LOCAL INFLUENCE INFERENCE: LEARNING SHARED FEATURE RANKINGS FROM SPATIAL NEIGHBORHOODS

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial analysis often seeks to identify which features of a local environment influence an outcome at a focal point. The target is a single feature ranking shared across locations, with influence that decays over distance. Predictive accuracy alone does not ensure recovery of this ranking. In a distance-weighted linear model, we show that bag-averaged permutation importance can reverse two positive influence coefficients even at the true predictor when neighborhood size co-varies with within-neighborhood feature variation. We formulate *Local Influence Inference* and propose LION, which represents the ranking as a shared model parameter. LION combines a learnable distance kernel, a feature-preserving normalizer, and a global influence vector through bilinear aggregation, using $G + 4$ trainable parameters. Under correct specification and identification conditions, the reduced maximum-likelihood estimator consistently recovers all strict feature comparisons. Across seven ranking benchmarks spanning synthetic data and real tissue geometry, LION leads or ties distance-weighted logistic regression in mean Spearman on every benchmark, with the largest gains on morphology-interaction ($+0.22$) and Dense ($+0.11$). LION ranks first in AUROC on five of seven benchmarks among the evaluated deep and interpretable MIL baselines, with 12 to 64 parameters against thousands for the deep baselines. On three Xenium/TCR-profiled tumors, LION increases the mean reactive fraction among recovered T cells and the mean bystander fraction among missed T cells over every baseline at matched predicted-positive rates. These results support a shared feature ranking as a useful model output for spatial inference, alongside bag-level prediction.

1 INTRODUCTION

Many spatial studies ask *which features of the surrounding environment influence an outcome*. Spatially resolved transcriptomics lets immunologists ask which genes expressed by nearby cells recruit T cells into tumor niches (Moffitt et al., 2022; Longo et al., 2021) (Figure 1); ecological surveys ask which traits of established plants determine where seedlings take root (Wright et al., 2017); spatial economics asks which neighborhood attributes predict where small businesses succeed (Glaeser et al., 2010). In each setting, the outcome at a focal point depends on nearby entities whose influence decays with distance. The scientific target is a single feature ranking shared across focal points.

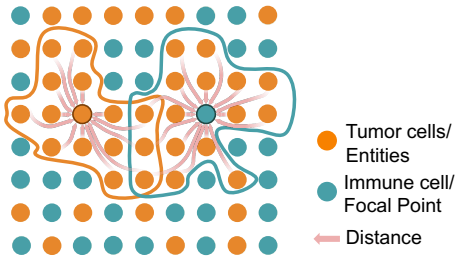

Figure 1: Illustration of local influence inference in immune cells. Each prediction is made over a bag of nearby entities with features and distances, with the goal of predicting the bag label and recovering a global feature influence vector.

We formalize this setting as *Local Influence Inference*. Each observation consists of a focal outcome and a neighborhood of entities with measured features and distances. The model predicts the outcome and estimates one feature ranking shared across all focal points. Standard prediction architectures do not generally expose this target directly: tabular learners require a fixed-size representation; graph neural networks (GNNs) (Kipf & Welling, 2016) aggregate neighbors through feature-mixing message passing; and common multiple-instance learning (MIL) architectures model the bag–instance

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.

structure but use per-prediction attention (Ilse et al., 2018; Lu et al., 2021; Li et al., 2021; Shao et al., 2021) or post-hoc explanations (Lundberg & Lee, 2017; Breiman, 2001; Sundararajan et al., 2017; Early et al., 2022). Distance-aware variants (Aftab et al., 2024; Kim et al., 2022) incorporate proximity into their encoders, but do not thereby identify a global influence coefficient for each observed feature. We show that bag-averaged within-bag permutation importance can reverse two positive coefficients on the probability scale even at the true distance-weighted linear predictor, when neighborhood size and within-neighborhood feature variation co-vary (Theorem 2); the counterexample uses two positive coefficients, so it does not depend on effect signs.

We then propose **LION** (Local Influence Inference Network), an architecture that represents the global ranking through a shared feature vector. LION combines a learnable distance kernel, a feature-preserving normalizer, and a global influence vector $\mathbf{S} \in \mathbb{R}^G$ through bilinear aggregation. The kernel weights nearby instances, the normalizer puts their features on a common scale, and the influence vector assigns one score to each feature. The model has $G + 4$ trainable parameters (Appendix U). Its scores admit a positive affine reparameterization that preserves their ordering. Under correct specification and identification conditions, the reduced maximum-likelihood estimator recovers all strict feature comparisons (Theorem 1). Our main contributions include:

- We define Local Influence Inference around a feature ranking shared across all bags, and propose LION, which represents this ranking through a shared influence vector and a separable distance kernel with $G + 4$ parameters.
- We establish consistency and asymptotic contrast intervals for LION's reduced maximum-likelihood estimator (Theorem 1), and construct an oracle example in which permutation importance reverses the coefficient ordering (Theorem 2).
- We separate direct coefficient ranking from magnitude-based attribution. LION leads or ties distance-weighted logistic regression in mean Spearman across all seven benchmarks, with the largest gains on morphology-interaction and Dense. LION ranks first in AUROC on five benchmarks among the evaluated deep and interpretable MIL baselines. A three-sample Xenium/TCR-seq study shows that similarly predictive models identify different T-cell populations.

2 Related Work

Bag-level prediction. Attention-based MIL (Ilse et al., 2018) predicts bags by pooling instance representations, and distance-aware variants (Kim et al., 2022; Aftab et al., 2024) incorporate proximity. LION makes a shared ranking over named observed features a model parameter. Appendix X expands these comparisons to other spatial architectures and causal MIL formulations.

Intrinsic explanations. Additive MIL makes region-level credit exact (Javed et al., 2022). SI-MIL fits a per-slide linear surrogate over handcrafted descriptors and aggregates slide-specific explanations for a cohort-level view (Kapse et al., 2024). Causal MIL formulations adjust bag prediction for confounding (IBMIL, CaMIL, CATTMIL) or identify instance-level latent factors (CausalMIL) (Lin et al., 2023; Chen et al., 2024; Wu et al., 2025; Zhang et al., 2022); none returns a ranking over named observed features. LION learns one feature vector shared by all neighborhoods; Theorem 1 states its identification and consistency conditions. We evaluate SI-MIL and Additive MIL as interpretable baselines; Appendix Q compares global readouts.

Post-hoc feature rankings. Aggregated explanations of trained predictors give a global ranking that depends on the attribution rule. For within-bag permutation importance (Breiman, 2001), Theorem 2 constructs an oracle distance-weighted linear example in which the bag-averaged probability-scale ranking reverses two positive coefficients as neighborhood size and feature variation co-vary. Appendix Q compares attribution readouts empirically. Further attribution methods are reviewed in Appendix X.

3 Problem Formulation: Local Influence Inference

Definition 1 (Local Influence Inference)**.** *Given a collection of bags* $\{(B_i, y_i)\}_{i=1}^{I}$*, where each bag* $B_i = \{(\mathbf{x}_{ij}, d_{ij})\}_{j=1}^{n_i}$ *consists of* n_i *instances with feature vectors* $\mathbf{x}_{ij} \in \mathbb{R}^G$ *and scalar distances*

$d_{ij} \in \mathbb{R}_{\geq 0}$ *to the bag center, and* $y_i \in \{0, 1\}$ *is a bag-level label, the* ***Local Influence Inference*** *problem requires:*

(i) ***Prediction:*** *learn a function* f *such that* $f(B_i) \approx y_i$.

(ii) ***Feature ranking recovery:*** *learn a unified vector* $\widehat{\mathbf{S}} \in \mathbb{R}^G$ *such that the ranking induced by* $\widehat{\mathbf{S}}$ *agrees with the ranking induced by the true (unknown) feature influence vector* $\mathbf{S}^*$.

The output $\widehat{\mathbf{S}}$ is a *single global vector* shared across all bags. The target is the coefficient ordering, $S_g^* > S_h^* \Rightarrow \widehat{S}_g > \widehat{S}_h$, rather than the ordering of absolute perturbation effects. We call this a signed ranking to distinguish larger coefficients from larger magnitudes; the sign of an individual score requires a specified reference because the scores admit a common shift.

Notation. All bags and instances share the same G named feature coordinates $\{\mathcal{F}_g\}_{g=1}^G$; $x_{ij,g}$ is the value of feature $\mathcal{F}_g$ for instance j in bag i.

Ranking target. Recovery means preserving every strict comparison $S_g^* > S_h^*$, with no ordering imposed within a tied group. This target is invariant to the positive affine reparameterizations of the model (Lemma 1).

Ranking as a model parameter. LION's shared influence vector is a native model parameter, $\widehat{\mathbf{S}} = \sigma(\widehat{\mathbf{s}})$. Post-hoc importance defines a separate population readout whose ordering can disagree with the coefficient ranking even with infinite data (Theorem 2). Section 5 derives the reduced parameterization for ranking identification and confidence intervals for feature contrasts.

4 LION: Local Influence Inference Network

4.1 Architecture

LION processes each bag through three modular components whose outputs are combined via bilinear aggregation, as illustrated in Figure 2. For a bag $B_i = \{(\mathbf{x}_{ij}, d_{ij})\}_{j=1}^{n_i}$:

Distance kernel. Spatial influence is modulated by a learnable kernel that assigns each instance an attention weight depending on its distance to the bag center,

$$w_{ij} = \frac{\exp(-e^a \cdot d_{ij})}{\sum_{k=1}^{n_i} \exp(-e^a \cdot d_{ik})},$$

where a is learnable and e^a is a positive decay rate in inverse-distance units, shared across bags. Its exponential form permits unconstrained optimization of a. Closer instances receive higher weight.

Feature-preserving normalizer. Within each instance, raw feature magnitudes are converted to a comparable distribution by a softmax with learnable inverse temperature,

$$E_{ij,g} = \frac{\exp\big(e^b \cdot x_{ij,g}\big)}{\sum_{g'=1}^{G} \exp(e^b \cdot x_{ij,g'})},$$

where b is a learnable scalar. This softmax normalization determines how strongly the feature magnitude hierarchy within each instance influences the prediction.

Global feature influence. Each feature g receives a single learnable score: $S_g = \sigma(s_g), \quad g = 1, \ldots, G$, where σ is the sigmoid function and s_g is a learnable parameter. The named feature coordinates remain separately scored, so the ordering of S_g gives the feature ranking shared across bags.

Bilinear aggregation and prediction. For each instance, we compute a weighted feature score and aggregate over the bag:

$$z_{ij} = \sum_{g=1}^{G} E_{ij,g} \cdot S_g, \qquad o_i = \sum_{j=1}^{n_i} w_{ij} \cdot z_{ij}.$$

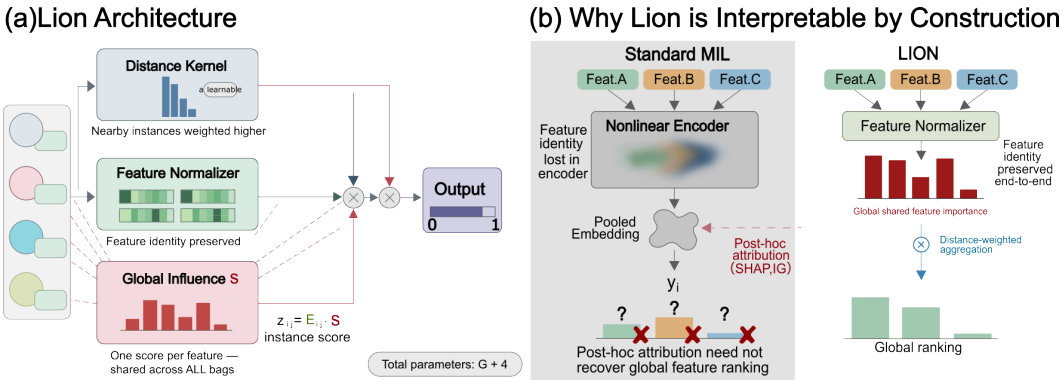

Figure 2: LION architecture and inference target. (a) Distance weighting, feature normalization, and a shared influence vector. (b) Shared feature scores and post-hoc attribution readouts. Theorem 2 analyzes within-bag permutation in a distance-weighted linear model.

The product $w_{ij}z_{ij}$ separates spatial weighting from feature scoring. The bag score is passed through a trainable logistic transformation: $\widehat{y}_i = \sigma(e^{\alpha} \cdot o_i + \beta)$, where α, β are learnable scalars. The total parameter count is $G + 4$ (the G-dimensional vector $\mathbf{s}$ plus scalars a, b, α, β). Related structural models and algorithmic extensions are discussed in Appendices X and W, respectively.

5 Theoretical Analysis

5.1 Generative Model

Definition 2 (Distance-Modulated Influence Model)**.** *Using the softmax modules of Section 4, let* $y_i \mid B_i \sim \text{Bernoulli}(p^*(B_i))$*, where*

$$p^*(B_i) = \sigma\left(e^{\alpha^*} \sum_{j=1}^{n_i} w_{ij}(a^*)\,\langle \mathbf{E}_{ij}(b^*), \mathbf{S}^*\rangle + \beta^*\right), \qquad \mathbf{S}^* = \sigma(\mathbf{s}^*),$$

5.2 Ranking Identification and Consistency

Define the bag profile $\mathbf{z}_i(a,b) = \sum_j w_{ij}(a)\mathbf{E}_{ij}(b)$. Because $\mathbf{1}^\top \mathbf{z}_i = 1$, the original scores have a common shift and scale freedom that leaves their ordering unchanged.

Lemma 1 (Affine invariance)**.** *Every transformation* $\mathbf{S} \mapsto k\mathbf{S} + m\mathbf{1}$ *with* $k > 0$ *and* $k\mathbf{S} + m\mathbf{1} \in (0,1)^G$ *can be absorbed by the logistic scale and intercept without changing any prediction. Define*

$$\mathbf{D} = e^{\alpha}(\mathbf{S} - \bar{S}\mathbf{1}), \qquad \widetilde{\beta} = \beta + e^{\alpha}\bar{S}, \qquad \bar{S} = G^{-1}\mathbf{1}^\top \mathbf{S}.$$

Then $\eta_i = \mathbf{z}_i^\top \mathbf{D} + \widetilde{\beta}$, $\mathbf{1}^\top \mathbf{D} = 0$, *and* $\text{sign}(S_g - S_h) = \text{sign}(D_g - D_h)$.

Let $\theta \in \mathbb{R}^{G+2}$ denote fixed independent linear coordinates for $(a, b, \mathbf{D}, \widetilde{\beta})$ with $\mathbf{1}^\top \mathbf{D} = 0$. Appendix D.1 gives these coordinates and the explicit affine transformation.

Theorem 1 (Ranking identification, consistency, and inference)**.** *Let* G *be fixed and let the bags and labels be i.i.d. with finite, nonempty bags and conditional law given by Definition 2. Assume:*

(A1) Identification at the truth: *for* $\theta \in \Theta$, $\eta_\theta(B) = \eta_{\theta^*}(B)$ *almost surely implies* $\theta = \theta^*$.

(A2) Regularity: $\Theta \subset \mathbb{R}^{G+2}$ *is compact,* θ^* *is in its interior, and* $\mathbb{E}[\max_j d_{ij}^2 + \max_j \|\mathbf{x}_{ij}\|_\infty^2] < \infty$.

(A3) Local information: *the joint Fisher information for the reduced coordinates is positive definite at* θ^*.

Let $\widehat{\theta}_I$ *be a global maximizer of the unpenalized likelihood over* Θ*. Under (A1)–(A2), all original parameter vectors with reduced coordinates in* Θ *that induce the true conditional law have the same*

$(a^*, b^*, \mathbf{D}^*, \widetilde{\beta}^*)$ *and the same ranking. Moreover,* $\widehat{\theta}_I \xrightarrow{p} \theta^*$ *and*

$$\Pr\Big[\widehat{S}_g > \widehat{S}_h \text{ for every } (g,h) \text{ with } S_g^* > S_h^*\Big] \longrightarrow 1,$$

where $\widehat{\mathbf{S}}$ *is any original representation of* $\widehat{\theta}_I$*. Under (A3) as well, the reduced estimator is asymptotically normal and yields pointwise asymptotic confidence intervals for* $D_g^* - D_h^*$*, whose sign equals that of* $S_g^* - S_h^*$*. The information matrix, limiting distribution, and interval formulas are given in Appendix D.3.*

Why ranking recovery follows. When strict comparisons exist, a sufficiently accurate reduced estimate has contrast errors smaller than the minimum nonzero contrast gap, preserving every strict comparison. This does not assert exact recovery of ties. Appendix D gives the likelihood argument and full proof.

Identification in practice. Joint identification of the distance, normalization, and feature parameters depends on the distribution of features and distances. Purely one-hot features require an additional reduction that preserves the ranking (Appendix D.4); Appendix F examines feature and distance degeneracy. The theorem concerns a global unpenalized MLE. Section 6.4 compares the actual AdamW rankings with unpenalized profile-likelihood refits on all seven benchmarks.

5.3 Post-Hoc Attribution: A Different Estimand

We analyze permutation importance in a distance-weighted linear model with an identity feature map:

$$\eta(B) = \sum_{j=1}^{n} \omega_j \mathbf{w}^\top \mathbf{x}_j + \beta, \qquad \omega_j = \frac{e^{-\lambda d_j}}{\sum_k e^{-\lambda d_k}}, \qquad f^*(B) = \sigma(\eta(B)).$$

Here f^* is the true conditional probability and $\lambda > 0$ is fixed. Conditional on bag size and composition, distances are i.i.d. from a common non-point-mass distribution F on $[0, r_{\max}]$, independently of features. Each feature column is either *varying*, with i.i.d. $\mathcal{N}(\mu, \tau^2)$ entries, or *bag-constant*, with one $\mathcal{N}(\mu, \tau^2)$ draw repeated across instances. Here $\tau > 0$ is common across features; columns are independent conditional on composition. Thus both types have identical per-instance marginals. For a uniform random within-bag permutation π of column g, define

$$R_g(B) = \mathbb{E}_\pi |f^*(B^{\pi,g}) - f^*(B)|, \qquad \bar{R}_g = \mathbb{E}_B R_g(B).$$

Theorem 2 (Size and composition can invert permutation importance)**.** *Under the preceding model:*

(a) Logit sensitivity. *For a varying feature, the expected absolute logit change equals* $|w_g|\tau$ *times a factor of order* $n^{-1/2}$ *as bag size grows. For a bag-constant feature the change is zero. With uniform weights, within-bag permutation importance vanishes on both logit and probability scales. Exact expressions and constants are in Appendix E.1.*

(b) Matched-composition controls on the logit scale. *If every feature varies in every bag, the global logit sensitivity is proportional to* $|w_g|$ *for any bag-size distribution. If bag size is fixed at* $n \geq 2$ *and two features vary in the same positive fraction of bags, their logit sensitivities have the same ordering as their coefficient magnitudes.*

(c) Inversion on the probability scale. *Consider two features* g, h *with* $w_g > w_h > 0$*. With probability* $1/2$ *a bag has size* $n_S \geq 2$*, with* g *bag-constant and* h *varying; otherwise its size is* $N > n_S$*, with the roles reversed. For every such coefficient pair there exists a finite* N_0 *such that* $\bar{R}_g < \bar{R}_h$ *for all* $N \geq N_0$*. Nevertheless, for each fixed* N*, a correctly specified logistic MLE of* (w_g, w_h, β) *with known* λ*, over a compact parameter set containing the truth in its interior, consistently recovers* $w_g > w_h$ *as the number of bags tends to infinity.*

Why the ordering reverses. The stronger coefficient's feature varies only in large bags, where its permutation signal decays with size. The weaker coefficient's feature varies in the small bags, retaining a strictly positive signal as N grows. This produces an inverted importance ordering even at the true predictor. Appendix E gives the perturbation calculation, full proof, and numerical check.

Interpretation. Within-bag permutation disrupts feature–distance alignment. Its effect depends on both the distance weights and the bags in which each feature varies. Matched composition preserves coefficient-magnitude order on the logit scale under the controls above; coupling composition with bag size can reverse the probability-scale ordering. This is a result for the stated linear oracle and within-bag permutation, not for arbitrary attribution methods or the softmax feature map. Appendix R examines size dependence empirically across several attribution methods.

6 Experiments

6.1 Experimental Setup

Datasets. We evaluate ranking recovery on seven benchmarks: three Spatial-MNIST settings, simulated spatial transcriptomics, and three H&E rules applied to real TCGA-LUAD nucleus detections. These benchmarks use constructed labels with exact ranking targets. VisiumHD and Xenium case studies evaluate prediction and biological specificity using observed assay labels, without exact ranking ground truth. Appendix G, Table 3, summarizes data provenance and label rules; construction details are in Appendices H–L.

Baselines and ranking targets. We compare with DeepSets+dist, AttentionMIL+dist, GNN+dist, SI-MIL (Kapse et al., 2024), and Additive MIL (Javed et al., 2022), all supplied with the same instance features and scalar distances. AttentionMIL+dist is identical to DistAttMIL and shares its entry. Our primary target is the signed coefficient ordering. Figure 3 therefore compares LION's native scores with unrestricted signed coefficients from distance-weighted logistic regression (DW-logistic). Because LION normalizes features nonlinearly, the logistic model can still be misspecified; the two models' coefficient meanings are not equated. The separate CV comparison in Figure 4(c) evaluates LION's native coefficients and the deep baselines' absolute permutation importance (Breiman, 2001) against the signed target; the latter measures perturbation magnitude. Additional readouts and feature-unit requirements are specified in Appendices P–R.

Metrics and protocols. The main comparison uses paired 5-fold cross-validation $\times$ 5 seeds, shared folds, and the default configuration. Figure 4 reports mean AUROC, AUPRC, and signed-target Spearman ρ, comparing LION with the best distance-equipped deep baseline for each benchmark and metric. In the figure, MNIST denotes Spatial-MNIST and ST denotes spatial transcriptomics. Full prediction metrics and numerical ranking results are in Appendix Tables 5 and 6. Evaluation pairs one positive with one negative bag, giving an AUPRC reference level of 0.50. The additional coefficient comparison, three-seed component ablation, and other diagnostics use separate protocols and are not pooled with the CV results. Appendices G and N give procedures, metric conventions, full results, and the single-split comparison.

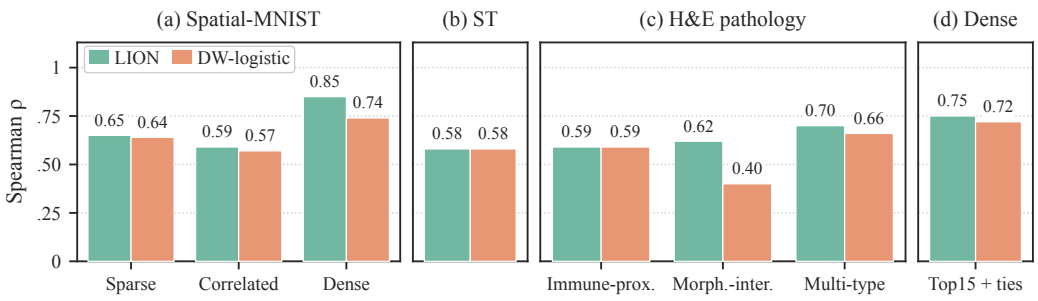

Figure 3: Direct recovery of signed feature rankings. Bars show mean Spearman ρ for LION's native ordering and DW-logistic's unrestricted signed coefficients. Panels (a–c) cover the full rankings on seven benchmarks; panel (d) evaluates the auxiliary Dense upper-coefficient subset (generator-defined Top15, including boundary ties). This additional comparison is reported separately from the CV results in Figure 4; details are in Appendix P. ST denotes spatial transcriptomics; Immune-prox. and Morph.-inter. denote the H&E immune-proximity and morphology-interaction rules.

6.2 Prediction and Ranking Recovery

Direct coefficient ranking. LION leads or ties DW-logistic in mean Spearman on all seven benchmarks (Figure 3). The largest gaps are on Dense full ranking (0.85 versus 0.74) and morphology-interaction (0.62 versus 0.40). Spatial transcriptomics and H&E immune-proximity tie at reported precision, and Sparse, Correlated, and multi-type differ by 0.01–0.04. The gap widens where the target has many coefficient levels: Dense orders 60 features across bag sizes 5–80, and morphology-interaction must rank a weaker morphology signal (`area`, $s = 3.0$) below two cell-type signals ($s = 5.0$). On the auxiliary Dense upper-coefficient subset the gap narrows to 0.75 versus 0.72 (Appendix P); small gaps are descriptive because paired uncertainty estimates are unavailable.

Prediction. LION leads the three distance-equipped deep baselines in AUROC and AUPRC on every benchmark (Figure 4(a,b)). In the separate interpretable-baseline protocol, mean-pooled Additive MIL has higher AUROC on H&E immune-proximity and spatial transcriptomics (Appendix R). Across these deep and interpretable MIL baselines, LION ranks first in AUROC on five of seven benchmarks. The effect of supplying distance to the baselines is reported separately in Appendix M.

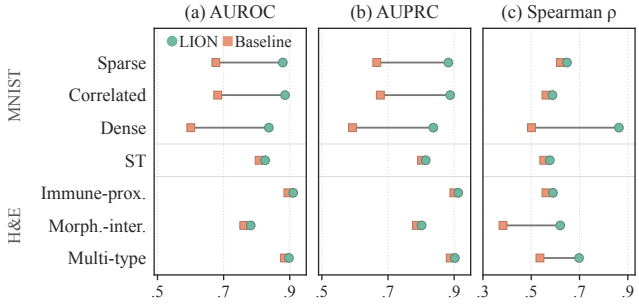

Figure 4: LION versus the best deep baseline in (a) AUROC, (b) AUPRC, and (c) signed-target Spearman ρ.

Attribution-based ranking. LION's native scores also have higher signed-target Spearman than the absolute permutation readouts on all seven CV benchmarks (Figure 4(c)), with the largest gaps on Dense (+0.362) and morphology-interaction (+0.237). These gaps partly reflect a difference in estimand, since a strong negative coefficient can correctly receive high absolute importance; the direct coefficient comparison in Figure 3 is free of this confound. The readout itself moves the ranking: the four readouts of one trained DeepSets model differ by 0.56 Spearman on Dense, and SI-MIL's three global readouts by 0.34–0.46 (Appendix Q).

6.3 Feature and Distance Recovery

Recovered features. Figure 5 places immune cells and macrophages highest under all three rules and `area` third under morphology-interaction. Blood ranks last under multi-type, consistent with its negative ground-truth effect ($s = -2.5$), and NDCG is 1.0 across all seeds and rules. On morphology-interaction, LION's Spearman reaches 88.8% of the tied-target ceiling of 0.698, against 54.8% for the best deep baseline. Absolute scores depend on their affine representation; ordering and contrasts against a specified reference carry the interpretation (Appendices J and D.5).

Distance and feature components. The learned distance curve tracks the ground truth ($r = 0.943 \pm 0.027$; Figure 6). Replacing the distance weights by uniform weights lowers AUROC on every benchmark by 0.073–0.261 while changing Dense Spearman by only −0.008; removing the influence vector changes it by −0.569 (Appendix T).



Figure 5: LION feature scores on H&E: (a) cell types and (b) morphology. Colors share a 0–1 scale; superscripts +/− mark positive/negative ground-truth effects.

Feature normalization. Under the three-seed ablation, removing feature normalization lowers ranking recovery on both displayed benchmarks (Figure 7): Dense Spearman falls from 0.844 to 0.717, and on morphology-interaction AUROC rises to 0.824 while Spearman drops from 0.610 to 0.332. Every variant ranks the two cell-type signals first, but only the five that retain the normalizer place `area` above the inactive features ($\rho = 0.565$–0.676); the two without it rank `area` 13th or 14th of 15. The default is not uniformly best: sum pooling reaches $\rho = 0.676$. Removing the normalizer retains the score constraint, so it is distinct from unrestricted DW-logistic (Appendix T).

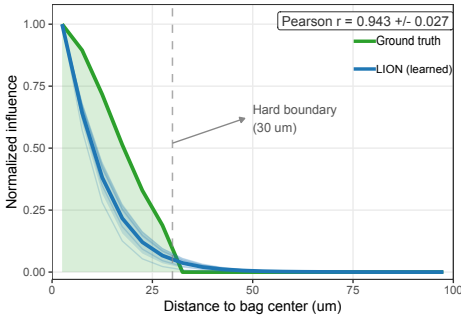

Figure 6: Learned versus ground-truth spatial decay ($r = 0.943 \pm 0.027$). Baseline distance responses are in Appendix S.

6.4 Robustness and Capacity

Connection to the likelihood estimator. Across seven benchmarks, AdamW and unpenalized profile-likelihood refits differ by at most 0.01 in ground-truth Spearman. This empirical agreement does not establish that AdamW reaches the global MLE required by Theorem 1; Appendix D.5 gives the estimator diagnostics.

Additional diagnostics. Oracle selection by ground-truth Spearman does not overtake fixed-configuration LION on the three tested benchmarks; SI-MIL's joint hyperparameter/readout oracle ties on saturated targets but remains below LION on Dense (Appendices O and R).

Neighborhood size. With the generating rule fixed, fitted attribution-attenuation exponents are 0.21–0.39 for the post-hoc readouts and 0.27 for SI-MIL's intrinsic coefficient, compared with -0.00 for LION. Sum-pooled Additive MIL also collapses to AUROC 0.500 in an increasing number of seeds ($0/5$ to $5/5$) as mean bag size grows from 27.7 to 144.4; mean pooling removes the collapse (Appendix R).

Capacity and scope. LION uses 12–64 parameters across the benchmarks. Increasing interactions costs 0.100 AUROC while leaving ranking nearly unchanged. At the hardest active-feature setting, LION and the strongest baseline are not significantly separated, with a Bayes ceiling of 0.645 (Appendices U and V).

6.5 Biological Evaluation

Independent biological evaluation. On three Xenium/TCR-profiled human tumors, models predict T-cell presence from surrounding tumor-cell expression and distances. TCR-defined reactive/bystander groups are reserved for evaluation; identity- and reactivity-defining genes and clonotype probes are excluded from model inputs. At matched predicted-positive rates, we compare the reactive fraction among true-positive (TP) T cells and the bystander fraction among false-negative (FN) T cells, paired within tumor (Appendix L).

Biological composition beyond prediction. LION improves AUROC and both composition measures in all 15 comparisons with the original five baselines, and exceeds DW-logistic in mean gains (Figure 8(a–c)). Against the five baselines, AUROC gains are 0.022–0.063, and the reactive-TP and bystander-FN fractions are 11.2–18.2 and 9.7–14.4 percentage points higher; against DW-logistic the mean gains are 0.023 AUROC and 10.3 and 9.2 points.

Removing the influence vector gives higher AUROC but lower reactive fractions among TP and bystander fractions among FN than removing the distance kernel, in all three tumors. Thus, predictive performance and biological composition need not improve together (Appendix L).

Gene-ranking support. On a shared gene set, NetMHC analysis (Andreatta & Nielsen, 2016) shows a pronounced concentration of predicted HLA binding in LION's top 1–5%, compared with

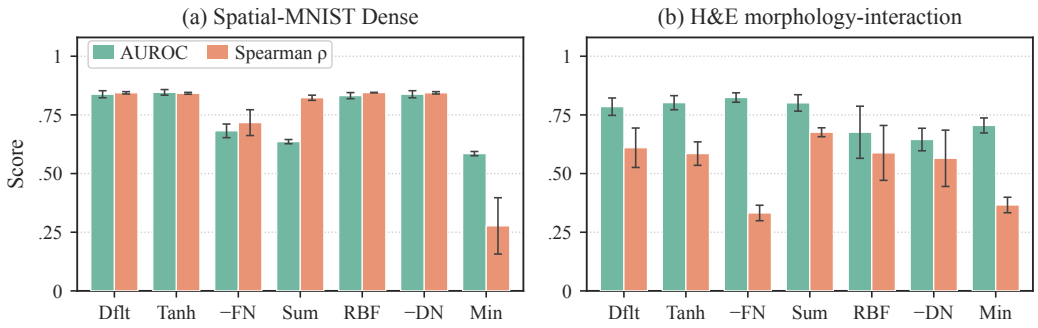

Figure 7: Component effects on prediction and ranking. Bars show mean AUROC and Spearman ρ; error bars are standard deviations over three seeds. Dflt is the default; $-$FN/$-$DN remove feature/distance normalization; Min removes both feature normalization and distance structure. This protocol (Appendix T) differs from the main CV.

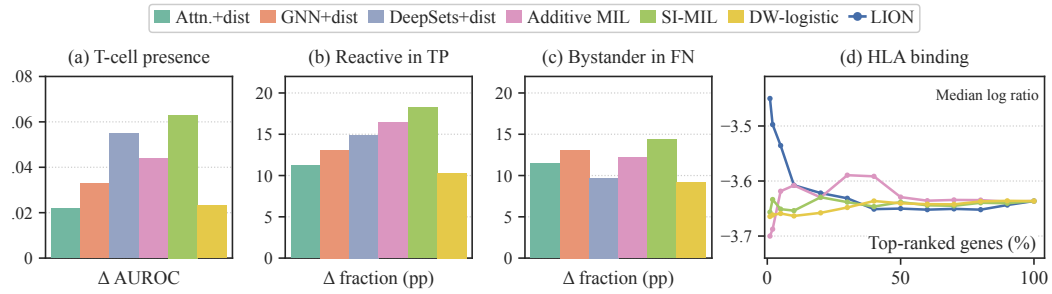

Figure 8: Xenium prediction and biological enrichment (Appendix L). Bars show mean paired gains (LION minus baseline) across three tumors in (a) T-cell-presence AUROC, (b) reactive fraction among true-positive T cells, and (c) bystander fraction among false-negative T cells. Panels (b,c) use TCR-defined groups, matched predicted-positive rates, and percentage-point gains. (d) Median $\log[(B_g+1)/(T_g+1)]$ over nested top-ranked gene sets; B_g and T_g count NetMHC-predicted binders and total peptides.

Additive MIL, SI-MIL, and DW-logistic (Figure 8(d)). This complements the cell-level evaluation; neither analysis supplies an exact or experimentally validated gene ranking.

Additional tissue setting: myocarditis. On VisiumHD mouse myocarditis data, LION predicts T-cell presence from ventricular-cardiomyocyte neighborhoods with AUROC 0.86 ± 0.11, versus 0.82 ± 0.14 for the strongest baseline and 0.78 ± 0.14 for DW-logistic. Preranked GSEA over the full gene ordering finds a mitochondrial respiratory set coupled to the TCA cycle at the non-infiltration end (BH-adjusted $P = 7.06 \times 10^{-4}$; endoplasmic-reticulum control $P = 0.643$), with analogous tests giving $P = 0.00237$ for macrophages and $P = 0.217$ for B cells (Appendix K).

7 Conclusion

LION represents a shared feature ranking with $G + 4$ parameters. Under correct specification and identification, the reduced global maximum-likelihood estimator consistently recovers strict comparisons. A linear oracle shows how bag size and composition can reverse two positive coefficients under within-bag permutation importance. LION leads or ties DW-logistic in mean Spearman across seven benchmarks, with the largest gains on morphology-interaction ($+0.22$) and Dense ($+0.11$), and ranks first in AUROC on five of seven benchmarks with 12 to 64 parameters; comparisons with absolute attribution readouts involve different estimands. On three Xenium/TCR-profiled tumors, LION raises the reactive fraction among recovered T cells and the bystander fraction among missed T cells over every baseline. These findings support learning shared rankings directly from spatial bag-level supervision. Future work will focus on extending the shared ranking to overlapping neighborhoods and explicit feature interactions.

REPRODUCIBILITY STATEMENT

All code used in this work is included in the supplementary materials. The model architecture is specified in Section 4, and the assumptions underlying the theoretical results are stated in Section 5, with complete proofs in Appendices D and E. Training configurations and evaluation protocols are documented in Appendix G. Benchmark construction, data processing, and case-study procedures are described in Appendices H–L, with additional evaluation details in Appendices N–R.

AI USE STATEMENT

AI tools were used primarily for language polishing and improving the clarity and presentation of the manuscript. The authors take responsibility for the final content of this work, including all text, claims, and artifacts produced with the assistance of AI.

REFERENCES

Rukhma Aftab, Qiang Yan, Juanjuan Zhao, Gao Yong, Yue Huajie, Zia Urrehman, and Faizi Mohammad Khalid. Neighborhood attention transformer multiple instance learning for whole slide image classification. *Frontiers in Oncology*, 14:1389396, 2024.

Massimo Andreatta and Morten Nielsen. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics*, 32(4):511–517, 2016.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Kaitao Chen, Shiliang Sun, and Jing Zhao. CaMIL: Causal multiple instance learning for whole slide image classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1120–1128, 2024.

David R Cox. Regression models and life-tables. *Journal of the royal statistical society: Series B (methodological)*, 34(2):187–202, 1972.

Thomas Duhen, Rebekka Duhen, Ryan Montler, Jake Moses, Tarsem Moudgil, Noel F De Miranda, Cheri P Goodall, Tiffany C Blair, Bernard A Fox, Jason E McDermott, et al. Co-expression of CD39 and CD103 identifies tumor-reactive CD8 T cells in human solid tumors. *Nature communications*, 9(1):2724, 2018.

Joseph Early, Christine Evers, and Sarvapali Ramchurn. Model agnostic interpretability for multiple instance learning. *arXiv preprint arXiv:2201.11701*, 2022.

Eran Eden, Roy Navon, Israel Steinfeld, Doron Lipson, and Zohar Yakhini. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC bioinformatics*, 10(1): 48, 2009.

Edward L Glaeser, William R Kerr, and Giacomo AM Ponzetto. Clusters of entrepreneurship. *Journal of urban economics*, 67(1):150–168, 2010.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pp. 2127–2136. PMLR, 2018.

Syed Ashar Javed, Dinkar Juyal, Harshith Padigela, Amaro Taylor-Weiner, Limin Yu, and Aaditya Prakash. Additive MIL: Intrinsically interpretable multiple instance learning for pathology. *Advances in Neural Information Processing Systems*, 35:20689–20702, 2022.

Suoqin Jin, Christian F Guerrero-Juarez, Lihua Zhang, Ivan Chang, Raul Ramos, Chen-Hsiang Kuan, Peggy Myung, Maksim V Plikus, and Qing Nie. Inference and analysis of cell-cell communication using CellChat. *Nature communications*, 12(1):1088, 2021.

Saarthak Kapse, Pushpak Pati, Srijan Das, Jingwei Zhang, Chao Chen, Maria Vakalopoulou, Joel Saltz, Dimitris Samaras, Rajarsi R Gupta, and Prateek Prasanna. SI-MIL: Taming deep MIL for self-interpretability in gigapixel histopathology. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11226–11237. IEEE, 2024.

Younghoon Kim, Tao Wang, Danyi Xiong, Xinlei Wang, and Seongoh Park. Multiple instance neural networks based on sparse attention for cancer detection using T-cell receptor sequences. *BMC bioinformatics*, 23(1):469, 2022.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14318–14328, 2021.

Tiancheng Lin, Zhimiao Yu, Hongyu Hu, Yi Xu, and Chang-Wen Chen. Interventional bag multi-instance learning on whole-slide pathological images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19830–19839, 2023.

Sophia K Longo, Margaret G Guo, Andrew L Ji, and Paul A Khavari. Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. *Nature Reviews Genetics*, 22(10): 627–644, 2021.

Ming Y Lu, Drew FK 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.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Andre Martins and Ramon Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International conference on machine learning*, pp. 1614–1623. PMLR, 2016.

Jeffrey R Moffitt, Emma Lundberg, and Holger Heyn. The emerging landscape of spatial profiling technologies. *Nature Reviews Genetics*, 23(12):741–759, 2022.

Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245, 1994.

Ross L Prentice and Ronald Pyke. Logistic disease incidence models and case-control studies. *Biometrika*, 66(3):403–411, 1979.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660, 2017.

Thummaluru Siddartha Reddy, Vempalli Naga Sai Saketh, and Mahesh Chandran. Interpretable and parameter efficient graph neural additive models with random Fourier features. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL `https://openreview.net/forum?id=yl9LxRL5tj`.

Ruichen Rong, Hudanyun Sheng, Kevin W Jin, Fangjiang Wu, Danni Luo, Zhuoyu Wen, Chen Tang, Donghan M Yang, Liwei Jia, Mohamed Amgad, et al. A deep learning approach for histology-based nucleus segmentation and tumor microenvironment characterization. *Modern Pathology*, 36 (8):100196, 2023.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021.

Yannick Simoni, Etienne Becht, Michael Fehlings, Chiew Yee Loh, Si-Lin Koo, Karen Wei Weng Teng, Joe Poh Sheng Yeong, Rahul Nahar, Tong Zhang, Hassen Kared, et al. Bystander CD8+ T cells are abundant and phenotypically distinct in human tumour infiltrates. *Nature*, 557(7706): 575–579, 2018.

Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the national academy of sciences*, 102(43):15545–15550, 2005.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The Cancer Genome Atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

Ian J Wright, Ning Dong, Vincent Maire, I Colin Prentice, Mark Westoby, Sandra Díaz, Rachael V Gallagher, Bonnie F Jacobs, Robert Kooyman, Elizabeth A Law, et al. Global climatic drivers of leaf size. *Science*, 357(6354):917–921, 2017.

Xiaochun Wu, Haitao Wang, and Hejun Wu. Causal ATTention multiple instance learning for whole slide image classification. In *Proceedings of the 16th Asian Conference on Machine Learning (ACML), PMLR 260*, pp. 984–999, 2025.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.

Weijia Zhang, Xuanhui Zhang, Min-Ling Zhang, et al. Multi-instance causal representation learning for instance label prediction and out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 35:34940–34953, 2022.

A LIMITATIONS

LION learns a shared ranking by fitting a constrained predictive model. Its instance scores are additive in normalized features; explicit pairwise interactions require an extension. The ranking-consistency result assumes correctly specified, independent bags, identification of the reduced parameters, and a global unpenalized maximum-likelihood estimator. It does not directly cover regularized AdamW training, overlapping spatial neighborhoods, or misspecified label rules. Exact ranking recovery is evaluated on benchmarks with constructed targets. The tissue studies instead assess prediction and biological associations, and the Xenium analysis covers three tumors. These results support hypothesis generation; broader sampling and independent biological validation are needed to establish how the learned rankings generalize.

B BROADER IMPACT AND RESPONSIBLE USE

Shared feature rankings can make spatial prediction models easier to inspect and help prioritize hypotheses for follow-up experiments. Applications to biology, ecology, or urban science would require validation in the intended setting. A ranking learned from observational spatial data is not, by itself, an estimate of a causal effect. Biased sampling, unrepresentative populations, and confounded features can also lead to misleading interpretations. LION-derived rankings should therefore be evaluated with domain expertise and independent evidence before informing clinical, policy, or intervention decisions.

C ETHICS STATEMENT

The human tumor samples profiled with Xenium and TCR sequencing were de-identified and were collected under a protocol approved by the institutional review board of the authors' institution. The VisiumHD myocarditis data were generated from mice under an approved institutional animal-use protocol. The remaining benchmarks use simulated data or public TCGA-LUAD whole-slide images with constructed labels. Learned feature rankings are hypothesis-generating and should be validated with domain-specific experiments before informing clinical or policy decisions (Appendix B).

D PROOF OF THEOREM 1

D.1 AFFINE INVARIANCE AND REDUCED COORDINATES

To remove the two affine degrees of freedom, fix an orthonormal basis $Q \in \mathbb{R}^{G\times(G-1)}$ of $\mathbf{1}^\perp$ and write $\mathbf{D} = Q\mathbf{u}$. The independent reduced coordinates in Theorem 1 are

$$\theta = (a, b, \mathbf{u}, \widetilde{\beta}) \in \mathbb{R}^{G+2}.$$

All derivatives and information matrices in this appendix use these coordinates. The distance weights and each normalized feature vector sum to one, so $\mathbf{1}^\top \mathbf{z}_i = \sum_j w_{ij} \mathbf{1}^\top \mathbf{E}_{ij} = 1$. Splitting $\mathbf{S}$ into its centered part and its mean therefore gives

$$e^\alpha \mathbf{z}_i^\top \mathbf{S} + \beta = \mathbf{z}_i^\top e^\alpha (\mathbf{S} - \bar{S}\mathbf{1}) + \beta + e^\alpha \bar{S} = \mathbf{z}_i^\top \mathbf{D} + \widetilde{\beta}.$$

For $k > 0$ and $m \in \mathbb{R}$ such that $k\mathbf{S} + m\mathbf{1} \in (0,1)^G$, the explicit transformation is

$$(\mathbf{S}, \alpha, \beta) \longmapsto (k\mathbf{S} + m\mathbf{1}, \alpha - \log k, \beta - e^\alpha m/k).$$

This transformation leaves both $\mathbf{D}$ and $\widetilde{\beta}$ unchanged. Conversely, at fixed a, b, any two original parameter vectors with the same $(\mathbf{D}, \widetilde{\beta})$ satisfy

$$\mathbf{S}' = k\mathbf{S} + m\mathbf{1}, \qquad k = e^{\alpha - \alpha'} > 0, \qquad m = \bar{S}' - k\bar{S},$$

and equality of $\widetilde{\beta}$ determines the corresponding intercept transformation. Each reduced parameter thus represents a two-dimensional positive affine orbit in the original parameterization.

Every centered $\mathbf{D}$ has an original representation: choose $e^\alpha > 2\|\mathbf{D}\|_\infty$, set $\mathbf{S} = \frac{1}{2}\mathbf{1} + e^{-\alpha}\mathbf{D}$ and $\beta = \widetilde{\beta} - e^\alpha/2$, and apply the componentwise logit to obtain $\mathbf{s}$. Consequently, maximizing over the

compact reduced set Θ is equivalent to maximizing over its preimage in the original coordinates. This equivalence does not extend to unconstrained optimization over all original parameters. Because $D_g - D_h = e^{\alpha}(S_g - S_h)$, every representation of the same reduced parameter preserves all strict comparisons and ties.

Coefficient ordering and raw-feature responses. The shared ordering describes the coefficients of normalized features. At fixed distances, the logit's response to a raw feature is

$$\frac{\partial \eta_i}{\partial x_{ij,g}} = e^{\alpha+b} w_{ij} E_{ij,g} \left(S_g - \sum_h E_{ij,h} S_h \right).$$

Its sign depends on the instance's normalized feature composition, so a shared coefficient ordering does not imply a fixed direction of response to increasing each raw feature. This distinction also applies to contrasts against a specified global reference group.

D.2 Identification and consistency

Assumption (A1) identifies the reduced parameter from the true conditional law on the observed bag distribution. Equal reduced parameters produce equal conditional laws; conversely, any parameter producing the true conditional law has the same logit almost surely because the sigmoid is injective. Assumption (A1) then forces that parameter to equal θ^*. This is identification at the truth, not a claim that every pair of distinct parameters induces distinct laws.

Write the per-bag log-likelihood as

$$\ell_\theta(B, y) = y\eta_\theta(B) - \log\Big(1 + e^{\eta_\theta(B)}\Big), \qquad L(\theta) = \mathbb{E}\ell_\theta(B, y).$$

Correct specification gives

$$L(\theta^*) - L(\theta) = \mathbb{E}_B[\mathrm{KL}(\mathrm{Bern}(p^*(B)) \,\|\, \mathrm{Bern}(\sigma(\eta_\theta(B))))] \,.$$

By (A1), the expected Kullback–Leibler divergence is strictly positive for $\theta \neq \theta^*$, so θ^* uniquely maximizes L. For every finite bag, the log-likelihood is continuous in θ. The simplex constraint on $\mathbf{z}$ and compactness of Θ give a constant M with $\sup_{\theta\in\Theta} |\eta_\theta(B)| \le M$ and hence $\sup_{\theta\in\Theta} |\ell_\theta(B, y)| \le \log 2 + 2M$. This uniform envelope yields a uniform law of large numbers; the argmax theorem then gives $\widehat{\theta}_I \xrightarrow{p} \theta^*$ (Newey & McFadden, 1994). Consistency uses compactness from (A2) and continuity of the model. The interior and moment conditions are needed for the inference argument below.

To obtain ranking recovery, suppose at least one true feature pair is strictly ordered and define $\Delta_* = \min\{|D_g^* - D_h^*| : D_g^* \neq D_h^*\} > 0$. On the event $\|\widehat{\mathbf{D}} - \mathbf{D}^*\|_\infty < \Delta_*/2$, every estimated pairwise contrast has error smaller than Δ_*, so all strict comparisons have the correct sign. Consistency makes this event's probability tend to one. If every feature ties, there are no strict comparisons to recover. The result neither requires exact equality of estimated tied coordinates nor selects a particular representative $\widehat{\mathbf{S}}$ from its affine orbit.

D.3 Asymptotic normality and contrast intervals

Inference must account jointly for the distance parameter, normalization parameter, feature contrasts, and intercept. In the coordinates $\theta = (a, b, \mathbf{u}, \widetilde{\beta})$ with $\mathbf{D} = Q\mathbf{u}$, the required first derivatives are

$$\partial_a w_{ij} = -e^a w_{ij} \left(d_{ij} - \sum_k w_{ik} d_{ik} \right), \qquad \partial_b E_{ij,g} = e^b E_{ij,g} \left(x_{ij,g} - \mathbf{E}_{ij}^\top \mathbf{x}_{ij} \right),$$

$$\partial_a \eta = \mathbf{D}^\top \partial_a \mathbf{z}, \qquad \partial_b \eta = \mathbf{D}^\top \partial_b \mathbf{z},$$

$$\nabla_{\mathbf{u}} \eta = Q^\top \mathbf{z}, \qquad \partial_{\widetilde{\beta}} \eta = 1.$$

On a compact neighborhood of θ^*, the first derivatives are bounded by $C(1 + \max_j d_{ij} + \max_j \|\mathbf{x}_{ij}\|_\infty)$ and the second derivatives by $C(1 + \max_j d_{ij}^2 + \max_j \|\mathbf{x}_{ij}\|_\infty^2)$, with mixed products bounded using $2uv \le u^2 + v^2$. The moment condition in (A2) therefore provides finite score variance

and an integrable uniform envelope for the likelihood Hessian. Writing $p_\theta = \sigma(\eta_\theta)$, these likelihood derivatives are

$$\nabla \ell_\theta = (y - p_\theta)\nabla \eta_\theta,$$
$$\nabla^2 \ell_\theta = (y - p_\theta)\nabla^2 \eta_\theta - p_\theta(1 - p_\theta)\nabla \eta_\theta \nabla \eta_\theta^\top.$$

At the truth the conditional mean of the first Hessian term is zero, giving the information equality. The joint Fisher information in assumption (A3) is

$$\mathcal{I}(\theta^*) = \mathbb{E}\big[\sigma'(\eta_{\theta^*}(B))\nabla_\theta \eta_{\theta^*}(B)\nabla_\theta \eta_{\theta^*}(B)^\top\big].$$

Consistency and interiority place the MLE in the interior of Θ with probability tending to one, where its score vanishes. Expanding that score around θ^* and applying the score central limit theorem, uniform Hessian convergence, and (A3) gives

$$\sqrt{I}(\widehat{\theta}_I - \theta^*) = \mathcal{I}(\theta^*)^{-1} \frac{1}{\sqrt{I}} \sum_{i=1}^{I} \nabla \ell_{\theta^*}(B_i, y_i) + o_p(1),$$

and the limiting score covariance $\mathcal{I}(\theta^*)$ yields

$$\sqrt{I}(\widehat{\theta}_I - \theta^*) \Rightarrow \mathcal{N}(0, \mathcal{I}(\theta^*)^{-1}).$$

A consistent estimate of the per-bag Fisher information is

$$\widehat{\mathcal{I}} = \frac{1}{I} \sum_{i=1}^{I} \sigma'(\eta_{\widehat{\theta}_I}(B_i))\nabla \eta_{\widehat{\theta}_I}(B_i)\nabla \eta_{\widehat{\theta}_I}(B_i)^\top.$$

For the contrast $D_g - D_h$, define $\mathbf{c}_{gh} = \mathbf{e}_g - \mathbf{e}_h$. Its estimated variance is

$$\widehat{\mathrm{se}}_{gh}^2 = \frac{1}{I} \mathbf{c}_{gh}^\top Q [\widehat{\mathcal{I}}^{-1}]_{\mathbf{uu}} Q^\top \mathbf{c}_{gh}.$$

The block is taken *after inverting the full joint information*, so uncertainty in $a, b, \widetilde{\beta}$ is included. The resulting pointwise asymptotic $(1 - \delta)$ confidence interval is

$$\widehat{D}_g - \widehat{D}_h \ \pm z_{1-\delta/2} \widehat{\mathrm{se}}_{gh}.$$

Here $z_{1-\delta/2}$ is the corresponding standard normal quantile. For asymptotic simultaneous coverage over $K = \binom{G}{2}$ pairs, a Bonferroni construction uses $z_{1-\delta/(2K)}$. These intervals target the identifiable contrasts $D_g^* - D_h^*$; the scale of $S_g^* - S_h^*$ remains unidentified. The manuscript reports no empirical assessment of interval coverage.

D.4 Degeneracies, misspecification, and optimization

Identification of the feature coefficients at fixed a, b is weaker than joint identification of all reduced parameters. At fixed a, b, the covariate vector is $(Q^\top \mathbf{z}, 1)$, and positive definite $\mathrm{Cov}(Q^\top \mathbf{z})$ identifies both the feature coefficients and the intercept. An uncentered second moment of $Q^\top \mathbf{z}$ alone is insufficient when the intercept is unknown. When a, b are also estimated, positive covariance does not exclude confounding with either nonlinear parameter. Assumptions (A1) and (A3) must therefore be checked for the full model and the relevant bag distribution.

If all instances within each bag are equidistant, $w_{ij} = 1/n_i$ for every a, so a is unidentified; feature contrasts may remain identifiable after removing a. Likewise, $\mathbf{D} = 0$ removes both a, b from the logit. With purely one-hot features $\mathbf{x}_{ij} = \mathbf{e}_{c_{ij}}$,

$$\mathbf{E}_{ij}(b)^\top \mathbf{D} = k(b) D_{c_{ij}}, \qquad k(b) = \frac{e^{e^b} - 1}{e^{e^b} + G - 1} > 0.$$

The likelihood depends on $b, \mathbf{D}$ only through $\mathbf{T} = k(b)\mathbf{D}$. If analogues of (A1)–(A3) hold for the further reduced coordinates $(a, Q^\top \mathbf{T}, \widetilde{\beta})$, the same proof applies there. Because $k(b) > 0$, this additional reduction preserves the feature ordering. Theorem 1 does not apply directly to the larger coordinate system in this case.

Under misspecification, compactness and uniform convergence still give consistency for the population likelihood maximizer $\theta^\dagger$, provided it is unique. This maximizer is the conditional Bernoulli Kullback–Leibler projection of the true law onto the fitted family. Its feature ordering need not match the scientific ground truth. The information equality may also fail, so inverse-Fisher intervals do not inherit the correctly specified model's coverage guarantee.

The two affine symmetries create at least two null directions in the original parameterization's population Fisher information. They also create null directions in the likelihood Hessian at an interior stationary point. Away from stationarity, the Hessian need not have two zero eigenvalues because the symmetry orbits are curved. Inference therefore uses the reduced coordinates.

Theorem 1 concerns a global unpenalized MLE over Θ. Fixed weight decay can change both the selected affine representative and the fitted conditional law. The profile refits below compare unpenalized estimation with finite-epoch AdamW empirically; ranking agreement does not establish that AdamW attains the theorem's global maximum.

D.5 Profile-likelihood comparison with AdamW

Unpenalized profile fit. We compare the training procedure with unpenalized likelihood estimation by performing a profile-likelihood refit on each of the seven ranking benchmarks. Using the bag profile $\mathbf{z}_i(a,b)$ from Section 5 and the basis Q from Appendix D.1, set

$$L_I^{\mathrm{prof}}(a,b) = \sup_{\mathbf{u},\widetilde{\beta}} \frac{1}{I}\sum_{i=1}^{I} \left[y_i\eta_i - \log(1+e^{\eta_i})\right], \qquad \eta_i = \mathbf{z}_i(a,b)^\top Q\mathbf{u} + \widetilde{\beta}.$$

For fixed (a,b), minimizing the inner negative log-likelihood is a convex logistic-regression problem. The corresponding log-likelihood supremum need not be attained under complete or quasi-complete separation. A finite solver iterate in that case is not a finite MLE; the existence guarantee in Theorem 1 instead concerns its compact domain Θ.

Only a, b require an outer search: we evaluate a grid and refine around its best candidates. Denote the resulting numerical refit by $\widehat{\theta}_{\mathrm{PL}}$, with $\widehat{\mathbf{D}}_{\mathrm{PL}} = Q\widehat{\mathbf{u}}_{\mathrm{PL}}$. The search does not certify a global optimum. The refits form a separate estimator diagnostic and are not pooled with the main cross-validation performance averages.

Numerical approximation and the theorem. Let $\mathcal{L}_I(\theta) = \sum_{i=1}^{I} \ell_\theta(B_i, y_i)$ be the sample log-likelihood. Consistency extends to approximate maximizers satisfying $\sup_{\theta\in\Theta} I^{-1}\mathcal{L}_I(\theta) - I^{-1}\mathcal{L}_I(\widetilde{\theta}_I) = o_p(1)$. Uniform convergence then gives $L(\theta^*) - L(\widetilde{\theta}_I) = o_p(1)$; uniqueness and separation of the population maximum imply $\widetilde{\theta}_I \xrightarrow{p} \theta^*$. On a compact product search domain containing the truth, an exhaustive (a,b) grid with vanishing mesh and vanishing inner-solver error provides such an approximation by continuity and uniform convergence. A fixed finite grid alone gives no such guarantee. Asymptotic normality requires a stronger optimization condition: for example, a consistent interior estimate with normalized likelihood score $o_p(I^{-1/2})$ has the same first-order limit under (A3) and the preceding regularity conditions.

Comparison quantities and local curvature. We compare $\sigma(\widehat{\mathbf{s}}_{\mathrm{AdamW}})$ directly with $\widehat{\mathbf{D}}_{\mathrm{PL}}$ by Spearman correlation and by signs of pairwise differences. Both comparisons are invariant to the admissible positive affine changes of the influence scores. To assess local curvature, we use the full reduced-coordinate observed information per bag,

$$\widehat{H}_{\mathrm{obs}} = -\frac{1}{I}\sum_{i=1}^{I} \nabla_\theta^2 \ell_{\widehat{\theta}_{\mathrm{PL}}}(B_i, y_i), \qquad \theta = (a, b, \mathbf{u}, \widetilde{\beta}),$$

with the same log-parameter definitions and an orthonormal Q across fits. This is the joint Hessian, not the information of the inner logistic problem with a, b held fixed. In particular it includes the residual term $-(y_i - p_i)\nabla^2\eta_i$ as well as $p_i(1-p_i)\nabla\eta_i\nabla\eta_i^\top$. The computed observed information is positive definite on all seven benchmarks (Table 1). This finite-sample local curvature check does not establish population nonsingularity in (A3) or identification in (A1). Moreover, eigenvalues depend on coordinate scaling; reporting only their positivity does not quantify conditioning or exclude weakly identified directions.

Ranking agreement. Define $\mathcal{R}_{.95}$ as the feature pairs whose reported two-sided 95% Wald reference intervals under the profile fit exclude zero. We call these pairs *resolved* solely as a diagnostic convention. AdamW and the profile fit agree on the contrast sign for 100% of these pairs on every benchmark. Their Spearman correlations with the ground truth differ by at most 0.01 across all seven benchmarks. This difference between two correlations with the truth is distinct from the direct AdamW–profile Spearman correlation, which is approximately 0.95 or higher on Spatial-MNIST Dense. Because the profile fit selects $\mathcal{R}_{.95}$, sign agreement on this set is descriptive and does not provide simultaneous confidence in all feature orderings or calibrated inference under misspecification.

Table 1: Profile-likelihood versus AdamW ranking diagnostics. Across all seven benchmarks, pairs whose profile-fit two-sided 95% Wald reference intervals exclude zero have matching signs, and the estimators' ground-truth Spearman correlations differ by at most 0.01. The final column reports only the sign of the minimum eigenvalue of the full reduced observed information per bag.

Benchmark	Additional ranking comparison	$\lambda_{\min}$
Spatial-MNIST Sparse	All 125 active–null pairs resolved; no null–null pair resolved	> 0
Spatial-MNIST Correlated	All resolved feature pairs have matching signs	> 0
Spatial-MNIST Dense	Direct AdamW–profile Spearman $\gtrsim 0.95$	> 0
Spatial transcriptomics	Both rank type 2 first	> 0
H&E immune-proximity	Both attain ground-truth Spearman 0.5901	> 0
H&E morphology-interaction	Both rank `area` third	> 0
H&E multi-type	Both rank blood last	> 0

On Spatial-MNIST Sparse, all $5 \times 25 = 125$ active–null comparisons are resolved, whereas none of the $\binom{25}{2} = 300$ null–null comparisons is. This agrees with the known active/null grouping, but an interval containing zero does not establish equality. The result therefore does not demonstrate recovery of the true ties. The Spearman ceiling for distinct estimated ranks against tied ground truth arises from the rank metric itself, independently of these interval diagnostics.

Sampling and the target parameter. If a bag is sampled with probability $q_y > 0$ depending only on its label, Bayes' rule gives the sampled conditional log odds

$$\operatorname{logit} p_{\text{sample}}(B) = \operatorname{logit} p_{\text{population}}(B) + \log(q_1/q_0).$$

Under correct specification, the sampled law remains a LION logistic model with the same $a, b, \mathbf{D}$ and a shifted $\widetilde{\beta}$, provided the shifted intercept lies in the admissible set. This case–control property (Prentice & Pyke, 1979) describes the unpenalized model target; it does not remove regularization bias in AdamW. Under misspecification, reweighting can change the Kullback–Leibler projection. The spatial transcriptomics and H&E refits therefore support empirical ranking agreement under their sampling distributions, not equality of population influence parameters before and after sampling.

Signed influence relative to a reference group. For the H&E multi-type rule, let $\mathcal{G}_0$ be the zero-influence features specified by the injected ground truth, and let $\bar{\widehat{D}}_0 = |\mathcal{G}_0|^{-1} \sum_{g \in \mathcal{G}_0} \widehat{D}_g$. The relative contrast is

$$\widehat{r}_{\text{blood}} = \frac{\widehat{D}_{\text{blood}} - \bar{\widehat{D}}_0}{\widehat{D}_{\text{immune}} - \bar{\widehat{D}}_0}.$$

The estimated ratio is -0.40, with a reported 95% model-based Wald reference interval of $[-0.60, -0.20]$; the injected coefficient ratio is $-2.5/5 = -0.5$. The interval lies below zero in the fitted model. Because the H&E generator is outside the fitted family, neither this interval nor its agreement with the injected ratio establishes calibrated coverage or recovery of effect magnitudes.

The ratio is invariant to the original affine representation of $\mathbf{S}$ and to the additional positive scale ambiguity in the one-hot case (Appendix D.4). It measures influence relative to a specified zero-influence reference group. In contrast, a raw sigmoid score that rounds to zero has no such signed interpretation, and $\widehat{D}_{\text{blood}} < 0$ by itself only places blood below the unweighted mean across features.

Uncertainty for contrasts and ratios. Write $v = \mathbf{e}_{\text{blood}} - m_0$ and $w = \mathbf{e}_{\text{immune}} - m_0$, where $(m_0)_g = \mathbf{1}\{g \in \mathcal{G}_0\}/|\mathcal{G}_0|$. For $r = (v^\top \mathbf{D})/(w^\top \mathbf{D})$ with denominator bounded away from zero, the delta-method gradient in reduced coordinates is

$$\nabla_\theta r = \left(0, 0, \frac{Q^\top (v - rw)}{w^\top \mathbf{D}}, 0\right)^\top.$$

The delta-method variance uses the covariance of the full joint estimator, including uncertainty in a, b. Conditioning on the selected grid point would omit that uncertainty. For independent bags under misspecification, the corresponding robust covariance estimate, subject to regularity and nonsingularity at the population optimum, is

$$\widehat{V}_{\text{sand}} = \frac{1}{I} \widehat{H}_{\text{obs}}^{-1} \widehat{J} \widehat{H}_{\text{obs}}^{-1}, \qquad \widehat{J} = \frac{1}{I} \sum_i \widehat{\psi}_i \widehat{\psi}_i^\top, \qquad \widehat{\psi}_i = \nabla_\theta \ell_{\widehat{\theta}_{\text{PL}}}(B_i, y_i).$$

For spatially dependent bags, the covariance must also account for their sampling dependence. By contrast, $\widehat{H}_{\text{obs}}^{-1}/I$ relies on the information equality for a correctly specified independent-bag likelihood. A denominator near zero can also invalidate the delta approximation and require ratio-specific inference. These formulas state the additional requirements for population inference; the reported blood diagnostic did not establish them. We therefore interpret its interval as local uncertainty within the fitted model and the resolved-pair agreement as a descriptive comparison of fitted rankings.

E Proof of Theorem 2

E.1 Exact perturbation and the size bound

For a fixed bag size n, define the distance-and-permutation factor under Theorem 2 as

$$\kappa(n) = \sqrt{2/\pi}\, \mathbb{E}_{\mathbf{d},\pi} \|\boldsymbol{\omega} - \boldsymbol{\omega} \circ \pi^{-1}\|_2.$$

For a varying feature g at this bag size, the exact logit sensitivity in part (a) is

$$\mathbb{E}|\eta(B^{\pi,g}) - \eta(B)| = |w_g| \tau \kappa(n).$$

For a bag-constant feature it is zero. The size bound and asymptotic constant are

$$0 \le \kappa(n) \le 2\sqrt{2/\pi}\, e^{\lambda r_{\max}/2} n^{-1/2}, \qquad \sqrt{n}\kappa(n) \longrightarrow \frac{2}{\sqrt{\pi}} \frac{\sqrt{\text{Var}(e^{-\lambda d})}}{\mathbb{E}[e^{-\lambda d}]} > 0.$$

To derive these expressions, condition on the distances and permutation and use the convention $x_{jg}^{\pi,g} = x_{\pi(j),g}$. Reindexing the perturbed sum gives

$$\Delta_g = \eta(B^{\pi,g}) - \eta(B) = w_g \sum_j (\omega_{\pi^{-1}(j)} - \omega_j) x_{jg}.$$

For a bag-constant column, the coefficient differences sum to zero, so $\Delta_g = 0$. For a varying column, independence from $(\mathbf{d}, \pi)$ and the same zero-sum identity make Δ_g conditionally Gaussian with mean zero and variance $w_g^2 \tau^2 \|\boldsymbol{\omega} \circ \pi^{-1} - \boldsymbol{\omega}\|_2^2$. Taking the Gaussian absolute first moment and then averaging over distances and permutations yields the stated sensitivity. The size bound follows from

$$\|\boldsymbol{\omega} - \boldsymbol{\omega} \circ \pi^{-1}\|_2 \le 2\|\boldsymbol{\omega}\|_2, \qquad \|\boldsymbol{\omega}\|_2^2 \le \max_j \omega_j \le e^{\lambda r_{\max}}/n,$$

while uniform weights are unchanged by every permutation. Their perturbation therefore vanishes on both the logit and probability scales, including when $n = 1$.

Exact asymptotic rate. Let $V_j = e^{-\lambda d_j}$, $\bar{V}_n = n^{-1} \sum_j V_j$, $\overline{V^2}_n = n^{-1} \sum_j V_j^2$, and $C_n = n^{-1} \sum_j V_j V_{\pi^{-1}(j)}$. Then

$$n\|\boldsymbol{\omega} - \boldsymbol{\omega} \circ \pi^{-1}\|_2^2 = \frac{2(\overline{V^2}_n - C_n)}{\bar{V}_n^2}.$$

Conditional on $V_1, \ldots, V_n$, a uniform permutation gives $\mathbb{E}_\pi C_n = \bar{V}_n^2$ and $\text{Var}_\pi(C_n) = (\overline{V^2}_n - \bar{V}_n^2)^2/(n-1)$ for $n \ge 2$. Boundedness of V_j and the law of large numbers imply $C_n \xrightarrow{p} (\mathbb{E}V)^2$. The last display therefore converges in probability to $2\,\text{Var}(V)/(\mathbb{E}V)^2$. The square root is uniformly bounded by $2e^{\lambda r_{\max}/2}$, so its expectation converges as well. Multiplying by $\sqrt{2/\pi}$ gives the stated limit for $\sqrt{n}\kappa(n)$. The limit is positive because $\lambda > 0$ and the non-point-mass distance law make $\text{Var}(V) > 0$.

E.2 CONTROLS AND PROBABILITY-SCALE INVERSION

Let A_g indicate that column g varies. Part (a) gives the global *logit-scale* sensitivity

$$\mathbb{E}|\Delta_g| = |w_g|\tau\,\mathbb{E}[\kappa(n)\mathbf{1}\{A_g = 1\}].$$

In the first control, $A_g = 1$ for every feature and bag, so the final factor is $\mathbb{E}\kappa(n)$ for every feature. In the second, fixed n makes it $\kappa(n)\Pr(A_g = 1)$, which is common to features with the same varying fraction. For $n \geq 2$, unequal distances and a permutation exchanging their weights occur with positive probability, implying $\kappa(n) > 0$. A positive varying fraction therefore preserves coefficient-magnitude order. Both controls use the common conditional distance law and feature variance in the theorem; neither asserts proportionality on the probability scale.

In the construction for part (c), permutation changes feature g only in the large bags, which have probability $1/2$. The sigmoid bound $|\sigma(u) - \sigma(v)| \leq |u - v|/4$ therefore gives

$$\bar{R}_g \leq \frac{1}{8} w_g \tau \kappa(N) \leq \frac{1}{4}\sqrt{2/\pi}\, w_g \tau e^{\lambda r_{\max}/2} N^{-1/2}.$$

Feature h changes only in the small bags, so $\bar{R}_h = \rho_h/2$, where $\rho_h = \mathbb{E}[R_h(B) \mid n = n_S]$ is independent of N. Moreover, $\rho_h > 0$: with positive probability, at least two distances are unequal and the permutation exchanges unequal weights. Conditional on such distances and permutation, Δ_h is a nondegenerate Gaussian. It is therefore nonzero almost surely, and the strictly increasing sigmoid produces a nonzero probability change. Consequently, any integer

$$N_0 > \max\left\{ n_S, \left(\frac{\sqrt{2/\pi}\, w_g \tau e^{\lambda r_{\max}/2}}{2\rho_h} \right)^2 \right\}$$

ensures $\bar{R}_g < \bar{R}_h$ for every $N \geq N_0$. Thus inversion occurs for any fixed finite coefficient ratio $w_g/w_h > 1$; the argument requires no assumption of sigmoid saturation.

Coefficient recovery in the same construction. For the consistency assertion in part (c), take bags and their Bernoulli labels to be i.i.d. draws from the stated two-feature mixture. With known λ, define the logistic-regression covariate $\mathbf{t}(B) = \sum_j \omega_j \mathbf{x}_j$. Conditional on the distances and bag type, its two coordinates are independent Gaussians. The varying coordinate has variance $\tau^2 \sum_j \omega_j^2 > 0$, and the bag-constant coordinate has variance $\tau^2 > 0$. The conditional covariate distribution therefore has full support in $\mathbb{R}^2$, so no nonzero affine function of $\mathbf{t}$ vanishes almost surely. Injectivity of the sigmoid then identifies (w_g, w_h, β).

On the stated compact parameter set, the log-likelihood has an integrable envelope $C(1 + \|\mathbf{t}\|)$. Uniform convergence and the argmax theorem therefore establish consistency. For each fixed $N \geq N_0$, the MLE recovers the strict coefficient order as $I \to \infty$, even though the oracle permutation ordering remains inverted. This is a result for logistic regression with identity features and known decay, not for jointly trained LION. Since $0 \leq R_g(B) \leq 1$, empirical averages of oracle permutation importance converge to $\bar{R}_g$; averaging additional bags converges to the same inverted ordering.

E.3 REPRODUCIBLE NUMERICAL CHECK

The numerical check evaluates the two-feature oracle with $\mathbf{w} = (2, 1)$, $\beta = 0$, $\mu = 0$, $\tau = 1$, $n_S = 5$, $d \sim \mathrm{Uniform}[0, 1]$, and $\lambda = 3$. Each setting uses 20,000 independent bags, equally likely small and large types, and five permutations per feature per bag; the random seed is 20260923. Table 2 reports probability-scale importance and bag-level Monte Carlo standard errors. Each bag, rather than each permutation, is the independent sampling unit. The equal-size control and increasing large-bag sizes illustrate the probability-scale inversion, while the fitted coefficients retain their order.

These numerical diagnostics are separate from the benchmark protocols. The check also evaluates affine invariance and zero within-bag permutation effects under uniform weights. It concerns the stated oracle linear predictor; extending the result to the softmax feature map requires separate analysis.

Table 2: Oracle permutation importance and coefficient recovery as large-bag size N increases. Importance values are means $\pm$ bag-level Monte Carlo standard errors; $N = 5$ is the equal-size control. The last two columns give linear-logistic coefficient estimates from newly sampled labels with known decay. These are theorem diagnostics, separate from the LION benchmarks.

N	$\bar{R}_g$ $(w_g = 2)$	$\bar{R}_h$ $(w_h = 1)$	$\widehat{w}_g$	$\widehat{w}_h$
5	0.0579 ± 0.0006	0.0241 ± 0.0003	2.070	0.976
10	0.0505 ± 0.0005	0.0240 ± 0.0003	1.981	0.972
20	0.0385 ± 0.0004	0.0244 ± 0.0003	2.007	0.989
50	0.0262 ± 0.0002	0.0244 ± 0.0003	1.953	0.964
100	0.0186 ± 0.0002	0.0240 ± 0.0003	2.051	1.011
200	0.0132 ± 0.0001	0.0242 ± 0.0003	2.010	1.006

F Sensitivity to Feature and Distance Degeneracy

We test how feature and distance degeneracy affect estimation on Spatial-MNIST Sparse, whose labels follow LION's forward pass. Both sweeps hold $\mathbf{S}^*$, bag size, sample count, and training settings fixed. The feature sweep measures the smallest eigenvalue $\lambda_{\min}$ of the bag-profile covariance restricted to $\mathbf{1}^\perp$; the distance sweep measures the within-bag coefficient of variation of kernel weights. These quantities describe variation relevant to identification. They do not establish identification in (A1), the joint Fisher condition in (A3), or the compactness, interiority, and moment conditions in (A2) of Theorem 1.

Distance degeneracy. Across the distance sweep, Spearman ρ remains at 0.644–0.647, at least 99.1% of the reported 0.6493 ceiling for distinct estimated ranks against tied truth, and top-k precision remains 1.00. At a kernel-weight coefficient of variation of 0.038, however, the learned log-decay parameter varies substantially ($\widehat{a} = 0.4469 \pm 0.7069$ versus $a^* = 1.0$). At the exactly equidistant limit, a disappears from the likelihood, although feature contrasts may remain identifiable in a smaller parameterization (Appendix D.4). Stable feature ranking therefore need not imply stable estimation of distance decay.

Feature degeneracy. In the feature sweep, ρ retains at least 98.6% of its reported ceiling when $\lambda_{\min}$ is reduced by a factor of 70 from the baseline 7.3×10^{-4}. At a factor of 287, ρ is still 0.504 ± 0.071; it falls to chance when $\lambda_{\min} = 0$ in this construction. Across both sweeps, LION attains 0.96–0.99 of the Bayes-optimal AUROC on the same splits. These are finite-sample observations, not verification of the theorem's asymptotic conclusions.

The feature sweep reduces the design isotropically, changing signal amplitude as well as the smallest covariance eigenvalue. It therefore does not isolate weakness in particular feature directions while holding the label signal fixed. The observed robustness should be read within this joint change in design variation and signal strength.

G Training Details and Protocols

Data sources and ranking targets. Table 3 separates the seven ranking benchmarks from the two tissue case studies. Spatial-MNIST uses the LION generative model, simulated spatial transcriptomics uses an external rule with a hard distance boundary, and H&E uses injected spatial rules on real tissue features and geometry. These constructed-label benchmarks provide exact ranking targets. VisiumHD and Xenium instead use observed assay labels and provide no exact ground-truth feature ranking. Appendices H–L describe their construction and evaluation.

Table 3: Data provenance and ranking supervision. Constructed-label benchmarks provide exact ranking targets; the tissue case studies evaluate prediction and biological associations without such targets.

Evaluation	Features / geometry	Label rule	Exact ranking target
Spatial-MNIST	simulated / simulated	LION-form	yes
Spatial transcriptomics	simulated / simulated	external hard boundary	yes
H&E pathology	real / real	injected spatial rule	yes
VisiumHD myocarditis	real / real	observed assay label	no
Xenium tumors + TCR-seq	real / real	observed assay label	no

Optimization. LION minimizes binary cross-entropy, jointly optimizing $\{s_1, \ldots, s_G, a, b, \alpha, \beta\}$ with AdamW. Mini-batches consist of k bags: one positive and $k-1$ negative. The unpenalized profile-likelihood refit in Appendix D.5 is a separate estimator diagnostic. It compares rankings from that refit with those from finite-epoch, regularized training; its Wald intervals apply to the profile fit, not to the AdamW parameters.

Main cross-validation protocol. Figure 4 uses paired 5-fold cross-validation with five seeds, giving $5 \times 5 = 25$ runs per model and benchmark. Folds are generated once with random seed 0 and shared across models. These runs use the default configuration: AdamW, learning rate 10^{-3}, weight decay 0.01, 300 epochs, and batch size 1. Hyperparameters are not selected using the evaluation folds.

Spatial-MNIST uses 5,000 bags per setting, with rule-specific neighborhood sizes described in Appendix H. The spatial transcriptomics benchmark uses 5,000 pseudo-cells with 500 focal points, a maximum radius of $100\,\mu$m, a ground-truth boundary at $30\,\mu$m, and Gaussian decay with $\sigma = 15\,\mu$m. The H&E benchmarks use 2,000 sampled tumor cells per rule, a 150-pixel neighborhood radius, a 60-pixel active zone, Gaussian decay with $\sigma = 30$ pixels, and label-noise standard deviation $\sigma = 0.3$.

Hyperparameter search. The baseline grid combines learning rates $\{3 \times 10^{-4}, 10^{-3}, 3 \times 10^{-3}\}$, weight decays $\{10^{-3}, 10^{-2}\}$, and hidden widths $\{32, 64, 128\}$, yielding 18 configurations per baseline. LION has no hidden width and uses the corresponding six learning-rate/weight-decay combinations. SI-MIL additionally sweeps $K \in \{8, 16, 32, \text{all}\}$ and $\lambda \in \{0.5, 1, 2\}$. Hyperparameter selection uses validation AUROC on an inner $80/20$ split of the training portion; the evaluation split is held out from selection.

The reported changes after tuning are ± 0.01–0.04 AUROC and do not reverse the model comparisons. Variation across the entire search grid can be larger: on Spatial-MNIST Dense, the AUROC ranges are 0.030 for AttentionMIL+dist, 0.068 for DeepSets+dist, and 0.057 for GNN+dist. DeepSets+dist is particularly sensitive, spanning 0.225 AUROC on H&E immune-proximity and 0.440 on spatial transcriptomics. These full-grid ranges describe configuration sensitivity, separately from the change after validation-based selection.

Separate comparisons and diagnostics. The additional analyses retain their own protocols and are not pooled with the main CV results. The direct coefficient comparison is specified in Appendix P; component ablations use three seeds (Appendix T). Spatial-transcriptomics score-gap and distance-recovery diagnostics use five seeds on a single split (Appendix I). The hyperparameter-oracle diagnostic uses the inner search split and two seeds (Appendix O), whereas the attribution-readout diagnostic compares readouts of the same trained models using five seeds on one $80/20$ split (Appendix Q). Additive MIL results are compared only within their stated protocol. Profile-likelihood ranking agreement and curvature summaries are reported separately in Appendix D.5.

Compute. The main sweep used a per-bag training loop at batch size 1 and was CPU-bound for the small models in that comparison (12 to 12.4k parameters). It ran on CPU workers, using approximately 547 core-hours for 1,835 runs. Development runs and the tissue case studies used a single NVIDIA A100 GPU.

H SPATIAL-MNIST BENCHMARK

H.1 BENCHMARK DESIGN

The three Spatial-MNIST settings test recovery of a known feature ordering under increasingly demanding coefficient and correlation structures. Labels follow the generative model in Definition 2, with known latent scores $\mathbf{s}^*$ and influence coefficients $\mathbf{S}^* = \sigma(\mathbf{s}^*)$. Features are derived from MNIST digit prototypes projected to lower dimensions, with structured noise. Approximately 30% of instances receive a "recruiter boost," an additive increase in generator-active features. Because the label rule weights nearby instances more strongly, these boosts create a feature–distance signal.

Sparse Continuous. This setting has $G = 30$ features. Five have latent scores $s_g \in [3.0, 5.0]$; the remaining 25 have $s_g = -6$, corresponding to sigmoid scores of approximately 0.0025. The task is to separate the five high-coefficient features from the low-coefficient group.

Correlated Groups. This setting also has $G = 30$ features, including 20 organized into four groups of five that share a latent factor. One feature per group has a high generator coefficient ($s_g \in [3.5, 5.0]$); the correlated low-coefficient features have $s_g = -6$. The task is to distinguish the four generator-active features from correlated alternatives.

Dense Gradient. This setting has $G = 60$ features. The latent scores s_g^* form four groups of 15: high positive, strongly negative, moderate, and near-zero values. Labels depend on $S_g^* = \sigma(s_g^*)$, so these descriptions refer to latent-score ranges, not to positive, negative, or null scientific effects. The sigmoid preserves their ordering. Bag sizes vary from 5 to 80, and the task is to recover the full ordering across heterogeneous neighborhoods. The auxiliary upper-coefficient diagnostic uses the largest generator coefficients (Appendix P); it does not classify features by effect sign.

H.2 DATA GENERATION

Label generation. Each setting uses LION's forward pass to generate a bag's positive-label probability:

$$p(y_i = 1 \mid B_i) = \sigma\left(\gamma \sum_{j=1}^{n_i} w_{ij} \sum_{g=1}^{G} E_{ij,g}\sigma(s_g^*) + \beta\right).$$

The weights w_{ij} and normalized features $E_{ij,g}$ use the distance kernel and feature normalizer of Section 4. The positive scale γ plays the role of e^{α^*}. The intercept β is calibrated so that the median bag has approximately 40% probability of a positive label. This calibration concerns the generator; the paired evaluation loader sets the evaluated class balance to 0.50 (Appendix N).

Feature generation. For Sparse Continuous, base features are i.i.d. $\mathcal{N}(0, 1)$. Approximately 30% of instances receive an additive recruiter boost on the active features, drawn from $\mathrm{Exp}(1.5)$ and scaled by feature rank. Bags with more boosted instances near their centers consequently receive larger scores.

Correlated Groups uses the same base features. Each recruiter instance receives a shared latent boost across its group and a private boost on the generator-active feature, producing within-group correlation of approximately $r = 0.6$. Dense Gradient also uses the same base features, with boosts on the 15 high-positive-latent-score features and bag sizes from 5 to 80. These generators use the softmax feature map and are distinct from the linear construction in Theorem 2.

Generator and training parameters. The generator uses $a^* = 1.0$, $b^* = 1.5$, and $\gamma = 25.0$ for Sparse Continuous and Correlated Groups, or $\gamma = 20.0$ for Dense Gradient. LION is initialized at $a_0 = 0.0$, $b_0 = 1.0$, $s_{g,0} = -1.0$, $\alpha_0 = 1.0$, and $\beta_0 = 0.0$. Training uses AdamW with learning rate 10^{-3}, weight decay 0.01, 300 epochs, and batch size 1 (one positive and one negative bag).

H.3 Experimental Setup

The main CV comparison includes DeepSets+dist (Zaheer et al., 2017), AttentionMIL+dist (Ilse et al., 2018), and GNN+dist (Kipf & Welling, 2016), with hidden width 64 in the default configuration. AttentionMIL+dist and DistAttMIL denote the same model. Each baseline receives the instance features and scalar distances specified in Section 6.1. Comparisons with SI-MIL (Kapse et al., 2024) and Additive MIL (Javed et al., 2022) follow their separately reported protocols. Distance-blind DeepSets and attention MIL appear in the distance-access ablation (Appendix M) and the separate readout diagnostic (Appendix Q).

For the deep baselines, global feature rankings use absolute within-bag permutation importance. For each evaluation bag and feature g, we shuffle that feature across instances five times, compute the absolute probability change, and average over permutations and bags. For probability output f, the population readout is

$$R_g = \mathbb{E}_{B,\pi} \left| f(B^{\pi,g}) - f(B) \right|.$$

This nonnegative readout measures perturbation magnitude, whereas the ranking target orders coefficients by value. Their correlation is therefore a comparison across different quantities. The signed alternative $\mathbb{E}[f(B^{\pi,g}) - f(B)]$ is not used: exchange symmetry can cause cancellation even for an influential feature. Integrated Gradients and KernelSHAP are evaluated separately in Appendix Q.

The reported metrics distinguish prediction from ranking recovery:

- **AUROC**: discrimination of positive and negative bags on held-out data.
- **Spearman** ρ: rank correlation between the estimated scores and the ground-truth coefficient ordering.
- **Top-k precision**: the fraction of ground-truth top-k features recovered in the estimated top-k set.
- **NDCG**: normalized discounted cumulative gain of the estimated ranking.

LION and the three deep baselines use 5,000 bags per dataset and the main paired 5-fold CV $\times$ 5-seed protocol, with 300 epochs of AdamW at learning rate 10^{-3}. Figure 4 summarizes prediction and ranking; Tables 5 and 6 give the full prediction metrics and numerical ranking results.

H.4 Analysis of Results

Prediction under the matched generator. LION achieves AUROC 0.879, 0.886, and 0.837 on Sparse, Correlated, and Dense, respectively, compared with 0.677, 0.682, and 0.601 for the best distance-equipped deep baseline on each setting (Figure 4(a)). The margins are 0.20–0.24. These results show that LION supports both prediction and ranking when the data-generating structure matches its architecture.

Ranking recovery. On Sparse and Correlated, the differences between LION and the best deep baseline are small: Spearman is 0.648 versus 0.621 and 0.588 versus 0.561, respectively, with near-perfect top-k precision. The tied ground truth limits Spearman for strictly ordered estimates. On Sparse, its ceiling under that convention is 0.6493; LION reaches 99.8% of this value and the best baseline 95.7%.

Dense requires recovery of a fuller ordering across 60 features and bag sizes from 5 to 80. LION reaches Spearman 0.863, compared with 0.501 for the best absolute permutation readout. Because the target orders coefficients by value, this gap alone does not establish failure to recover magnitude-based importance. The separate coefficient comparison yields 0.85 for LION versus 0.74 for DW-logistic (Figure 3).

Scope of the Dense result. Permutation importance estimates 60 feature-specific readouts, each subject to Monte Carlo error, across bags of heterogeneous sizes. Theorem 2 establishes that size and composition can reverse permutation rankings in a related linear model. It does not identify the cause of the Dense gap: this benchmark uses a softmax feature map and nonlinear baselines. The theorem's two coefficients are positive, so its counterexample does not rely on a sign–magnitude mismatch; that separate mismatch remains relevant to the empirical attribution comparison. Dense also requires

resolving many moderate contrasts. The advantage of native scores on this benchmark is an empirical observation under these combined conditions.

I Simulated Spatial Transcriptomics Benchmark

I.1 Benchmark Design

This benchmark tests feature and distance recovery under a spatial rule outside LION's generative family. Labels depend on the local abundance of one simulated cell type, restricted to a fixed active radius and weighted by Gaussian decay. The model must identify type 2 among eight feature dimensions, rank the other four cell types and three morphology features below it, and approximate the spatial decay (Figures 4(c) and 6). Ranking separation, rather than proximity of fitted scores to zero, is the relevant feature-recovery criterion because LION's scores admit positive affine reparameterizations.

The benchmark differs from Spatial-MNIST in three respects. Its features combine one-hot cell types with continuous morphology measurements; its label rule has a hard boundary at $30\,\mu\mathrm{m}$; and its raw score is a direct weighted sum without LION's softmax normalization or sigmoid output. It therefore evaluates performance under a specific form of model misspecification, alongside recovery of a known feature ordering.

I.2 Data Generation

Canvas and features. We scatter $N = 5{,}000$ pseudo-cells uniformly over a $1000 \times 1000\,\mu\mathrm{m}$ canvas. Each cell is assigned uniformly to one of $K = 5$ types and represented by a one-hot type vector. We concatenate $D = 3$ morphology features drawn independently from $\mathcal{N}(0, 1)$, yielding $G = K + D = 8$ instance features.

Bags. We sample $M = 500$ focal cells without replacement. For each focal cell i, all other cells within Euclidean distance $R_{\max} = 100\,\mu\mathrm{m}$ form bag B_i. Bags containing fewer than five instances are discarded.

Labels. The noiseless score is

$$\mathrm{score}_i = \sum_{j \in B_i} \mathbb{1}[c_j = 2]\,\mathbb{1}[d_{ij} < 30]\,\exp\left(-\frac{d_{ij}^2}{2 \cdot 15^2}\right),$$

where c_j is the cell type and d_{ij} is the distance in $\mu\mathrm{m}$. Only type-2 cells inside the $30\,\mu\mathrm{m}$ boundary contribute, with Gaussian decay scale $\sigma = 15\,\mu\mathrm{m}$; all other types and morphology features have zero coefficients in this rule. We add Gaussian noise with standard deviation $\sigma_\epsilon = 0.3$ and threshold the resulting score at a calibrated cutoff targeting 40% positive labels.

I.3 Distance Recovery Evaluation

We probe the learned distance response with synthetic bags containing one target-type cell at distance $d \in [0, R_{\max}]$ and 20 background cells. At each distance, we compute the target cell's effective contribution, its distance weight multiplied by its instance score, and compare the resulting curve with

$$\mathbb{1}[d < 30]\exp\left(-d^2/450\right).$$

Pearson correlation summarizes agreement between the learned and ground-truth curves.

I.4 Analysis of Results

The score-gap and distance-recovery results below use five seeds on a single diagnostic split, as specified in Appendix G. They are separate from the paired CV comparison in Figure 4.

Feature identification. LION ranks the target cell type first in all five seeds. Its reported target score is 0.973 on the 0–1 scale, with a gap of 0.886 to the next type. DistAttMIL also identifies the target type, with reported score 0.393 and gap 0.259. These numerical gaps are not comparable effect sizes: the readouts use different scales, and LION's raw gap depends on its affine representation. DeepSets and AttentionMIL identify the target type in only two of five seeds.

Distance response. The learned curve has Pearson correlation $r = 0.943 \pm 0.027$ with the ground-truth curve across five seeds (Figure 6). The reported per-curve $p < 10^{-10}$ values are not used as evidence from independent replicates: each curve evaluates the same fitted model over a distance grid. We interpret the correlations as descriptive measures of curve agreement. The ground truth follows Gaussian decay within the $30\,\mu\mathrm{m}$ active zone and drops discontinuously to zero at the boundary, whereas LION's exponential kernel $\exp(-e^a d)$ decays smoothly. The correlation therefore measures approximation of the distance response, not recovery of the hard cutoff. The reported residual influence is below 5% of its peak at $32.5\,\mu\mathrm{m}$ and below 1% by $50\,\mu\mathrm{m}$.

LION learns this response from bag labels without supervision on distance weights or instance scores. The original deep baselines do not expose the same standalone kernel; their probed responses are reported in Appendix S. DW-logistic also uses an explicit exponential kernel, $k_\lambda(d) = \exp(-\lambda d)$, so this structural property is shared by both models. No distance-recovery correlation for DW-logistic is reported here.

J H&E Pathology Benchmark

J.1 Benchmark Design

This semi-synthetic benchmark combines real tissue features and geometry with known spatial label rules. Nucleus detections from H&E-stained whole-slide images provide cell-type annotations and morphology measurements, including area, eccentricity, and solidity. Injecting labels into these observed neighborhoods makes the ranking target known while retaining tissue-specific feature distributions and spatial organization. Figure 9 illustrates nucleus detection, neighborhood construction, and the resulting ground-truth score map on a TCGA-LUAD tissue patch.

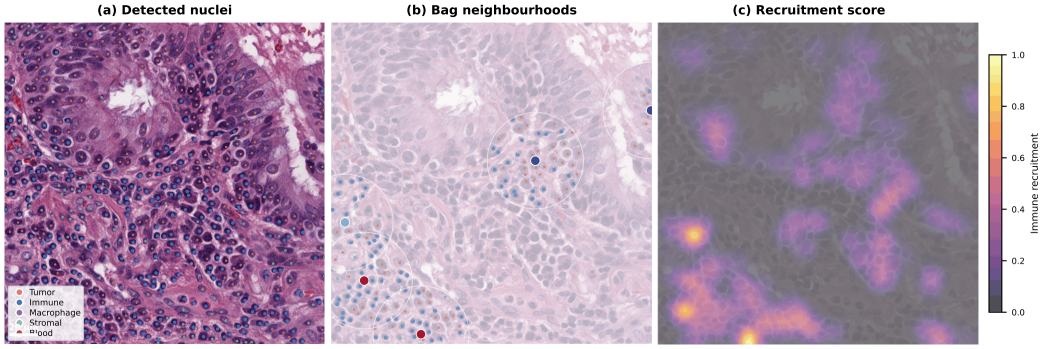

Figure 9: Construction of the H&E benchmark. (a) Detected nuclei in a TCGA-LUAD patch, colored by cell type: tumor (coral), immune (blue), stromal (teal), and blood (red); macrophages are also identified. (b) Five example neighborhoods; white outer circles mark the 150-pixel bag radius and dashed inner circles the 60-pixel active zone. Center-dot colors encode ground-truth recruitment scores (blue: high; red: low). (c) Gaussian-smoothed, normalized immune-proximity scores at tumor-cell locations; yellow marks high-score regions.

Three label rules. The immune-proximity rule uses the Gaussian-weighted count of immune cells and macrophages within a 60-pixel boundary, with decay scale $\sigma = 30$ pixels. It tests cell-type ranking. The morphology-interaction rule additionally weights neighboring immune cells by their area, so larger cells contribute more; it tests recovery of cell-type and morphology signals. The multi-type rule assigns positive contributions to immune cells, negative contributions to blood cells, and weak positive contributions to stromal cells, testing a ranking with effects in both directions.

J.2 Data Generation

Nucleus detection and features. We apply HD-YOLO (Rong et al., 2023) to H&E whole-slide images. Each detected nucleus has spatial coordinates, one of seven cell-type labels—tumor, stromal, immune, blood, macrophage, dead, or other—and eight morphology descriptors. Concatenating the one-hot cell type and z-scored morphology descriptors yields $G = 15$ features per instance.

Bags and labels. Each sampled tumor cell is a bag center. Its bag contains cells within $R_{\max} = 150$ pixels, while the label-generating rule acts within $d_{\text{boundary}} = 60$ pixels. The three rules share this geometry and differ in which cell types and morphology measurements contribute to the score. Their training configuration and label-noise scale are specified in Appendix G.

Illustrated tissue patch. Figure 9 uses a 1000×1000-pixel TCGA-LUAD patch (Weinstein et al., 2013), corresponding to $250 \times 250\,\mu\text{m}$ at $0.25\,\mu\text{m/px}$ in the nucleus-detection coordinates. The H&E image is resampled from its native microns-per-pixel resolution to align with these coordinates. The patch contains 608 nuclei, including 236 tumor, 246 immune, and 110 stromal nuclei. Panel (b) shows two high-score, one mid-score, and two low-score bags. The background is faded and nuclei outside these bags are dimmed, while cells inside retain their cell-type colors. The patch illustrates the construction; it does not define the full benchmark sample.

J.3 Experimental Setup

We use the distance-equipped deep baselines described in Section 6.1, including the GNN with distance-weighted message passing. Every model receives the instance features and scalar distances. For each rule, models are trained on 2,000 sampled tumor-cell neighborhoods with binary cross-entropy and AdamW (learning rate 10^{-3}, weight decay 0.01). The main comparison uses the default configuration and paired 5-fold CV with five seeds; folds are fixed by seed 0 and shared across models. Comparisons with additional interpretable MIL models retain their separate protocols.

J.4 Analysis of Results

Ranking against the signed target. LION achieves Spearman 0.590, 0.620, and 0.698 on immune-proximity, morphology-interaction, and multi-type, respectively. The corresponding best distance-equipped deep baseline, GNN+dist, reaches 0.561, 0.383, and 0.536 on the same folds (Figure 4(c)). LION's immune-proximity result reaches the ceiling for strictly ordered estimates under the tied-target convention (Appendix T). Its NDCG is 1.0 across all seeds and rules. This describes agreement at the top of the coefficient ordering; it does not measure recovery of every active feature, because a negative-effect feature can correctly rank at the bottom. The comparison also retains the distinction between LION's native coefficients and the baselines' absolute attribution readouts.

Feature-level scores. Figure 5 shows all 15 feature scores for the three rules on a shared color scale, with superscripts marking positive and negative ground-truth effects. Immune cells and macrophages have the highest scores under every rule (at least 0.92). All other scores are at most 0.25, except for area under morphology-interaction (0.28). Under immune-proximity, immune cells (0.99) and macrophages (1.00) rank highest, while every morphology score is at most 0.14.

Under morphology-interaction, area is the highest-scoring morphology feature and ranks third overall, consistent with its positive ground-truth weight ($s = 3.0$), below the cell-type weights ($s = 5.0$). Under multi-type, blood ranks last and its score rounds to 0.00, consistent with its negative ground-truth weight ($s = -2.5$). Stromal has a weak positive ground-truth effect and fitted score 0.07. The score's absolute level depends on its positive affine representation (Lemma 1); ordering and identifiable reduced contrasts carry the interpretation, rather than distance from zero (Appendix D.5).

Prediction. LION also leads the three distance-equipped deep baselines in AUROC: 0.910 versus 0.894 on immune-proximity, 0.782 versus 0.761 on morphology-interaction, and 0.898 versus 0.884 on multi-type. These predictive margins are modest. The signed-target Spearman margins, 0.029, 0.237, and 0.162, compare native coefficients with absolute attribution readouts; they are not prediction gains and do not isolate how much the difference in the quantities being ranked contributes to the gap. For scale, the attention and GNN models in Table 11 use 3.2k–5.2k parameters, whereas

LION uses $G + 4 = 19$; the table also includes the larger DeepSets model. The separate interpretable-baseline comparison includes a higher AUROC for mean-pooled Additive MIL on immune-proximity (Appendix R).

Morphology interaction. This rule distinguishes recovery of a weaker third feature (`area`, $s = 3.0$) from recovery of the two stronger cell-type features ($s = 5.0$). Its target has more distinct coefficient levels than immune-proximity. LION's Spearman of 0.620 is 88.8% of the tied-target ceiling 0.6982, compared with 0.383 and 54.8% for the best deep baseline. The separate direct coefficient comparison also favors LION (0.62 versus 0.40), so the observed advantage is not confined to the absolute-attribution comparison. Nevertheless, LION's bilinear form does not represent the generating interaction exactly (Appendix A).

Distance access and scope. Supplying scalar distance improves baseline AUROC on most benchmark–architecture combinations (Appendix M). Once distance is available to every deep model, LION still has higher signed-target Spearman across the seven benchmarks, by 0.026–0.362 (Figure 4(c)). Equal access to distance does not equate coefficient rankings with attribution magnitudes, nor does this comparison isolate a single architectural cause of the gap.

The H&E results extend the benchmark evidence to observed tissue features and geometry while retaining constructed labels. Their agreement with the fully synthetic benchmarks supports ranking recovery under these semi-synthetic conditions; it does not establish biological validity for the injected rules. The tissue case studies in Appendices K and L address observed biological labels separately.

K Case Study: T-Cell Infiltration in Mouse Myocarditis

This case study asks which ventricular-cardiomyocyte (VCM) expression programs are associated with reduced T-cell infiltration in myocarditis. LION predicts T-cell presence from the expression and spatial arrangement of nearby VCMs, then supplies a gene ordering for biological analysis. The observed labels support prediction and association analyses; they do not provide an exact gene-ranking target.

K.1 Data and Bag Construction

Data processing. We generated VisiumHD data from *Pdcd1*$^{-/-}$*Ctla4*$^{+/-}$ mice. Cell types were annotated using gene-signature scores projected onto a UMAP embedding. Leiden clusters were assigned biological identities from their signature enrichment and spatial organization. VCM expression was library-size normalized and log-transformed, and the 500 most highly expressed VCM genes were selected as input features.

Bags and supervision. Each bag contains VCMs within a radius of a focal spatial location. Instance features are VCM gene-expression values; distances are measured from each VCM to that location. The binary label indicates T-cell presence: spots with T-cell infiltration are positive bags. LION is trained with binary cross-entropy to predict this label.

K.2 Predictive Performance

LION has the highest reported mean AUROC among the compared methods, including DW-logistic and the neural MIL baselines (Table 4). Its mean AUROC is 0.86, compared with 0.82 for GNN and 0.78 for DW-logistic. The table reports substantial variability, so the ordering of means alone does not establish a statistically significant difference.

K.3 Pathway Associations

Gene Ontology enrichment analysis (Eden et al., 2009) of genes ranked toward the non-infiltration end identifies ATP- and TCA-cycle-related terms (Figure 10(a)). We then compare gene-set expression in cardiomyocyte regions with and without T-cell infiltration. ATP-related and mitochondrial gene sets are more highly expressed in non-infiltrated regions. The Wilcoxon tests for ATP, general metabolism,

Table 4: Prediction of T-cell presence from VCM neighborhoods in VisiumHD myocarditis data. Entries are mean AUROC ± standard deviation.

Method	AUROC
LION	$\mathbf{0.86 \pm 0.107}$
GNN	0.82 ± 0.141
DW-logistic	0.78 ± 0.14
AttentionMIL	0.75 ± 0.102
DistAttMIL	0.74 ± 0.213
DeepSets	0.71 ± 0.143
Linear regression	0.65 ± 0.168

and mitochondrial gene sets give $P < 1 \times 10^{-3}$, $P = 0.018$, and $P < 1 \times 10^{-3}$, respectively (Figure 10(b)). The general-metabolism set provides a comparison for assessing the breadth of the expression pattern.

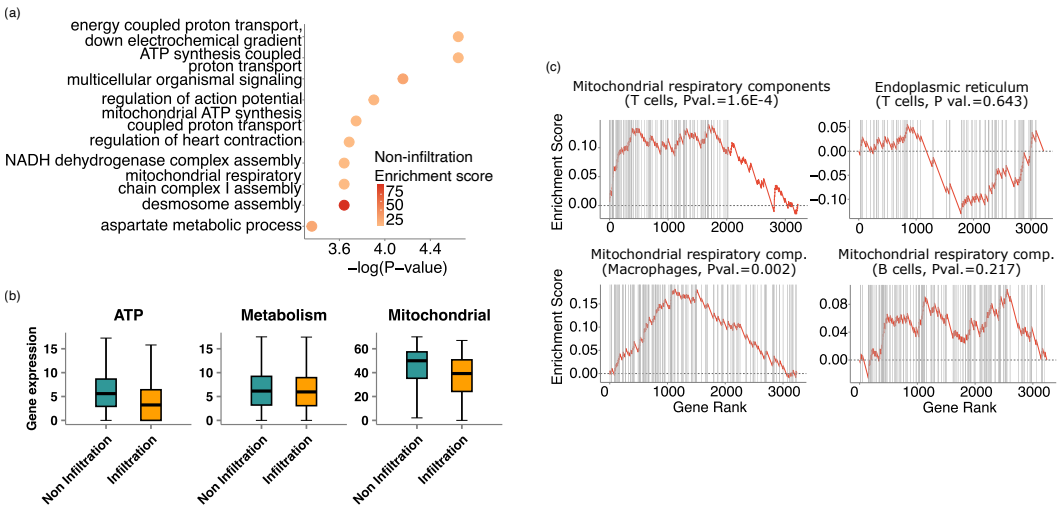

Figure 10: Pathway associations in the myocarditis case study. (a) GO enrichment for VCM genes ranked toward T-cell non-infiltration. (b) ATP-, metabolism-, and mitochondria-related expression in infiltrated and non-infiltrated regions. (c) Preranked GSEA for the T-cell infiltration target using mitochondrial respiratory and endoplasmic-reticulum gene sets, followed by mitochondrial respiratory analyses for macrophage and B-cell infiltration targets.

K.4 Enrichment Controls

Preranked GSEA (Subramanian et al., 2005) uses the full gene ordering rather than a selected top-gene subset. A mitochondrial respiratory set coupled to the TCA cycle is enriched at the non-infiltration end of the T-cell analysis (nominal $P = 1.63 \times 10^{-4}$; BH-adjusted $P = 7.06 \times 10^{-4}$), whereas the endoplasmic-reticulum control gives $P = 0.643$. The corresponding mitochondrial respiratory analyses give $P = 0.00237$ for macrophages and $P = 0.217$ for B cells (Figure 10(c)). Figure 11 reports the additional mitochondrial outer-membrane and nucleoid controls for all three infiltration targets.

Together, the analyses associate the learned ordering with particular mitochondrial programs and show different enrichment patterns across the three cell-lineage targets. They do not establish that the ranked genes cause immune-cell exclusion or myocardial tissue damage.

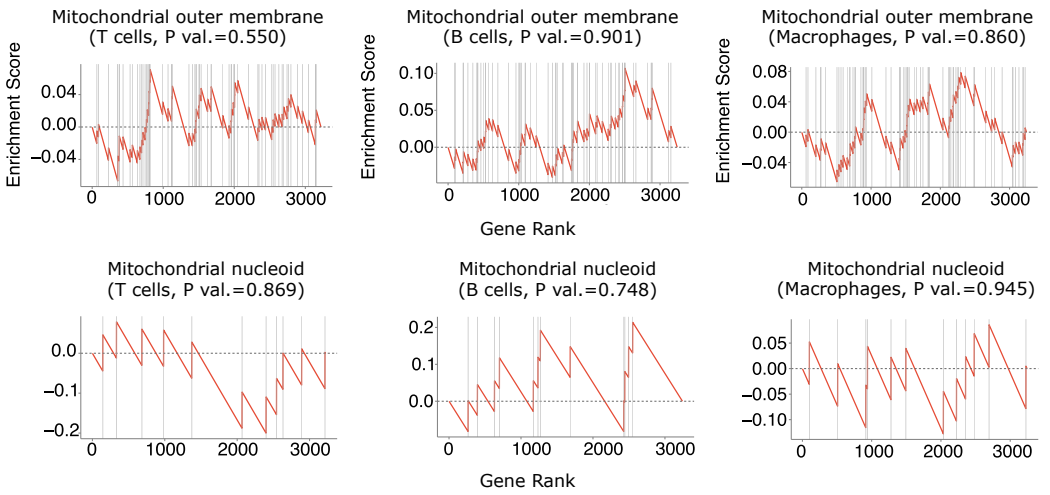

Figure 11: Additional GSEA controls for myocarditis. Columns correspond to T-cell, B-cell, and macrophage infiltration targets; rows test mitochondrial outer-membrane and nucleoid gene sets. Each panel shows its enrichment curve and reported P value.

L Biological Evaluation with Xenium and TCR Sequencing

Objective and scale. Human tumors contain both tumor-reactive and bystander T cells (Simoni et al., 2018), a distinction relevant to antitumor immunity (Duhen et al., 2018). We evaluate three human tumors profiled with Xenium and TCR sequencing, comprising approximately 670,000 cells in total. This number counts all profiled cells, not T cells or independent tissue samples. The experiment asks whether models trained on T-cell presence preferentially recover an independently defined reactive subgroup. It provides a biological evaluation of predictions, without an exact ground-truth gene ranking.

Annotations and supervision. Cell-type annotation determines the binary training target: $y_i = 1$ for an annotated focal T cell and $y_i = 0$ for a non-T focal cell. TCR clonotype information separately assigns annotated T cells to operationally defined reactive and bystander groups. These assignments are fixed independently of model predictions and are reserved for evaluation; models are not trained to classify reactive versus bystander T cells.

Model inputs. Each bag contains surrounding tumor cells, represented by gene-expression features and physical distances to the focal cell. The model maps these neighboring instances to the focal-cell T-cell-presence label. The focal T cell and its TCR information are not inputs. Genes and clonotype probes used to define T-cell identity or reactivity are excluded from every model's feature set. Thus, the models share the same prediction task and input restrictions, while TCR status stratifies their test-set outcomes.

Metrics. Let $\widehat{y}_i$ be a thresholded test-set prediction and define the sets of true-positive and false-negative T cells as

$$\mathcal{T}_{\mathrm{TP}} = \{i : y_i = 1, \widehat{y}_i = 1\}, \qquad \mathcal{T}_{\mathrm{FN}} = \{i : y_i = 1, \widehat{y}_i = 0\}.$$

Let $\mathcal{R}$ and $\mathcal{B}$ denote the preassigned TCR-defined reactive and bystander groups. Alongside AUROC for T-cell presence, we evaluate

$$q_{\mathrm{reactive}|\mathrm{TP}} = \frac{|\mathcal{T}_{\mathrm{TP}} \cap \mathcal{R}|}{|\mathcal{T}_{\mathrm{TP}}|}, \qquad q_{\mathrm{bystander}|\mathrm{FN}} = \frac{|\mathcal{T}_{\mathrm{FN}} \cap \mathcal{B}|}{|\mathcal{T}_{\mathrm{FN}}|}.$$

The first fraction describes the composition of recovered T cells and uses all true-positive T cells as its denominator. The second describes missed T cells and uses all false-negative T cells as its denominator. Neither denominator is the full set of predicted positives or the full bystander population; in particular, $q_{\mathrm{bystander}|\mathrm{FN}}$ is not a bystander-specific rejection rate. A fraction is undefined when its denominator is zero.

Paired comparison. Models are evaluated at a matched predicted-positive rate within each tumor. Figure 8(a–c) reports mean within-tumor differences across the three tumors. For the original five baselines, the direction of the differences is also available for all 15 tumor-by-baseline comparisons. The additional DW-logistic comparison provides mean differences only: 0.0231 in AUROC, 0.103 in $q_{\mathrm{reactive|TP}}$, and 0.092 in $q_{\mathrm{bystander|FN}}$. Its tumor-level differences are unavailable, so the 15-comparison consistency statement does not extend to 18 comparisons. Pairing compares models in the same tumor context; the large number of cells does not create additional independent tumor-level replicates.

Across the original five baselines, mean paired AUROC gains range from 0.022 to 0.063. The corresponding mean increases in reactive fractions among TP T cells and bystander fractions among FN T cells are 11.2–18.2 and 9.7–14.4 percentage points, respectively. Each baseline-level mean averages the within-tumor differences across three tumors; the DW-logistic means are reported separately above.

Biological composition. LION improves AUROC and both composition measures in all 15 comparisons with the original five baselines and exceeds DW-logistic in the reported means. Higher reactive fractions among recovered T cells, together with higher bystander fractions among missed T cells, support preferential recovery of the operationally defined reactive group. This is biological specificity within the T-cell-presence task, not performance of a reactive-versus-bystander classifier. TCR-defined reactivity is not a direct functional assay of tumor recognition, and three retrospective tumors do not establish clinical utility or a gold-standard gene ranking.

Component comparison. Removing the influence vector yields higher AUROC but less favorable biological composition than removing the distance kernel: the AUROC difference is $+0.021$, while the reactive fraction among TP changes by -0.090 and the bystander fraction among FN by -0.082. All three differences have the same direction in every tumor. This comparison is between the two ablated models, not between either ablation and full LION. It shows that better T-cell-presence discrimination need not improve the TCR-defined composition of recovered and missed cells.

NetMHC gene-set construction. We intersect the gene lists from LION, Additive MIL, SI-MIL, and DW-logistic, retaining genes with available peptide totals. Each model orders this shared gene set by descending signed score. Duplicate gene entries are collapsed by retaining the score with the largest absolute magnitude.

From the supplied NetMHC table, we retain peptide records whose rank-field value is at most 0.05 and deduplicate them by protein identifier and peptide sequence. Protein identifiers are mapped to genes, and retained peptide counts are summed over proteins assigned to each gene to obtain B_g. The denominator T_g is the maximum peptide-total entry recorded for that gene. A gene with a known denominator and no retained binder record receives $B_g = 0$.

Ranking summary. For each gene, we compute $h_g = \log[(B_g + 1)/(T_g + 1)]$. Figure 8(d) plots the median h_g among the top 1, 2, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100% of the shared gene set, omitting cutoffs that contain fewer than five genes. The sets are nested, so points along each curve are dependent. We therefore report neither Kendall-test p values nor a calibrated between-model significance test. Such inference would require a null procedure preserving the dependence, for example recomputing each full curve after permuting the gene ranking. The measured quantity concerns predicted HLA binding; experimental antigen presentation and T-cell recognition remain untested by this analysis.

M Distance Parity and the Information-Budget Ablation

The distance-access ablation measures how much scalar distance contributes to bag prediction. With the architecture and other inputs held fixed, concatenating distance increases mean AUROC by $+0.062$ for AttentionMIL (7/7 benchmarks), $+0.057$ for DeepSets (7/7), and $+0.041$ for the GNN (6/7). Every architecture benefits on average. Accordingly, all baselines in the main comparison receive the same scalar distances as LION.

`AttentionMIL+dist` is architecturally identical to DistAttMIL and produces numerically identical outputs on all seven benchmarks; the two names therefore refer to one baseline entry. With distance supplied, LION still exceeds the strongest deep baseline's absolute permutation readout in signed-target Spearman by 0.026 to 0.362 across the seven benchmarks (Figure 4(c)). This comparison matches the available inputs but compares different ranking quantities: native coefficients and perturbation magnitudes. Figure 3 provides the separate directional coefficient comparison, and Appendix T reports the corresponding ablation of LION's distance channel.

N Full Prediction Metrics and Cross-Validation

Ranking convention. Spearman ρ is evaluated against the signed ground-truth coefficient ordering. When the truth contains ties, the reported ceilings assume strictly ordered estimates and average ranks for the tied truth. These ceilings describe that evaluation convention; an estimator that reproduces the true ties can attain $\rho = 1$.

Figure 4(a,b) compares LION with the strongest distance-equipped deep baseline for each prediction metric and benchmark. Table 5 gives all deep-baseline and LION results under the main paired 5-fold cross-validation $\times$ 5-seed protocol. AUROC and AUPRC summarize the scores across thresholds; accuracy and macro-F1 use the fixed threshold 0.5.

The evaluation loader pairs one positive bag with one negative bag, so the evaluated positive fraction is 0.50 on every benchmark, regardless of the generator's raw positive rate. The corresponding AUPRC reference level is therefore 0.50. A nearly all-positive classifier can attain a relatively high positive-class F_1 despite poor discrimination. We report fixed-threshold accuracy and macro-F1 to make this behavior visible rather than selecting a separate threshold for each model. LION's F_1-maximizing thresholds lie in 0.358–0.438 across benchmarks, whereas those of the weakest baselines lie in 0.000–0.024.

Table 6 gives the mean ranking correlations and baseline identities summarized in Figure 4(c). Both native coefficients and absolute permutation importance are evaluated against the signed coefficient target; their difference therefore combines ranking performance with a difference in the quantities being ranked. These results use the main cross-validation protocol. The direct DW-logistic coefficient comparison in Figure 3 is a separate evaluation. Reported differences Δ retain the source precision and may differ from subtracting the displayed rounded means. Spearman standard deviations are unavailable in this summary.

Cross-validation versus a single split. The reported model-by-benchmark comparisons remain within one reported standard deviation of the earlier single-split results. The largest AUROC discrepancy is -0.029 for DeepSets+dist on H&E multi-type; LION's largest discrepancy is -0.014 on spatial transcriptomics. The broad performance pattern is retained under cross-validation, although this agreement does not establish invariance to the choice of split.

O Hyperparameter Oracle Diagnostic

The oracle diagnostic tests whether hyperparameter selection using the evaluation target can reverse the ranking comparison. On the three benchmarks with a complete grid search, each baseline may select the configuration that maximizes ground-truth Spearman, while LION uses one fixed configuration and its native readout. This gives the baselines access to the target against which they are evaluated and is not a deployable selection procedure. None exceeds LION in Table 7.

The table separates prediction-oriented and ranking-oriented selection. "Oracle AUROC" is the best AUROC in the searched grid, and the next column gives Spearman at that configuration; "Oracle ρ" selects the configuration by ground-truth Spearman. Values use the inner search split and two seeds. They are comparable across models within this diagnostic and are not pooled with the main cross-validation results.

Table 5: Prediction metrics under paired 5-fold CV $\times$ 5 seeds (mean $\pm$ standard deviation). Accuracy and macro-F1 use threshold 0.5; the evaluated positive fraction and AUPRC reference level are 0.50. Figure 4(a,b) summarizes AUROC and AUPRC.

Model	AUROC	AUPRC	Accuracy	macro-F1
Spatial-MNIST Sparse				
LION	$\mathbf{0.8790 \pm 0.0115}$	$\mathbf{0.8826 \pm 0.0128}$	$\mathbf{0.794 \pm 0.013}$	$\mathbf{0.794 \pm 0.013}$
AttentionMIL+dist	0.6720 ± 0.0197	0.6609 ± 0.0192	0.625 ± 0.019	0.623 ± 0.019
DeepSets+dist	0.6766 ± 0.0209	0.6661 ± 0.0244	0.625 ± 0.019	0.623 ± 0.021
GNN+dist	0.5987 ± 0.0269	0.5803 ± 0.0247	0.570 ± 0.022	0.569 ± 0.021
Spatial-MNIST Correlated				
LION	$\mathbf{0.8857 \pm 0.0110}$	$\mathbf{0.8882 \pm 0.0126}$	$\mathbf{0.801 \pm 0.013}$	$\mathbf{0.801 \pm 0.013}$
AttentionMIL+dist	0.6746 ± 0.0198	0.6604 ± 0.0256	0.625 ± 0.017	0.623 ± 0.017
DeepSets+dist	0.6821 ± 0.0186	0.6769 ± 0.0171	0.627 ± 0.016	0.625 ± 0.016
GNN+dist	0.6113 ± 0.0141	0.5957 ± 0.0165	0.579 ± 0.014	0.578 ± 0.014
Spatial-MNIST Dense				
LION	$\mathbf{0.8368 \pm 0.0177}$	$\mathbf{0.8370 \pm 0.0184}$	$\mathbf{0.753 \pm 0.018}$	$\mathbf{0.752 \pm 0.018}$
AttentionMIL+dist	0.5831 ± 0.0203	0.5748 ± 0.0226	0.559 ± 0.016	0.557 ± 0.016
DeepSets+dist	0.6007 ± 0.0178	0.5925 ± 0.0213	0.569 ± 0.016	0.566 ± 0.018
GNN+dist	0.5576 ± 0.0213	0.5419 ± 0.0222	0.543 ± 0.017	0.542 ± 0.017
Spatial transcriptomics				
LION	$\mathbf{0.8255 \pm 0.0465}$	$\mathbf{0.8142 \pm 0.0572}$	$\mathbf{0.751 \pm 0.047}$	$\mathbf{0.750 \pm 0.047}$
AttentionMIL+dist	0.7936 ± 0.0451	0.7798 ± 0.0554	0.724 ± 0.046	0.723 ± 0.046
DeepSets+dist	0.8012 ± 0.0439	0.7920 ± 0.0515	0.731 ± 0.044	0.730 ± 0.045
GNN+dist	0.8075 ± 0.0412	0.8011 ± 0.0468	0.738 ± 0.043	0.737 ± 0.043
H&E immune-proximity				
LION	$\mathbf{0.9104 \pm 0.0145}$	$\mathbf{0.9127 \pm 0.0191}$	$\mathbf{0.822 \pm 0.019}$	$\mathbf{0.821 \pm 0.019}$
AttentionMIL+dist	0.8812 ± 0.0153	0.8801 ± 0.0204	0.796 ± 0.020	0.795 ± 0.020
DeepSets+dist	0.8886 ± 0.0148	0.8897 ± 0.0193	0.805 ± 0.019	0.804 ± 0.019
GNN+dist	0.8944 ± 0.0139	0.8992 ± 0.0181	0.812 ± 0.019	0.811 ± 0.019
H&E morphology-interaction				
LION	$\mathbf{0.7820 \pm 0.0229}$	$\mathbf{0.8018 \pm 0.0197}$	$\mathbf{0.713 \pm 0.026}$	$\mathbf{0.709 \pm 0.027}$
AttentionMIL+dist	0.7471 ± 0.0224	0.7405 ± 0.0273	0.688 ± 0.024	0.687 ± 0.024
DeepSets+dist	0.7543 ± 0.0257	0.7748 ± 0.0248	0.696 ± 0.023	0.694 ± 0.024
GNN+dist	0.7610 ± 0.0249	0.7862 ± 0.0261	0.704 ± 0.023	0.702 ± 0.024
H&E multi-type				
LION	$\mathbf{0.8977 \pm 0.0179}$	$\mathbf{0.9023 \pm 0.0189}$	$\mathbf{0.810 \pm 0.020}$	$\mathbf{0.808 \pm 0.021}$
AttentionMIL+dist	0.8712 ± 0.0180	0.8691 ± 0.0190	0.784 ± 0.022	0.783 ± 0.022
DeepSets+dist	0.8784 ± 0.0254	0.8810 ± 0.0261	0.795 ± 0.024	0.793 ± 0.025
GNN+dist	0.8837 ± 0.0155	0.8882 ± 0.0164	0.802 ± 0.019	0.800 ± 0.019

Table 6: Signed-target ranking under paired 5-fold CV $\times$ 5 seeds (mean Spearman ρ). LION uses native coefficients; the comparator is the strongest distance-equipped deep baseline using absolute permutation importance on each benchmark. Δ is LION minus baseline.

Benchmark	LION	Perm. baseline	Baseline model	Δ
Spatial-MNIST Sparse	0.648	0.621	DeepSets+dist	+0.027
Spatial-MNIST Correlated	0.588	0.561	AttentionMIL+dist	+0.027
Spatial-MNIST Dense	0.863	0.501	DeepSets+dist	+0.362
Spatial transcriptomics	0.577	0.552	GNN+dist	+0.026
H&E immune-proximity	0.590	0.561	GNN+dist	+0.029
H&E morphology-interaction	0.620	0.383	GNN+dist	+0.237
H&E multi-type	0.698	0.536	GNN+dist	+0.162

Table 7: Hyperparameter-oracle diagnostic on the inner search split (two seeds). "Oracle AUROC" and "Oracle ρ" select configurations by AUROC and ground-truth Spearman, respectively; the middle column reports Spearman at the AUROC-selected configuration. LION uses one fixed configuration and native readout.

Benchmark	Model	Oracle AUROC	ρ at that config	Oracle ρ
Spatial-MNIST Dense	**LION**	**0.8470**	0.8515	**0.8515**
	AttentionMIL+dist	0.6007	0.3347	0.4003
	DeepSets+dist	0.6127	0.5033	0.5634
	GNN+dist	0.5860	−0.0653	0.2964
H&E immune-proximity	**LION**	**0.9240**	0.5901	**0.5901**
	AttentionMIL+dist	0.8952	0.5105	0.5521
	DeepSets+dist	0.9027	0.5362	0.5576
	GNN+dist	0.9080	0.5624	0.5688
Spatial transcriptomics	**LION**	**0.8418**	0.5774	**0.5774**
	AttentionMIL+dist	0.8154	0.5421	0.5536
	DeepSets+dist	0.8238	0.5487	0.5562
	GNN+dist	0.8069	0.5510	0.5581

P Coefficient Ranking and Attribution Magnitude

Direct coefficient baseline. For a prespecified feature representation $\widetilde{x}_{ij,g}$, define

$$t_{ig}(\lambda) = \sum_{j=1}^{n_i} \frac{\exp(-\lambda d_{ij})}{\sum_{k=1}^{n_i} \exp(-\lambda d_{ik})} \widetilde{x}_{ij,g}, \qquad p_i = \sigma\left(\beta + \sum_{g=1}^{G} c_g t_{ig}(\lambda)\right), \qquad c_g \in \mathbb{R}.$$

Distance-weighted logistic regression (DW-logistic) ranks the unrestricted coefficients c_g directly, retaining their signs. Mean-pool logistic regression is the $\lambda = 0$ special case; no separate result for that case is reported here. This family provides a directional coefficient comparator, but its coefficients need not have the same interpretation as LION's scores: LION normalizes features nonlinearly, and DW-logistic can be misspecified for LION-form generators or the morphology-interaction rule.

Protocol and units. A matched coefficient comparison requires common outer folds, seeds, instance features, and scalar distances. Distance decay and regularization must be selected within the training split using a prespecified validation loss or AUROC criterion, without access to test rankings. Feature scales must be fixed in advance or estimated from training data only. If $\widetilde{x}_g = (x_g - \mu_g)/q_g$, the corresponding coefficient in the original units is c_g/q_g; the fitted coefficients and ground truth must be compared in the same units. An exact compositional constraint additionally requires an explicit convention for identifying coefficient contrasts.

Benchmark-level comparison. In the direct coefficient comparison (Figure 3), LION and DW-logistic tie at the reported precision on spatial transcriptomics and H&E immune-proximity. LION's mean Spearman advantage on Sparse, Correlated, and multi-type is 0.01–0.04. These small differences are descriptive because paired uncertainty estimates are unavailable. This additional comparison is reported separately from the main cross-validation results in Figure 4.

Dense upper-coefficient diagnostic. The primary Dense metric uses all 60 features. The auxiliary metric uses

$$\mathcal{G}_{\text{upper}} = \text{Top15}\{S_g^*\}_{g=1}^{60}, \qquad S_g^* = \sigma(s_g^*),$$

and evaluates Spearman between each model's scores and the ground truth restricted to that set. The selection rule uses generator coefficients, not model predictions, and specifies a shared set for all models, folds, and training seeds of a generated dataset. The specified rule includes boundary ties, so the set may contain more than 15 features. The specified aggregation is over 25 fold–seed evaluations; Figure 3 reports means only. Standard deviations and the realized set size are not available in the supplied summary. Spearman is undefined if either restricted vector is constant.

The reported means are 0.75 for LION and 0.72 for DW-logistic. This subset selects high coefficients, not a positive-effect group: sigmoid preserves ordering, but the sign of s_g^* alone does not define a scientific effect. In particular, $s_g^* < 0$ denotes a negative effect only if the generator specifies $s^* = 0$ as a neutral reference. The smaller gap on this subset does not identify the source of the full-ranking permutation gap; that would require permutation readouts on the same subset and runs.

Magnitude and active-feature targets. A coefficient-magnitude analysis requires a reference level and answers a different question from signed ordering. If the generator supplies a null set $\mathcal{G}_0$, an anchored ground-truth magnitude and its corresponding LION score can be defined as

$$m_g^* = \left| S_g^* - \frac{1}{|\mathcal{G}_0|} \sum_{h \in \mathcal{G}_0} S_h^* \right|, \qquad \widehat{m}_g = \left| \widehat{S}_g - \frac{1}{|\mathcal{G}_0|} \sum_{h \in \mathcal{G}_0} \widehat{S}_h \right|.$$

A positive affine transformation of $\widehat{\mathbf{S}}$ scales every $\widehat{m}_g$ by the same positive factor and therefore preserves its ordering. Since sigmoid scores are positive, ranking $|\widehat{S}_g|$ is identical to ranking $\widehat{S}_g$; neither measures coefficient magnitude relative to a specified null. The generator-supplied null set would serve only as an evaluation reference, without entering training. An active-feature AUROC analysis could compare these anchored LION scores with nonnegative permutation importance using the generator's active/null labels. The two scores would still estimate different quantities because perturbation magnitude also depends on feature variation and distances. We report neither coefficient-strength nor active-feature AUROC results here, and do not impose a null reference on benchmarks that lack one.

Signed attribution is not signed coefficient estimation. For a chosen output F (probability or logit), reference X_i^0, and fixed distances d_i, signed IG has the form

$$A_{ijg}^{\mathrm{IG}} = (x_{ijg} - x_{ijg}^0) \int_0^1 \frac{\partial F(X_i^0 + u(X_i - X_i^0), d_i)}{\partial x_{ijg}} \, du.$$

A global sum readout is $I^{-1} \sum_i \sum_j A_{ijg}$, whereas a bag-mean readout is $I^{-1} \sum_i n_i^{-1} \sum_j A_{ijg}$. Both retain signs, but they weight bags differently when bag sizes vary and depend on the reference and feature values. For SHAP, the background distribution and masking unit define the contribution being estimated. Instance-level values require the same sum-versus-mean distinction; bag-level values ϕ_{ig} can instead be aggregated as $I^{-1} \sum_i \phi_{ig}$. SI-MIL's plain coefficients likewise require specified feature units and cross-bag aggregation and must be distinguished from $|\beta| \, \mathrm{std}$ and permutation readouts. Retaining signs alone therefore does not establish consistency for a signed coefficient ranking.

The available readout summaries do not fully specify the reference, output scale, masking, sign treatment, and aggregation. Those choices cannot be inferred from the reported correlations. Accordingly, Table 8 supports a readout-sensitivity analysis but not a fully specified comparison of signed coefficient estimators.

Q Attribution Readouts

This diagnostic measures how the reported global ranking changes with the readout applied to a fixed trained model. It uses five seeds on a single $80/20$ split, separately from the main cross-validation protocol (Appendix G). AttentionMIL and DeepSets are distance-blind in this diagnostic; only AttentionMIL+dist receives the distance channel. The results therefore do not establish a distance-matched advantage across all models. The corresponding distance-equipped comparison is Figure 4(c).

Table 8 reports all evaluated readouts on three benchmarks. On Spatial-MNIST Dense, DeepSets permutation importance has the highest baseline mean Spearman, 0.518 ± 0.070, compared with 0.847 ± 0.045 for LION's native scores. KernelSHAP reaches -0.038 for DeepSets and 0.081 for AttentionMIL+dist. These correlations describe agreement with the signed coefficient target; they do not establish that the attribution methods fail to explain their own prediction functions. As detailed in Appendix P, the available IG/SHAP summaries do not fully specify the reference, scale, masking,

Table 8: Readout sensitivity on a fixed trained model: Spearman ρ against the signed coefficient target (mean $\pm$ standard deviation over five seeds on a single $80/20$ split). †Distance-blind model. "IG mean" denotes mean-aggregated Integrated Gradients; LION uses its native coefficient ordering.

Spatial-MNIST Dense				
Readout	LION	AttentionMIL+dist	AttentionMIL†	DeepSets†
Permutation	N/A	0.313 ± 0.068	0.289 ± 0.104	0.518 ± 0.070
IG	N/A	0.299 ± 0.093	0.316 ± 0.094	0.150 ± 0.103
IG mean	N/A	0.336 ± 0.074	0.302 ± 0.111	0.171 ± 0.095
KernelSHAP	N/A	0.081 ± 0.130	0.119 ± 0.115	-0.038 ± 0.182
Native	0.847 ± 0.045	N/A	N/A	N/A
Spatial transcriptomics				
Readout	LION	AttentionMIL+dist	AttentionMIL†	DeepSets†
Permutation	N/A	0.552 ± 0.019	-0.016 ± 0.414	-0.049 ± 0.358
IG	N/A	0.538 ± 0.026	0.082 ± 0.606	-0.313 ± 0.300
IG mean	N/A	0.542 ± 0.024	0.082 ± 0.606	-0.313 ± 0.300
KernelSHAP	N/A	0.530 ± 0.031	0.082 ± 0.606	-0.247 ± 0.165
Native	0.577 ± 0.000	N/A	N/A	N/A
H&E immune-proximity				
Readout	LION	AttentionMIL+dist	AttentionMIL†	DeepSets†
Permutation	N/A	0.517 ± 0.041	-0.118 ± 0.123	-0.118 ± 0.114
IG	N/A	0.499 ± 0.079	0.091 ± 0.064	-0.091 ± 0.096
IG mean	N/A	0.499 ± 0.064	0.045 ± 0.102	-0.136 ± 0.085
KernelSHAP	N/A	0.536 ± 0.020	0.281 ± 0.117	-0.054 ± 0.165
Native	0.590 ± 0.000	N/A	N/A	N/A

and aggregation needed to identify their estimands. These empirical comparisons are also distinct from Theorem 2, which analyzes within-bag permutation in an identity-feature linear model.

Readout choice changes the ranking even when the trained model is held fixed. On Spatial-MNIST Dense, the range of mean Spearman across the four readouts is 0.20 for AttentionMIL, 0.26 for AttentionMIL+dist, and 0.56 for DeepSets. On H&E immune-proximity, the range is 0.40 for AttentionMIL. These differences reflect choices in how predictions are summarized as feature importance, rather than changes in the fitted predictor.

SI-MIL shows the same dependence. Table 9 compares three global readouts of each trained model: $|\beta| \cdot \mathrm{std}$, plain β, and permutation importance. Only plain β is potentially aligned with a directional coefficient target, subject to feature units and cross-bag aggregation. On the benchmarks with more finely resolved ranking targets, the spread is 0.34–0.46 and remains substantial after tuning. Small spreads on coarsely tied targets must be interpreted in light of their limited metric resolution. Readout disagreement changes the reported feature ordering, but does not by itself prove statistical nonidentifiability. LION specifies one native ordering, whose identification is conditional on Theorem 1.

R Interpretable MIL Baselines

SI-MIL (Kapse et al., 2024) and Additive MIL (Javed et al., 2022) provide a per-slide linear surrogate and per-instance additive contributions, respectively. Neither natively returns the shared ordering of named features studied here, so we evaluate global readouts derived from their formulations. Each receives the common 18-configuration grid, with an additional sub-sweep over SI-MIL's K and λ.

Ranking and prediction. Under a joint hyperparameter-and-readout oracle, SI-MIL selects both choices using the ground-truth ranking against which it is scored. It does not exceed LION on any benchmark: it ties on the two saturated targets and reaches 0.7458 on Spatial-MNIST Dense, compared with 0.8471 for LION. Prediction gives a different comparison. Under Additive MIL's

Table 9: SI-MIL readout spread before and after tuning (five seeds). Spread is the maximum minus minimum Spearman across $|\beta| \cdot \mathrm{std}$, plain β, and permutation readouts of the same trained model. N/A marks an unreported entry.

Benchmark	Spread (untuned)	Spread (tuned)
Spatial-MNIST Dense	0.449	**0.457**
H&E multi-type	N/A	**0.428**
H&E morphology-interaction	0.456	**0.338**
H&E immune-proximity	0.145	0.136
Spatial-MNIST Sparse	N/A	0.031
Spatial-MNIST Correlated	N/A	0.003
Spatial transcriptomics	0.000	0.000

separate protocol, its mean-pooled variant has higher AUROC than LION on H&E immune-proximity (0.9131 versus 0.9096) and spatial transcriptomics (0.8614 versus 0.8520). Across the evaluated deep and interpretable MIL prediction baselines, LION ranks first in mean AUROC on 5/7 benchmarks. This statement excludes the additional DW-logistic coefficient comparison and does not pool values across protocols.

Neighborhood-size dependence. This diagnostic varies bag size while holding the generating rule fixed. SI-MIL's intrinsic β readout attenuates with neighborhood size, with fitted exponent $+0.27$, within the $+0.21$ to $+0.39$ range of the post-hoc readouts; LION's fitted exponent is -0.00. Intrinsic explanations therefore remain sensitive to bag size in this setting.

The pooling comparison shows a related predictive effect. As mean bag size increases from 27.7 to 144.4, sum-pooled Additive MIL reaches AUROC 0.500 in an increasing number of seeds, from 0/5 to 5/5. Mean pooling removes this collapse in the reported runs, and all prediction comparisons use that variant. The diagnostic thus identifies a dependence on pooling as well as on attribution, without treating the sum-pooled failure as a result for the evaluated mean-pooled model.

S Distance Response of the Baselines

Figure 6 evaluates the distance-response probe defined in Appendix I: one target-type cell is moved across $d \in [0, 100]\,\mu\mathrm{m}$ in a bag with 20 fixed background cells. The measured LION response is its normalized distance weight multiplied by its instance score, rather than the unnormalized kernel alone. Its correlation with the ground-truth spatial decay rule is $r = 0.943 \pm 0.027$ across retrains. Because weights are normalized within each bag, this effective response depends on the background configuration.

LION separately exposes a standalone exponential kernel, $k(d) = \exp(-e^a d)$, as does DW-logistic, $k_\lambda(d) = \exp(-\lambda d)$. These kernels make the distance component explicit, but their normalized weights still depend on other distances in the bag. Distance recovery for DW-logistic is well-defined but has not been reported in this diagnostic. Distance-blind DeepSets and AttentionMIL receive no distance input; their distance-equipped variants instead incorporate distance through nonlinear encoders and do not expose an analogous standalone kernel.

For distance-aware attention, the reported probe similarly sweeps a target cell over $d \in [0, 100]\,\mu\mathrm{m}$ against 20 fixed background cells and records its attention weight multiplied by its linear contribution to the bag logit. This response decays smoothly, remains nonzero up to $100\,\mu\mathrm{m}$, and becomes slightly negative beyond the generator's hard $30\,\mu\mathrm{m}$ boundary. Its correlation with the ground-truth curve is $r = 0.842 \pm 0.115$; the reported standard deviation is approximately four times LION's. Both models' evaluated responses depend on the fixed background, and LION's exponential kernel has no hard cutoff either. These results characterize the specified effective-influence probes, not a comparison of two standalone learned kernels. The deep encoders lack a native standalone kernel; DW-logistic has an explicit kernel, but its distance-recovery result is unreported here.

T ABLATION STUDY: WHICH ARCHITECTURAL CONSTRAINTS MATTER?

We compare the default LION configuration with six architectural variants on Spatial-MNIST Dense ($G = 60$), H&E immune-proximity, and H&E morphology-interaction. Table 10 reports mean $\pm$ standard deviation over three seeds; Figure 7 shows the Dense and morphology-interaction results. This ablation protocol is separate from the main cross-validation protocol. Each variant changes one component except LION-minimal, which removes both feature normalization and distance structure.

LION-tanh replaces sigmoid by tanh for the influence scores, giving $S_g \in (-1, 1)$. Because the normalized feature vectors and distance weights each sum to one, the shift and positive scale relating these score ranges can be absorbed by the final intercept and positive scale. With both normalizers retained and that final scale and intercept free, the tanh and sigmoid versions define the same unpenalized prediction family and preserve the feature ordering. This variant tests a different parameterization; differences after finite-epoch AdamW training can reflect optimization or regularization rather than added expressivity. LION-no-feat-norm passes raw features in place of the softmax-normalized features while retaining the sigmoid constraint on scores. It is therefore distinct from unrestricted DW-logistic: equivalence would also require matching coefficient freedom, feature units, distance treatment, regularization, and the fitting protocol (Figure 3). LION-RBF-dist replaces the exponential kernel $\exp(-e^a \cdot d)$ with the Gaussian kernel $\exp\left(-e^a \cdot d^2\right)$. LION-sum-pool uses unnormalized exponential weights and sums instance contributions. LION-unnorm-dist removes the distance softmax normalization. LION-minimal removes both feature normalization and distance structure, using identity mappings in those components.

Feature normalization. Removing the feature normalizer reduces Dense Spearman from 0.844 to 0.717, approximately a 15% relative decrease, and lowers AUROC by approximately 16 percentage points. This comparison shows sensitivity to normalization within LION's constrained parameterization.

Tied targets and metric resolution. The immune-proximity rule assigns equal positive coefficients to 2 of its 15 features and zero to the remaining 13. With strictly ordered predictions and average ranks for the tied truth, Spearman has 27 attainable values and a maximum of 0.5901. The maximum is attained exactly when the two active features occupy the top two predicted ranks, as confirmed by enumerating all $\binom{15}{2}$ placements. Every variant meets this condition in every seed, yielding standard deviation 0, although AUROC ranges from 0.746 to 0.927. The constant column therefore reflects saturation of this ranking metric. Dense, with a signed target over 60 features, distinguishes many more orderings.

To reproduce the enumeration, let $r_1, r_2 \in \{1, \ldots, 15\}$ be the predicted ranks of the two tied high-coefficient features, with rank 15 largest. Their ground-truth average rank is 14.5, and each null feature has average rank 7. The integer numerator $15(r_1 + r_2 - 16)$ takes 27 values over $\binom{15}{2} = 105$ placements. For morphology-interaction, let r_m be the distinct predicted rank of the intermediate feature. The three ground-truth average ranks are 14.5, 13, and 6.5, giving numerator $16(r_1 + r_2 - 16) + 13(r_m - 8)$. Enumerating the $\binom{15}{2} \times 13 = 1365$ placements yields 397 distinct values. Each numerator is proportional to Spearman with a fixed positive denominator, so integer arithmetic avoids ambiguity from floating-point deduplication.

The morphology-interaction rule adds an intermediate coefficient to the H&E target: two cell types have $s = 5.0$, `area` has $s = 3.0$, and twelve features have zero coefficients. Under strictly ordered predictions, Spearman then has 397 attainable values and a ceiling of 0.6982, compared with 27 values and 0.5901 for immune-proximity. These ceilings use average ranks for tied truth. Allowing ties in the predictions changes the attainable set and permits $\rho = 1$ when the two rank vectors agree.

The seven variants separate on this target (Table 10, morphology-interaction panel), with mean Spearman ranging from 0.332 to 0.676. The five variants that retain feature normalization lie between 0.565 and 0.676; LION-no-feat-norm and LION-minimal reach 0.332 and 0.366. These results support feature normalization for mean ranking recovery under this rule. All five retained-normalizer means remain below the 0.6982 ceiling, so retaining normalization does not ensure the correct ordering of the three active features above the inactive features in every seed. The aggregate correlations neither determine each feature's position in an individual run nor establish statistical equivalence among the retained-normalizer variants.

Table 10: Architectural ablations under the three-seed protocol (mean ± standard deviation where available). LION-minimal removes feature normalization and distance structure; other variants modify one component. Bold numerical entries mark the largest mean in each nonsaturated metric column; underlining marks variants without feature normalization. Figure 7 plots the Dense and morphology-interaction results. The immune-proximity ranking metric is saturated under strictly ordered predictions; its tie convention is explained below.

Spatial-MNIST Dense Variant	AUROC	Spearman ρ
LION (default)	0.838 ± 0.015	0.844 ± 0.005
LION-tanh	$\mathbf{0.846 \pm 0.012}$	0.842 ± 0.004
LION-no-feat-norm	0.682 ± 0.029	$\underline{0.717 \pm 0.055}$
LION-sum-pool	0.636 ± 0.009	0.823 ± 0.011
LION-RBF-dist	0.832 ± 0.013	$\mathbf{0.845 \pm 0.001}$
LION-unnorm-dist	0.838 ± 0.015	0.844 ± 0.005
LION-minimal	0.585 ± 0.009	$\underline{0.277 \pm 0.120}$
H&E immune-proximity **Variant**	**AUROC**	**Spearman ρ**
LION (default)	0.894 ± 0.036	0.590
LION-tanh	0.918 ± 0.033	0.590
LION-no-feat-norm	$\mathbf{0.927 \pm 0.028}$	0.590
LION-sum-pool	0.910 ± 0.025	0.590
LION-RBF-dist	0.821 ± 0.107	0.590
LION-unnorm-dist	0.749 ± 0.055	0.590
LION-minimal	0.746 ± 0.024	0.590
H&E morphology-interaction **Variant**	**AUROC**	**Spearman ρ**
LION (default)	0.785 ± 0.037	0.610 ± 0.084
LION-tanh	0.802 ± 0.030	0.585 ± 0.050
LION-no-feat-norm	$\mathbf{0.824 \pm 0.020}$	$\underline{0.332 \pm 0.033}$
LION-sum-pool	0.801 ± 0.035	$\mathbf{0.676 \pm 0.019}$
LION-RBF-dist	0.676 ± 0.111	0.588 ± 0.117
LION-unnorm-dist	0.645 ± 0.048	0.565 ± 0.120
LION-minimal	0.705 ± 0.032	$\underline{0.366 \pm 0.033}$

Prediction–ranking trade-offs. The default configuration does not dominate this ablation. On morphology-interaction, LION-sum-pool has higher mean Spearman, 0.676 ± 0.019, than the default, 0.610 ± 0.084, with AUROC 0.801 versus 0.785. LION-no-feat-norm instead achieves the highest mean AUROC on both H&E rules (0.824 on morphology-interaction and 0.927 on immune-proximity), while giving the lowest ranking correlation on morphology-interaction. An assessment based only on AUROC would therefore favor a change that impairs the target feature ordering. This distinction between prediction and the explanation-related evaluation also appears in the separate Xenium component comparison (§6); the two studies measure different explanation-related outcomes.

Symmetric ablation of the distance channel. To complement the baseline distance-access ablation (Appendix M), LION-no-dist replaces the distance weights by uniform weights while retaining the feature normalizer and influence vector. AUROC falls on every benchmark by 0.073 to 0.261, with the largest decrease on spatial transcriptomics and a -0.122 change on H&E immune-proximity. Dense Spearman changes by only -0.008, compared with -0.569 when the influence vector is removed. In these diagnostics, distance weighting has a larger effect on prediction, while the influence vector has a larger effect on Dense ranking recovery; the results do not imply that their roles are independent in every setting.

Why the main results use a single configuration. The cross-validated comparison in Figure 4 uses LION's default configuration on all seven benchmarks. Figure 7 evaluates alternatives under the

separate three-seed protocol, including rules on which a variant improves on the default. Holding the main configuration fixed tests whether one specification transfers across the benchmark suite without per-benchmark architecture selection. The ablation identifies the costs and benefits of changing that specification; it does not retroactively select a different configuration for each main result.

U Model Compactness Across Benchmarks

LION uses $G + 4$ trainable parameters: one shared score per feature and four scalars controlling distance weighting, feature normalization, and the final logistic transformation. Table 11 compares this parameterization with the evaluated baseline architectures on representative benchmarks. These counts describe model size rather than training time or a matched capacity budget.

Table 11: Trainable parameter counts for the listed configurations. Deep baselines use hidden width 64. N/A denotes an unreported count; distance-blind and distance-aware attention are listed separately.

Model	Spatial-MNIST Sparse	Spatial-MNIST Dense	Spatial Tx ($G = 8$)	H&E ($G = 15$)
LION	**34**	**64**	**12**	**19**
DeepSets	10,369	12,289	8,961	9,409
AttentionMIL	4,162	6,082	2,754	3,202
DistAttMIL	4,226	6,146	2,818	3,266
GNN	N/A	N/A	N/A	5,249

On Spatial-MNIST Sparse, LION has 34 parameters, compared with 4,162 for AttentionMIL, 4,226 for DistAttMIL, and 10,369 for DeepSets. On spatial transcriptomics, LION has 12 parameters while the listed baselines have thousands. The reduction follows from scoring named input features directly instead of learning a high-dimensional instance encoder.

This structure exposes the shared ranking without an auxiliary explanation module. Compactness alone, however, neither proves ranking identification nor establishes sufficient predictive capacity. Identification depends on the conditions in Theorem 1, and Appendix V tests how specific departures from the additive model affect prediction and ranking.

V Capacity and Expressiveness

We ran 475 experiments to test how LION's $G + 4$ parameterization responds to three changes: the number of simultaneously active features, feature-correlation strength, and the fraction of instance-score variance contributed by within-instance interactions. The comparisons include distance-equipped DeepSets, attention MIL, and a GNN (4.2k–10.4k parameters), together with three augmented LION variants containing 41, 58, and 469 parameters. Each augmentation retains a shared base feature-score vector; that vector need not summarize all contributions of the extended prediction function.

Prediction and targeted extensions. No baseline exceeds LION's reported mean AUROC at any tested point, including settings with 29 of 30 features active and interactions contributing 83% of the instance-score variance. At the hardest active-feature setting, LION reaches 0.582 ± 0.036 and the strongest 10.4k-parameter baseline reaches 0.554 ± 0.043; the reported comparison does not detect a difference ($p = 0.30$). The Bayes ceiling is 0.645 in this setting, indicating limited achievable discrimination, but the nonsignificant test does not establish equality or identify the source of the remaining error.

Increasing interaction strength reduces LION's AUROC by 0.100. Richer spatial kernels and a more flexible link recover only $+0.000$ and $+0.001$, respectively. Adding the omitted pairwise terms recovers $+0.026$ ($p = 0.014$), approximately one quarter of the observed loss. This variant, LION-pair, has $G + G(G-1)/2 + 4 = 469$ parameters at $G = 30$ and retains $\mathbf{S}$ as its base feature ranking. At the extreme of the active-feature sweep, the monotone-link variant LION-mono instead

recovers $+0.049$ AUROC ($p = 0.0002$). Its reported total of 58 parameters is 24 more than the base model's $G + 4 = 34$ at $G = 30$; pairwise terms do not help in this setting. These results favor targeted extensions for the tested departures rather than establishing that parameter count can never constrain performance.

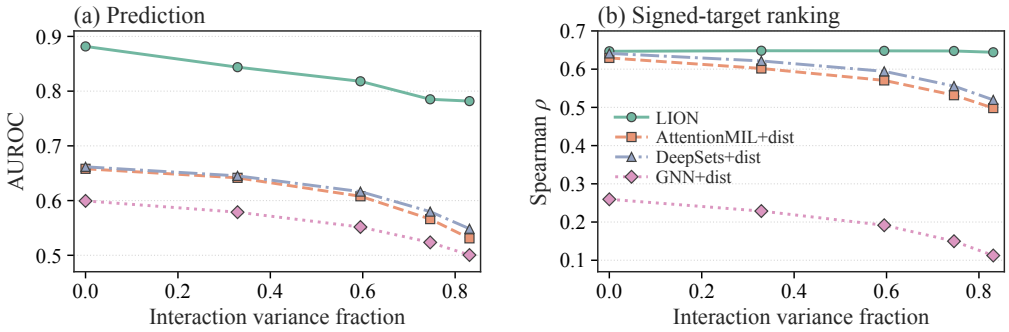

Figure 12: Prediction and ranking across the interaction-strength sweep. Points show five-seed means of (a) AUROC and (b) signed-target Spearman ρ for four models. The horizontal axis is the fraction of instance-score variance due to interactions. Table 12 gives the exact means; uncertainty estimates are unavailable in the reported summary.

Ranking under interactions. Figure 12 separates the predictive loss from recovery of the benchmark's coefficient ordering. As the interaction-variance fraction increases from 0 to 0.83, LION's Spearman changes from 0.6468 to 0.6442 (-0.0026), whereas each distance-equipped baseline declines monotonically by 0.12 to 0.15. Thus, LION's ordering remains nearly unchanged over this sweep even as AUROC decreases. This finding concerns the specified coefficient target, not recovery of the added interaction effects or a general guarantee of interpretability under misspecification. Table 12 retains the exact reported means.

Table 12: Interaction-strength sweep (5 seeds). Columns give the fraction of instance-score variance due to interactions, from 0 to 0.83. AUROC measures prediction and Spearman ρ measures agreement with the benchmark coefficient ordering.

Model	Metric	0.000	0.329	0.595	0.746	0.831
LION	AUROC	**0.8817**	**0.8438**	**0.8180**	**0.7851**	**0.7818**
	ρ	**0.6468**	**0.6482**	**0.6479**	**0.6476**	**0.6442**
AttentionMIL+dist	AUROC	0.6581	0.6416	0.6079	0.5662	0.5310
	ρ	0.6294	0.6018	0.5706	0.5319	0.4982
DeepSets+dist	AUROC	0.6617	0.6452	0.6164	0.5798	0.5486
	ρ	0.6412	0.6215	0.5943	0.5557	0.5196
GNN+dist	AUROC	0.5994	0.5788	0.5517	0.5236	0.5005
	ρ	0.2595	0.2287	0.1916	0.1498	0.1124

W Algorithm Extensions

The following are proposed extensions of the basic architecture. They are not evaluated in the reported experiments, and the identification and consistency results for the current model do not automatically extend to them.

W.1 Sparsemax normalization

Replacing softmax by sparsemax (Martins & Astudillo, 2016) in the distance or feature module could yield exact zero weights, selecting a subset of instances or features for each bag or instance. This changes the normalization map and would require separate analysis of optimization, identification, and ranking recovery.

W.2 CONTEXT-DEPENDENT PARAMETERIZATION

The scalar parameters $\{a, b, \alpha, \beta\}$ could be predicted from bag-level covariates, such as metadata or global context descriptors, while the feature vector $\mathbf{s}$ remains shared across bags. This would allow spatial decay, feature normalization, and the prediction map to vary with context while retaining a common parameter ordering. Covariates would need to be available at inference and constructed without outcome leakage. Restricting conditioning to bag-level information alone does not guarantee valid attribution or transfer the present identification result.

W.3 HIERARCHICAL MULTI-SAMPLE LEARNING

A hierarchical or federated extension could share a global influence vector across datasets while allowing dataset-specific parameters $\{a, b, \alpha, \beta\}$. Local feature vectors $\mathbf{s}^{(p)}$ could be coupled to a common vector $\mathbf{s}^{\text{global}}$ through the penalty $\frac{\mu}{2}\|\mathbf{s}^{(p)} - \mathbf{s}^{\text{global}}\|^2$. The penalty encourages agreement but does not enforce identical local rankings. Such an extension would require a specified synchronization procedure and an explicit choice of which vector defines the reported ranking.

X ADDITIONAL RELATED WORK

Bag prediction and spatial representations. Attention-based MIL (Ilse et al., 2018), CLAM (Lu et al., 2021), DSMIL (Li et al., 2021), and TransMIL (Shao et al., 2021) learn bag-level predictors from instance representations. Their explanations typically locate cells or regions contributing to a prediction. Distance-aware variants (Kim et al., 2022; Aftab et al., 2024) incorporate spatial proximity into attention, while graph neural networks (Kipf & Welling, 2016) and point cloud methods (Qi et al., 2017) encode spatial relations through message passing or geometric layers. Turning these instance- or region-level explanations into a shared ordering of named observed features requires an additional readout.

Post-hoc attribution methods. SHAP (Lundberg & Lee, 2017), Integrated Gradients (Sundararajan et al., 2017), permutation importance (Breiman, 2001), and MIL-specific attribution (Early et al., 2022) explain trained predictors. Aggregating their outputs can produce global feature rankings, but the resulting quantity depends on the attribution rule, reference, and aggregation. Appendix Q examines this readout dependence empirically; it does not treat all attribution methods as estimators of the same signed coefficient target.

Causal formulations of multiple-instance learning. IBMIL, CaMIL, and CATTMIL use causal adjustment to improve bag prediction (Lin et al., 2023; Chen et al., 2024; Wu et al., 2025). CausalMIL studies identification of instance-level latent factors for instance-label prediction and out-of-distribution generalization (Zhang et al., 2022), including settings with variable bag size. LION instead studies a shared ranking of named observed features with a separable physical-distance kernel. Identification of that ranking under the stated predictive model does not by itself identify causal effects.

Architectural constraints and identification. Structural assumptions can expose a common parameter across observations. Cox proportional hazards regression (Cox, 1972), for example, uses proportional hazards to define shared hazard ratios. Reddy et al. (2025) obtain intrinsic interpretability through graph neural additive models, while cell–cell communication tools can fix a biologically named feature space using curated interaction databases (Jin et al., 2021). LION uses a factored bilinear architecture to expose a shared feature ordering under bag-level supervision. Theorem 1 specifies the conditions under which that ordering is identified and consistently recovered.