-2}))$	$M_7^{env}(-0.48, 9.37\times10^{-3})$	$M_6^{env}(+1.12, 1.31\times10^{-1})$
	$M_8^{env}(+0.32, 1.01\times10^{-2}\ (4.14\times10^{-4}))$	$M_8^{env}(-1.08, 3.15\times10^{-3})$	$M_7^{env}(+0.87, 1.31\times10^{-1})$
	$M_9^{env}(+0.38, 5.21\times10^{-3}\ (6.39\times10^{-2}))$	$M_9^{env}(-0.98, 4.12\times10^{-3})$	
	$M_{10}^{env}(+0.32, 1.08\times10^{-1}\ (1.66\times10^{-2}))$	$M_{10}^{env}(+0.42, 1.03\times10^{-2})$	
	$M_{11}^{env}(+0.42, 4.76\times10^{-3}\ (8.45\times10^{-2}))$	$M_{11}^{env}(-0.62, 7.53\times10^{-2})$	
Tis.		$M_1^{tis}(-0.82, 4.63\times10^{-9})$	
		$M_2^{tis}(+0.52, 4.27\times10^{-7})$	
	$M_1^{tis}(-0.47, 4.83\times10^{-2}(3.13\times10^{-2}))$	$M_3^{tis}(-1.71, 1.56\times10^{-2})$	$M_1^{tis}(+0.73, 3.17\times10^{-2})$
	$M_2^{tis}(+0.43, 2.09\times10^{-2}\ (1.14\times10^{-6}))$	$M_4^{tis}(-1.82, 1.12\times10^{-3})$	$M_2^{tis}(+0.56, 3.65\times10^{-2})$
	$M_3^{tis}(+0.36, 4.58\times10^{-2}\ (2.32\times10^{-2}))$	$M_5^{tis}(-1.22, 1.73\times10^{-1})$	$M_3^{tis}(-0.39, 1.59\times10^{-2})$
		$M_6^{tis}(-1.66, 3.15\times10^{-3})$	
		$M_7^{tis}(-1.34, 1.11\times10^{-2})$	
T-C	$M_{1\text{-}1}^{\text{T-C}}(-5.54, 5.31\times10^{-10}\ (2.85\times10^{-6}))$	$M_{1\text{-}1}^{\text{T-C}}(-1.75, 2.46\times10^{-16})$	
	$M_{1\text{-}2}^{\text{T-C}}(+1.40, 5.94\times10^{-9}\ (2.65\times10^{-5}))$	$M_{1\text{-}2}^{\text{T-C}}(+0.64, 1.15\times10^{-27})$	
	$M_{1\text{-}3}^{\text{T-C}}(+1.30, 1.87\times10^{-1}\ (1.04\times10^{-14}))$	$M_{2\text{-}1}^{\text{T-C}}(-0.56, 2.49\times10^{-21})$	
	$M_{1\text{-}4}^{\text{T-C}}(+0.80, 1.87\times10^{-1}\ (5.39\times10^{-8}))$	$M_{2\text{-}2}^{\text{T-C}}(-0.73, 1.29\times10^{-8})$	
	$M_{2\text{-}1}^{\text{T-C}}(-2.64, 8.33\times10^{-10}\ (6.96\times10^{-6}))$	$M_{2\text{-}3}^{\text{T-C}}(+0.27, 4.03\times10^{-26})$	$M_{1\text{-}1}^{\text{T-C}}(+0.60, 1.51\times10^{-2})$
	$M_{2\text{-}2}^{\text{T-C}}(-4.88, 7.06\times10^{-2}\ (1.48\times10^{-3}))$	$M_{2\text{-}4}^{\text{T-C}}(-0.58, 1.48\times10^{-3})$	$M_{1\text{-}2}^{\text{T-C}}(+0.73, 3.49\times10^{-2})$
	$M_{2\text{-}3}^{\text{T-C}}(-3.57, 3.58\times10^{-2}\ (3.34\times10^{-2}))$	$M_{2\text{-}5}^{\text{T-C}}(-0.78, 3.34\times10^{-2})$	$M_{1\text{-}3}^{\text{T-C}}(+0.53, 3.14\times10^{-2})$
	$M_{2\text{-}4}^{\text{T-C}}(+0.60, 9.89\times10^{-7}\ (3.26\times10^{-2}))$	$M_{3\text{-}1}^{\text{T-C}}(-1.46, 8.47\times10^{-2})$	
	$M_{3\text{-}1}^{\text{T-C}}(-1.59, 6.43\times10^{-8}\ (1.54\times10^{-1}))$	$M_{3\text{-}2}^{\text{T-C}}(+1.04, 3.94\times10^{-2})$	
	$M_{3\text{-}2}^{\text{T-C}}(+0.56, 3.12\times10^{-10}\ (9.61\times10^{-2}))$	$M_{3\text{-}3}^{\text{T-C}}(-0.51, 1.72\times10^{-1})$	
	$M_{3\text{-}3}^{\text{T-C}}(+1.41, 4.84\times10^{-2}\ (2.53\times10^{-1}))$	$M_{3\text{-}4}^{\text{T-C}}(+0.38, 3.94\times10^{-2})$	

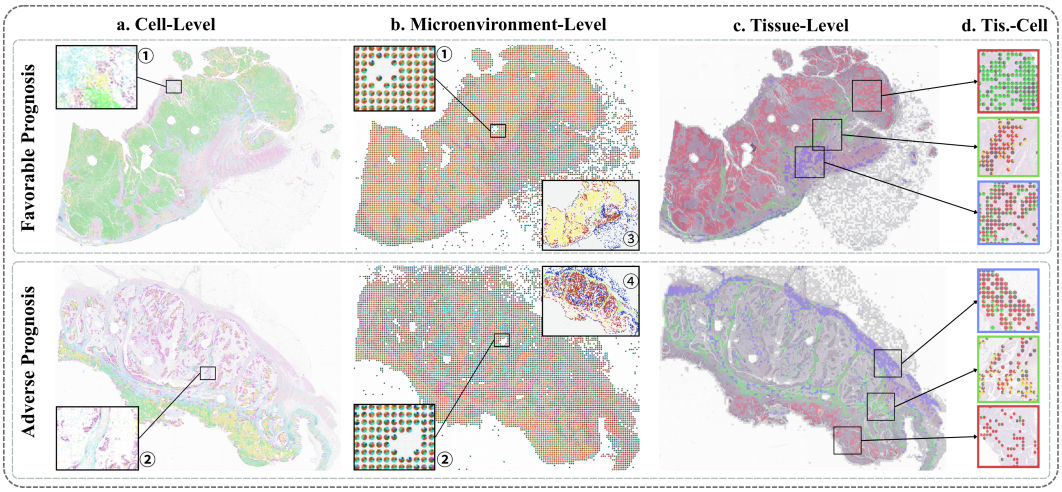

Figure 6: **Spatial back-tracing of biomarker candidates.** A representative whole-slide image (WSI) from the favorable-prognosis group is shown in the top row, and one from the adverse-prognosis group in the bottom row. (a) Cell level. Individual cells are colored by discovered candidate components, highlighting cell-level compositional differences between prognostic groups. (b) Microenvironment level. Each patch is represented by a pie chart (①,②), with slices indicating the relative expression of protein channels within the local neighborhood. Insets (③,④) show the spatial localization of microenvironment-level candidates back-traced to the WSI. (c) Tissue level. Tissue-level candidates are shown in different colors, revealing their spatial organization and distribution across the WSI. (d) T–C level. Conditioned on the tissue-level candidates in (c), cell-level discovery is performed within each tissue region to reveal fine-grained cellular differences within the same tissue context.

colors identify visualized groups and should not be interpreted using the enrichment-frame colors of Figure 4.

D Implementation Parameters

Table 8 summarizes the representation and search parameters. The CPU configuration applies to the mining benchmarks. The realized bounds apply to the Orion-CRC cell-level configuration with 19 features, 120,000 entities per outcome group, and seed 42; other views estimate their own bounds. Boundary and fallback neighborhood support denote entity counts within an individual query-center neighborhood; they do not refer to graph-component size or the number of supporting WSIs.

Sampling checks. The sampling experiment evaluates spatial coverage and sampled-count balance (Table 9). Spatial-grid sampling covers all evaluated tiles, and cohort balancing reduces the sampling-count coefficient of variation from 0.420 to 0.

E Extended Efficiency and Sensitivity Results

Measurement scope. Table 10 reports the end-to-end mining measurements. These experiments measure distinct workloads: single-query latency, search-only runtime, component-ablation runtime, and end-to-end mining. The end-to-end measurements cover mining from extracted features, including feature loading.

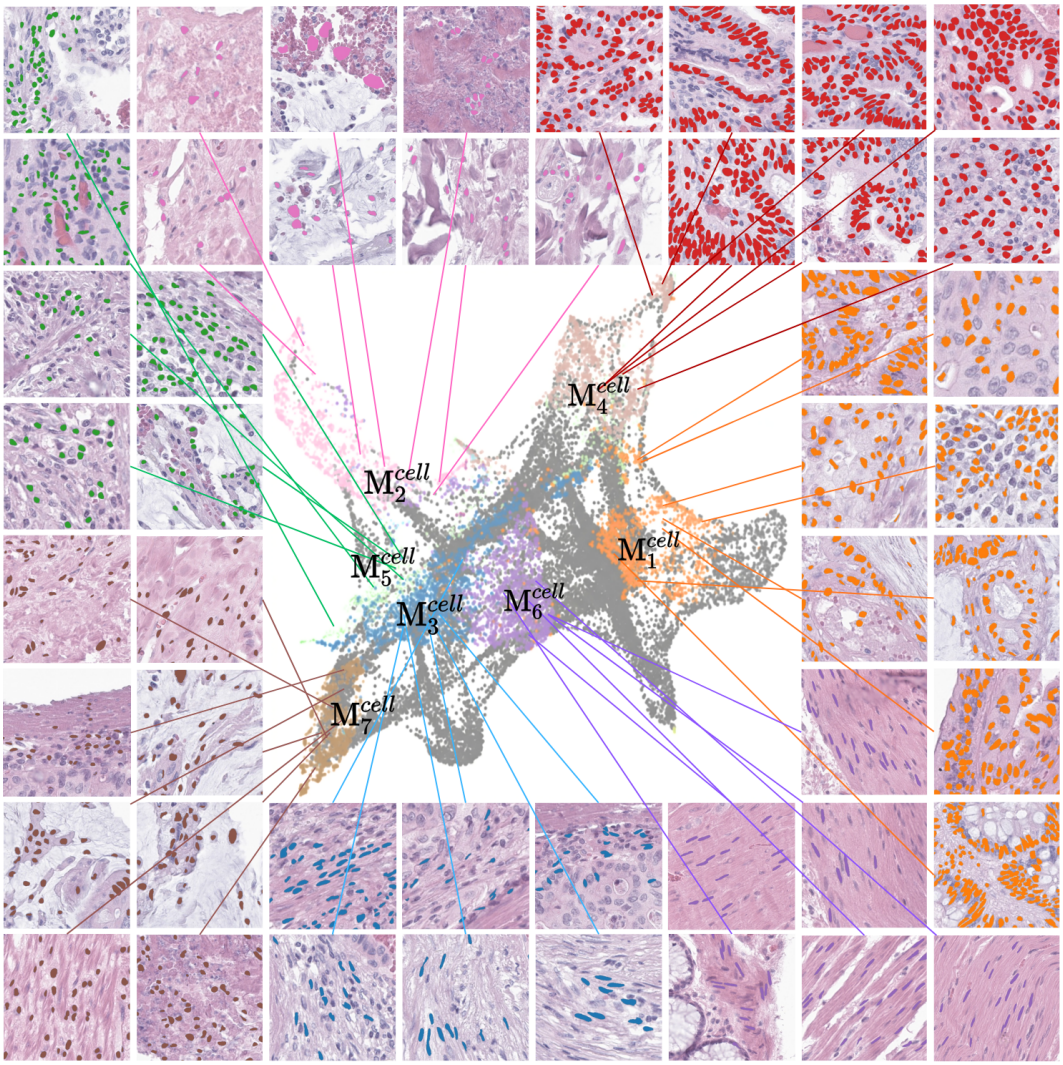

Figure 7: **Cell-level representation and morphological examples.** The central UMAP (Becht et al., 2019) highlights seven labeled components, M_1^{cell}–M_7^{cell}. Peripheral H&E fields show representative cellular instances with colored overlays, and connecting lines indicate their corresponding component membership. Colors distinguish candidate components. The paired representation and morphology views illustrate the cellular patterns associated with each component.

Table 8: **Representation and search parameters.**

Parameter	Setting
Morphology feature dimension d_1	1000 (morphology descriptor)
Aligned dimension K	19 (colorectal), 22 (bile duct), 10 (breast)
mIF aggregation radius r_{cell}	8.5 pixels at 20×
Intensity threshold v_{th}	25
Patch sampling	1,500 patches per WSI for multiscale discovery
Abundance threshold θ_p	0.05
Radius initialization	Eq. 6; up to 10,000 sampled entities, mean five-neighbor distances and tenth percentile
Radius fallback	$r_{\max} = 5r_{\min}$ if $r_{\min} \geq r_{\max}$
Radius sequence	70 linearly spaced radii, Eq. 7
Orion-CRC realized bounds	$r_{\min} = 0.027523$, $r_{\max} = 0.071968$ (cell-level configuration below)
Stopping trajectory	$U_{q,t} = s_q \widetilde{L}_{q,t}$; signed LogFC retained for outputs
Stopping persistence	$\tau = 1$
Graph neighbors	$k = 3$
Boundary neighborhood support	$n_{\min} = 200$ entities
Fallback neighborhood support	$n_{\text{fb}} = 20$ entities
Fallback response threshold	$\delta_{\text{fb}} = 0.1$
Random seed	42
Smoothing window	5
Negative patch tokens	$T = 2048$ for contrastive alignment
Pseudo-count	$\psi_{\text{base}} = 5.0$; $\psi(r) = \psi_{\text{base}} r_{\min}/(r + \epsilon)$
Numerical stabilizer	$\epsilon = 10^{-6}$
CPU platform	Two Intel Xeon Platinum 8358 CPUs; 128 logical threads

Table 9: **Sampling checks.** Spatial-grid sampling and cohort balancing address different properties of the sampled data. Coverage refers to the tiles evaluated by the sampling experiment. CV denotes the coefficient of variation of sampled counts.

Sampling component	Evaluation	Reported result
Spatial-grid sampling	Tile coverage	Complete
Cohort balancing	Sampling-count CV before balancing	0.420
Cohort balancing	Sampling-count CV after balancing	0

Table 10: End-to-end scalability of AutoMarker under increasing numbers of nodes per outcome group.

Nodes / group	Total nodes	**KD-tree (s)**	Query + LogFC (s)	**Prune/Test (s)**	Total time (s)	Peak memory (MB)
$10K$	$20K$	0.006	54.6	0.25	54.9	818
$50K$	$100K$	0.041	412.5	2.40	414.9	970
$100K$	$200K$	0.083	1486.2	7.10	1493.4	1102
$500K$	$1M$	0.795	5580.3	348.8	6260.8	2114
$1M$	$2M$	1.494	12197.6	1053.3	13587.6	3588

The runtime is decomposed into KD-tree construction, adaptive range query with LogFC computation, and topology pruning with WSI-level abundance screening. Total time includes feature loading overhead. Peak memory denotes the maximum memory footprint during the end-to-end mining process.

Retrieval-backend substitution. Replacing KD-tree retrieval with HNSW on 240,000 cells with 19 features improves query efficiency while closely preserving the discovered outputs (Table 13). This experiment compares retrieval efficiency and output agreement across indexing backends. In Algorithm 1, a query is valid if it returns an informative candidate through either the curvature-

and-support condition or the maximum-radius fallback, before graph construction and abundance screening.

Table 11: **Search-only runtime on 200,000 entities.** The pairwise and scan values are lower bounds. This workload differs from the end-to-end and component-ablation runs.

Search method	Runtime (s)
Pairwise	$> 10{,}000$
Scan	$> 10{,}000$
Window	1826.4
Graph	641.8
AutoMarker	314.5

Table 12: Representative configurations from the radius-step sensitivity study.

Radius Steps	**Smooth Window**	**Avg. Signal (%)**	**Time (s)**
30	5	100	4.8
50	5	106	8.5
70	5	**108**	11.4

Table 13: **Retrieval-backend substitution on 240,000 cells with 19 features.** Query times have the scope of this backend benchmark; they are not full mining times. Radius tolerance is one grid step. Top-region Jaccard is distinct from whole-candidate recovery.

Metric	KD-tree	HNSW
Build time (s)	0.145	2.147
Query time (s)	4.888	2.544
Recall@200	1.000000	0.999921
Valid queries (%)	94.414	94.413
Retained regions	13	13
Radius within one grid step (%)	100	99.978
Top-region Jaccard vs. exact	1.000000	0.999999

Search implementations. Table 11 compares search-only runtimes under the shared workload. Pairwise materializes all pairwise distances; Scan retrieves range neighborhoods by brute force; Window evaluates a fixed-window search; Graph uses sparse k-NN neighborhoods. The AutoMarker search combines indexed range retrieval and adaptive expansion. The comparison uses $N = 2 \times 100{,}000$ entities and a shared feature representation and outcome-scoring rule. Values above 10,000 seconds are lower bounds. The speedups in this table compare methods within this search-only workload; the 314.5-second result is not the query-stage timing of the separate 200,000-entity end-to-end run in Table 10.

The total adaptive search cost depends on the number of radii visited and the entities returned at every step. For Q query centers, let J_q denote the number of radii visited by query q on the shared per-view radius grid $\{r_t\}_{t=0}^{J-1}$. A suitable accounting expression is

$$T_{\text{search}} = T_{\text{build}} + \sum_{q=1}^{Q} \sum_{t=0}^{J_q-1} \left[T_{\text{range}}(N, D, r_t) + O(m_{q,t}) \right], \tag{27}$$

where $m_{q,t}$ is the retrieved neighborhood size. Graph construction and refinement add separate costs. Materializing a complete distance matrix uses quadratic storage; this is distinct from storing only an $N \times D$ feature matrix. This accounting separates index construction, range retrieval, returned-neighborhood processing, and graph refinement.

Scaling and sensitivity. Figure 8 summarizes the query workload, data-size scaling, and radius-step sensitivity. The latency study reports mean latency near 1.5 ms, P95 near 5.6 ms, and throughput around 700 queries per second as the number of query centers increases. A logarithmic fit to the measured database-size latency trend yielded $R^2 = 0.968$. At two million entities, the full mining run uses 3588 MB peak memory. The runtime decomposition includes additional loading overhead in the total, so the listed computational stages do not always sum to that total.

Table 12 reports the radius-step sensitivity measurements. With the smoothing window fixed to five, 30, 50, and 70 radius steps give normalized average signals of 100%, 106%, and 108%, at 4.8, 8.5, and 11.4 seconds in the sensitivity workload. Seventy steps gives the largest signal among these tabulated configurations, at greater cost. Signal denotes the average differential strength of retained candidate regions. The tabulated signal is normalized to 100% for the 30-step configuration.

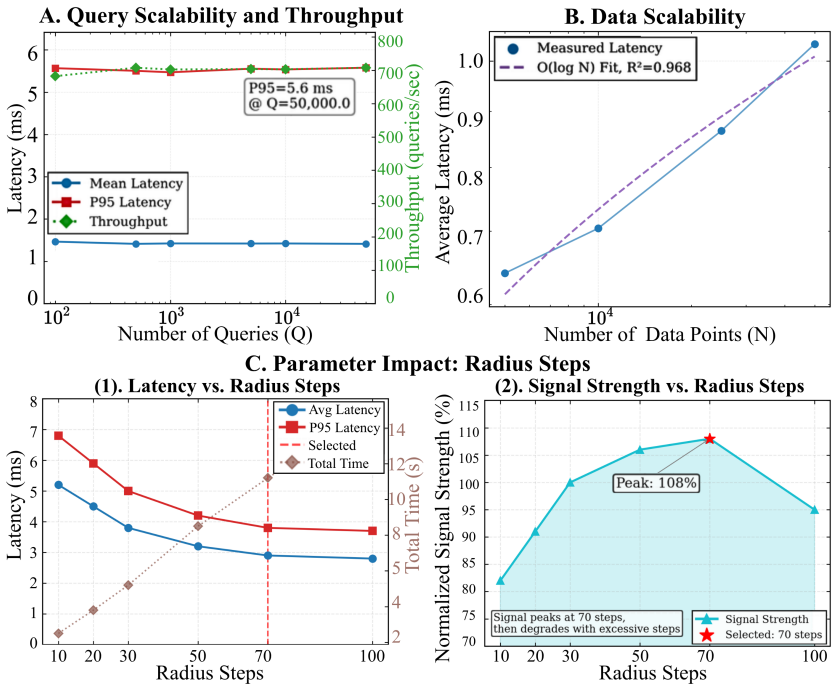

Figure 8: **Query workload, data-size scaling, and radius-step sensitivity.** (A) Query latency and throughput as query-center count varies. (B) Empirical latency trend as database size increases. (C) Radius-step sensitivity. Query-level measurements are distinct from the full mining timings in Table 10.

E.1 Supplementary component measurements

A separate Orion-CRC cell-level ablation analysis uses 240,000 entities (Table 14). It uses 120,000 entities per outcome group, 19 features, 70 radius steps, a five-step smoothing window, and $\tau = 1$. Table 4 instead reports the full search-and-refinement ablation workload.

Removing direction consistency expands the median component from 39 to 110,470.5 entities and its median WSI support from 11 to all 24 slides (Table 15). WSI-level abundance screening reduces 95 intermediate components to 13 retained candidates.

HNSW reduces query time by approximately $1.92\times$ in its backend benchmark but requires more construction time; the memory use and serialized index size are also larger. The backend comparison quantifies the trade-off between index construction and query time.

Patient-aggregation sensitivity and candidate traceability. This analysis concerns the 13 retained candidates from the separate Orion-CRC cell-level workload with 240,000 entities and 19

Table 14: **Boundary selection and component ablations.** Raw p-values; radius steps are averaged over all queries.

A. Boundary selection				B. Component ablations			
Rule	# Outputs	Median $-\log_{10}(p)$	Radius steps	Ablation	# Outputs	Median $-\log_{10}(p)$	Median $\lvert\text{LogFC}\rvert$
Adaptive (ours)	13	1.784	42.79	Full	13	1.784	0.1640
Fixed radius	1	2.613	1.00	w/o sign pruning	1	1.680	0.00130
k-distance	3	3.707	1.00	w/o screening	95	0.445	0.1827
Kneedle	14	1.609	58.04	w/o direction	2	4.012	0.8087

Table 15: **Support of screened components.** WSI support and prevalence refer to the 24-slide supplementary experiment. Size, supporting slides, and prevalence are medians over components.

Configuration	Entities	Supporting WSIs	WSI prevalence
Full	39	11	45.83%
w/o direction consistency	110,470.5	24	100%

features. The original two-sided Mann–Whitney U comparisons used 24 WSI observations. Averaging WSI abundances within each patient yielded 23 observations (11 favorable and 12 adverse). The fixed retained candidates were then evaluated using patient-mean abundances. Eleven had nominal $p < 0.05$, and eight had BH-adjusted $q < 0.05$ within this selected 13-candidate set. All 13 retained the enrichment direction indicated by the original LogFC. This sensitivity analysis does not replace the primary WSI-level results, and the within-set BH adjustment does not account for the preceding candidate selection.

For these 13 candidates, the WSI-level abundance vectors reproduce the saved original p-values exactly (maximum absolute discrepancy $= 0$). Candidate-level records link every retained component to its query center and source WSIs, including member/query count, center and median component radii, supporting WSI and patient counts, patient prevalence, score, and LogFC. These records refer to the retained graph-component entities/query points. The sensitivity and traceability results are specific to this cell-level workload.

E.2 Outcome-label reversal and direction-normalized boundary selection

Protocol. This experiment used the frozen Orion-CRC cell representation with 240,000 entities and 11 favorable versus 12 adverse patients. Entity features, query centers, radius grid, random seed, graph construction, neighborhood-support thresholds, scoring, and post-search filtering were held fixed; only the two outcome labels were reversed. The legacy signed-curvature rule was compared with the direction-normalized rule used in the final method. The latter fixes each query's direction from its first valid nonzero response and computes curvature on $U_{q,t} = s_q \widetilde{L}_{q,t}$, retaining signed LogFC for enrichment interpretation and graph pruning.

Results. Under label reversal, the legacy signed rule changed both query boundaries and downstream outputs (Table 16). The direction-normalized rule gave 100% agreement in query validity, executed steps, and stopping radii. Neighborhoods, components, rankings, and final candidates were unchanged. Both label assignments produced 90 intermediate components and 16 final candidates, with mean and median matched-component Jaccard equal to 1.000. At fixed queries, saved LogFC values were exactly antisymmetric: $\max_q |L_{q,\text{reversed}} + L_{q,\text{original}}| = 0$. Enrichment signs reversed as expected.

Interpretation and scope. If label reversal negates the signed trajectory, it also negates its initial direction: $s'_q = -s_q$ and $\widetilde{L}'_{q,t} = -\widetilde{L}_{q,t}$ give $U'_{q,t} = U_{q,t}$. The curvature test then receives the same oriented trajectory. The experiment verifies this behavior for the tested 11:12 patient configuration; it does not establish superior biological discovery accuracy or universal invariance for every imbal-

Table 16: **Outcome-label reversal on the frozen Orion-CRC cell-level representation.** Agreement compares each rule with itself after reversing outcome labels. Paired counts list original/reversed label assignments. The legacy rule is retained only as the comparison condition.

Metric	Legacy signed rule	Direction-normalized rule
Valid-query agreement	98.69%	100%
Executed-step agreement	35.42%	100%
Exact stopping-radius agreement	32.55%	100%
Median absolute radius difference	0.000644	0
Median relative radius difference	1.32%	0
LogFC sign reversal	100%	100%
Components, original/reversed labels	95/96	90/90
Mean matched-component Jaccard	0.728	1.000
Median matched-component Jaccard	0.963	1.000
Final candidates, original/reversed labels	13/16	16/16
Final-candidate match rate at Jaccard ≥ 0.5	43.75%	100%

ance pattern. The 90-component, 16-candidate counts belong to this label-reversal workload. The separate component-ablation, retrieval-backend, and patient-aggregation workloads retain their own reported counts and should not be numerically pooled with this experiment.

E.3 Complementary patient-level baseline evaluations

The following analyses address different levels of the pipeline. Milo evaluates differential graph neighborhoods; HIPT evaluates hierarchical representations on matched Orion-CRC regions. Their outputs are summarized alongside the separate AutoMarker cell-level sensitivity workload in Table 17. This table describes output structure and evidence scope, not a common-task ranking.

Milo neighborhood outputs. The patient-level Milo analysis tested 21,941 neighborhoods. Differentially abundant neighborhoods were defined using the prespecified spatially weighted FDR threshold $\mathrm{SpatialFDR} < 0.05$: 922 neighborhoods passed, and the minimum SpatialFDR was 0.000929. Milo's spatial FDR weighting accounts for neighborhood connectivity density through the distance to the kth nearest neighbor; it is distinct from a raw p-value threshold or an ordinary unweighted BH adjustment. Among the 922 neighborhoods, median patient support was five, and none was supported by at least 20 patients. All ten patient-label permutation runs produced neighborhoods meeting the identical $\mathrm{SpatialFDR} < 0.05$ threshold. This records whether each permutation produced any significant neighborhood; it does not specify a permutation distribution for the observed count, an empirical p-value for 922, or an estimated false-discovery proportion. The finding is specific to this workload and does not establish a general failure of Milo's inferential model. Neighborhood overlap, patient support, and component consolidation describe different properties of a discovery output.

HIPT representation comparison. The matched Orion-CRC region classification experiment used leave-one-patient-out evaluation and compared HIPT with the aligner tissue representation. Confidence intervals were estimated using 5,000 patient-level bootstrap resamples stratified by outcome group. The same resampled patients were used for both representations to form the paired AUROC difference, defined throughout as aligner minus HIPT. HIPT achieved an AUROC of 0.758 (95% CI, 0.530–0.932), and aligner achieved 0.697 (95% CI, 0.455–0.909). The paired difference was -0.061 (95% CI, -0.235 to 0.129). Thus HIPT had the higher point estimate, but the difference interval crossed zero; this does not establish equivalence. This experiment evaluates the two named tissue representations, separately from the CellID-based candidate-generation comparison in Table 3.

Table 17: **Complementary patient-level analyses.** Each row addresses a different output or task; counts and AUROC values are not directly comparable across rows. AutoMarker results refer only to the separate 13-candidate Orion-CRC cell-level workload.

Analysis	Main output	Patient-level evidence	Interpretation
AutoMarker	Consolidated candidate components	All 13 retained candidates preserved direction after patient aggregation	Candidate traceability and directional sensitivity
Milo	Overlapping graph neighborhoods	922/21,941 had $\mathrm{SpatialFDR} < 0.05$; median support 5 patients; permutation positives in 10/10 runs at the same threshold	Local differential neighborhoods with workload-specific support
HIPT	Hierarchical representation	AUROC 0.758 versus 0.697 for aligner; aligner minus HIPT: -0.061 (95% paired CI, -0.235 to 0.129)	Competitive representation baseline

F Limitations and Discussion

Representations and interpretation. Representation quality may vary with registration accuracy, protein-panel composition, and tissue-domain shift. H&E projections do not measure protein abundance in the target cohort. Tissue interpretations are hypotheses associated with the candidate components, not mechanistic assignments established by intervention. Retrospective associations may reflect clinical covariates or tissue sampling, and prospective evaluation is needed before clinical use.

Statistical scope. Repeated WSIs from a patient share the outcome label and may have correlated abundances. The primary WSI-level comparison does not model that dependence. The patient-aggregation sensitivity analysis covers only the separate Orion-CRC cell-level workload, not all main-table candidates. Likewise, applying a raw screening threshold does not provide false-discovery-rate control over outcome-guided candidate search. Adjusting the selected 13 candidates in the sensitivity analysis does not correct the preceding adaptive selection. Fixed-candidate external testing and within-cohort discovery should therefore be interpreted separately. Directional replication additionally requires the external enrichment sign, which cannot be inferred from the two-sided p-value alone.

Feature geometry and robustness scope. The data-adaptive radius range remains dependent on the feature metric and scaling; using a scalar global standard deviation does not make Euclidean search invariant to rescaling individual dimensions. The direction-normalized rule removes the observed label-naming dependence in the tested configuration, but label-reversal invariance does not establish biological discovery accuracy or robustness to every group-imbalance pattern (Appendix E.2).

Benchmark scope. The adapted baselines compare candidate-generation strategies under a shared representation, not complete original-method implementations or calibrated discovery rates. Absolute threshold-passing counts need the total number of tested candidates to become rates. Appendix E.3 adds patient-level Milo and HIPT analyses at their respective output and representation levels. These analyses do not provide independent-cohort evaluation of every method's frozen candidate set or a common-task ranking of discovery quality. Runtime ratios are interpretable within each benchmark: equal entity counts across tables do not establish equal numbers of queries, radii visited, or returned neighbors.

Appendix References

Etienne Becht, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel W H Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W Newell. Dimensionality reduction for visualizing

single-cell data using UMAP. *Nature Biotechnology*, 37(1):38–44, 2019. doi: 10.1038/nbt.4314.

M. Büttner, J. Ostner, C. L. Müller, F. J. Theis, and B. Schubert. scCODA is a Bayesian model for compositional single-cell data analysis. *Nature Communications*, 12(1), 2021. doi: 10.1038/s41467-021-27150-6.

Giovanni Palla, Hannah Spitzer, Michal Klein, David Fischer, Anna Christina Schaar, Louis Benedikt Kuemmerle, Sergei Rybakov, Ignacio L. Ibarra, Olle Holmberg, Isaac Virshup, Mohammad Lotfollahi, Sabrina Richter, and Fabian J. Theis. Squidpy: a scalable framework for spatial omics analysis. *Nature Methods*, 19(2):171–178, 2022. doi: 10.1038/s41592-021-01358-2.

Lukas M. Weber, Malgorzata Nowicka, Charlotte Soneson, and Mark D. Robinson. diffcyt: Differential discovery in high-dimensional cytometry via high-resolution clustering. *Communications Biology*, 2(1), 2019. doi: 10.1038/s42003-019-0415-5.