KA-MIL: KNOWLEDGE-ANCHORED MULTIPLE INSTANCE LEARNING FOR PATIENT CLASSIFICATION FROM SINGLE-CELL TRANSCRIPTOMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Patient classification from single-cell transcriptomics requires aggregating highly heterogeneous cell populations under limited patient-level supervision. Standard data-driven multiple instance learning (MIL) pooling lacks an explicit mechanism for organizing cellular contributions by biological concepts. We introduce Knowledge-Anchored Multiple Instance Learning (KA-MIL), an architecture that utilizes embeddings of biological descriptions as semantic queries for attention-based cell aggregation. Within this framework, a geometry regularizer aligns the pairwise similarities of projected semantic queries with those of the original text embeddings, while learnable residual queries provide complementary task-adaptive capacity. Using Gene Ontology (GO) descriptions with cohort-wide, label-free vocabulary construction, KA-MIL achieves the highest mean macro-F1 in all five primary evaluation settings: four clinical cohorts and a reduced-training setting. On the COVID-19 cohort, it improves macro-F1 over Hier-MIL by 7.04 percentage points. Beyond prediction, our COVID-19 analysis reveals a recurring relative humoral–myeloid attention contrast between plasmablast and CD14 monocyte summaries in 54 of 58 higher-severity test patients across five folds. Together, these findings demonstrate that knowledge-anchored aggregation effectively couples accurate patient classification with the structured, description-guided inspection of cellular summaries.

1 INTRODUCTION

Single-cell profiling resolves cellular heterogeneity that can be masked by bulk measurements. Translating this resolution into patient-level phenotype prediction, however, poses a weak-supervision challenge: each sample contains thousands of measured cells, whereas supervision is available only at the patient-level and does not identify which cells are associated with the phenotype. Predictive signals may be concentrated in specific cell states or rare subpopulations, motivating methods that selectively aggregate information across heterogeneous cells (Arvaniti & Claassen, 2017; He et al., 2022). Multiple instance learning (MIL) provides a natural framework for this task by representing each patient sample as a bag of cellular instances and learning from patient-level labels (Ilse et al., 2018; Engelmann et al., 2024; Jeong et al., 2026).

Within the MIL framework, a central component is the aggregation module, which combines cellular evidence into a patient representation. It determines which cells to emphasize and how to integrate information across heterogeneous populations. Although attention can learn these weights from patient labels, limited cohort sizes can make such models susceptible to overfitting. Biomedical descriptions offer complementary semantic priors by providing shared references for summarizing each patient's cells. The challenge is to adapt these references for prediction while preserving their text-derived semantic relationships and accommodating predictive signals beyond the supplied vocabulary.

To address this challenge, we introduce *knowledge-anchored multiple instance learning* (KA-MIL), which incorporates a *knowledge-anchored aggregation* (KAA) module into MIL. External text descriptions are encoded once into a shared vocabulary of anchors. Each projected anchor queries a patient's cells to produce a separate cellular summary; patient-specific anchor weights combine

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.

these summaries for prediction. KAA operates on cell features and text embeddings without requiring cell-level semantic annotations. KA-MIL couples knowledge-anchored aggregation with a task-adaptable cell encoder, using the shared-encoder teacher–student training framework of scMILD (Jeong et al., 2026).

KAA uses trainable projections to map text anchors and cell features into a shared attention space. To retain text-derived semantic relationships during task adaptation, we regularize the pairwise cosine similarities of projected knowledge queries toward those of the fixed text anchors. This encourages geometric consistency while allowing the queries to adapt to prediction. Learnable residual queries further accommodate predictive signals beyond the supplied descriptions.

In our experiments, we instantiate KAA with Gene Ontology (GO) biological-process descriptions (Ashburner et al., 2000; Gene Ontology Consortium, 2026), which provide a standardized vocabulary for interpreting query-associated cellular summaries. Under cohort-wide, label-free GO construction, KA-MIL achieves the highest mean macro-F1 on each of four clinical cohorts and the highest mean AUC and macro-F1 in the reduced-training Lupus-10 setting with 16 labeled training patients per split. Patient-level COVID analysis further reveals a recurring humoral–myeloid query contrast across all five folds.

The contributions are threefold: (i) a knowledge-anchored aggregation module that uses external text descriptions to guide cell-to-patient aggregation, combining task-adapted semantic queries, geometry regularization, and learnable residual capacity; (ii) consistent mean macro-F1 leadership and top-tier AUC across four clinical cohorts and the reduced-training setting, with ablation studies validating the efficacy of geometry regularization and residual query capacity; and (iii) a patient-level analysis of query-associated cell populations in COVID-19, demonstrating how semantic queries support biological interpretation through a humoral–myeloid contrast.

2 Related Work

Aggregation of patient cell sets. Patient-level prediction requires combining heterogeneous cells into a sample representation. Attention-based MIL and set architectures provide learned pooling mechanisms (Ilse et al., 2018; Zaheer et al., 2017; Lee et al., 2019), while MixMIL combines MIL with generalized linear mixed models (Engelmann et al., 2024). In single-cell applications, an important design choice is how cellular information is organized for aggregation. ProtoCell4P uses cell-type information to guide prototypes, and Hier-MIL aggregates within and across annotated cell types (Xiong et al., 2023; Do & Lähdesmäki, 2025). scCap instead derives cellular groups through clustering refined in pretrained scGPT representations (Noh et al., 2026). CELLECTION and scMILD learn attention-based patient representations without requiring a cell-type partition (Hu et al., 2025; Jeong et al., 2026). While their attention weights can highlight cell populations emphasized by the model, their aggregation is not explicitly guided by external biological knowledge. By using text-derived descriptions as shared queries to guide the attention mechanism itself, KA-MIL naturally yields interpretable cellular summaries with overlapping cell contributions.

Biological knowledge and concept-guided aggregation. Biological knowledge can inform molecular representations through textual descriptions or curated gene modules. GenePT incorporates textual gene information, while VEGA organizes latent representations through biological gene modules (Chen & Zou, 2025; Seninge et al., 2021). In histopathology, ConcepPath combines textual and learnable concepts for two-stage MIL aggregation, starting from image–text correspondences established through pathology vision–language pretraining (Zhao et al., 2024).

In contrast, KA-MIL eliminates the need for prior cross-modal alignment by learning query–cell interactions directly from patient-level supervision. The architecture maps separately pretrained text representations and task-adapted cell features into a shared attention space. During this adaptation, geometry regularization leverages the intrinsic relationships among text-derived queries as a structural prior. Consequently, KA-MIL achieves description-guided aggregation without requiring any paired cell–text pretraining data.

Relational regularization and model inspection. Maintaining the semantic structure of learned queries during task adaptation is a core challenge. Similarity-preserving and relational knowledge distillation address structural transfer by matching relationships among representations rather than

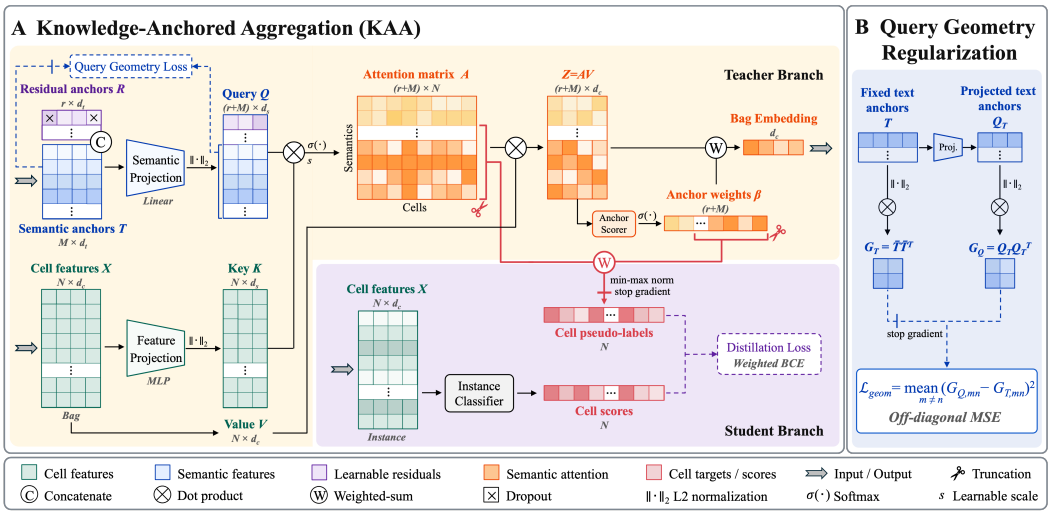

Figure 1: **KAA and query geometry regularization. (A)** Projected semantic and residual anchors attend to cell features, producing summaries combined by patient-specific anchor weights β. Only semantic attention contributes directly to the detached cell targets used for student training. **(B)** Geometry regularization penalizes differences between the off-diagonal entries of the pairwise cosine similarity matrices of projected semantic queries and fixed text anchors. Gradients are stopped through the reference matrix.

absolute coordinates (Tung & Mori, 2019; Park et al., 2019). KA-MIL adapts this relational principle to semantic anchors. While supervised concept bottlenecks require explicit concept annotations (Koh et al., 2020), our geometry regularizer guides task adaptation using only the pairwise cosine similarities of the fixed text embeddings. Specifically, geometry regularization preserves the relative relationships among semantic queries, while the patient-level objective drives their interaction with cellular features. This dual mechanism structurally regularizes the model and extracts interpretable query activations for direct patient-level biological inspection.

3 Method

3.1 Problem formulation and framework

In the context of single-cell MIL, each patient i is a bag of cells with expression matrix $C_i \in \mathbb{R}_{\geq 0}^{N_i \times G}$ and binary label y_i. Here, N_i indicates the number of cells for the i-th patient, and G denotes the number of input genes. A cell encoder produces cell features $X_i = f_\theta(C_i) \in \mathbb{R}^{N_i \times d_c}$. We initialize it with a count autoencoder (AE) pretrained on training cells and keep it trainable during MIL. Adopting the shared-encoder teacher–student architecture from scMILD (Jeong et al., 2026), KA-MIL introduces knowledge-anchored aggregation (KAA) at the teacher branch. Specifically, the teacher aggregates cells via semantic and residual queries, structurally guided by a query geometry regularization (Figure 1).

A vocabulary of M GO biological-process descriptions is encoded offline using frozen BioLORD (Remy et al., 2024), yielding text anchors $T \in \mathbb{R}^{M \times d_t}$. Within each run, these anchors are fixed and shared across patient bags; their construction is specified in Appendix B.4. KAA learns to aggregate cell features using these anchors without cell-level semantic targets. The full training pipeline and network specifications are provided in Figure 3 and Appendix C.

3.2 Knowledge-Anchored Aggregation

To bridge biological knowledge with cellular states, KAA projects semantic anchors T and cell features X_i into a shared space using CLIP-style similarity scaling (Radford et al., 2021). To cap-

ture phenotype-associated signals beyond the fixed text-anchor vocabulary, we introduce learnable residual anchors $R \in \mathbb{R}^{r \times d_t}$ to provide task-adaptive query capacity.

Anchor–cell cross-attention. We concatenate the semantic anchors with the dropout-regularized residual anchors and project them into queries using a shared linear layer, while an multilayer perceptron (MLP) maps the cell features to keys:

$$\begin{aligned} Q &= \mathrm{rownorm}(\mathrm{Linear}_Q([T; \mathrm{Dropout}(R)]))\,, \\ K_i &= \mathrm{rownorm}(\mathrm{MLP}_K(X_i))\,. \end{aligned} \tag{1}$$

Here, $[\cdot;\cdot]$ denotes row-wise concatenation, and $\mathrm{rownorm}$ applies row-wise ℓ_2 normalization.

Both projections have output dimension d_s. Following CLIP's learnable scaling of cosine similarities (Radford et al., 2021), we scale the query–key similarities by a bounded learnable factor s and normalize them over cells to obtain the attention matrix A_i. Using the encoded cell features X_i as values, the queries attend to the cells to form the summaries Z_i:

$$A_i = \mathrm{softmax}_{\mathrm{cells}}(sQK_i^\top), \qquad Z_i = A_i X_i. \tag{2}$$

Each row of $Z_i \in \mathbb{R}^{(M+r)\times d_c}$ represents a query-specific summary, computed as an attention-weighted sum of the patient's cells.

Patient-specific anchor weighting. To aggregate the query-specific summaries z_{im} (the m-th row of Z_i), an MLP scorer evaluates each summary to determine its relative contribution. Writing the unnormalized score as $u_{im} = \mathrm{MLP}_{\mathrm{score}}(z_{im})$, we compute the final patient representation $h_i \in \mathbb{R}^{d_c}$ as a convex combination:

$$\beta_{im} = \frac{\exp(u_{im})}{\sum_{k=1}^{M+r} \exp(u_{ik})}, \qquad h_i = \sum_{m=1}^{M+r} \beta_{im} z_{im}. \tag{3}$$

The scorer assigns patient-specific weights β_{im} to both the semantic and residual summaries z_{im}, allowing their joint contribution to the final representation h_i to vary dynamically with each patient's unique cellular composition.

Network dimensions, dropout, and scale parameterization are specified in Appendix C.2; permutation invariance and further properties are discussed in Appendix G.

3.3 Query geometry regularization

The learned query projection enables task adaptation but can distort the relationships encoded by the text embeddings. Thus, we regularize this adaptation using their pairwise cosine similarities as a fixed reference. Let $\bar{T} = \mathrm{rownorm}(T)$ and let Q_T contain the first M normalized query rows, corresponding to semantic anchors. The off-diagonal penalty is

$$\mathcal{L}_{\mathrm{geom}} = \frac{1}{M(M-1)} \sum_{m \neq n} \left[(Q_T Q_T^\top)_{mn} - (\bar{T}\bar{T}^\top)_{mn} \right]^2. \tag{4}$$

This soft penalty encourages semantic queries to retain text-derived similarities while allowing task adaptation. Specifically, it constrains relative relationships rather than absolute coordinates and supplies no direct per-cell semantic supervision. To maintain flexibility, the learnable residual anchors are explicitly excluded from this penalty, retaining their unconstrained capacity.

3.4 Training objectives and inference

The teacher branch maps h_i through an MLP to predict the positive-class probability $\hat{y}_i$. Given true labels $y_i \in \{0, 1\}$ for a minibatch $\mathcal{B}$, the teacher minimizes:

$$\mathcal{L}_{\mathrm{teacher}} = \mathcal{L}_{\mathrm{bag}} + \lambda \mathcal{L}_{\mathrm{geom}}, \qquad \mathcal{L}_{\mathrm{bag}} = -\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \left[y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i) \right]. \tag{5}$$

Here, standard cross-entropy ($\mathcal{L}_{\mathrm{bag}}$) provides task supervision, while λ scales the geometric regularization. Minimizing $\mathcal{L}_{\mathrm{teacher}}$ updates the shared cell encoder, KAA, and classifier end-to-end.

To further refine the shared encoder, we periodically update it via the auxiliary student branch every k teacher epochs. To construct cell-level pseudo-targets, we aggregate the teacher's detached semantic attention maps according to their anchor weights, using only semantic attention maps while normalizing weights over all queries. Because these targets are detached, the student loss fine-tunes the shared encoder without backpropagating through the teacher. The exact student loss formulation follows scMILD and is detailed in Appendix C.3.

During inference, the student branch is discarded. Patient prediction relies exclusively on the cell encoder, the KAA module (with precomputed text embeddings and disabled dropout), and the teacher classifier.

4 Experiments

4.1 Experimental Setup

Datasets and evaluation. We evaluate our framework on four clinical single-cell cohorts covering complementary patient-level prediction tasks: (1) disease severity in COVID-19 peripheral blood mononuclear cells (PBMCs) (**COVID**; 90 patients; 32 asymptomatic/mild cases [lower severity] vs. 58 moderate/severe/critical cases [higher severity]); (2) infection status in nasal swabs (**NS**; 50 donors; 15 controls vs. 35 with COVID-19); (3) intestinal inflammatory status in ulcerative colitis (**UC**; 30 patients; 12 healthy vs. 18 inflamed); and (4) autoimmune disease status in PBMCs (**Lupus**; 169 individuals; 50 healthy vs. 119 with systemic lupus erythematosus [SLE]).

For the four standard cohort settings, we use five-fold patient-level cross-validation with approximately 60%/20%/20% train/validation/test splits. To assess performance with fewer training labels, we also evaluate **Lupus-10**. **Lupus-10** reuses the Lupus cohort with five repeated stratified splits of approximately 10%/10%/80%, allowing overlapping test sets (Appendix B.2).

All methods share the same partitions and validation-based checkpoint selection. Thresholds maximize validation macro-F1 and are fixed for testing. We report test AUC and macro-F1 as mean $\pm$ sample standard deviation (SD) across the five splits. Dataset details, baseline configurations, additional metrics, and evaluation procedures appear in Appendices B and D.

Knowledge anchor construction. We construct 50 GO anchors per cohort using frozen BioLORD embeddings (Remy et al., 2024), evaluated under two knowledge-construction protocols. The primary cohort-wide transductive protocol selects the vocabulary based on cohort-wide expression statistics—including held-out patients—without accessing disease labels, while keeping supervised model fitting strictly confined to training patients. To assess sensitivity to this data-access assumption, a secondary split-specific protocol restricts vocabulary construction exclusively to the training patients within each training split, retaining identical optimization parameters. Preprocessing details for both protocols are provided in Appendix B.4.

Baselines. We compare KA-MIL with six baseline configurations. **ABMIL** (Ilse et al., 2018) uses attention-based pooling with bag-level supervision and no student branch. **scMILD** (Jeong et al., 2026) retains its attention pooler and shared-encoder teacher–student framework; we evaluate both its orthogonal projection loss (**OPL**)**-enabled** version and an **OPL-disabled** control, matching KA-MIL's OPL setting. These local pipelines share AE initialization and cohort-specific optimization profiles. **Hier-MIL (HA)** (Do & Lähdesmäki, 2025) performs hierarchical attention over annotated cell types using frozen scGPT features, with validation-based hyperparameter selection. **scGPT + MIL** and **SE + MIL** combine frozen scGPT (Cui et al., 2024) and State Embedding (SE) (Adduri et al., 2026) features, respectively, with the attention-based patient predictor from our scMILD implementation.

scMILD (–OPL) provides the matched encoder-and-training comparison for KAA, while the frozen-feature baselines evaluate alternative representation–aggregation pipelines. Implementation details are provided in Appendix D.2. We additionally evaluate KAA on frozen scGPT and SE features in Appendix F.

Table 1: **Patient-level classification.** Results are reported as the test mean ± SD over five patient-level folds. **Bold** marks the highest unrounded mean, including ties. OPL is enabled (+) or disabled (–) as indicated; KA-MIL omits OPL, using cohort-wide, label-free GO anchors.

	COVID		NS	
Model	AUC	Macro-F1	AUC	Macro-F1
ABMIL	0.7292 ± 0.1236	0.6061 ± 0.1013	$\mathbf{0.9399 \pm 0.0395}$	0.8013 ± 0.0843
scMILD (+OPL)	0.6327 ± 0.0861	0.5612 ± 0.0799	0.8923 ± 0.1068	0.8061 ± 0.1576
scMILD (–OPL)	0.7213 ± 0.1286	0.5789 ± 0.0814	0.8798 ± 0.1091	0.7823 ± 0.1317
scGPT + MIL	0.7750 ± 0.1290	0.6233 ± 0.1188	0.7905 ± 0.0783	0.6892 ± 0.1853
SE + MIL	0.7606 ± 0.1324	0.6339 ± 0.1133	0.8363 ± 0.1187	0.8112 ± 0.1491
Hier-MIL (HA)	0.7578 ± 0.0938	0.6371 ± 0.1505	0.7429 ± 0.2511	0.6298 ± 0.2755
KA-MIL	$\mathbf{0.7773 \pm 0.1015}$	$\mathbf{0.7075 \pm 0.0657}$	0.9274 ± 0.0834	$\mathbf{0.8679 \pm 0.1407}$
	UC		Lupus	
Model	AUC	Macro-F1	AUC	Macro-F1
ABMIL	$\mathbf{1.0000 \pm 0.0000}$	0.9333 ± 0.1491	0.9960 ± 0.0057	0.9582 ± 0.0280
scMILD (+OPL)	0.9778 ± 0.0497	0.8527 ± 0.0852	0.9991 ± 0.0020	0.9849 ± 0.0338
scMILD (–OPL)	$\mathbf{1.0000 \pm 0.0000}$	0.8971 ± 0.0939	0.9992 ± 0.0019	0.9841 ± 0.0219
scGPT + MIL	0.9306 ± 0.1011	0.7879 ± 0.1380	0.9964 ± 0.0036	0.9566 ± 0.0443
SE + MIL	$\mathbf{1.0000 \pm 0.0000}$	0.9314 ± 0.0939	0.9985 ± 0.0034	0.9737 ± 0.0266
Hier-MIL (HA)	0.9056 ± 0.1297	0.8240 ± 0.1777	0.9939 ± 0.0050	0.9374 ± 0.0544
KA-MIL	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{0.9657 \pm 0.0767}$	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{0.9914 \pm 0.0193}$

Table 2: **Evaluation on the reduced-training Lupus-10 setting.** Results are reported as test mean ± SD across five repeated stratified splits.

Metric	ABMIL	scMILD (+OPL)	scMILD (–OPL)	scGPT + MIL	SE + MIL	Hier-MIL (HA)	KA-MIL
AUC	0.9616 ± 0.0276	0.9832 ± 0.0208	0.9788 ± 0.0128	0.9056 ± 0.0888	0.9693 ± 0.0410	0.9221 ± 0.0415	$\mathbf{0.9870} \pm 0.0132$
Macro-F1	0.8725 ± 0.0528	0.9213 ± 0.0528	0.8684 ± 0.0398	0.8328 ± 0.0504	0.8787 ± 0.0949	0.7781 ± 0.0806	$\mathbf{0.9301} \pm 0.0466$

4.2 PATIENT-LEVEL CLASSIFICATION

Consistent macro-F1 leadership across clinical cohorts. With cohort-wide GO anchors, KA-MIL achieves the highest mean macro-F1 across all four cohorts (Table 1). Under matched encoder and optimization settings, KA-MIL improves mean macro-F1 over scMILD (–OPL) by 12.86 percentage points on COVID and 8.56 points on NS. Against scMILD (+OPL), the corresponding gains are 14.63 and 6.18 points, supporting the benefit of KAA within the shared teacher–student training framework. On NS, KA-MIL achieves higher mean macro-F1 than ABMIL (0.8679 vs. 0.8013), with comparable mean AUC (0.9274 vs. 0.9399). On COVID, it likewise improves mean macro-F1 over scGPT + MIL (0.7075 vs. 0.6233) while maintaining similar mean AUC (0.7773 vs. 0.7750). These results show that KA-MIL combines competitive ranking performance with higher mean macro-F1 under the shared validation-based threshold-selection protocol.

Performance ceilings and label efficiency. KA-MIL achieves near-perfect performance on UC and Lupus cohorts. To evaluate performance with fewer patient-level training labels, we use Lupus-10, a reduced-training setting with 16 training patients per split.

With 16 training patients per split, KA-MIL achieves the highest mean AUC (0.9870) and macro-F1 (0.9301) on Lupus-10. Its macro-F1 exceeds scMILD (–OPL) by 6.17 points and scMILD (+OPL) by 0.88 points.

Table 3: **Component ablations and encoder controls.** Test AUC and macro-F1, mean ± SD over five splits. OPL is disabled throughout. KAA ablations retain the remaining full-model settings; frozen-encoder controls use the indicated scMILD or KA-MIL architecture.

	COVID		NS		UC		Lupus-10	
Setting	AUC	Macro-F1	AUC	Macro-F1	AUC	Macro-F1	AUC	Macro-F1
KA-MIL (control)	0.7773 ±0.1015	0.7075 ±0.0657	0.9274 ±0.0834	0.8679 ±0.1407	1.0000 ±0.0000	0.9657 ±0.0767	0.9870 ±0.0132	0.9301 ±0.0466
Random anchors (fixed)	0.7372 ±0.1041	0.7020 ±0.0790	0.8804 ±0.1458	0.7817 ±0.1270	1.0000 ±0.0000	0.9333 ±0.1491	0.9505 ±0.0872	0.8943 ±0.1001
Random anchors (learned)	0.7149 ±0.1106	0.6074 ±0.0867	0.8613 ±0.1402	0.7817 ±0.1270	0.9750 ±0.0559	0.9314 ±0.0939	0.9789 ±0.0176	0.9123 ±0.0549
No geometry penalty	0.7230 ±0.1267	0.5950 ±0.1037	0.9208 ±0.0919	0.7623 ±0.1487	1.0000 ±0.0000	0.8463 ±0.1591	0.9701 ±0.0252	0.8926 ±0.0568
No residual queries	0.6624 ±0.1781	0.6409 ±0.1359	0.8798 ±0.0982	0.7677 ±0.1292	0.9028 ±0.1626	0.7879 ±0.2288	0.9863 ±0.0113	0.8887 ±0.0640
No residual dropout	0.7724 ±0.1182	0.6692 ±0.1317	0.8798 ±0.0790	0.8065 ±0.1162	0.9778 ±0.0497	0.8463 ±0.1591	0.9649 ±0.0478	0.8868 ±0.0806
Frozen encoder (scMILD –OPL)	0.6673 ±0.0832	0.5933 ±0.0480	0.4631 ±0.3501	0.4631 ±0.2271	0.7722 ±0.2724	0.5860 ±0.2534	0.9046 ±0.0929	0.7910 ±0.1347
Frozen encoder (KA-MIL)	0.6870 ±0.1286	0.6623 ±0.0713	0.3970 ±0.2451	0.3834 ±0.1646	0.6111 ±0.3849	0.5000 ±0.3281	0.9005 ±0.0521	0.7522 ±0.1226

4.3 Effect of GO vocabulary construction scope

Under split-specific protocol—where GO anchors are constructed exclusively from training patients within each fold—KA-MIL sustains consistent predictive efficacy across the majority of cohorts. Notably, aggregate test metrics on the UC cohort remain identical (macro-F1 0.9657) under both construction scopes. Furthermore, KA-MIL preserves a higher mean macro-F1 than the matched scMILD (–OPL) control on COVID (0.6333) and Lupus (0.9845). On Lupus-10, it achieves a macro-F1 of 0.9152, compared with 0.8684 and 0.9213 for scMILD (–OPL) and scMILD (+OPL), respectively.

On NS, cohort-wide GO construction yields a mean macro-F1 of 0.8679 and a mean AUC of 0.9274, compared with 0.7649 and 0.8512 under split-specific construction. Cohort-wide expression statistics may provide a stabilizing prior in NS, where relatively few cells are available per patient (Appendix B). Differences in vocabulary composition and variability in model fitting and selection may also contribute to the observed performance difference.

4.4 Ablation study: Contributions of model components

To dissect the performance gains of KA-MIL, we assess query content, geometry regularization, and residual capacity under fixed cohort-specific profiles and cohort-wide GO construction (Table 3). Appendix E defines each intervention and reports additional encoder controls.

Text content and geometry constraint. Removing geometry regularization lowers mean macro-F1 in all four settings. GO anchors also outperform both random-anchor controls in mean macro-F1 across all four settings, with the largest gaps on NS, UC, and Lupus-10. Both controls retain geometry regularization against their initial random reference. On COVID, GO anchors show their clearest advantage over fixed random anchors in AUC (0.7773 vs. 0.7372), alongside macro-F1 of 0.7075 vs. 0.7020. These matched comparisons support the contributions of text-derived anchors and geometry regularization.

Residual capacity and encoder adaptation. Removing residual queries, disabling their dropout, or freezing the local AE encoder lowers KA-MIL's mean macro-F1 in all four settings. These controls support complementary contributions from residual capacity, its regularization, and encoder adaptation. Appendix F additionally evaluates KAA on frozen scGPT and SE features.

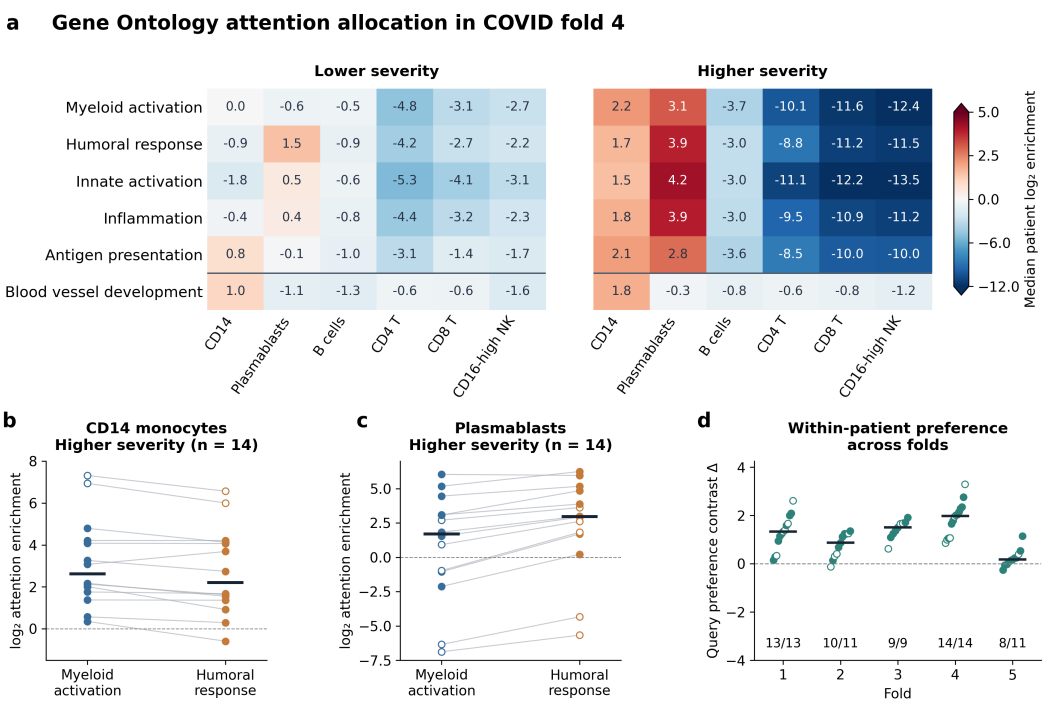

Figure 2: **Biologically coherent attention allocation in COVID-19. (a)** Query attention enrichment across key cell populations. Heatmap shows median patient-level $\log_2 E$ (restricted to patients with $\geq$ 20 cells per type), where red indicates enrichment above the proportional baseline ($E = 1$). **(b, c)** Myeloid vs. humoral attention in CD14 monocytes and plasmablasts. Connected patient lines demonstrate KAA's ability to consistently show distinct query preferences within patients. **(d)** Within-patient query preference (Δ) across five folds. Positive Δ indicates a stronger relative humoral-versus-myeloid preference in plasmablasts than in CD14 monocytes. *Plotting notes:* Bars denote medians; dashed lines mark neutral baselines. Open points indicate sparse populations (< 20 cells; see Appendix H).

5 Query-Associated Cellular Evidence in COVID

5.1 A patient-level view of attention allocation

To investigate how attention allocation maps to distinct clinical phenotypes, we examine patient-level attention in fold 4 (14 patients in the higher severity group and 4 in the lower severity group) and assess the consistency of within-patient query preferences across all five folds.

For query m and a cell type t represented by $n_{it} > 0$ cells in patient i, we define attention enrichment as:

$$E_{imt} = \frac{N_i}{n_{it}} \sum_{j \in t} A_{imj}. \tag{6}$$

This statistic quantifies the attention assigned to a cell type relative to its share of the patient's cells. The baseline $E = 1$ ($\log_2 E = 0$) indicates abundance-proportional allocation; values above or below one indicate greater or smaller relative attention. By mapping the patient-median $\log_2 E$, Figure 2a summarizes query-specific cellular attention in the fold-4 test patients, showing different group-median allocation patterns between the two severity groups.

5.2 Distinct query preferences within patient profiles

Figure 2a contrasts five immune queries against a stable blood-vessel-development reference. While the reference maintains a consistently low-contrast allocation across patient groups, the immune queries actively concentrate attention on CD14 monocytes and plasmablasts in patients in the higher

severity group (e.g., median plasmablast $\log_2 E = 3.87$ vs. 1.46 in the lower severity group for the humoral query). This demonstrates a highly query-dependent cellular organization.

Crucially, KAA captures distinct query preferences *within the same patient*. When comparing the humoral immune response (GO:0006959) with myeloid leukocyte activation (GO:0002274) in fold 4, CD14 enrichment favors the myeloid query in 13 of 14 patients in the higher severity group, while plasmablast enrichment favors the humoral query in 13 of 14 (Figure 2b,c). We formally quantify this relative distinction as:

$$\Delta_i = \log_2 \frac{E_{i,\text{humoral,plasmablast}}}{E_{i,\text{myeloid,plasmablast}}} - \log_2 \frac{E_{i,\text{humoral,CD14}}}{E_{i,\text{myeloid,CD14}}}. \tag{7}$$

The cell-type abundance normalization cancels within each query ratio. Positive Δ_i indicates a stronger relative humoral preference in plasmablasts than in CD14 cells. All 14 patients in the higher severity group of fold 4 have positive Δ_i (median 1.98), including all eight with at least 20 cells of both types. Across five folds, 54 of 58 patient observations in the higher severity group show the same direction, and every fold has a positive median. This recurring contrast demonstrates that KAA distinguishes query-specific cellular preferences even when multiple queries attend strongly to the same populations.

5.3 Biological context of the query-associated summaries

The observed distinction is consistent with prominent plasmablast responses and heterogeneous monocyte states reported in COVID (Wilk et al., 2020; Stephenson et al., 2021), including myeloid dysregulation and extrafollicular B-cell responses associated with severe disease (Schulte-Schrepping et al., 2020; Woodruff et al., 2020). Documented changes in CD8 T-cell states and NK-cell responses further motivate the lymphoid comparisons (Stephenson et al., 2021; Witkowski et al., 2021).

These biological references place the observed cellular preferences within established COVID immunology. Interpreting E as relative attention and Δ as within-patient query preference makes KAA's aggregation accessible to patient-level inspection. Together, literature-consistent cellular associations and recurring contrasts across folds support the biological coherence and query specificity of the learned summaries. Complete patient coverage and presentation details are provided in Appendix H.

6 Discussion and Conclusion

Under cohort-wide, label-free GO construction, KA-MIL achieves the highest mean macro-F1 across all five primary settings. By integrating semantic queries, geometry regularization, and residual capacity with a task-adaptable cell encoder, the architecture connects patient classification with structured, description-guided inspection of cellular attention. In the COVID-19 analysis, a recurring relative humoral–myeloid attention contrast distinguishes plasmablast and CD14 monocyte summaries in 54 of 58 test patients in the higher severity group across five folds.

Several limitations define the scope of these findings. The primary evaluation uses held-out expression statistics for vocabulary construction, and the scope comparison shows that performance depends on the data available for constructing the vocabulary, particularly on NS. The split-specific GO comparison also retains cohort-wide feature selection, leaving fully inductive evaluation of the complete pipeline open. In addition, KAA does not consistently improve over attention MIL on frozen foundation-model features. The query-associated cellular patterns provide descriptive biological associations, while their correspondence to pathway activity and causal disease mechanisms requires independent validation.

Future work should evaluate the complete pipeline with preprocessing and vocabulary construction confined to training patients, examine how query design interacts with frozen and adaptable cell representations, and test query-associated summaries against independent gene-expression and perturbation evidence. These directions build on the central contribution of KA-MIL: integrating structured biological priors into patient-level prediction while making cellular aggregation accessible to inspection.

Reproducibility Statement

To support reproducibility, an anonymized implementation of KA-MIL accompanies this submission. Appendices B–D detail the experimental configurations, including datasets, patient-level partitions, model hyperparameters, and evaluation protocols. The construction of the GO-based vocabulary is outlined in Appendix B.4. Together, the provided codebase and comprehensive appendices document the evaluation framework and facilitate the reproduction of the reported findings.

AI Use Statement

Generative AI tools (GPT 5.6 Sol, GPT 6.0 Astra, and Gemini 3.1 Pro) were utilized to assist with language refinement, manuscript review, and coding assistance, including hyperparameter-search scripts and code refactoring. While these tools provided editorial feedback and phrasing suggestions, all experimental conceptualization, core model implementation, original drafting, and final editorial decisions were executed by the human authors. No generative models were utilized to synthesize patient data or experimental performance metrics. The authors assume full responsibility for the integrity of the final text, claims, references, and all associated artifacts.

References

Abhinav K. Adduri, Dhruv Gautam, Beatrice Bevilacqua, et al. Predicting cellular responses to perturbation across diverse contexts with State. *Cell*, 189(19):5914–5931.e20, 2026. doi: 10.1016/j.cell.2026.07.052. URL https://www.sciencedirect.com/science/article/pii/S0092867426009219.

Eirini Arvaniti and Manfred Claassen. Sensitive detection of rare disease-associated cell subsets via representation learning. *Nature Communications*, 8(1):14825, 2017. doi: 10.1038/ncomms14825. URL https://www.nature.com/articles/ncomms14825.

Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, et al. Gene Ontology: Tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000. doi: 10.1038/75556. URL https://www.nature.com/articles/ng0500_25.

Yiqun Chen and James Zou. Simple and effective embedding model for single-cell biology built from ChatGPT. *Nature Biomedical Engineering*, 9(4):483–493, 2025. doi: 10.1038/s41551-024-01284-6. URL https://www.nature.com/articles/s41551-024-01284-6.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8):1470–1480, 2024. doi: 10.1038/s41592-024-02201-0. URL https://www.nature.com/articles/s41592-024-02201-0.

F. S. Czepluch, M. Bernhardt, H. Kuschicke, R. Gogiraju, M. R. Schroeter, J. Riggert, G. Hasenfuss, and Katrin Schäfer. In vitro and in vivo effects of human monocytes and their subsets on new vessel formation. *Microcirculation*, 21(2):148–158, 2014. doi: 10.1111/micc.12100. URL https://doi.org/10.1111/micc.12100.

Chau Do and Harri Lähdesmäki. Incorporating hierarchical information into multiple instance learning for patient phenotype prediction with single-cell RNA-sequencing data. *Bioinformatics*, 41(Supplement_1):i96–i104, 2025. doi: 10.1093/bioinformatics/btaf241. URL https://academic.oup.com/bioinformatics/article/41/Supplement_1/i96/8199355.

Jan P. Engelmann, Alessandro Palma, Jakub M. Tomczak, Fabian Theis, and Francesco Paolo Casale. Mixed models with multiple instance learning. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3664–3672. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/engelmann24a.html.

Gene Ontology Consortium. The Gene Ontology knowledgebase in 2026. *Nucleic Acids Research*, 54(D1):D1779–D1792, 2026. doi: 10.1093/nar/gkaf1292. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC12807639/.

Bryan He, Matthew Thomson, Meena Subramaniam, Richard Perez, Chun Jimmie Ye, and James Zou. CloudPred: Predicting patient phenotypes from single-cell RNA-seq. In *Pacific Symposium on Biocomputing 2022*, volume 27, pp. 337–348. World Scientific, 2022. doi: 10.1142/9789811250477_0031. URL https://psb.stanford.edu/psb-online/proceedings/psb22/he.pdf.

Hongru Hu, Siddhant Sanghi, and Gerald Quon. Predicting emergent phenotypes from single cell populations using CELLECTION. *bioRxiv*, 2025. doi: 10.1101/2025.09.02.673886. URL https://www.biorxiv.org/content/early/2025/09/07/2025.09.02.673886. Preprint.

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

Kyeonghun Jeong, Jinwook Choi, and Kwangsoo Kim. scMILD: Single-cell multiple instance learning for sample classification and associated subpopulation discovery. *iScience*, 29(4): 115284, 2026. doi: 10.1016/j.isci.2026.115284. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC13019583/.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 5338–5348, 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set Transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 3744–3753, 2019. URL https://proceedings.mlr.press/v97/lee19d.html.

Igor Mandric, Tommer Schwarz, Arunabha Majumdar, Kangcheng Hou, Leah Briscoe, Richard Perez, Meena Subramaniam, Christoph Hafemeister, Rahul Satija, Chun Jimmie Ye, Bogdan Pasaniuc, and Eran Halperin. Optimized design of single-cell RNA sequencing experiments for cell-type-specific eQTL analysis. *Nature Communications*, 11(1):5504, 2020. doi: 10.1038/s41467-020-19365-w. URL https://doi.org/10.1038/s41467-020-19365-w.

Janghyun Noh, Yoobin Shin, Min Kim, and Minsik Oh. Annotation-free phenotype prediction using knowledge-augmented clustering from single-cell rna sequencing data. *Briefings in Bioinformatics*, 27(4):bbag395, 07 2026. ISSN 1477-4054. doi: 10.1093/bib/bbag395. URL https://doi.org/10.1093/bib/bbag395.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3967–3976, 2019. doi: 10.1109/CVPR.2019.00409. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Park_Relational_Knowledge_Distillation_CVPR_2019_paper.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

François Remy, Kris Demuynck, and Thomas Demeester. BioLORD-2023: semantic textual representations fusing large language models and clinical knowledge graph insights. *Journal of the American Medical Informatics Association*, 31(9):1844–1855, 2024. doi: 10.1093/jamia/ocae029. URL https://academic.oup.com/jamia/article/31/9/1844/7614965.

Jonas Schulte-Schrepping, Nico Reusch, Daniela Paclik, Kevin Baßler, Stephan Schlickeiser, Bowen Zhang, et al. Severe COVID-19 is marked by a dysregulated myeloid cell compartment. *Cell*, 182(6):1419–1440.e23, 2020. doi: 10.1016/j.cell.2020.08.001. URL https://www.sciencedirect.com/science/article/pii/S0092867420309922.

Lucas Seninge, Ioannis Anastopoulos, Hongxu Ding, and Joshua Stuart. VEGA is an interpretable generative model for inferring biological network activity in single-cell transcriptomics. *Nature Communications*, 12(1):5684, 2021. doi: 10.1038/s41467-021-26017-0. URL https://www.nature.com/articles/s41467-021-26017-0.

Christopher S. Smillie, Moshe Biton, Jose Ordovas-Montanes, Keri M. Sullivan, Grace Burgin, Daniel B. Graham, et al. Intra- and inter-cellular rewiring of the human colon during ulcerative colitis. *Cell*, 178(3):714–730.e22, 2019. doi: 10.1016/j.cell.2019.06.029. URL https://doi.org/10.1016/j.cell.2019.06.029.

Emily Stephenson, Gary Reynolds, Rachel A. Botting, Fernando J. Calero-Nieto, Michael D. Morgan, Zewen Kelvin Tuong, et al. Single-cell multi-omics analysis of the immune response in COVID-19. *Nature Medicine*, 27(5):904–916, 2021. doi: 10.1038/s41591-021-01329-2. URL https://www.nature.com/articles/s41591-021-01329-2.

Yapeng Su, Daniel Chen, Dan Yuan, Christopher Lausted, Jongchan Choi, Chengzhen L. Dai, et al. Multi-omics resolves a sharp disease-state shift between mild and moderate COVID-19. *Cell*, 183(6):1479–1495.e20, 2020. doi: 10.1016/j.cell.2020.10.037. URL https://doi.org/10.1016/j.cell.2020.10.037.

Fred Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1365–1374, Los Alamitos, CA, USA, 2019. IEEE Computer Society. doi: 10.1109/ICCV.2019.00145. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Tung_Similarity-Preserving_Knowledge_Distillation_ICCV_2019_paper.html.

Aaron J. Wilk, Arjun Rustagi, Nancy Q. Zhao, Jonasel Roque, Giovanny J. Martínez-Colón, Julia L. McKechnie, et al. A single-cell atlas of the peripheral immune response in patients with severe COVID-19. *Nature Medicine*, 26(7):1070–1076, 2020. doi: 10.1038/s41591-020-0944-y. URL https://www.nature.com/articles/s41591-020-0944-y.

Mario Witkowski, Caroline Tizian, Marta Ferreira-Gomes, Daniela Niemeyer, Terry C. Jones, Frederik Heinrich, et al. Untimely TGFβ responses in COVID-19 limit antiviral functions of NK cells. *Nature*, 600(7888):295–301, 2021. doi: 10.1038/s41586-021-04142-6. URL https://doi.org/10.1038/s41586-021-04142-6.

Matthew C. Woodruff, Richard P. Ramonell, Doan C. Nguyen, Kevin S. Cashman, Ankur Singh Saini, Natalie S. Haddad, et al. Extrafollicular B cell responses correlate with neutralizing antibodies and morbidity in COVID-19. *Nature Immunology*, 21(12):1506–1516, 2020. doi: 10.1038/s41590-020-00814-z. URL https://www.nature.com/articles/s41590-020-00814-z.

Guangzhi Xiong, Stefan Bekiranov, and Aidong Zhang. ProtoCell4P: an explainable prototype-based neural network for patient classification using single-cell RNA-seq. *Bioinformatics*, 39 (8):btad493, 08 2023. ISSN 1367-4811. doi: 10.1093/bioinformatics/btad493. URL https://doi.org/10.1093/bioinformatics/btad493.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep Sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/f22e4747da1aa27e363d86d40ff442fe-Abstract.html.

Weiqin Zhao, Ziyu Guo, Yinshuang Fan, Yuming Jiang, Maximus C. F. Yeung, and Lequan Yu. Aligning knowledge concepts to whole slide images for precise histopathology image analysis. *npj Digital Medicine*, 7(1):383, 2024. doi: 10.1038/s41746-024-01411-2. URL https://www.nature.com/articles/s41746-024-01411-2.

Carly G. K. Ziegler, Vincent N. Miao, Anna H. Owings, Andrew W. Navia, Ying Tang, Joshua D. Bromley, et al. Impaired local intrinsic immunity to SARS-CoV-2 infection in severe COVID-19. *Cell*, 184(18):4713–4733.e22, 2021. doi: 10.1016/j.cell.2021.07.023. URL https://doi.org/10.1016/j.cell.2021.07.023.

A KA-MIL MODEL PIPELINE

Figure 3 summarizes the architecture and training pipeline of KA-MIL. Precomputed GO text embeddings provide semantic anchors, while a shared cell encoder produces cellular representations (A). The Knowledge-Anchored Aggregation (KAA) module uses semantic and residual queries to form patient-level representations, and teacher attention supplies detached cell targets for student training (B). Training alternates between teacher and student optimization, with both steps updating the shared encoder and query-geometry regularization applied during the teacher step (C). Inference uses only the encoder and teacher.

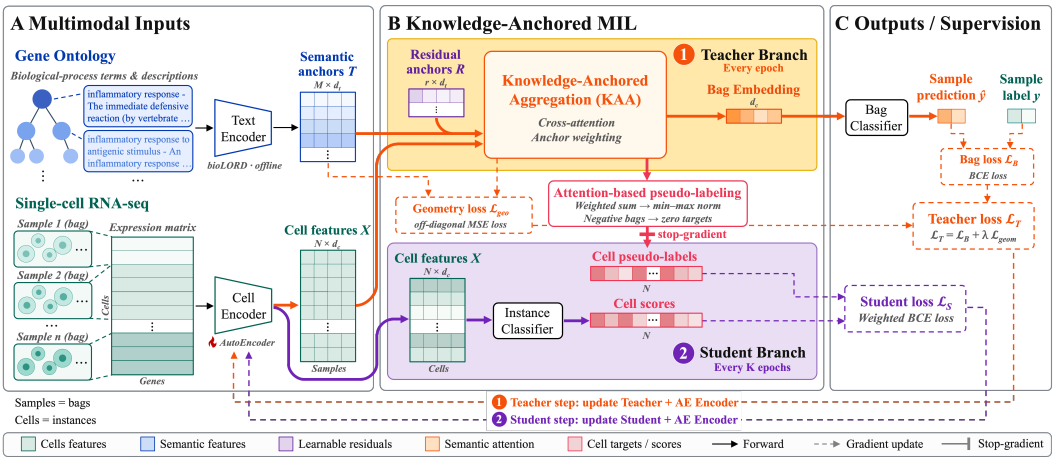

Figure 3: **KA-MIL pipeline.** **(A)** Precomputed GO text embeddings provide semantic anchors, while a shared encoder extracts cell features. **(B)** KAA aggregates cellular evidence using semantic and residual queries. Teacher attention provides detached cell targets for student training. **(C)** The teacher is optimized with patient classification and query geometry losses, while the student uses a weighted cell classification loss. Both training steps update the shared encoder. Inference uses only the encoder and teacher.

B DATA, SPLITS, AND KNOWLEDGE ANCHORS

B.1 COHORTS AND PREDICTION TARGETS

Table 4 summarizes the datasets evaluated in our study, and Table 5 lists their source publications and public data repositories. For the four clinical single-cell cohorts, we use the quality-controlled gene expression matrices and metadata provided by the source repositories. Within the multiple instance learning framework, cells are grouped into bags based on their patient-level sample identifiers.

For the COVID cohort, we use the data released by Stephenson et al. (2021) under accession E-MTAB-10026 and retain cells annotated with `Status = Covid`. Motivated by the mild-to-moderate disease-state transition reported by Su et al. (2020), we construct a binary severity classification task using the cohort's clinical severity annotations: asymptomatic and mild cases form the negative class (0), while moderate, severe, and critical cases form the positive class (1). We refer to these groups as lower and higher severity, respectively. For the remaining three cohorts, we strictly follow the original clinical metadata to define binary labels: healthy controls versus COVID-19 for NS, healthy versus inflamed for UC, and healthy versus systemic lupus erythematosus (SLE) for Lupus.

B.2 PATIENT-LEVEL DATA SPLITS

For the four standard cohort settings, we use five-fold patient-level cross-validation with `StratifiedGroupKFold`, enabling shuffling with a random seed of 0. Within each fold, 25% of the remaining patients are assigned to validation using stratified sampling with the fold index

Table 4: **Summary of the evaluated clinical single-cell cohorts.** Note that the **Lupus-10** variant is an alternative train/validation/test split of the main Lupus cohort and is thus omitted from this table. N/P denotes negative/positive classes

Cohort	Genes	Total Cells	Cells/bag (mean ± SD)	Bags (N/P)	Prediction targets (N/P)
COVID	24,929	527,286	5858.73 ± 3206.51	32 / 58	Lower / higher severity
NS	28,696	26,947	538.94 ± 500.61	15 / 35	Controls / COVID-19
UC	21,784	235,229	7840.97 ± 5845.97	12 / 18	Healthy / inflamed
Lupus	32,738	834,096	4935.48 ± 1731.70	50 / 119	Healthy / SLE

Table 5: **Dataset references and public data repositories.** Cohort sizes and prediction targets in Table 4 refer to the subsets analyzed in this study.

Cohort	Dataset reference	Repository / accession
COVID	Stephenson et al. (2021)	ArrayExpress: E-MTAB-10026
NS	Ziegler et al. (2021)	Single Cell Portal: SCP1289
UC	Smillie et al. (2019)	Single Cell Portal: SCP259
Lupus	Mandric et al. (2020)	`yelabucsf/lupus_1M_cells_clean`

(1–5) as the random seed. This yields approximately 60%/20%/20% training/validation/test partitions. All cells from a patient remain in the same partition. The same patient partitions are used for autoencoder (AE) pretraining, MIL training, and comparisons across methods.

To examine performance with fewer training labels, we additionally evaluate the reduced-training **Lupus-10** setting, which reuses the same 169-patient Lupus cohort with nominal training/validation/test proportions of 10%/10%/80%. For each random seed in 1–5, two successive stratified splits first separate the test patients and then divide the remaining patients into training and validation sets, yielding 16 training, 17 validation, and 136 test patients (Table 6). Unlike the standard cross-validation protocol, the test sets overlap across repetitions.

Table 6: Class composition of the patient-level partitions. Each entry reports class 0 / class 1 bag counts. Rows represent cross-validation folds for the standard settings and repeated stratified splits for Lupus-10. Standard test folds are disjoint; Lupus-10 test sets overlap across repetitions.

	COVID			NS			UC			Lupus			Lupus-10		
Split	Train	Val	Test	Train	Val	Test	Train	Val	Test	Train	Val	Test	Train	Val	Test
1	20/34	7/11	5/13	9/21	3/7	3/7	8/10	2/4	2/4	28/73	10/24	12/22	5/11	5/12	40/96
2	19/35	6/12	7/11	10/20	3/7	2/8	7/11	3/3	2/4	30/71	10/24	10/24	5/11	5/12	40/96
3	17/37	6/12	9/9	9/21	3/7	3/7	7/11	3/3	2/4	31/70	10/24	9/25	5/11	5/12	40/96
4	21/33	7/11	4/14	8/22	3/7	4/6	7/11	2/4	3/3	29/72	10/24	11/23	5/11	5/12	40/96
5	19/35	6/12	7/11	9/21	3/7	3/7	7/11	2/4	3/3	32/70	10/24	8/25	5/11	5/12	40/96

B.3 Feature extraction and input dimensions

For the local AE pipelines, we use non-negative count tensors and select highly variable features. The retained input dimensions are set to 1,000 for COVID, 3,000 for NS, 2,000 for UC, and 3,000 for both the Lupus and Lupus-10 settings.

For the foundation model baselines, we extract frozen cell representations via the pretrained pipelines of scGPT (Cui et al., 2024) and State Embedding (SE) (Adduri et al., 2026), utilizing the `scGPT_human` and `se600m_epoch16.ckpt` checkpoints, respectively. This deterministic feature extraction is performed without downstream fine-tuning, yielding embedding dimensions of 512 for scGPT and 2,058 for SE.

B.4 CONSTRUCTION OF SEMANTIC KNOWLEDGE ANCHORS

The KAA module takes a semantic knowledge matrix $T \in \mathbb{R}^{M \times d_t}$ as input, where each row represents a dense embedding of a biological process concept. While this multi-query interface is agnostic to the specific knowledge base, we ground it in Gene Ontology (GO) biological processes to construct our semantic anchors.

GO selection relies on label-free, patient-level expression variability across the entire cohort—including individuals subsequently assigned to validation and test partitions. Crucially, disease labels are strictly excluded from this vocabulary construction process. This transductive access to held-out expression statistics is entirely decoupled from supervised model fitting: the encoder and classifiers are optimized exclusively on training patients, utilizing standard validation-based model selection. We select $M = 50$ terms per cohort and reuse the standard Lupus vocabulary for the cohort-wide Lupus-10 experiments. The specific selection procedure is as follows:

1. **Data Aggregation and Variable Gene Selection:** Single-cell counts are aggregated into patient-level pseudo-bulk profiles, excluding samples with < 10 cells or zero counts, with at least three valid samples required per cohort. Pseudo-bulk counts are counts per million (CPM)-normalized (using pre-filtering library sizes) and $\log_2(1 + \mathrm{CPM})$ transformed. Genes require counts ≥ 10 in at least two valid samples. To account for the mean–variance relationship, genes are stratified into up to 20 equal-frequency bins by mean expression. Bin-wise log variances are robustly standardized via the median and scaled median absolute deviation (reverting to standard deviation if near zero). Finally, we select up to 2,000 top-ranked variable genes, explicitly requiring a ≥ 0.5 difference between the 90th and 10th expression percentiles across samples.
2. **GO Term Enrichment and Filtering:** Candidate biological-process (BP) terms span GO-slim roots (level 0) and their descendants up to level 2. The background set comprises expression-filtered genes with valid BP annotations. Candidates must possess a non-empty definition, map ≥ 10 but $\leq 50\%$ of background genes, maintain $\geq 10\%$ coverage of their annotated genes, intersect with ≥ 3 patient-variable genes, and exhibit a variable-gene enrichment factor ≥ 1.5. Overrepresentation is evaluated via a hypergeometric test. Finally, Benjamini–Hochberg correction is computed globally across all eligible BP terms (explicitly including those beyond level 2), retaining candidates at false discovery rate (FDR) ≤ 0.05.
3. **Redundancy-Aware Greedy Selection:** To balance candidate quality and gene-set redundancy, we employ a greedy selector that, at each iteration, maximizes $0.6 \times$ Quality $- 0.4 \times$ Maximum Jaccard Overlap. Here, the Quality score is a transformed FDR defined as $\frac{-\log_{10}(\mathrm{FDR})}{-\log_{10}(\mathrm{FDR})+2.0}$. Jaccard overlap is computed between the terms' gene sets restricted to the background, taking the maximum overlap with any previously selected term (zero for the first selection). Terms with a maximum Jaccard overlap ≥ 0.85 are excluded, and selection continues until 50 terms have been selected or no eligible candidates remain.

To construct the matrix T, each selected term is formatted as a textual prompt "*name: definition*". We extract text embeddings using the BioLORD model, applying attention-mask-weighted mean pooling to the last hidden state of the transformer, followed by ℓ_2 normalization. The text is tokenized with padding and truncation up to 512 tokens and processed in batches of 16. This yields a precomputed anchor matrix of size 50×768 (i.e., $d_t = 768$). The BioLORD text encoder remains frozen during the downstream MIL training.

Crucially, our query-geometry regularization (Equation 4) operates dynamically on these continuous text embeddings rather than fixed graph distances from the GO hierarchy. By avoiding a strict per-cell annotation loss, this design allows related terms to naturally develop correlated attention profiles. Consequently, the selected GO descriptions define complementary queries for cellular aggregation. Their text-derived similarity structure provides the fixed reference for geometry regularization, allowing related queries to form overlapping cellular summaries.

split-specific GO vocabulary control. For the control in Appendix D.4, the expression-dependent steps above—patient-variable gene selection, enrichment background, term ranking, and vocabulary selection—are restricted to the training patients of each split. Selected descriptions are encoded with

the same frozen BioLORD model, and the resulting anchors are held fixed for that split's validation and test sets. Lupus-10 uses only its 16 training patients per repetition rather than the cohort-wide vocabulary. Hyperparameters remain as in Table 7.

Since this control retains global cell-feature preprocessing (e.g., cohort-wide selection of highly variable genes), it does not establish a fully inductive pipeline. We therefore strictly refer to this setting as *split-specific GO construction*.

C Implementation and Optimization

C.1 AE pretraining and network architectures

For each data split, we first pretrain a negative-binomial (NB) AE on raw single-cell counts from the training partition. The downstream MIL encoder then uses these pretrained weights as its starting point. The encoder architecture follows $G \to 512 \to 256 \to 128 \to 128$, applying sigmoid activations after the three hidden layers and producing a linear 128-dimensional output representation. All pretraining settings utilize the Adam optimizer with a learning rate of 10^{-3}, a training cell batch size of 128, a maximum of 250 epochs, and early stopping with a patience of 15 epochs based on validation NB loss. The pretraining objective minimizes the NB negative log-likelihood:

$$\mathcal{L}_{\mathrm{NB}} = -\frac{1}{G \sum_{i \in \mathcal{I}_{\mathrm{tr}}} N_i} \sum_{i \in \mathcal{I}_{\mathrm{tr}}} \sum_{j=1}^{N_i} \sum_{g=1}^{G} \log \mathrm{NB}(C_{ijg}; \mu_{ijg}, \vartheta_{ijg}). \tag{8}$$

The decoder predicts means and inverse dispersions using exponential and softplus transformations, appropriately bounded to ensure numerical stability. Pretraining fits the training cells, and optimal checkpoints are selected based on the validation NB negative log-likelihood, averaged over validation minibatches. During downstream MIL training, the pretrained encoder is fine-tuned, while the decoder and reconstruction objective are discarded.

In the evaluated configuration, the shared projection dimension for the GO queries and cell keys is set to 256. The GO vocabulary size is $M = 50$, and five residual anchors are allocated with an element-wise dropout probability of 0.5. The anchor scorer employs a 128-dimensional $\tanh$ hidden layer. Both the teacher and student classifier consist of two 64-dimensional $\tanh$ hidden layers yielding two output logits. The learnable log-scale parameter ℓ is initialized as $\log(1/0.07)$. The query projection is fully learnable and includes a bias term. We report scMILD with OPL enabled and disabled; OPL is disabled in all KA-MIL configurations. Fixed text anchors utilize the precomputed BioLORD matrix. Finally, the encoder and the teacher–student training backbones inherit their core architectures from scMILD (Jeong et al., 2026).

C.2 KAA architecture and forward computation

Unless otherwise specified in the ablations, the full KA-MIL model uses $M = 50$ fixed semantic text anchors generated by BioLORD (Remy et al., 2024) alongside $r = 5$ randomly initialized, learnable residual anchors. To regularize task adaptation, element-wise dropout ($p = 0.5$) is applied exclusively to the residual anchors prior to concatenation. The unified queries are generated via a shared linear projection (Linear_Q) that includes a learnable bias but no non-linear activation. The cellular key network (MLP_K) consists of two linear layers with dimensions $d_c \to d_s \to d_s$, interleaved with a $\tanh$ activation, where the latent attention dimension is set to $d_s = 256$.

For the cross-attention mechanism (Equation 2), the inverse-temperature scale is parameterized as $s = \min(e^{\ell}, 100)$ to ensure numerical stability. Following CLIP (Radford et al., 2021), the globally trainable log-scale ℓ is initialized to $\log(1/0.07)$. When computing the attention matrix A_i, the encoded cell features serve directly as values. The anchor scorer ($\mathrm{MLP}_{\mathrm{score}}$) is structured as a two-layer network with dimensions $d_c \to 128 \to 1$ and a $\tanh$ hidden activation, which maps each cellular summary to a scalar score. For the final predictive stages, both the teacher and student classifiers share identical architectures, each comprising two 64-dimensional $\tanh$ hidden layers and terminating in two output logits.

The text anchors and their reference similarity matrix remain fixed during training. The query and key projections, residual anchors, anchor scorer, shared cell encoder, and classifiers are train-

able. KA-MIL uses query-geometry regularization; scMILD's cell-level OPL is evaluated separately through the +OPL and –OPL controls.

C.3 Exact auxiliary student target

For cell j, let X_j denote its shared-encoder representation and $q_j = \mathrm{softmax}(g_\omega(X_j))_1$ denote the student's positive-class probability. For an input set S, let A_{Smj} denote the attention of query m to cell j, normalized over the cells in S as in Equation 2. Let z_{Sm} be the corresponding query summary and $u_{Sm} = \mathrm{MLP}_{\mathrm{score}}(z_{Sm})$ its anchor score. The semantic cell score is

$$a_{Sj} = \sum_{m=1}^{M} \beta_{Sm} A_{Smj}, \qquad \beta_{Sm} = \frac{\exp(u_{Sm})}{\sum_{k=1}^{M+r} \exp(u_{Sk})}. \tag{9}$$

The weights β_{Sm} are normalized over all $M + r$ queries and are not renormalized over the semantic subset. Only the first M attention maps contribute directly to a_{Sj}.

During epoch e, the teacher training pass collects these cell scores for the training bags. Let $\mathcal{S}_e$ be the collection of scores recorded from the forward passes used for successive teacher updates. The scaling bounds are

$$l_e = \min \mathcal{S}_e, \qquad u_e = \max \mathcal{S}_e.$$

During the student pass, the teacher processes each cell minibatch B as a single input set, which may contain cells from multiple patients. It recomputes the attention maps and anchor weights on B using its current parameters. Writing these recomputed values as $\widetilde{A}_{Bmj}$ and $\widetilde{\beta}_{Bm}$, the cell score is $\widetilde{a}_{Bj} = \sum_{m=1}^{M} \widetilde{\beta}_{Bm} \widetilde{A}_{Bmj}$, with the same normalization over all queries as in Equation 9. With $i(j)$ denoting the patient of origin of cell j, the auxiliary target is

$$t_{Bj} = y_{i(j)} \,\mathrm{clip}\left(\frac{\widetilde{a}_{Bj} - l_e}{u_e - l_e + 10^{-6}}, 10^{-5}, 1 - 10^{-5} \right). \tag{10}$$

The same bounds are used throughout the student pass. Multiplication by the patient label occurs after clipping: cells from negative bags receive zero targets, while cells from positive bags retain the clipped scores. With $\epsilon = 10^{-5}$, the weighted cell-level objective is

$$\mathcal{L}_{\mathrm{cell}} = -\frac{1}{|B|} \sum_{j \in B} \left[\eta \big(1 - \mathrm{sg}(t_{Bj})\big) \log(1 - q_j + \epsilon) + (1 - \eta) \mathrm{sg}(t_{Bj}) \log(q_j + \epsilon) \right], \tag{11}$$

where η weights the negative-class term and sg denotes stop gradient. OPL is disabled in the KA-MIL configuration described here.

The teacher, including KAA, and the shared encoder are updated using the bag-plus-geometry objective each epoch. After the teacher pass at epochs $e \in \{0, k, 2k, \ldots\}$, the student and shared encoder are updated using $\mathcal{L}_{\mathrm{cell}}$. Target generation is performed without gradient tracking while the teacher remains in training mode. Residual-anchor dropout therefore remains active during target recomputation. Although residual attention maps are excluded from the cell-score sum, residual queries participate in the shared softmax denominator of β; their dropout can thus affect the semantic cell scores indirectly through the anchor weights. Gradients through q_j update the student and shared encoder, and the same encoder optimizer and its state are used for both training phases. This procedure combines epoch-level scaling bounds with cell scores recomputed on each student minibatch to provide cell-level auxiliary supervision.

C.4 Selected training profiles and hyperparameter configurations

Table 7 specifies the cohort-specific profiles used for the local AE pipelines. These profiles are shared by ABMIL (attention-based multiple instance learning), both scMILD settings, and KA-MIL; the geometry constraint applies only to KA-MIL. The split-specific GO experiment retains the same profile and geometry weight as its cohort-wide counterpart, without an additional hyperparameter search. Hier-MIL uses the separate validation-based search described in Appendix D.3.

The bag-batch sizing rule is $\max(2, \lfloor n_{\text{training bags}}/10 \rfloor)$, resulting in sizes of 5, 3, 2, 10, and 2 for COVID, NS, UC, Lupus, and Lupus-10, respectively. Matching scMILD and ABMIL baselines strictly adhere to these same computed sizes.

Table 7: Selected hyperparameter configurations across datasets. η weights the negative-class term of the cell-level loss, k specifies the student update period, and learning rates are ordered as encoder/teacher/student. The maximum epoch limit is capped at 500 on all AE-based pipelines. Feature counts refer to the input dimensionality of the local encoders.

Cohort	Features	Cell batch	k	η	Patience	Learning rates	λ
COVID	1,000	256	3	0.3	15	$10^{-4}/5\times10^{-4}/10^{-3}$	1.00
NS	3,000	256	5	0.3	15	$10^{-3}/10^{-4}/10^{-3}$	0.75
UC	2,000	256	4	0.3	15	$5\times10^{-4}/5\times10^{-4}/10^{-3}$	0.50
Lupus	3,000	512	5	0.1	45	$10^{-3}/5\times10^{-4}/10^{-3}$	0.75
Lupus-10	3,000	512	5	0.1	45	$10^{-3}/5\times10^{-4}/10^{-3}$	0.75

D Evaluation Details and Complete Split-level Results

D.1 Checkpoint selection and decision rules

The model checkpoint is selected by minimizing the sum of the validation cross-entropy and $1 - \mathrm{AUC}$. For the chosen checkpoint, the optimal decision threshold t maximizes the validation macro-F1 score over the candidate set:

$$\mathcal{T} = \{p_v\}_v \cup \left\{0.5, \mathrm{nextafter}(\max_v p_v, +\infty)\right\} \tag{12}$$

where p_v is a validation patient's positive-class probability. Undefined class-specific F1 scores are set to zero. Threshold ties within 10^{-12} are resolved by proximity to 0.5, then by choosing the smaller threshold. The chosen threshold is applied unchanged to the test set using $\hat{y} = \mathbb{I}(p_{\mathrm{test}} \geq t)$.

Tables 10–11 detail the validation and split-level test results. For controlled comparisons, the scMILD and ABMIL baselines share our local optimization settings, while the foundation models (scGPT and SE) utilize identical downstream architectures, differing only in their frozen backbones. Note that all reported aggregate metrics are computed as the arithmetic mean across splits without pooling fold predictions, and the standard deviations for Lupus-10 reflect variation over overlapping partitions.

D.2 Baseline implementations and OPL control

All local pipelines (ABMIL, scMILD, and KA-MIL) share the identical local AE initialization, cohort-specific optimization profiles (Table 7), and saved patient partitions (Table 6). For ABMIL, we utilize patient-level attention training without student updates. Both scMILD variants retain their original attention pooler, shared encoder, and teacher–student optimization scheme. In scMILD (+OPL), the auxiliary orthogonal projection loss is enabled exclusively during student-branch optimization, configured with a weight of 0.5, $\gamma = 0.5$, and two a Gaussian mixture model (GMM) components. scMILD (–OPL) and all KA-MIL variants disable this auxiliary objective.

For the foundation model baselines (scGPT + MIL and SE + MIL), the respective pretrained cell embeddings are strictly frozen and not fine-tuned. The downstream patient-level attention predictors share the identical architecture and validation-based checkpoint/threshold selection rules as our local implementations.

D.3 Reproduction of the Hier-MIL baseline

We reproduce Hier-MIL (hierarchical attention, HA) using the authors' model implementation (Do & Lähdesmäki, 2025), with frozen 512-dimensional scGPT embeddings and the provided cell-type annotations. The HA network is trained from random initialization on the same patient level train/validation/test partitions used for the other methods.

For each saved split, we conduct 30 Optuna tree-structured Parzen estimator (TPE) trials (seed 0) over the authors' search space: hidden-layer counts $\{1, 2\}$, hidden widths $\{32, 64, 128\}$, dropout rates $\{0, 0.3, 0.5, 0.7\}$, learning rates $\{10^{-3}, 5 \times 10^{-3}\}$, weight decay values $\{10^{-4}, 10^{-3}, 10^{-2}\}$, and epoch caps $\{100, 500, 1000\}$. Configurations are selected using the same validation criterion as checkpoints.

D.4 Results on GO construction scope

Table 8: **GO construction scope.** Test mean ± sample SD over five splits, with OPL disabled in both KA-MIL settings. Cohort-wide construction uses all patients' expression without disease labels; split-specific construction uses only training patients for vocabulary selection.

GO selection	AUC	Macro-F1	ACC	Sens.	Spec.
COVID					
Cohort-wide	0.7773 ± 0.1015	0.7075 ± 0.0657	0.7444 ± 0.0633	0.8132 ± 0.1689	0.6497 ± 0.2329
split-specific	0.7761 ± 0.1160	0.6333 ± 0.0746	0.6667 ± 0.0878	0.7321 ± 0.1981	0.5889 ± 0.2365
NS					
Cohort-wide	0.9274 ± 0.0834	0.8679 ± 0.1407	0.9000 ± 0.1000	0.9179 ± 0.0753	0.8333 ± 0.2357
split-specific	0.8512 ± 0.0924	0.7649 ± 0.1541	0.8000 ± 0.1414	0.8036 ± 0.1646	0.7667 ± 0.2236
UC					
Cohort-wide	1.0000 ± 0.0000	0.9657 ± 0.0767	0.9667 ± 0.0745	0.9500 ± 0.1118	1.0000 ± 0.0000
split-specific	1.0000 ± 0.0000	0.9657 ± 0.0767	0.9667 ± 0.0745	0.9500 ± 0.1118	1.0000 ± 0.0000
Lupus					
Cohort-wide	1.0000 ± 0.0000	0.9914 ± 0.0193	0.9939 ± 0.0136	1.0000 ± 0.0000	0.9750 ± 0.0559
split-specific	1.0000 ± 0.0000	0.9845 ± 0.0215	0.9881 ± 0.0164	0.9917 ± 0.0186	0.9750 ± 0.0559
Lupus-10					
Cohort-wide	0.9870 ± 0.0132	0.9301 ± 0.0466	0.9397 ± 0.0427	0.9396 ± 0.0568	0.9400 ± 0.0285
split-specific	0.9845 ± 0.0147	0.9152 ± 0.0618	0.9250 ± 0.0578	0.9167 ± 0.0824	0.9450 ± 0.0274

Table 8 details the complete performance metrics comparing cohort-wide against split-specific GO construction. Incorporating cohort-wide statistical priors yields equal or higher mean AUC and macro-F1 across all five evaluation settings. The two metrics reveal distinct aspects of this comparison. On COVID, mean AUC remains comparable across the two scopes (0.7773 versus 0.7761), while macro-F1 is higher under cohort-wide construction (0.7075 versus 0.6333). On NS, cohort-wide construction achieves substantially higher values for both AUC (0.9274 versus 0.8512) and macro-F1 (0.8679 versus 0.7649).

Both construction scopes achieve a mean AUC of 1.0000 on UC and Lupus, maintaining identical aggregate test metrics on UC. On Lupus-10, cohort-wide construction reaches an AUC of 0.9870 and a macro-F1 of 0.9301, compared with 0.9845 and 0.9152 under the split-specific protocol. Together, these results characterize how localized vocabulary constraints influence ranking capacity and thresholded classification across diverse cohorts.

On NS, the two GO construction protocols yield matching aggregate validation AUC and macro-F1 scores (Table 10), while cohort-wide construction achieves higher mean test AUC and macro-F1. The comparison holds patient partitions, optimization hyperparameters, and checkpoint- and threshold-selection rules fixed, with vocabulary construction scope as the experimental intervention. Changing the vocabulary changes the semantic anchors and can affect the fitted model; matching aggregate validation metrics do not imply identical predicted probabilities, selected checkpoints, or numerical thresholds. These results support the contribution of cohort-wide vocabulary construction to NS classification under the shared training and selection protocol.

D.5 Complete aggregate and split-level results

Table 9 details the complete suite of test metrics, including accuracy (ACC), sensitivity (Sens.), and specificity (Spec.). Table 10 reports validation performance across all eight configurations and five evaluation settings used for model selection. Finally, Table 11 provides the split-level AUC and macro-F1 results.

Table 9: **Complete test metrics.** Results are reported as the mean ± sample SD across five fits. Standard cohorts are evaluated on distinct patient-level folds, whereas Lupus-10 utilizes overlapping repeated splits. KA-MIL employs cohort-wide GO anchors in this comparison. The highest unrounded means for each metric are marked in **bold**.

Model	AUC	Macro-F1	ACC	Sens.	Spec.
COVID					
ABMIL	0.7292 ± 0.1236	0.6061 ± 0.1013	0.6556 ± 0.0994	0.7682 ± 0.1779	0.5159 ± 0.3070
scMILD (+OPL)	0.6327 ± 0.0861	0.5612 ± 0.0799	0.6222 ± 0.0824	0.7691 ± 0.1535	0.4287 ± 0.2534
scMILD (–OPL)	0.7213 ± 0.1286	0.5789 ± 0.0814	0.6222 ± 0.0994	0.6058 ± 0.2895	0.6270 ± 0.2533
scGPT + MIL	0.7750 ± 0.1290	0.6233 ± 0.1188	0.6444 ± 0.1217	0.6423 ± 0.2329	0.6683 ± 0.2016
SE + MIL	0.7606 ± 0.1324	0.6339 ± 0.1133	0.6556 ± 0.1204	0.6361 ± 0.1770	0.6968 ± 0.2191
Hier-MIL (HA)	0.7578 ± 0.0938	0.6371 ± 0.1505	0.6556 ± 0.1542	0.6130 ± 0.2965	$\mathbf{0.7627 \pm 0.1546}$
KA-MIL	$\mathbf{0.7773 \pm 0.1015}$	$\mathbf{0.7075 \pm 0.0657}$	$\mathbf{0.7444 \pm 0.0633}$	$\mathbf{0.8132 \pm 0.1689}$	0.6497 ± 0.2329
NS					
ABMIL	$\mathbf{0.9399 \pm 0.0395}$	0.8013 ± 0.0843	0.8400 ± 0.0548	0.8560 ± 0.1021	0.7667 ± 0.2236
scMILD (+OPL)	0.8923 ± 0.1068	0.8061 ± 0.1576	0.8400 ± 0.1342	0.8607 ± 0.1751	0.8000 ± 0.2981
scMILD (–OPL)	0.8798 ± 0.1091	0.7823 ± 0.1317	0.8200 ± 0.1095	0.8321 ± 0.1208	0.7667 ± 0.2236
scGPT + MIL	0.7905 ± 0.0783	0.6892 ± 0.1853	0.7200 ± 0.1789	0.7250 ± 0.2976	0.7167 ± 0.1826
SE + MIL	0.8363 ± 0.1187	0.8112 ± 0.1491	0.8400 ± 0.1517	0.8571 ± 0.2474	0.7833 ± 0.2173
Hier-MIL (HA)	0.7429 ± 0.2511	0.6298 ± 0.2755	0.7600 ± 0.1949	0.9143 ± 0.1278	0.3667 ± 0.4150
KA-MIL	0.9274 ± 0.0834	$\mathbf{0.8679 \pm 0.1407}$	$\mathbf{0.9000 \pm 0.1000}$	$\mathbf{0.9179 \pm 0.0753}$	$\mathbf{0.8333 \pm 0.2357}$
UC					
ABMIL	$\mathbf{1.0000 \pm 0.0000}$	0.9333 ± 0.1491	0.9333 ± 0.1491	0.9000 ± 0.2236	$\mathbf{1.0000 \pm 0.0000}$
scMILD (+OPL)	0.9778 ± 0.0497	0.8527 ± 0.0852	0.8667 ± 0.0745	0.9000 ± 0.1369	0.8333 ± 0.2357
scMILD (–OPL)	$\mathbf{1.0000 \pm 0.0000}$	0.8971 ± 0.0939	0.9000 ± 0.0913	0.9000 ± 0.1369	0.9333 ± 0.1491
scGPT + MIL	0.9306 ± 0.1011	0.7879 ± 0.1380	0.8000 ± 0.1394	0.8000 ± 0.2739	0.8333 ± 0.2357
SE + MIL	$\mathbf{1.0000 \pm 0.0000}$	0.9314 ± 0.0939	0.9333 ± 0.0913	$\mathbf{0.9500 \pm 0.1118}$	0.9333 ± 0.1491
Hier-MIL (HA)	0.9056 ± 0.1297	0.8240 ± 0.1777	0.8333 ± 0.1667	0.8333 ± 0.1559	0.8333 ± 0.2357
KA-MIL	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{0.9657 \pm 0.0767}$	$\mathbf{0.9667 \pm 0.0745}$	$\mathbf{0.9500 \pm 0.1118}$	$\mathbf{1.0000 \pm 0.0000}$
Lupus					
ABMIL	0.9960 ± 0.0057	0.9582 ± 0.0280	0.9645 ± 0.0246	0.9655 ± 0.0352	0.9568 ± 0.0603
scMILD (+OPL)	0.9991 ± 0.0020	0.9849 ± 0.0338	0.9882 ± 0.0263	0.9920 ± 0.0179	0.9778 ± 0.0497
scMILD (–OPL)	0.9992 ± 0.0019	0.9841 ± 0.0219	0.9881 ± 0.0164	$\mathbf{1.0000 \pm 0.0000}$	0.9550 ± 0.0622
scGPT + MIL	0.9964 ± 0.0036	0.9566 ± 0.0443	0.9645 ± 0.0383	0.9670 ± 0.0543	0.9528 ± 0.0648
SE + MIL	0.9985 ± 0.0034	0.9737 ± 0.0266	0.9765 ± 0.0246	0.9731 ± 0.0404	$\mathbf{0.9800 \pm 0.0447}$
Hier-MIL (HA)	0.9939 ± 0.0050	0.9374 ± 0.0544	0.9469 ± 0.0482	0.9416 ± 0.0692	0.9528 ± 0.0648
KA-MIL	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{0.9914 \pm 0.0193}$	$\mathbf{0.9939 \pm 0.0136}$	$\mathbf{1.0000 \pm 0.0000}$	0.9750 ± 0.0559
Lupus-10					
ABMIL	0.9616 ± 0.0276	0.8725 ± 0.0528	0.8897 ± 0.0474	0.8938 ± 0.0669	0.8800 ± 0.0925
scMILD (+OPL)	0.9832 ± 0.0208	0.9213 ± 0.0528	0.9353 ± 0.0445	0.9604 ± 0.0450	0.8750 ± 0.0952
scMILD (–OPL)	0.9788 ± 0.0128	0.8684 ± 0.0398	0.8912 ± 0.0300	0.9146 ± 0.0727	0.8350 ± 0.1607
scGPT + MIL	0.9056 ± 0.0888	0.8328 ± 0.0504	0.8735 ± 0.0346	$\mathbf{0.9625 \pm 0.0291}$	0.6600 ± 0.1207
SE + MIL	0.9693 ± 0.0410	0.8787 ± 0.0949	0.9029 ± 0.0777	0.9437 ± 0.0876	0.8050 ± 0.1956
Hier-MIL (HA)	0.9221 ± 0.0415	0.7781 ± 0.0806	0.7926 ± 0.0884	0.7479 ± 0.1483	0.9000 ± 0.0884
KA-MIL	$\mathbf{0.9870 \pm 0.0132}$	$\mathbf{0.9301 \pm 0.0466}$	$\mathbf{0.9397 \pm 0.0427}$	0.9396 ± 0.0568	$\mathbf{0.9400 \pm 0.0285}$

Table 10: **Complete validation metrics.** Results are reported as the mean ± sample SD across five fits. All results use the validation partitions of the five saved patient-level splits. Validation sets can overlap across splits; these metrics are used for model selection.

Model	AUC	Macro-F1	ACC	Sens.	Spec.
COVID					
ABMIL	0.7055 ± 0.0467	0.7206 ± 0.0518	0.7667 ± 0.0465	0.8773 ± 0.1321	0.5571 ± 0.2073

Continued on next page

Table 10 (continued)

Model	AUC	Macro-F1	ACC	Sens.	Spec.
scMILD (+OPL)	0.8127 ± 0.0731	0.8154 ± 0.0854	0.8444 ± 0.0724	0.9288 ± 0.1183	0.6762 ± 0.1787
scMILD (–OPL)	0.7930 ± 0.0738	0.7608 ± 0.0633	0.7889 ± 0.0609	0.8455 ± 0.1600	0.6905 ± 0.2393
scGPT + MIL	0.7415 ± 0.1036	0.7452 ± 0.0839	0.7667 ± 0.0913	0.7955 ± 0.1936	0.7143 ± 0.2027
SE + MIL	0.7363 ± 0.1261	0.7425 ± 0.1267	0.7667 ± 0.1204	0.8318 ± 0.1549	0.6524 ± 0.1343
Hier-MIL (HA)	0.8437 ± 0.1062	0.8349 ± 0.0703	0.8444 ± 0.0609	0.8106 ± 0.0347	0.9000 ± 0.1491
KA-MIL	0.8058 ± 0.0941	0.7805 ± 0.0793	0.8111 ± 0.0633	0.8788 ± 0.1405	0.6810 ± 0.2100
KA-MIL (split-specific)	0.7463 ± 0.1220	0.7556 ± 0.0846	0.7889 ± 0.0724	0.8591 ± 0.1035	0.6524 ± 0.2501
NS					
ABMIL	0.9714 ± 0.0426	0.9560 ± 0.0602	0.9600 ± 0.0548	0.9429 ± 0.0782	1.0000 ± 0.0000
scMILD (+OPL)	0.9714 ± 0.0426	0.9560 ± 0.0602	0.9600 ± 0.0548	0.9429 ± 0.0782	1.0000 ± 0.0000
scMILD (–OPL)	0.9905 ± 0.0213	0.9780 ± 0.0491	0.9800 ± 0.0447	0.9714 ± 0.0639	1.0000 ± 0.0000
scGPT + MIL	0.9143 ± 0.0782	0.8900 ± 0.1050	0.9000 ± 0.1000	0.8857 ± 0.1565	0.9333 ± 0.1491
SE + MIL	0.9048 ± 0.1010	0.8861 ± 0.1250	0.9000 ± 0.1225	0.9143 ± 0.1917	0.8667 ± 0.1826
Hier-MIL (HA)	0.9524 ± 0.0673	0.9467 ± 0.0730	0.9600 ± 0.0548	1.0000 ± 0.0000	0.8667 ± 0.1826
KA-MIL	0.9905 ± 0.0213	0.9780 ± 0.0491	0.9800 ± 0.0447	0.9714 ± 0.0639	1.0000 ± 0.0000
KA-MIL (split-specific)	0.9905 ± 0.0213	0.9780 ± 0.0491	0.9800 ± 0.0447	0.9714 ± 0.0639	1.0000 ± 0.0000
UC					
ABMIL	0.9306 ± 0.1011	0.9314 ± 0.0939	0.9333 ± 0.0913	0.8833 ± 0.1624	1.0000 ± 0.0000
scMILD (+OPL)	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000
scMILD (–OPL)	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000
scGPT + MIL	0.9083 ± 0.0940	0.8971 ± 0.0939	0.9000 ± 0.0913	0.8833 ± 0.1624	0.9333 ± 0.1491
SE + MIL	0.9778 ± 0.0497	0.9657 ± 0.0767	0.9667 ± 0.0745	1.0000 ± 0.0000	0.9333 ± 0.1491
Hier-MIL (HA)	0.9750 ± 0.0559	0.9657 ± 0.0767	0.9667 ± 0.0745	0.9500 ± 0.1118	1.0000 ± 0.0000
KA-MIL	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000
KA-MIL (split-specific)	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000
Lupus					
ABMIL	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000
scMILD (+OPL)	0.9967 ± 0.0035	0.9789 ± 0.0193	0.9824 ± 0.0161	0.9833 ± 0.0228	0.9800 ± 0.0447
scMILD (–OPL)	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000
scGPT + MIL	0.9958 ± 0.0072	0.9862 ± 0.0189	0.9882 ± 0.0161	0.9833 ± 0.0228	1.0000 ± 0.0000
SE + MIL	0.9775 ± 0.0503	0.9781 ± 0.0490	0.9824 ± 0.0395	0.9917 ± 0.0186	0.9600 ± 0.0894
Hier-MIL (HA)	0.9992 ± 0.0019	0.9931 ± 0.0154	0.9941 ± 0.0132	0.9917 ± 0.0186	1.0000 ± 0.0000
KA-MIL	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000
KA-MIL (split-specific)	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000
Lupus-10					
ABMIL	0.9467 ± 0.0519	0.8967 ± 0.0817	0.9059 ± 0.0789	0.8833 ± 0.1264	0.9600 ± 0.0894
scMILD (+OPL)	0.9833 ± 0.0289	0.9582 ± 0.0630	0.9647 ± 0.0526	0.9667 ± 0.0456	0.9600 ± 0.0894
scMILD (–OPL)	0.9767 ± 0.0435	0.9539 ± 0.0718	0.9647 ± 0.0526	0.9833 ± 0.0373	0.9200 ± 0.1789
scGPT + MIL	0.8500 ± 0.1599	0.8467 ± 0.1410	0.8941 ± 0.0967	0.9833 ± 0.0373	0.6800 ± 0.3033
SE + MIL	0.9333 ± 0.0773	0.9129 ± 0.0975	0.9294 ± 0.0767	0.9500 ± 0.0456	0.8800 ± 0.1789
Hier-MIL (HA)	0.9233 ± 0.0855	0.9006 ± 0.1107	0.9059 ± 0.1069	0.8667 ± 0.1514	1.0000 ± 0.0000
KA-MIL	0.9800 ± 0.0274	0.9448 ± 0.0589	0.9529 ± 0.0492	0.9500 ± 0.0456	0.9600 ± 0.0894
KA-MIL (split-specific)	0.9767 ± 0.0279	0.9448 ± 0.0589	0.9529 ± 0.0492	0.9500 ± 0.0456	0.9600 ± 0.0894

Table 11: Split-level test AUC and macro-F1 for all eight configurations. Values are on the 0–1 scale. Standard cohorts use disjoint test folds; the five Lupus-10 columns denote repeated splits with overlapping test sets. Test scores use validation-selected checkpoints and thresholds.

Setting	Metric	Split 1	Split 2	Split 3	Split 4	Split 5	Mean	SD
ABMIL								
COVID	AUC	0.9231	0.5844	0.6790	0.7321	0.7273	0.7292	0.1236
COVID	Macro-F1	0.7662	0.5500	0.5000	0.5846	0.6296	0.6061	0.1013
NS	AUC	0.9048	0.9375	0.9048	1.0000	0.9524	0.9399	0.0395
NS	Macro-F1	0.7619	0.6875	0.8667	0.8990	0.7917	0.8013	0.0843

Continued on next page

Table 11 (continued)

Setting	Metric	Split 1	Split 2	Split 3	Split 4	Split 5	Mean	SD
UC	AUC	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000
UC	Macro-F1	1.0000	1.0000	0.6667	1.0000	1.0000	0.9333	0.1491
Lupus	AUC	1.0000	0.9917	1.0000	0.9881	1.0000	0.9960	0.0057
Lupus	Macro-F1	0.9684	0.9328	1.0000	0.9328	0.9569	0.9582	0.0280
Lupus-10	AUC	0.9865	0.9422	0.9229	0.9802	0.9763	0.9616	0.0276
Lupus-10	Macro-F1	0.9393	0.8300	0.8254	0.9185	0.8491	0.8725	0.0528
scMILD (+OPL)								
COVID	AUC	0.7385	0.6364	0.5062	0.6071	0.6753	0.6327	0.0861
COVID	Macro-F1	0.6727	0.5786	0.4582	0.5786	0.5179	0.5612	0.0799
NS	AUC	0.7143	0.9375	0.9048	1.0000	0.9048	0.8923	0.1068
NS	Macro-F1	0.6000	0.8667	0.8667	1.0000	0.6970	0.8061	0.1576
UC	AUC	1.0000	1.0000	1.0000	1.0000	0.8889	0.9778	0.0497
UC	Macro-F1	0.8286	0.7778	0.8286	1.0000	0.8286	0.8527	0.0852
Lupus	AUC	1.0000	1.0000	0.9956	1.0000	1.0000	0.9991	0.0020
Lupus	Macro-F1	1.0000	1.0000	0.9244	1.0000	1.0000	0.9849	0.0338
Lupus-10	AUC	0.9883	0.9945	0.9464	0.9943	0.9927	0.9832	0.0208
Lupus-10	Macro-F1	0.8952	0.9281	0.8450	0.9728	0.9656	0.9213	0.0528
scMILD (–OPL)								
COVID	AUC	0.8615	0.6364	0.5679	0.6964	0.8442	0.7213	0.1286
COVID	Macro-F1	0.5556	0.5786	0.4582	0.6296	0.6727	0.5789	0.0814
NS	AUC	0.7143	0.8750	0.8571	1.0000	0.9524	0.8798	0.1091
NS	Macro-F1	0.7619	0.6875	0.6703	1.0000	0.7917	0.7823	0.1317
UC	AUC	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000
UC	Macro-F1	1.0000	0.8286	0.8286	1.0000	0.8286	0.8971	0.0939
Lupus	AUC	1.0000	0.9958	1.0000	1.0000	1.0000	0.9992	0.0019
Lupus	Macro-F1	1.0000	0.9635	1.0000	1.0000	0.9569	0.9841	0.0219
Lupus-10	AUC	0.9854	0.9701	0.9609	0.9919	0.9854	0.9788	0.0128
Lupus-10	Macro-F1	0.8161	0.8980	0.8545	0.8564	0.9172	0.8684	0.0398
scGPT + MIL								
COVID	AUC	0.9692	0.7403	0.6914	0.6429	0.8312	0.7750	0.1290
COVID	Macro-F1	0.6625	0.4985	0.6099	0.5418	0.8036	0.6233	0.1188
NS	AUC	0.7143	0.7500	0.7619	0.9167	0.8095	0.7905	0.0783
NS	Macro-F1	0.4949	0.5238	0.8667	0.8901	0.6703	0.6892	0.1853
UC	AUC	0.8750	1.0000	1.0000	1.0000	0.7778	0.9306	0.1011
UC	Macro-F1	0.6667	0.7778	0.6667	1.0000	0.8286	0.7879	0.1380
Lupus	AUC	1.0000	0.9917	0.9956	1.0000	0.9950	0.9964	0.0036
Lupus	Macro-F1	1.0000	0.9014	0.9244	1.0000	0.9569	0.9566	0.0443
Lupus-10	AUC	0.9341	0.7539	0.9047	0.9701	0.9654	0.9056	0.0888
Lupus-10	Macro-F1	0.8349	0.7525	0.8282	0.8857	0.8627	0.8328	0.0504
SE + MIL								
COVID	AUC	0.9077	0.7143	0.6420	0.6429	0.8961	0.7606	0.1324
COVID	Macro-F1	0.5556	0.6000	0.5556	0.6296	0.8286	0.6339	0.1133
NS	AUC	0.8571	0.6875	0.7619	0.8750	1.0000	0.8363	0.1187
NS	Macro-F1	0.6000	0.8039	0.7619	0.8901	1.0000	0.8112	0.1491
UC	AUC	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000
UC	Macro-F1	1.0000	1.0000	0.8286	1.0000	0.8286	0.9314	0.0939
Lupus	AUC	0.9924	1.0000	1.0000	1.0000	1.0000	0.9985	0.0034
Lupus	Macro-F1	0.9377	0.9635	1.0000	0.9671	1.0000	0.9737	0.0266
Lupus-10	AUC	0.9987	0.9302	0.9190	1.0000	0.9984	0.9693	0.0410
Lupus-10	Macro-F1	0.8120	0.8189	0.7978	0.9825	0.9820	0.8787	0.0949
Hier-MIL (HA)								
COVID	AUC	0.7846	0.7922	0.6049	0.7500	0.8571	0.7578	0.0938
COVID	Macro-F1	0.4375	0.7500	0.6000	0.5786	0.8194	0.6371	0.1505
NS	AUC	0.5238	1.0000	0.7143	1.0000	0.4762	0.7429	0.2511
NS	Macro-F1	0.6000	0.8039	0.4118	1.0000	0.3333	0.6298	0.2755

Continued on next page

Table 11 (continued)

Setting	Metric	Split 1	Split 2	Split 3	Split 4	Split 5	Mean	SD
UC	AUC	1.0000	0.7500	1.0000	1.0000	0.7778	0.9056	0.1297
UC	Macro-F1	0.8286	0.6250	1.0000	1.0000	0.6667	0.8240	0.1777
Lupus	AUC	0.9962	0.9917	0.9867	1.0000	0.9950	0.9939	0.0050
Lupus	Macro-F1	0.9684	0.8712	0.8904	1.0000	0.9569	0.9374	0.0544
Lupus-10	AUC	0.9133	0.8531	0.9437	0.9487	0.9516	0.9221	0.0415
Lupus-10	Macro-F1	0.8179	0.6570	0.8245	0.8545	0.7364	0.7781	0.0806
KA-MIL								
COVID	AUC	0.8308	0.8052	0.6173	0.7500	0.8831	0.7773	0.1015
COVID	Macro-F1	0.8036	0.6667	0.7143	0.7231	0.6296	0.7075	0.0657
NS	AUC	0.8095	0.8750	1.0000	1.0000	0.9524	0.9274	0.0834
NS	Macro-F1	0.7619	0.6875	1.0000	1.0000	0.8901	0.8679	0.1407
UC	AUC	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000
UC	Macro-F1	1.0000	1.0000	0.8286	1.0000	1.0000	0.9657	0.0767
Lupus	AUC	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000
Lupus	Macro-F1	1.0000	1.0000	1.0000	1.0000	0.9569	0.9914	0.0193
Lupus-10	AUC	0.9974	0.9974	0.9698	0.9758	0.9945	0.9870	0.0132
Lupus-10	Macro-F1	0.9567	0.9732	0.8522	0.9376	0.9311	0.9301	0.0466
KA-MIL (split-specific GO)								
COVID	AUC	0.9077	0.7532	0.6173	0.7321	0.8701	0.7761	0.1160
COVID	Macro-F1	0.6099	0.6990	0.5500	0.5846	0.7231	0.6333	0.0746
NS	AUC	0.7619	0.8750	0.8095	1.0000	0.8095	0.8512	0.0924
NS	Macro-F1	0.7619	0.6875	0.5833	1.0000	0.7917	0.7649	0.1541
UC	AUC	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000
UC	Macro-F1	1.0000	1.0000	0.8286	1.0000	1.0000	0.9657	0.0767
Lupus	AUC	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.0000
Lupus	Macro-F1	1.0000	0.9656	1.0000	1.0000	0.9569	0.9845	0.0215
Lupus-10	AUC	0.9932	0.9982	0.9698	0.9672	0.9940	0.9845	0.0147
Lupus-10	Macro-F1	0.9476	0.9732	0.8332	0.8656	0.9560	0.9152	0.0618

E Matched Ablation Experiments

E.1 Scope, controls, and configurations

Our ablation experiments evaluate the COVID, NS, UC, and Lupus-10 settings, using five patient-level splits per variant. The standard Lupus dataset is excluded because its inherent ceiling effect obscures component-level differences. All configurations utilize identical dataset-specific optimization profiles, fixed patient splits, and threshold selection protocols. KAA ablations retain the cohort-wide GO protocol.

The random-anchor controls utilize Gaussian vectors rescaled to match the row norms of the true reference anchors, initialized using a fold-dependent seed. In the fixed-random-anchor configuration, these anchors are frozen while the query projection remains learnable. In the learnable-random-anchor setting, both the anchor matrix and the projection are jointly optimized, with the geometry penalty computed relative to the saved initial random matrix. Consequently, these controls actively enforce geometric regularization rather than merely disabling the semantic reference loss. Finally, the `go_no_geometry` ablation disables the geometry penalty ($\lambda = 0$) while maintaining the learnable query projection.

F Exploratory Analysis: KAA on Frozen Foundation Embeddings

The frozen-AE ablations in Section 4.4 motivate a further question: can knowledge-guided aggregation make effective use of informative cell representations without adapting their encoder? Although freezing the local AE reduces performance in several settings, frozen scGPT and SE representations coupled with ordinary attention MIL often outperform the frozen-AE baseline (Tables 1 and 3). This suggests that these fixed representations contain task-relevant information accessible to a learned aggregation head.

Table 12: **Definitions of the seven ablation and encoder-control configurations.** Unless otherwise specified, each ablation variant inherits the complete KA-MIL architecture and optimization profile.

Configuration	Implementation
Fixed random anchors (`go_random_fixed`)	Use fixed Gaussian directions with reference-matched row norms; retain geometry relative to this random reference.
Learnable random anchors (`go_random_learnable`)	Train the random anchor matrix; retain its initial, detached matrix as the geometry reference.
Frozen encoder + scMILD (`frozen_ae_scMILD`)	Freeze the pretrained cell encoder; retain the original attention teacher and student.
Frozen encoder + KA-MIL (`frozen_ae_KA-MIL`)	Freeze the pretrained cell encoder; retain the full knowledge pooler and student.
Without residual queries (`no_residual`)	Set the number of residual anchors to zero, retaining the 50 semantic queries.
Without residual dropout (`no_residual_dropout`)	Keep five residual anchors but set their element-wise dropout probability to zero.
Without geometry loss (`go_no_geometry`)	Set the geometry weight to zero while retaining a learned query projection, GO anchors, residuals, and student.

Driven by this observation, we investigate whether KAA can use GO-guided queries to reweight and aggregate frozen cell representations for patient-level prediction, without fine-tuning the foundation-model backbone. The downstream KAA layers and patient classifier remain trainable. For each backbone, we compare KAA with ordinary attention MIL using identical pre-extracted embeddings and patient partitions, evaluating the aggregation pipelines within a fixed representation space. Comparisons with the trainable local-AE pipeline additionally involve differences in encoder architecture, pretraining data, feature dimensionality, and encoder adaptation.

All evaluations use the five patient-level splits from the primary experiments, including repeated stratified splits for Lupus-10. Within the downstream module, KAA's cell-key projection, semantic-query projection, residual anchors, anchor scorer, and geometry objective remain fully active and trainable. By definition, the frozen input features preclude gradient propagation to the encoder via the auxiliary cell objective. Consequently, this configuration effectively isolates the aggregation capabilities of KAA independently of end-to-end feature refinement. Finally, in Table 13, the "+ KAA" notation strictly denotes the integration of our trainable aggregation module atop the respective frozen feature space.

The frozen-AE controls support the contribution of encoder adaptation within the local KA-MIL pipeline; the foundation-model comparisons assess aggregation within fixed representation spaces. Together, these results support combining knowledge-guided aggregation with task-adaptable cell representations and motivate further study of this interaction across pretrained encoders.

G Properties of the Pooler

Permutation invariance. For a cell-permutation matrix P, row-wise encoding gives $X_i' = PX_i$ and $K_i' = PK_i$. Consequently $A_i' = A_iP^\top$, so $Z_i' = A_iP^\top PX_i = Z_i$. The anchor weights, patient representation, and prediction are unchanged. Per-cell outputs follow the permutation.

What geometry preservation constrains. For any orthogonal matrix U, $(Q_TU)(Q_TU)^\top = Q_TQ_T^\top$. Equation 4 therefore regularizes pairwise relationships while leaving a shared rotation unconstrained. It does not specify which regions of cell-feature space correspond to a named process. The task loss supplies the learned interaction with cell keys. The off-diagonal loss also avoids counting the normalized diagonal, which is fixed at one apart from numerical tolerance.

Cell duplication and enrichment. Duplicating every cell in a bag r times repeats its query logits and divides the attention assigned to each copy by r. Query summaries and the patient representation are unchanged. Summing attention over a type and dividing by its cell fraction gives Equation 6, which is also unchanged under uniform duplication. In the ratio of humoral to myeloid enrichment for the same patient and cell type, the N_i/n_{it} factor cancels. The contrast in Equation 7 there-

Table 13: **Exploratory evaluation on frozen foundation embeddings.** Validation and test metrics are reported as mean ± SD across five patient-level splits. To isolate the effect of the aggregation module, each "+ KAA" setting is compared with its corresponding standard attention baseline ("+ MIL") using the same frozen backbone.

		Validation		Test	
Setting	Feature + Pipeline	AUC	Macro-F1	AUC	Macro-F1
COVID	scGPT + MIL	0.742 ± 0.104	0.745 ± 0.084	0.775 ± 0.129	0.623 ± 0.119
	scGPT + KAA	0.867 ± 0.087	0.842 ± 0.041	0.796 ± 0.125	0.627 ± 0.022
	SE + MIL	0.736 ± 0.126	0.743 ± 0.127	0.761 ± 0.132	0.634 ± 0.113
	SE + KAA	0.828 ± 0.072	0.790 ± 0.061	0.708 ± 0.134	0.668 ± 0.087
NS	scGPT + MIL	0.914 ± 0.078	0.890 ± 0.105	0.790 ± 0.078	0.689 ± 0.185
	scGPT + KAA	0.933 ± 0.054	0.903 ± 0.056	0.723 ± 0.251	0.708 ± 0.203
	SE + MIL	0.905 ± 0.101	0.886 ± 0.125	0.836 ± 0.119	0.811 ± 0.149
	SE + KAA	0.819 ± 0.071	0.807 ± 0.076	0.771 ± 0.056	0.634 ± 0.182
UC	scGPT + MIL	0.908 ± 0.094	0.897 ± 0.094	0.931 ± 0.101	0.788 ± 0.138
	scGPT + KAA	0.903 ± 0.104	0.887 ± 0.105	0.856 ± 0.221	0.721 ± 0.248
	SE + MIL	0.978 ± 0.050	0.966 ± 0.077	1.000 ± 0.000	0.931 ± 0.094
	SE + KAA	0.950 ± 0.112	0.956 ± 0.099	1.000 ± 0.000	0.921 ± 0.109
Lupus	scGPT + MIL	0.996 ± 0.007	0.986 ± 0.019	0.996 ± 0.004	0.957 ± 0.044
	scGPT + KAA	0.979 ± 0.019	0.958 ± 0.029	0.972 ± 0.027	0.887 ± 0.060
	SE + MIL	0.978 ± 0.050	0.978 ± 0.049	0.998 ± 0.003	0.974 ± 0.027
	SE + KAA	0.998 ± 0.004	0.993 ± 0.016	0.994 ± 0.009	0.933 ± 0.041
Lupus-10	scGPT + MIL	0.850 ± 0.160	0.847 ± 0.141	0.906 ± 0.089	0.833 ± 0.050
	scGPT + KAA	0.813 ± 0.217	0.810 ± 0.152	0.866 ± 0.090	0.787 ± 0.071
	SE + MIL	0.933 ± 0.077	0.913 ± 0.098	0.969 ± 0.041	0.879 ± 0.095
	SE + KAA	0.970 ± 0.067	0.967 ± 0.073	0.955 ± 0.036	0.835 ± 0.057

fore summarizes relative query allocation rather than directly measuring a difference in cell-type abundance.

Computational scope. For a fixed vocabulary, attention requires $(M + r)N_i$ query–cell scores rather than an $N_i \times N_i$ cell self-attention matrix. The active patient path contains the cell encoder, key/query projections, anchor scorer, and teacher classifier. Precomputed anchors remove the text encoder from this path; the reconstruction decoder and auxiliary student are not required for patient-only inference.

H COVID Figure Construction and Patient Coverage

The case study uses the COVID cohort with hyperparameters in Table 7, with fold-4 query exports for the detailed visualizations and all five folds for the patient-level preference analysis. Heatmaps use per-query attention A. For each patient and query, attention enrichment is computed as a cell type's share of query attention divided by its share of the patient's cells (Equation 6). Figure 2 connects cellular allocation patterns to within-patient query preferences; Figure 4 displays all 50 GO queries and all 18 annotated cell types in the same focal fold.

Display selection and reference query. The focal display combines five immune queries—myeloid leukocyte activation, humoral immune response, activation of innate immune response, inflammatory response, and antigen processing and presentation—with blood vessel development (GO:0001568). The six populations are CD14 monocytes, plasmablasts, B cells, CD4 T cells, CD8 T cells, and CD16-high NK cells (NK_16hi). Selection combined literature-supported comparisons with exploratory inspection of attention patterns and patient coverage.

Blood vessel development is included as a low-contrast reference query selected from the displayed query set. Notably, despite this overall low contrast, the reference consistently maintains strong CD14 enrichment in both patient groups. This CD14 enrichment is consistent with reported contri-

butions of monocytes to new vessel formation (Czepluch et al., 2014), providing biological context for the observed attention allocation.

Aggregation and sparse populations. To ensure robust signal extraction, the heatmap entry (Figure 2) is the median of patient-level $\log_2 E$, strictly requiring at least 20 cells of that type within a patient. Each eligible patient contributes once, with equal weight. In the displayed column order (CD14, plasmablast, B cell, CD4 T, CD8 T, NK_16hi), eligible counts are 4, 3, 4, 4, 4, 4 for patients in the lower severity group and 12, 9, 14, 13, 13, 13 for patients in the higher severity group.

The focal and complete heatmaps share a red–white–blue normalization with endpoints -12 and 5, centered at zero. Values outside these limits receive endpoint colors while retaining their exact numerical values in annotations and exported tables. Numerical logarithms incorporate a 10^{-8} floor at zero. Absent or ineligible observations are strictly omitted from group summaries.

Pair definition and point symbols. Humoral immune response (GO:0006959) and myeloid leukocyte activation (GO:0002274) are retrieved by GO ID. Each patient in the higher severity group contributes its humoral-versus-myeloid preference in plasmablasts and CD14 monocytes; their difference is Δ in Equation 7. To ensure valid contrasts, an observed patient requires both cell types.

Figure 2b and c retain all 14 fold-4 patients in the higher severity group to demonstrate within-patient contrast. Filled points explicitly indicate at least 20 cells of the displayed type; open points indicate 1–19 cells. CD14 meets this robust threshold in 12 patients and plasmablasts in 9. Lines connect the same patient across queries. The reported median E values use all 14 patients; the plotted values and median bars use the $\log_2 E$ scale.

In Figure 2d, filled points require at least 20 cells of both types, and open points indicate fewer than 20 cells of either type. Horizontal bars show median Δ, and counts give positive/observed patients.

Consistency across folds. All 58 test-patient observations in the higher severity group across the five folds contain both populations. Positive Δ occurs in 54 of 58 test patients in this group, with a positive median in every fold. Among observations with at least 20 cells of both types, 36 of 39 confirm this directional preference. Table 14 reports each fold separately, detailing the patient-level distribution underlying the recurring humoral–myeloid attention contrast.

Table 14: **Directional consistency of the within-patient humoral–myeloid attention contrast.** Positive Δ confirms a stronger relative humoral preference in plasmablasts than in CD14 cells. The final column restricts the observation counts to patients having at least 20 cells of both types.

Fold	Higher severity patients	Positive Δ	Median Δ	Positive/eligible, ≥ 20 cells
1	13	13	1.334	8/8
2	11	10	0.874	7/7
3	9	9	1.504	6/6
4	14	14	1.982	8/8
5	11	8	0.178	7/10
Total	58	54	—	36/39



Figure 4: **Complete GO hierarchy and cellular attention across the COVID-19 query set.** Displayed for all 50 GO queries and 18 annotated populations in fold 4, utilizing the patient-median statistic, the 20-cell eligibility threshold, and the color normalization of Figure 2a. The figure presents the GO hierarchy on the left, group heatmaps in the center, and full GO names on the right, with bold text marking the six focal queries. Depths D0–D3 denote display-tree levels, where solid links indicate direct parent-child relations, dashed links bridge unshown terms, and dotted links represent additional parents. Colors show median patient $\log_2 E$, with zero marking attention proportional to cell abundance. Counts below each heatmap indicate the number of eligible patients per cell type. Table 15 expands all cell-population labels.

Table 15: **Cell-population labels in the COVID attention heatmaps.** Labels follow the left-to-right column order in Figure 4. The source study uses the broader term hematopoietic stem and progenitor cells (HSPCs) (Stephenson et al., 2021).

Figure label	Full cell-population name
B cells	B cells
CD14	CD14 monocytes
CD16	CD16 monocytes
CD4	CD4 T cells
CD8	CD8 T cells
DCs	Dendritic cells
HSC	Hematopoietic stem cells
Lymph_prolif	Proliferating lymphocytes
MAIT	Mucosal-associated invariant T cells
Mono_prolif	Proliferating monocytes
CD16-high NK	CD16-high natural killer cells
CD56-high NK	CD56-high natural killer cells
Plasmablasts	Plasmablasts
Platelets	Platelets
RBC	Red blood cells
Treg	Regulatory T cells
gdT	$\gamma\delta$ T cells
pDC	Plasmacytoid dendritic cells