scOLAR: Ontology-Anchored Open-Set Annotation of Single-Cell RNA-seq Data

Anonymous authors
Paper under double-blind review

Abstract

Reference-based single-cell annotation becomes an open-set problem when target data contain cell types absent from the reference. Existing approaches can reject unfamiliar cells, but the evidence used to distinguish known from unknown cells usually does not explicitly model cell-type hierarchy. This is limiting because an unseen fine-grained type may share broad transcriptional programs with a represented lineage, causing lineage similarity to support an incorrect known identity. Rejected cells must also be organized without knowing the number of unseen types. We present scOLAR, an ontology-anchored framework that separates fine-type evidence from broader-lineage affinity. scOLAR learns Cell Ontology-indexed prototypes with hierarchy-aware objectives, applies a reference-calibrated rule for selective label acceptance, groups rejected cells without a supplied novel-class count, and assigns broader ontology context post hoc. Across benchmarks, scOLAR achieves an AUROC of 0.973 for novelty detection, the highest published-protocol Overall accuracy on all five datasets, and the highest Novel accuracy on four. Under a separate common evaluator, scOLAR achieves a 7.2-percentage-point advantage in macro novel recovery together with higher known-cell accuracy. Matched ontology controls show stage-specific differences, with the largest observed changes in routing and post-rejection lineage context rather than novelty ranking or rejected-cell grouping. scOLAR turns open-set rejection into structured analysis of unresolved populations. Code is available in an anonymous repository.

1 Introduction

Reference atlases have made cell-type annotation scalable in single-cell RNA sequencing (scRNA-seq). Resources such as the Human Cell Atlas, Tabula Muris, and developmental atlases link gene-expression profiles to curated cell identities across tissues, organisms, and developmental stages (Regev et al., 2017; Tabula Muris Consortium, 2018; Cao et al., 2020). Reference-based methods transfer these annotations rather than requiring each dataset to be interpreted from scratch.

The central difficulty is that reference coverage is necessarily incomplete. Tissue specificity, developmental state, disease, and limited sampling can introduce target cell types with no reference label (Regev et al., 2017; Cao et al., 2020). A closed-set annotator must still choose among available labels. This is especially difficult when an unseen fine-grained type shares transcriptional programs with a represented lineage. Biologically meaningful similarity can then support a confident but incorrect fine label. Closely related unknown classes are likewise difficult in open-set recognition (OSR) (Scheirer et al., 2013; Vaze et al., 2022b; Lang et al., 2024).

Recent single-cell methods relax the closed-set assumption by identifying unmatched cells or coupling annotation with unseen-population recovery (Brbić et al., 2020; Zhai et al., 2023; 2024; Chechekhina et al., 2026). This setting is related to generalized category discovery (GCD), which recognizes known classes while organizing unseen categories (Vaze et al., 2022a). Rejection alone, however, does not organize the rejected branch. The number of reference-missing populations is generally unknown at inference time, and the rejected population can contain both target-private cells and represented cells rejected in error. Grouping this complete unresolved population therefore differs from clustering an oracle-filtered set of truly novel cells.

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.

Single-cell annotation also has external structure. The Cell Ontology (CL) links fine cell types to broader ancestors through explicit ontology `is_a` relations (Diehl et al., 2016; Tan et al., 2026). Prior work uses ontology structure for unseen-label prediction, representation learning, hierarchical rejection, and hierarchy-aware classification (Wang et al., 2021; Yuan et al., 2024; Theunissen et al., 2024; Cultrera di Montesano et al., 2026). These approaches show that hierarchy can guide prediction, but broad lineage affinity is not evidence for a particular fine identity.

These limitations call for separating fine-label support from how rejected cells are subsequently handled. We propose scOLAR (Single-Cell Ontology-guided Lineage-Aware open-set Recognition), an ontology-anchored framework for open-set annotation of single-cell RNA-seq data. scOLAR associates learnable prototypes with Cell Ontology terms and uses hierarchy-aware source objectives to distinguish fine-type evidence from broader-lineage affinity. A calibrated maximum-logit rule determines whether a mapped label is accepted. Rejected cells enter PREDNOVEL and are grouped without access to the target-private class count. The Lineage Consistency Check (LCC) is applied only after this partition is fixed, adding broader ontology context without altering routing or grouping. This decomposition lets us evaluate novelty ranking, selective routing, accepted-label accuracy, unresolved-population organization, and lineage context as distinct stages of the open-set workflow. We make three contributions.

- We formulate incomplete-reference annotation as three linked but distinct problems: selective fine-label recognition, class-count-free organization of the complete unresolved population, and post-hoc lineage interpretation.
- We develop scOLAR, combining ontology-indexed prototypes and hierarchy-aware objectives with reference-calibrated label acceptance, no-K grouping, and lineage context that cannot overwrite the original routing decision.
- Across five incomplete-reference benchmarks containing 271,125 cells, we evaluate scOLAR through native deployment, published-protocol breadth, a common downstream evaluator, and matched ontology controls, localizing the largest observed ontology-associated differences to routing and post-rejection context rather than uniform gains across all stages.

2 RELATED WORK

Open-set annotation and category discovery. Incomplete-reference single-cell methods identify unmatched cells or jointly annotate represented types and recover unseen populations (Brbić et al., 2020; Zhai et al., 2023; 2024; Chechekhina et al., 2026). MARS uses learned landmarks; scGAD and scBOL couple annotation with novel-population recovery and category-count estimation. scParadise combines multilevel annotation with unknown-cell detection, but its hierarchy structures annotation rather than the unknown score. This setting connects open-set recognition, which decides whether a known label is supported, with generalized category discovery, which also organizes unseen categories (Scheirer et al., 2013; Vaze et al., 2022a;b). Organizing unseen categories does not itself provide biological context for unresolved single-cell populations. Related open-world methods improve class compactness, multimodal prototype geometry, or dynamic novel-category modeling (Ming et al., 2023; Lu et al., 2024; Li et al., 2026). scOLAR instead organizes the complete rejected population, including false rejections, without a supplied novel-class count and uses biological hierarchy to separate fine-type evidence from broader-lineage affinity.

Ontology-aware and hierarchical prediction. The Cell Ontology links fine cell types to broader ancestors in a directed acyclic graph (Diehl et al., 2016; Tan et al., 2026). Here, lineage denotes CL ancestry rather than experimentally traced developmental lineage. OnClass and ontology-guided representation methods use this structure for unseen-term prediction or hierarchy-aware classification (Wang et al., 2021; Yuan et al., 2024; Cultrera di Montesano et al., 2026); other approaches enforce hierarchical consistency or canonical cell-type outputs (Magre et al., 2026; Song et al., 2025). Hierarchical rejection can instead back off to an ancestral label when fine-level confidence is insufficient (Theunissen et al., 2024). scOLAR keeps calibrated fine-label acceptance separate from post-rejection interpretation: ontology structure regularizes source representations, rejected cells are grouped first, and LCC then summarizes broader community-level lineage affinity for downstream interpretation without changing routing or the fixed PREDNOVEL partition.

3 METHOD

3.1 PROBLEM FORMULATION

Let $D_r = \{(x_i, y_i)\}_{i=1}^{n_r}$ be a labeled reference and $D_t = \{x_i\}_{i=1}^{n_t}$ an unlabeled target, where $x_i \in \mathbb{R}^G$ denotes the expression profile of a cell over G genes, with type sets $\mathcal{C}_r$ and $\mathcal{C}_t$. Shared types are $\mathcal{C}_s = \mathcal{C}_r \cap \mathcal{C}_t$; target-private and reference-private types are $\overline{\mathcal{C}}_t = \mathcal{C}_t \setminus \mathcal{C}_s$ and $\overline{\mathcal{C}}_r = \mathcal{C}_r \setminus \mathcal{C}_s$. Each target cell receives an accepted mapped label or enters PREDNOVEL, whose partition $\Pi_{\mathrm{PN}} = \{G_1, \ldots, G_M\}$ is constructed without the identities or cardinality of $\overline{\mathcal{C}}_t$. PREDNOVEL is a predicted rejection set and is not synonymous with the ground-truth target-private set $\overline{\mathcal{C}}_t$. Communities need not correspond one-to-one to biological types. Learning is transductive: target expression enters preprocessing and training, while target labels, class count, and metrics are used only after outputs are fixed.

3.2 ONTOLOGY-ANCHORED REPRESENTATION AND LEARNING OBJECTIVES

Figure 1 summarizes the complete scOLAR workflow and its main information flow. We begin with the shared representation because routing, grouping, and lineage interpretation all depend on the learned cell–prototype geometry. Ontology structure enters this geometry through the source-side hierarchy-aware objectives described below.

Representation. A shared encoder maps standardized expression to a latent code in the prototype model (Snell et al., 2017). With $\mathcal{O} = \{1, \ldots, N\}$ and $\mathcal{K} \subset \mathcal{O}$, let $z_i = f_{\mathrm{enc}}(x_i)$, $\hat{z}_i = z_i / \|z_i\|_2$, and $M = [\mu_1^\top; \ldots; \mu_N^\top] \in \mathbb{R}^{N \times d}$ with $\hat{\mu}_j = \mu_j / \|\mu_j\|_2$. Scoring is

$$\ell_{i,j} = s\langle \hat{z}_i, \hat{\mu}_j \rangle, \qquad s > 0. \tag{1}$$

Only $\mathcal{K}$ rows can return Known labels; all ontology-indexed rows enter classification normalization and novelty scoring. Unobserved rows compete in the full classifier head but are neither zero-shot cell-type predictions nor a validated semantic geometry.

Reconstruction and classification. A ZINB decoder (Lopez et al., 2018; Eraslan et al., 2019) reconstructs raw counts from both domains, while source cross-entropy anchors mapped cell types in the ontology-indexed head. The decoder predicts a library-adjusted mean, dispersion, and dropout probability; Appendix B gives the complete probability mass. With $\mathcal{V}$ indexing source cells with mapped labels,

$$\mathcal{L}_{\mathrm{rec}} = -\mathbb{E}_{i,g} \log p(x_{ig}^{\mathrm{raw}}), \qquad \mathcal{L}_{\mathrm{CE}} = -\frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \log \frac{e^{\ell_{i,y_i}}}{\sum_{j \in \mathcal{O}} e^{\ell_{i,j}}}. \tag{2}$$

Hierarchy and fine-type evidence. Let $A_{ak} = 1$ when ontology term k is an ancestor of source term a, including a itself. Hierarchical centripetal loss (HCL) selects triplets only among mapped source terms, using shared ancestry rather than current prototype similarity to define positive and negative relationships:

$$\phi(a, b) = \sum_k A_{ak} A_{bk}, \qquad \mathcal{L}_{\mathrm{HCL}} = \frac{1}{|\mathcal{T}|} \sum_{(a,p,n) \in \mathcal{T}} [\langle \hat{\mu}_a, \hat{\mu}_n \rangle - \langle \hat{\mu}_a, \hat{\mu}_p \rangle + m_t]_+. \tag{3}$$

HCL supplies relational structure that source classification alone does not encode; it organizes mapped source prototypes according to shared ancestry rather than using ontology distance as a calibrated metric. Positive candidates satisfy a fixed shared-ancestor threshold; negatives do not. Mining selects the least-related positive and most-related negative with scheduled margin m_t (Schroff et al., 2015). HCL acts between source prototypes, whereas decision-boundary regularization (DBR) acts at the cell–prototype boundary and favors fine-type evidence over a detached coarse-lineage reference.

Each $\hat{\mu}_c^{\mathrm{coarse}}$ is the normalized centroid of represented descendants of coarse term c:

$$\begin{aligned} S_i^{\mathrm{coarse}} &= \max_c \langle \hat{z}_i, \hat{\mu}_c^{\mathrm{coarse}} \rangle, \\ \mathcal{L}_{\mathrm{DBR}} &= \mathbb{E}_{i \in \mathcal{V}} \Big[\mathbf{1}[\hat{y}_i = y_i] \big[(\beta - \eta) - (\ell_{i,y_i}/s - \mathrm{sg}(S_i^{\mathrm{coarse}})) \big]_+ \Big]. \end{aligned} \tag{4}$$

In DBR, $\hat{y}_i = \arg\max_{j \in \mathcal{O}} \ell_{i,j}$ denotes the source classification prediction. Here, sg stops gradients through the entire coarse affinity, so DBR gradients act only through the fine branch. DBR therefore serves as a training regularizer rather than an additional acceptance test.

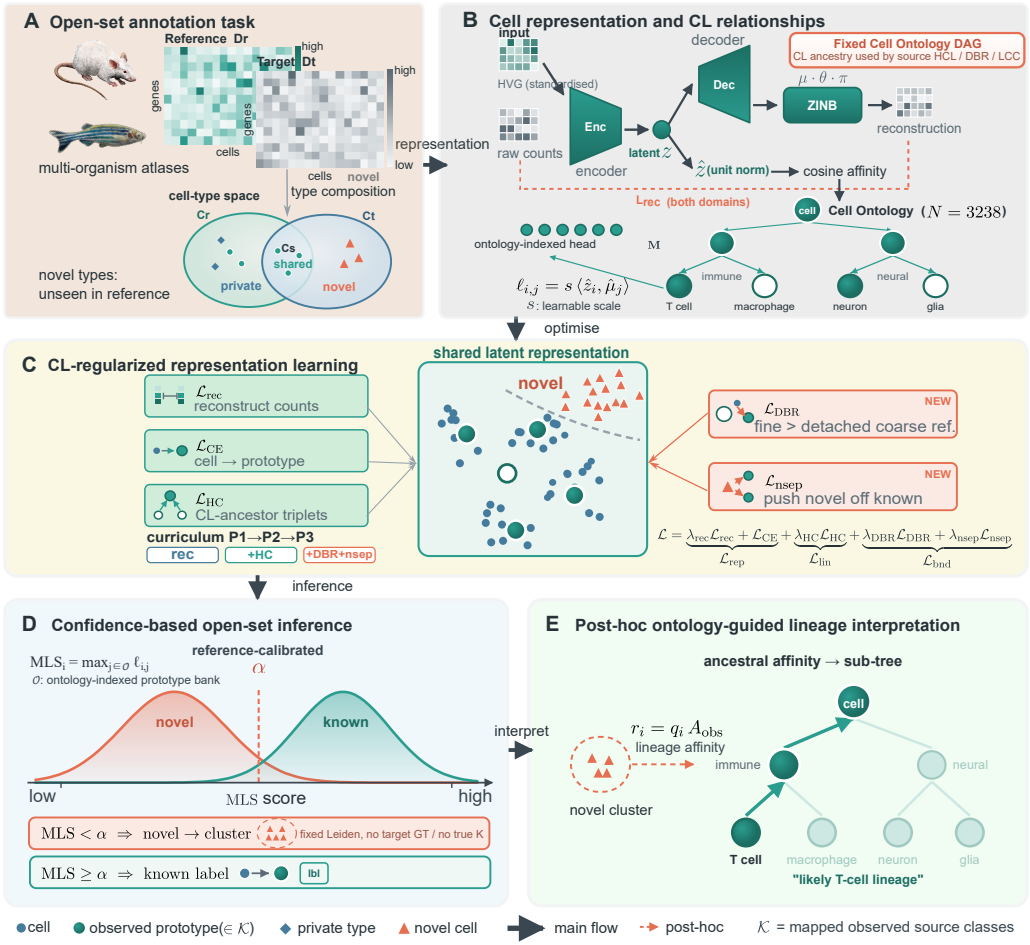

Figure 1: scOLAR overview. Reference labels and unlabeled target expression jointly train the prototype learner. Full-head maximum-logit scoring routes target cells to mapped Known labels or PREDNOVEL; the latter is grouped without a supplied class count. Ontology relations enter source HCL and DBR during training and optional post-hoc LCC context after grouping.

Novel separation and combined objective. Novel separation (nSep) complements this source-side regularization by reducing mapped-source affinity for target cells with high detached novelty weights. Using the full-head MLS defined below, nSep further repels likely-novel target cells from the mapped source-class prototypes:

$$w_i = \text{sg}(\sigma(\kappa(\alpha - \text{MLS}_i))), \qquad \mathcal{L}_{\text{nsep}} = \frac{\sum_{i \in D_t} w_i [\max_{k \in \mathcal{K}} \langle \hat{z}_i, \hat{\mu}_k \rangle - \gamma]_+}{\sum_{i \in D_t} w_i}. \tag{5}$$

It uses no ancestry or coarse membership. Let $\mathcal{L}_{\text{rep}} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{CE}}$, $\mathcal{L}_{\text{lin}} = \lambda_{\text{HCL}}\mathcal{L}_{\text{HCL}}$, and $\mathcal{L}_{\text{bnd}} = \lambda_{\text{DBR}}\mathcal{L}_{\text{DBR}} + \lambda_{\text{nsep}}\mathcal{L}_{\text{nsep}}$. The full objective is

$$\mathcal{L} = \mathcal{L}_{\text{rep}} + \mathcal{L}_{\text{lin}} + \mathcal{L}_{\text{bnd}}. \tag{6}$$

A fixed curriculum activates representation, hierarchy, then boundary objectives. Architecture, weights, and schedules are in Appendix B.

3.3 SELECTIVE ANNOTATION AND CLASS-COUNT-FREE ORGANIZATION

Selective annotation separates confidence estimation from label selection. For each target cell, the full-head maximum logit over all ontology-indexed prototypes serves as the acceptance statistic.

Conditional on acceptance, the highest-scoring mapped source type is returned,

$$\mathrm{MLS}_i = \max_{j \in \mathcal{O}} \ell_{i,j}, \qquad \hat{y}_i = \begin{cases} \arg\max_{k \in \mathcal{K}} \ell_{i,k}, & \mathrm{MLS}_i \geq \alpha, \\ \textsc{PredNovel}, & \mathrm{MLS}_i < \alpha. \end{cases} \tag{7}$$

where $\mathcal{O}$ indexes the full prototype bank and $\mathcal{K} \subset \mathcal{O}$ contains the mapped source types. The confidence score therefore reflects competition from every ontology-indexed row during target routing, whereas an accepted label can only come from $\mathcal{K}$. Prototype rows without reference examples participate in the learned geometry and confidence calculation, but they are not interpreted as zero-shot cell-type predictions.

The threshold α is estimated from reference-only class subsampling. After the final checkpoint is reloaded, α is recalibrated under the frozen plan and fixed before target inference (Appendix C). It defines a reference operating point rather than a guaranteed target-known retention rate.

Cells below this threshold enter PREDNOVEL. PREDNOVEL is a model output rather than an oracle set of truly novel cells: it can contain both target-private types and represented types rejected by the open-set decision. scOLAR therefore organizes exactly the population that would be available at deployment. A fixed nearest-neighbor graph is constructed from its learned embeddings and partitioned with Leiden clustering (Traag et al., 2019),

$$G_{\mathrm{PN}} = k\mathrm{NN}(\{\hat{z}_i : \hat{y}_i = \textsc{PredNovel}\}), \qquad \Pi_{\mathrm{PN}} = \mathrm{Leiden}(G_{\mathrm{PN}}). \tag{8}$$

yielding Π_{PN} without a supplied target-private class count. The number of communities is thus an output of the grouping procedure rather than an input derived from ground truth. These communities are unresolved organizational units and need not correspond one-to-one to biological cell types.

3.4 ONTOLOGY CONTEXT FOR REJECTED COMMUNITIES

Rejection establishes that a fine reference label is not sufficiently supported, but it does not remove all biological information about the cell. A rejected population may still show coherent affinity to represented members of a broader lineage. Assigning the nearest fine type at this stage would undo the open-set decision, so scOLAR instead summarizes this residual information only after the PREDNOVEL partition has been fixed.

For each rejected cell, logits over the mapped source types are converted to a softened affinity vector and propagated to their ontology ancestors,

$$q_{i,k} = \frac{\exp(\ell_{i,k}/T)}{\sum_{u \in \mathcal{K}} \exp(\ell_{i,u}/T)}, \quad r_i = q_i A_{\mathrm{obs}}, \quad \bar{r}_G = \frac{1}{|G|} \sum_{i \in G} r_i. \tag{9}$$

where A_{obs} contains ancestry relations from mapped source types to ontology ancestors. Because a fine type can contribute affinity to multiple ancestors along the ontology graph, r_i is not a normalized probability distribution over mutually exclusive lineage labels. Instead, it is an overlapping lineage-affinity profile that records how strongly the cell's source-type evidence propagates through different parts of the hierarchy. Averaging these profiles within a PREDNOVEL community gives $\bar{r}_G$, so the resulting description reflects lineage evidence shared across the community rather than the nearest source type of any individual cell.

Near-root terms are excluded because many unrelated cell types share them, providing little discriminative lineage information. The fixed depth-filtered Top-3 coverage rule returns a broader ontology description only when the remaining affinity is concentrated; otherwise the community remains ambiguous. Appendix C gives the full rule. LCC provides post-hoc lineage context for biological review, not a replacement fine-type prediction. Applied after grouping, it changes neither Known/PREDNOVEL routing nor the Leiden partition.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP AND EVALUATION VIEWS

We evaluate Cao (Cao et al., 2017), Quake_10x and Quake_Smart-seq2 (Tabula Muris Consortium, 2018), Wagner (Wagner et al., 2018), and Zeisel_2018 (Zeisel et al., 2018). Together, these five

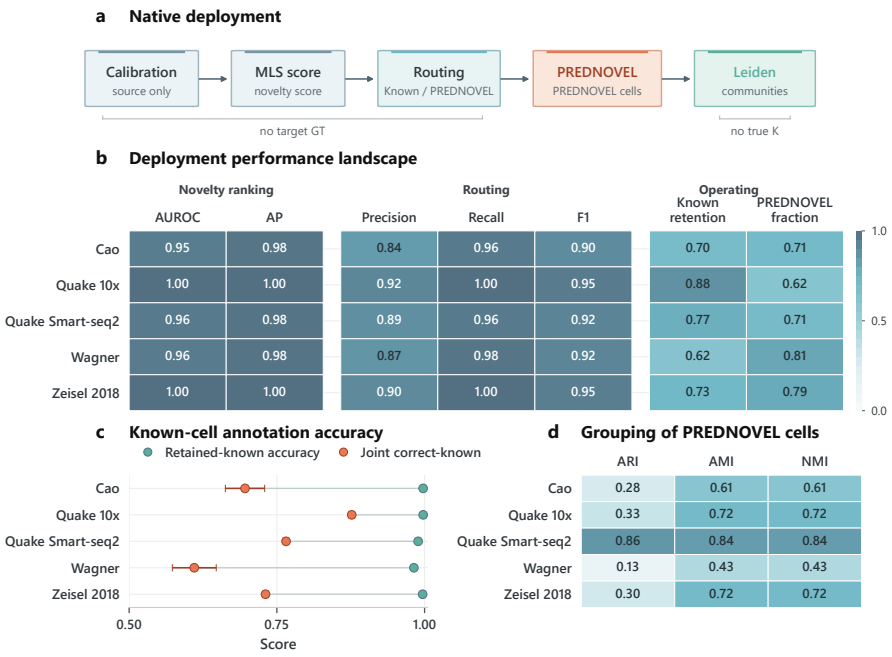

Figure 2: Native open-set performance across five benchmarks. (a) Deployment; (b) novelty ranking and calibrated routing; (c) retained-known accuracy together with joint correct-known; (d) fixed-resolution grouping of the complete PREDNOVEL population. In panel (c), points are four-run means and horizontal bars are sample SDs. Conditional accuracy applies only to accepted known cells.

benchmarks contain 271,125 cells across multiple organisms and tissues. Shared, reference-private, and target-private type sets follow the established open-set benchmark of Zhai et al. (2024), with disjoint reference and target cells. Training is transductive: target expression is available, but target labels, the target-private class count, and evaluation metrics are withheld from training, calibration, checkpoint selection, routing, and grouping. The epoch-100 checkpoint, reference-only calibration, and Leiden resolution are fixed before target labels are used for scoring. Results summarize four computational runs; datasets receive equal weight within each run, and uncertainty is the sample SD across runs. Dataset partitions, preprocessing, optimization, metrics, and complete summaries are in Appendices A–D.

Published comparisons cover seven prior methods under the established benchmark protocol, including the closest open-set comparator, scBOL (Brbić et al., 2020; Hu et al., 2020; Kimmel & Kelley, 2021; Lotfollahi et al., 2022; Wang et al., 2022; Gan et al., 2023; Zhai et al., 2024). Native evaluation separates novelty ranking, calibrated routing, selective known-cell annotation, and fixed-resolution grouping of the complete PREDNOVEL population. The published-protocol view provides breadth under established method-specific settings. The common-evaluator view applies common downstream scoring to scOLAR and strict scBOL while preserving each method's training procedure. In Table 2, Known is accuracy on all ground-truth-known cells, Novel weighted is cell-weighted novel recovery, Novel macro averages per-type recovery, ARI measures novel-partition agreement, and OverallJ is the common evaluator's joint global-matching score. Matched ontology controls compare real CL with depth-preserving shuffled and generic-star structures under the same architecture, schedule, calibration, grouping, and evaluator.

4.2 NATIVE OPEN-SET ROUTING AND CLASS-COUNT-FREE ORGANIZATION

Across five benchmarks, full-head MLS provides strong novelty ranking, with mean AUROC 0.973 (Figure 2). The reference-calibrated operating point favors recovery of unfamiliar cells: novelty recall is 0.978, while known-cell retention is 0.741. Among retained known cells, label accuracy remains 0.993, showing that the principal cost at this operating point is routing represented cells to PREDNOVEL rather than assigning an incorrect accepted label.

Table 1: Published-protocol comparison (percentages). K: Known; N: Novel; O: Overall; SS2: Smart-seq2. scOLAR entries are four-run means; scOLAR SDs are reported in Appendix D, Table 9. Comparator point estimates retain method-specific benchmark settings. Bold marks the highest value per column. This protocol is distinct from native evaluation and common-evaluator scoring (Table 2).

	Cao			Quake_10x			Quake_SS2			Wagner			Zeisel_2018		
Method	K	N	O	K	N	O	K	N	O	K	N	O	K	N	O
scCNC	55.8	42.8	28.7	84.4	65.3	68.5	58.0	35.3	38.5	83.1	57.4	59.8	64.9	79.1	70.1
scDECL	51.4	41.6	26.3	31.7	45.9	26.1	22.0	30.3	25.9	32.5	48.8	35.0	55.8	69.4	49.5
MARS	92.7	58.2	63.1	96.4	49.8	67.5	88.1	78.9	78.3	78.0	53.0	54.8	89.8	87.0	83.9
ItClust	3.4	45.0	48.0	54.3	43.3	53.2	10.9	62.0	65.9	29.4	30.8	36.0	32.1	69.3	63.7
scNym	98.5	63.1	61.2	98.5	48.1	53.7	95.3	69.6	65.9	93.9	44.6	43.5	99.3	62.2	62.4
scArches	78.0	45.4	57.7	90.3	57.3	70.1	64.0	56.1	58.1	65.2	41.0	46.8	73.1	63.2	63.6
scBOL	96.5	74.6	77.4	98.0	65.8	77.5	96.2	**82.1**	82.5	94.9	54.7	62.5	96.5	91.7	89.1
scOLAR	**99.2**	**78.1**	**89.5**	**99.3**	**68.6**	**87.2**	**97.8**	78.3	**88.3**	**99.9**	**58.0**	**72.1**	**99.6**	**99.3**	**99.4**

Table 2: Common-evaluator comparison (four matched runs; mean $\pm$ SD). Each method retains its training procedure; bold marks the higher mean.

Method	Known	Novel weighted	Novel macro	ARI	OverallJ
scOLAR	**0.9682** $\pm$ 0.0006	0.6389 $\pm$ 0.0225	**0.6586** $\pm$ 0.0187	0.5093 $\pm$ 0.0277	0.7549 $\pm$ 0.0143
strict scBOL	0.9584 $\pm$ 0.0064	**0.6656** $\pm$ 0.0319	0.5867 $\pm$ 0.0197	**0.5280** $\pm$ 0.0382	**0.7673** $\pm$ 0.0232

The complete PREDNOVEL branch is grouped without target labels, oracle filtering, or true K; fixed-resolution Leiden reaches AMI 0.661 (Figure 2).

4.3 PUBLISHED-PROTOCOL BREADTH AND COMMON-EVALUATOR COMPARISON

Table 1 places scOLAR against a broad set of prior methods under the established published protocol. Table 2 complements this breadth comparison by evaluating the closest comparator, strict scBOL, under common downstream scoring while preserving each method's training procedure.

Under the published protocol, scOLAR has the highest Known and Overall score on all five datasets and the highest Novel score on four. Its Overall margin over the strongest reported comparator ranges from 5.8 to 12.1 percentage points; Quake_Smart-seq2 is the only dataset on which its Novel score is not highest. These results establish broad performance across the five benchmark settings.

Under the common evaluator, scOLAR has higher known accuracy and macro novel recovery, including a 7.2-point macro advantage, whereas strict scBOL has higher weighted novel recovery, ARI, and OverallJ. The endpoints therefore expose complementary strengths: class-balanced recovery favors scOLAR, while abundance-weighted and global clustering summaries favor strict scBOL. Section 4.5 examines a major abundance-driven contribution to this weighted–macro disagreement.

4.4 MATCHED ONTOLOGY CONTROLS ACROSS THE PIPELINE

Ontology-dependent variation is concentrated more in routing and post-rejection context than in novelty ranking or complete-PREDNOVEL organization. Relative to the depth-preserving shuffle, real CL changes joint correct-known by +4.35 percentage points in Cao, +3.95 in Zeisel_2018, little in either Quake dataset, and -1.79 in Wagner, for an equal-dataset mean of +1.34 points (Figure 3). By contrast, the mean effects on novelty AP (+0.00005) and full-PREDNOVEL AMI (-0.00564) are near zero; the generic-star comparison shows the same lack of a consistent ranking or grouping gain.

Together, the controls indicate a stage-specific rather than uniform role for ontology structure. The largest observed differences occur where hierarchy distinguishes fine-label acceptance from broader lineage context, while novelty ranking and complete-PREDNOVEL organization remain comparatively stable across datasets. None of the six prespecified primary contrasts passes multiplicity-adjusted exact testing; Appendix E reports full intervals and paired effects.

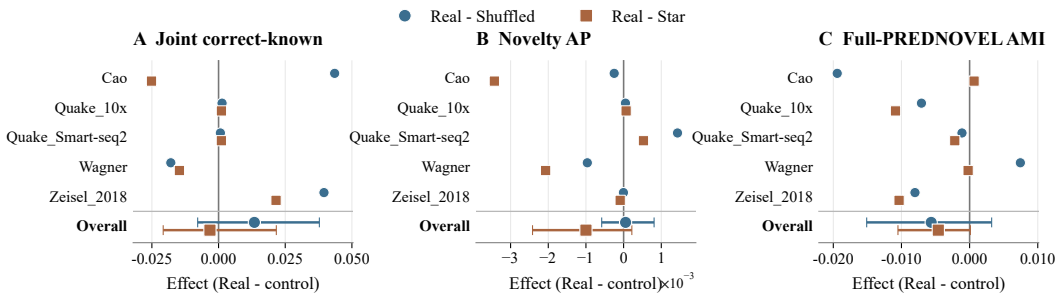

Figure 3: Paired effects for three prespecified ontology-control endpoints. Each panel compares Real CL with the depth-preserving shuffled and generic-star controls. Dataset rows are means across four matched computational runs; Overall is the equal-dataset mean with a 95% hierarchical-bootstrap interval. Effects are Real CL minus the corresponding control.

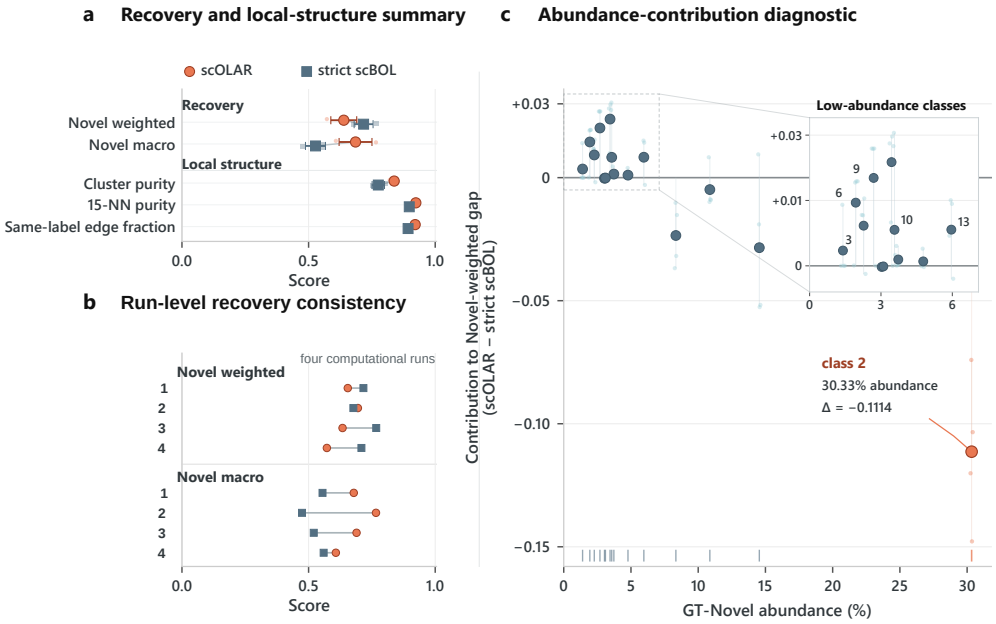

Figure 4: Post-hoc Quake_Smart-seq2 diagnostic. scOLAR has lower cell-weighted and higher class-balanced novel recovery; local-structure summaries remain consistently high across runs. In panel (a), large markers and horizontal bars show four-run means and sample SDs, while small points show individual runs. Panel (b) pairs methods within each run. In panel (c), large points are four-run means, small points are run-level values, and vertical ranges span the four runs. The abundant numeric class 2 contributes strongly to the observed weighted-recovery gap.

4.5 Quake Smart-seq2 metric diagnostic

Quake_Smart-seq2 identifies a major abundance-driven contribution to the weighted–macro disagreement. Under the common evaluator, scOLAR has lower Novel weighted recovery than strict scBOL (0.6389 versus 0.7172) but higher Novel macro recovery (0.6854 versus 0.5272). Numeric class 2 contains 4,394 cells, or 30.33% of the ground-truth-novel population, and contributes approximately -0.1114 to the four-run weighted difference (Figure 4).

Local purity, 15-nearest-neighbor agreement, and same-label graph-edge agreement remain high for scOLAR, showing that local organization and globally matched recovery capture different aspects of the representation. The diagnostic identifies an abundance-sensitive source of the aggregate pattern. Appendix F provides the corresponding numerical details and provenance.

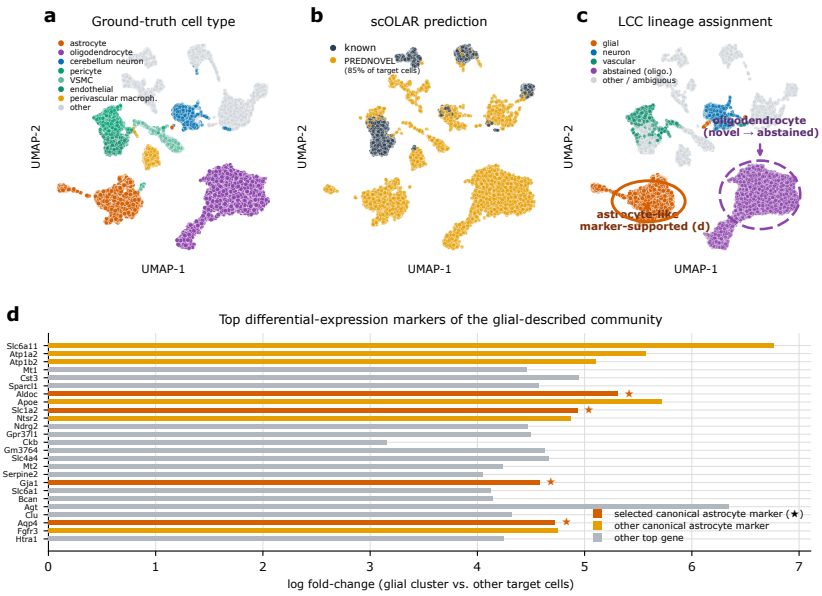

Figure 5: Illustrative post-hoc context for a Zeisel_2018 PREDNOVEL community of 9,156 benchmark astrocytes from a represented type. LCC assigns the broader ontology term *glial cell*, while same-dataset expression markers support the finer astrocyte-like interpretation.

4.6 POST-HOC LINEAGE CONTEXT

LCC adds broader ontology context after rejection without changing routing or grouping. In a Zeisel_2018 PREDNOVEL community containing 9,156 benchmark astrocytes from a represented type, LCC assigns the broader term *glial cell*. Same-dataset differential expression identifies *Aqp4*, *Slc1a2*, *Aldoc*, and *Gja1* (Figure 5). The fixed Top-3 coverage score is approximately 0.491, exceeding the 0.40 description threshold.

This example illustrates scOLAR's intended fallback behavior after an unsupported fine-label decision. Fine-label rejection need not erase biological structure: community aggregation concentrates residual source-type evidence on the broader glial branch, while the concordant expression program provides an expression-level consistency check. Because this context is attached only after the PREDNOVEL partition is fixed, it supplements rather than overwrites the selective decision. Quantitative ontology effects are evaluated separately in Section 4.4.

5 LIMITATIONS

The current evaluation follows a transductive setting. The ontology controls assess pipeline-level effects, and PREDNOVEL communities are treated as analytical units for downstream organization and interpretation. The Zeisel analysis provides same-dataset support for broader lineage context, while independent biological validation remains a direction for future study.

6 CONCLUSION

In this work, we propose scOLAR, an ontology-anchored framework for open-set single-cell annotation with incomplete references. By separating fine-label acceptance from the treatment of rejected cells, scOLAR uses Cell Ontology structure to shape source representations while preserving a class-count-free PREDNOVEL branch for downstream organization and lineage-level interpretation. This prevents broader lineage affinity from becoming an unsupported fine label while retaining it as useful post-rejection context. More broadly, scOLAR points toward open-set annotation that preserves evidence across biological resolutions, allowing unresolved fine-type assignments to retain structured lineage context rather than collapsing into forced labels or undifferentiated unknowns.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, manuscript organization, figure scripting and formatting, code review and implementation support, consistency checks, and methodological and experimental feedback. All AI-assisted code, analyses, figures, citations, and manuscript content were reviewed and verified by the authors, who take full responsibility for the reported scientific content and resulting conclusions.

REPRODUCIBILITY STATEMENT

Appendices A-H detail data processing, methods, results, ontology controls, and the computational environment. An anonymous supplementary package and repository provide executable code, the locked protocol, split metadata, reproduction scripts, and machine-readable results. Public datasets and the Cell Ontology are downloaded separately.

REFERENCES

Maria Brbić, Marinka Zitnik, Sheng Wang, Angela O. Pisco, Russ B. Altman, Spyros Darmanis, and Jure Leskovec. MARS: Discovering novel cell types across heterogeneous single-cell experiments. *Nat. Methods*, 17(12):1200–1206, 2020. doi: 10.1038/s41592-020-00979-3.

Junyue Cao, Jonathan S. Packer, Vijay Ramani, Darren A. Cusanovich, Chau Huynh, Riza Daza, Xiaojie Qiu, Choli Lee, Scott N. Furlan, Frank J. Steemers, Andrew Adey, Robert H. Waterston, Cole Trapnell, and Jay Shendure. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, 357(6352):661–667, 2017. doi: 10.1126/science.aam8940.

Junyue Cao, Diana R. O'Day, Hannah A. Pliner, Paul D. Kingsley, Mei Deng, Riza M. Daza, Michael A. Zager, Kimberly A. Aldinger, Ronnie Blecher-Gonen, Fan Zhang, Malte Spielmann, James Palis, Dan Doherty, Frank J. Steemers, Ian A. Glass, Cole Trapnell, and Jay Shendure. A human cell atlas of fetal gene expression. *Science*, 370(6518):eaba7721, 2020. doi: 10.1126/science.aba7721.

Elizaveta Chechekhina, Liya Shcherbakova, Maksim Vigovskiy, Amir Tukhvatulin, Denis Logunov, Dmitry Tychinin, Dmitry Svetlichnyy, Konstantin Kulebyakin, Vsevolod Tkachuk, and Vadim Chechekhin. scParadise: tunable, highly accurate multi-level cell type annotation, unknown cell type identification, and modality imputation. *Nucleic Acids Res.*, 54(11):gkag612, 2026. doi: 10.1093/nar/gkag612.

Sebastiano Cultrera di Montesano, Davide D'Ascenzo, Srivatsan Raghavan, Ava P. Amini, Peter S. Winter, and Lorin Crawford. Improving atlas-scale single-cell annotation models with hierarchical cross-entropy loss. *Nature Computational Science*, 6(3):243–249, 2026. doi: 10.1038/s43588-025-00945-z.

Alexander D. Diehl, Terrence F. Meehan, Yvonne M. Bradford, Matthew H. Brush, Wasila M. Dahdul, David S. Dougall, Yongqun He, David Osumi-Sutherland, Alan Ruttenberg, Sirarat Sarntivijai, Ceri E. Van Slyke, Nicole A. Vasilevsky, Melissa A. Haendel, Judith A. Blake, and Christopher J. Mungall. The Cell Ontology 2016: Enhanced content, modularization, and ontology interoperability. *J. Biomed. Semantics*, 7(1):44, 2016. doi: 10.1186/s13326-016-0088-7.

Gökcen Eraslan, Lukas M. Simon, Maria Mircea, Nikola S. Mueller, and Fabian J. Theis. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.*, 10(1):390, 2019. doi: 10.1038/s41467-018-07931-2.

Yanglan Gan, Yuhan Chen, Guangwei Xu, Wenjing Guo, and Guobing Zou. Deep enhanced constraint clustering based on contrastive learning for scRNA-seq data. *Brief. Bioinform.*, 24(4):bbad222, 2023. doi: 10.1093/bib/bbad222.

Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joseph Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling Out-of-Distribution Detection for Real-World Settings. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pp. 8759–8773. PMLR, 2022.

Jian Hu, Xiangjie Li, Gang Hu, Yafei Lyu, Katalin Susztak, and Mingyao Li. Iterative transfer learning with neural network for clustering and cell type classification in single-cell RNA-seq analysis. *Nat. Mach. Intell.*, 2(10):607–618, 2020. doi: 10.1038/s42256-020-00233-7.

Jacob C. Kimmel and David R. Kelley. Semisupervised adversarial neural networks for single-cell classification. *Genome Res.*, 31(10):1781–1793, 2021. doi: 10.1101/gr.268581.120.

H. W. Kuhn. The Hungarian method for the assignment problem. *Nav. Res. Logist. Q.*, 2(1-2):83–97, 1955. doi: 10.1002/nav.3800020109.

Nico Lang, Vésteinn Snæbjarnarson, Elijah Cole, Oisin Mac Aodha, Christian Igel, and Serge Belongie. From coarse to fine-grained open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17804–17814, 2024.

Chunming Li, Shidong Wang, and Haofeng Zhang. Adaptive gaussian expansion for on-the-fly category discovery. In *International Conference on Learning Representations*, 2026.

Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nat. Methods*, 15(12):1053–1058, 2018. doi: 10.1038/s41592-018-0229-2.

Mohammad Lotfollahi, Mohsen Naghipourfar, Malte D. Luecken, Matin Khajavi, Maren Büttner, Marco Wagenstetter, Žiga Avsec, Adam Gayoso, Nir Yosef, Marta Interlandi, Sergei Rybakov, Alexander V. Misharin, and Fabian J. Theis. Mapping single-cell data to reference atlases by transfer learning. *Nat. Biotechnol.*, 40(1):121–130, 2022. doi: 10.1038/s41587-021-01001-7.

Haodong Lu, Dong Gong, Shuo Wang, Jason Xue, Lina Yao, and Kristen Moore. Learning with mixture of prototypes for out-of-distribution detection. In *International Conference on Learning Representations*, 2024.

Nimish Magre, Ebtisam Alshehri, Fedor Grab, Yerdos Ordabayev, Steven A. McCarroll, Mehrtash Babadi, and Stephen J. Fleming. A supervised ontology-aware cell annotation method for single-cell transcriptomic data. *bioRxiv*, pp. 2026.01.13.699356, January 2026. ISSN 2692-8205. doi: 10.64898/2026.01.13.699356.

Yifei Ming, Yiyou Sun, Ousmane Dia, and Yixuan Li. How to exploit hyperspherical embeddings for out-of-distribution detection? In *International Conference on Learning Representations*, 2023.

Aviv Regev, Sarah A. Teichmann, Eric S. Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, et al. The Human Cell Atlas. *eLife*, 6:e27041, 2017. doi: 10.7554/eLife.27041.

Walter J. Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult. Toward Open Set Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(7):1757–1772, 2013. doi: 10.1109/TPAMI.2012.256.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823, 2015. doi: 10.1109/CVPR.2015.7298682.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 4080–4090, Red Hook, NY, USA, 2017. Curran Associates Inc.

Yiran Song, Muyao Tang, Qi Liu, Haofei Wang, Li Qian, Fei Zou, and Wenpin Hou. GPTAnno: Ontology-tree-guided hierarchical cell type annotation based on GPT models for single-cell data. *bioRxiv*, pp. 2025.11.27.690951, December 2025. ISSN 2692-8205. doi: 10.1101/2025.11.27.690951.

Tabula Muris Consortium. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*, 562(7727):367–372, 2018. doi: 10.1038/s41586-018-0590-4.

Shawn Zheng Kai Tan, Aleix Puig-Barbe, Damien Goutte-Gattat, Caroline Eastwood, Brian Aevermann, Alida Avola, James P. Balhoff, Ismail Ugur Bayindir, Jasmine Belfiore, Anita Reane Caron, et al. The Cell Ontology in the age of single-cell omics. *Sci. Data*, 13(1):946, April 2026. ISSN 2052-4463. doi: 10.1038/s41597-026-07173-8.

Lauren Theunissen, Thomas Mortier, Yvan Saeys, and Willem Waegeman. Uncertainty-aware single-cell annotation with a hierarchical reject option. *Bioinformatics*, 40(3):btae128, March 2024. ISSN 1367-4811. doi: 10.1093/bioinformatics/btae128.

V. A. Traag, L. Waltman, and N. J. van Eck. From Louvain to Leiden: Guaranteeing well-connected communities. *Sci. Rep.*, 9(1):5233, 2019. doi: 10.1038/s41598-019-41695-z.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized Category Discovery. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7482–7491, 2022a. doi: 10.1109/CVPR52688.2022.00734.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. In *International Conference on Learning Representations (ICLR)*, pp. 1–13, 2022b.

Daniel E. Wagner, Caleb Weinreb, Zach M. Collins, James A. Briggs, Sean G. Megason, and Allon M. Klein. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science*, 360(6392):981–987, 2018. doi: 10.1126/science.aar4362.

Hai-Yun Wang, Jian-Ping Zhao, Chun-Hou Zheng, and Yan-Sen Su. scCNC: A method based on capsule network for clustering scRNA-seq data. *Bioinformatics*, 38(15):3703–3709, 2022. doi: 10.1093/bioinformatics/btac393.

Sheng Wang, Angela Oliveira Pisco, Aaron McGeever, Maria Brbić, Marinka Zitnik, Spyros Darmanis, Jure Leskovec, Jim Karkanias, and Russ B. Altman. Leveraging the Cell Ontology to classify unseen cell types. *Nat. Commun.*, 12(1):5556, 2021. doi: 10.1038/s41467-021-25725-x.

F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol.*, 19(1):15, 2018. doi: 10.1186/s13059-017-1382-0.

Xinyu Yuan, Zhihao Zhan, Zuobai Zhang, Manqi Zhou, Jianan Zhao, Boyu Han, Yue Li, and Jian Tang. Cell-ontology guided transcriptome foundation model. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0204.

Amit Zeisel, Hannah Hochgerner, Peter Lönnerberg, Anna Johnsson, Fatima Memic, Job van der Zwan, Martin Häring, Emelie Braun, Lars E. Borm, Gioele La Manno, et al. Molecular Architecture of the Mouse Nervous System. *Cell*, 174(4):999–1014.e22, 2018. doi: 10.1016/j.cell.2018.06.021.

Yuyao Zhai, Liang Chen, and Minghua Deng. scGAD: A new task and end-to-end framework for generalized cell type annotation and discovery. *Brief. Bioinform.*, 24(2):bbad045, 2023. doi: 10.1093/bib/bbad045.

Yuyao Zhai, Liang Chen, and Minghua Deng. scBOL: A universal cell type identification framework for single-cell and spatial transcriptomics data. *Brief. Bioinform.*, 25(3):bbae188, 2024. doi: 10.1093/bib/bbae188.

A Datasets and Preprocessing

Table 3: Primary benchmarks; type counts are shared/reference-private/target-private.

Dataset	Cells	Types	Organism / platform
Cao	30,960	6/4/6	*C. elegans* / sci-RNA-seq
Quake_10x	53,587	12/12/12	mouse / 10x
Quake_Smart-seq2	41,526	15/15/15	mouse / Smart-seq2
Wagner	34,348	5/4/5	zebrafish / inDrop
Zeisel_2018	110,704	6/5/6	mouse / 10x

The type partitions follow the published benchmark (Zhai et al., 2024). Each shared-class cell enters the reference with probability 0.5 and otherwise enters the target; reference-private cells remain in the reference and target-private cells remain in the target. The resulting cell partitions are disjoint and shared by matched conditions. Cells outside the benchmark registry are excluded before processing. Fixed type withholding does not imply identical cell assignments across computational runs.

Reference and target expression are joined for unsupervised feature processing. Counts are normalized to 10,000 per cell and log transformed. Scanpy selects 2,000 highly variable genes (Wolf et al., 2018); each gene is standardized over the combined reference-target matrix and clipped to $[-10, 10]$. Raw counts for those genes supply the reconstruction loss. No target labels enter preprocessing. This joint, unlabeled processing is one reason the method is correctly described as transductive.

The CL release dated 2025-12-17 contains 3,238 compiled terms and 4,493 `is_a` edges. The ancestor closure used by HCL, DBR, and LCC is computed from this compiled `is_a` graph. Dataset labels are mapped by a shared lookup over names, synonyms, and CL identifiers, with no dataset-specific ontology curation or manual remapping.

A.1 Preprocessing implementation

The benchmark cell and gene sets are already quality-controlled. Integer counts are taken from `adata.layers['counts']` when available, otherwise `adata.X`, and preserved for reconstruction. Scanpy's `normalize_total` uses a target sum of 10^4, followed by `log1p`. The highly-variable-gene call sets `n_top_genes=2000` with `subset=False`; its Boolean mask selects the encoder input and corresponding raw-count genes. Gene means and variances for standardization use the combined reference–target matrix. The count-scale decoder uses each cell's raw library size relative to the median, clipped to $[0.1, 10]$. These operations use target expression but no target annotations at any preprocessing stage.

A.2 Ontology compilation and label mapping

Obsolete terms and terms outside the `is_a` closure of `CL:0000000` are removed. The compiled term order indexes both the ancestor matrix and prototype rows. The implementation raises an error if a source benchmark label has no usable mapping. Target-private labels are not exposed to training. Ancestor closure includes self-membership. The graph supplies source HCL relationships, source DBR memberships, and post-hoc LCC propagation; reconstruction, nSep, score thresholding, and Leiden do not directly consume ontology edges.

Table 4: Method families in Table 1. Entries retain their published benchmark implementations, not a common retraining configuration.

Method	Benchmark output	Novel grouping in the published protocol
MARS (Brbić et al., 2020)	Known or unassigned	k-means on unassigned cells
ItClust (Hu et al., 2020)	Known or unassigned	k-means on unassigned cells
scNym (Kimmel & Kelley, 2021)	Known or unassigned	k-means on unassigned cells
scArches (Lotfollahi et al., 2022)	Known or unassigned	k-means on unassigned cells
scCNC (Wang et al., 2022)	Unsupervised clusters	Clustering output
scDECL (Gan et al., 2023)	Unsupervised clusters	Clustering output
scBOL (Zhai et al., 2024)	Joint known/novel	Bipartite prototype alignment

A.3 Published comparator implementations

The separate common-evaluator comparison and its benchmark scorer are specified in Appendix D; this analysis does not convert the published table into a matched-information experiment.

B Model and Optimization

B.1 ZINB reconstruction likelihood

The decoder predicts relative mean $\bar{\nu}$, dispersion $\theta > 0$, and dropout probability π. Raw counts x_{ig}^{raw} supply the likelihood. With library size L_i, define $\rho_i = \mathrm{clip}(L_i / \operatorname{median}_k L_k, 0.1, 10)$ and $\nu_i = \rho_i \bar{\nu}_i$. The negative-binomial and zero-inflated log probability masses are

$$\begin{aligned} \log \mathrm{NB}(x \mid \nu, \theta) = {} & \log \Gamma(x+\theta) - \log \Gamma(\theta) - \log \Gamma(x+1) \\ & + \theta \log \frac{\theta}{\nu+\theta} + x \log \frac{\nu}{\nu+\theta}, \end{aligned} \tag{10}$$

$$\log p(x) = \begin{cases} \log\big(\pi + (1-\pi)\mathrm{NB}(0 \mid \nu, \theta)\big), & x = 0, \\ \log(1-\pi) + \log \mathrm{NB}(x \mid \nu, \theta), & x > 0. \end{cases} \tag{11}$$

The main-text reconstruction objective averages this likelihood over genes and over reference and target cells; no target label enters reconstruction.

B.2 Network architecture

The same model is used for every dataset. The encoder receives standardized expression; its unnormalized latent code also enters the ZINB decoder. Normalized codes and prototypes enter cosine scoring in Eq. (1). Unobserved ontology rows receive no positive source supervision but remain in the cross-entropy denominator and full-head MLS. HCL, DBR, the nSep repulsive similarity, accepted labels, and LCC use mapped source rows. This distinction does not confer semantic predictions on unobserved rows.

Table 5: Network configuration. LN: layer normalization; activation: LeakyReLU with slope 0.2.

Module	Operation	Output dimension
Encoder input	Standardized HVGs	2,000
Hidden 1	Linear, LN, activation, dropout 0.2	512
Hidden 2	Linear, LN, activation	256
Latent	Linear, LN	128
Decoder	Two linear/activation layers	$256 \to 512$
Mean head	Linear, softmax over genes	2,000
Dispersion head	Linear, softplus	2,000
Dropout head	Linear, sigmoid	2,000
Prototype matrix	Xavier-uniform initialization	3238×128
Logit scale	Learned positive scalar, initial value 8.0	1

B.3 Triplet construction and boundary objectives

For each mapped source anchor present in a source mini-batch, define

$$P(a) = \{b \in \mathcal{K} : b \neq a,\ \phi(a,b) \geq 3\}, \qquad N(a) = \{b \in \mathcal{K} : \phi(a,b) < 3\}. \tag{12}$$

Mining selects the positive with the fewest shared ancestors and the negative with the most, sampling ties without current prototype similarities. Triplets are resampled per batch. The shared-ancestor count determines a binary relationship, not a calibrated ontology distance; deeper nodes generally have more ancestors. The HCL margin decreases linearly from 0.3 at activation to 0.05 at epoch 50, then remains fixed.

DBR coarse terms have depth at most 3 and average observed descendant prototypes with equal weights before normalization. In Eq. (4), stop-gradient blocks both the cell-code and prototype paths through the complete coarse affinity. A correctly classified source mask gates a hinge with $\beta = 0.3$ and slack $\eta = 0.05$. Target HCL, target DBR, and target-neighborhood consistency are disabled.

nSep starts only after a finite threshold is available. Its detached weights use full-head MLS with $\kappa = 5$; its repulsive similarity uses mapped source prototypes with tolerance $\gamma = 0.1$. Normalization by the sum of weights prevents loss magnitude from simply scaling with the number of likely-novel cells. Neither ancestry nor coarse memberships enter this objective.

B.4 Curriculum and hyperparameters

Specifically, at epoch $t \in \{1, \ldots, 100\}$,

$$\mathrm{lr}_t = 10^{-5} + \tfrac{1}{2}(10^{-3} - 10^{-5}) \left[1 + \cos\left(\pi \frac{t-1}{99}\right)\right]. \tag{13}$$

Mini-batches with non-finite activations are skipped to prevent invalid states. This safeguard does not alter run definition. The four repeats quantify stochastic training variability rather than biological replication. Appendix H gives the complete training and final-prediction sequence, including scheduled calibration refreshes, checkpoint reload, output freezing, and evaluation from saved outputs.

Table 6: Fixed optimization settings shared across all datasets and computational runs. The curriculum, loss weights, learning-rate schedule, calibration epochs, and final checkpoint are prespecified; no target validation metric is used to select the schedule or checkpoint.

Item	Setting
Training	100 epochs; final epoch checkpoint
Epochs 1–19	Reconstruction and source classification
Epochs 20–49	Add source HCL; margin $0.3 \to 0.05$
Epochs 50–100	Add source DBR; reconstruction weight becomes 0.5
Epoch 60 onward	Add nSep; refresh calibration at 60, 70, 80, 90, 100
Reconstruction weight	1.0 at epochs 1–49; 0.5 at 50–79; linear ramp to zero at 80–100
Other loss weights	CE: 1; HCL: 0.1; DBR: 0.05; nSep: 0.01
Optimizer	AdamW; weight decay 10^{-4}
Batch size	1,024; reference and target loaders cycled jointly
Gradient clipping	Global ℓ_2 norm at 5
Learning rate	Fixed cosine schedule, $10^{-3} \to 10^{-5}$

C Calibration, Routing, Grouping, and Context

C.1 Reference-only calibration

MLS is the maximum logit over the complete head (Vaze et al., 2022b; Hendrycks et al., 2022). For $K = |\mathcal{K}|$ mapped source classes, the exclusion count is $m = \min(10, K - 1)$. Twenty deterministic source-class draws are fixed before training; excluding a class removes its cells from the calibration sample, not its prototype from scoring. Thus at least one class remains, including when $K \leq 10$. Training-time refreshes reuse this plan at epochs 60, 70, 80, 90, and 100 for nSep. Final prediction has a separate sequence: finish epoch 100, save and reload that checkpoint, recompute α with Algorithm 1, freeze it, then score target cells. The fifth-percentile rule targets reference retention; distribution and composition changes prevent interpreting it as a 95% target-known guarantee.

Algorithm 1 Reference-only threshold calibration

Require: Reference D_r, full head $\mathcal{O}$, fixed draws $H_1, \ldots, H_{20} \subset \mathcal{K}$, each of size $\min(10, K-1)$
1: **for** $r = 1, \ldots, 20$ **do**
2: $\quad S_r \leftarrow \{\max_{j \in \mathcal{O}} \ell_{i,j} : (x_i, y_i) \in D_r,\ y_i \notin H_r\}$
3: $\quad \alpha_r \leftarrow \mathrm{quantile}(S_r, 0.05)$
4: **end for**
5: **return** $\alpha \leftarrow \mathrm{median}(\alpha_1, \ldots, \alpha_{20})$

C.2 Full rejected-population partition

Target cells with MLS below α enter PredNovel. For a retained cell, the returned label maximizes the mapped-source logits, even though MLS uses all rows. Grouping first row-normalizes the embeddings, then constructs a Euclidean 15-nearest-neighbor graph (clipped to the available cell

count). Leiden uses resolution 1.0, random state 0, the igraph implementation, an undirected graph, and two iterations. With fewer than two cells, the partition is trivial. All rejected cells are included; native grouping neither filters ground-truth-known cells nor supplies true K. TP-only diagnostic scores apply a true-novel mask only after this partition is fixed.

C.3 LCC ASSIGNMENT RULE

LCC requires at least ten rejected cells in the run; communities smaller than five cells are omitted. Eq. (9) uses source-only softmax at $T = 2$. For each eligible G, entries of $\bar{r}_G$ corresponding to ontology terms at depth < 2 are set to zero, yielding $\tilde{r}_G$. Let $t_{(1)}, t_{(2)}, t_{(3)}$ be its highest-affinity terms. Assignment requires

$$\frac{\sum_{j=1}^{3} \tilde{r}_G[t_{(j)}]}{\sum_t \tilde{r}_G[t]} \geq 0.40. \tag{14}$$

The deepest of these three terms is emitted; communities without sufficient supported affinity remain ambiguous. The software status *Consistent* denotes this concentration criterion, not biological accuracy or validated lineage identity.

Algorithm 2 Native prediction and post-hoc community context

Require: Final epoch-100 checkpoint, reference calibration plan, target expression
1: Reload checkpoint; compute and freeze α with Algorithm 1
2: Compute embeddings and all head logits for target cells
3: Apply Eq. (7); collect all rejected cells
4: Construct the graph and fixed Leiden partition in Eq. (8)
5: **for** each eligible rejected community G **do**
6: Compute q_i, r_i, and $\bar{r}_G$ by Eq. (9)
7: Filter depths and apply Eq. (14)
8: Emit deepest Top-3 term if the rule is met; otherwise mark ambiguous
9: **end for**
10: Save labels, scores, partition, and descriptions before target evaluation

D EVALUATION DEFINITIONS AND SUPPORTING RESULTS

D.1 POPULATIONS AND NATIVE METRICS

Let B be the ground-truth-known target cells, U the target-private cells, R the accepted cells, and P the complete PREDNOVEL population. Define $C = \{i \in B \cap R : \hat{y}_i = y_i\}$. Then

$$\text{Retention} = \frac{|B \cap R|}{|B|}, \quad \text{ConditionalAccuracy} = \frac{|C|}{|B \cap R|}, \quad \text{JointCorrectKnown} = \frac{|C|}{|B|}. \tag{15}$$

False-novel rate is $1 - \text{Retention}$. Joint correct-known is computed cellwise, not by multiplying separately aggregated retention and accuracy. Novel precision is $|P \cap U|/|P|$, recall is $|P \cap U|/|U|$, and F1 is their harmonic mean. Novel prevalence is $|U|/|D_t|$; predicted-novel fraction is $|P|/|D_t|$. AUROC and AP use $-\text{MLS}$ as the novelty-oriented score.

ARI, AMI, and NMI compare the single fixed Leiden partition of all P against benchmark labels, so these native grouping scores include false-rejected represented cells as well as target-private cells. TP-only diagnostics do not recluster an oracle-filtered subset; they reuse the identical partition and restrict scoring to $P \cap U$ only after grouping is complete. This preserves the deployed partition while separating the effect of false rejections in a secondary diagnostic. Degenerate grouping endpoints remain undefined rather than being assigned zero. Within each run, datasets receive equal weight, with uncertainty reported as sample SD over four computational runs. Paired summaries use complete run pairs, so paired means need not equal marginal-mean differences.

These quantities describe different stages of the deployment decision. Known retention measures the fraction of represented cells that remain in the accepted branch, conditional accuracy measures label correctness among retained represented cells, and joint correct-known measures the unconditional fraction that are both retained and correctly labeled. A high conditional accuracy can therefore

coexist with substantial false rejection. For scOLAR, the gap between retained-known accuracy and joint correct-known is driven primarily by routing represented cells to PREDNOVEL rather than by assigning incorrect labels after acceptance.

Novelty ranking, routing, and grouping are likewise distinct. AUROC and AP use the continuous novelty score before thresholding; precision, recall, and F1 evaluate thresholded PREDNOVEL membership; ARI, AMI, and NMI evaluate the single fixed Leiden partition of that complete rejected population. TP-only scores are post-hoc diagnostics on the same partition after restricting evaluation to $P \cap U$; Leiden is not rerun on an oracle-filtered subset.

Table 7: Deployment and supplementary diagnostic summaries. Each cell is mean ± sample SD followed by valid runs/4. AUROC and AP for the comparator are supplementary ranking diagnostics based on $-$`source_max_logit`; cross-system deltas are therefore omitted for these two endpoints. Routing metrics retain each method's native outputs. Values are run-level equal-dataset summaries.

Deployment metric	scOLAR	strict scBOL	Paired delta
AUROC	0.9726 ± 0.0012 [4/4]	0.2741 ± 0.0591 [4/4]	—
AUPRC (AP)	0.9871 ± 0.0006 [4/4]	0.5660 ± 0.0262 [4/4]	—
Novel prevalence	0.6564 ± 0.0018 [4/4]	0.6564 ± 0.0018 [4/4]	—
Precision	0.8837 ± 0.0052 [4/4]	0.2500 ± 0.2516 [4/4]	0.6338 ± 0.2483 [4/4]
Recall	0.9782 ± 0.0015 [4/4]	0.0430 ± 0.0319 [4/4]	0.9352 ± 0.0323 [4/4]
F1	0.9284 ± 0.0023 [4/4]	0.0664 ± 0.0477 [4/4]	0.8621 ± 0.0469 [4/4]
Known retention	0.7411 ± 0.0145 [4/4]	1.0000 ± 0.0000 [4/4]	-0.2589 ± 0.0145 [4/4]
PREDNOVEL fraction	0.7273 ± 0.0062 [4/4]	0.0288 ± 0.0209 [4/4]	0.6985 ± 0.0233 [4/4]
All-PREDNOVEL ARI	0.3802 ± 0.0033 [4/4]	0.0426 ± 0.0603 [2/4]	0.5279 ± 0.3635 [2/4]
All-PREDNOVEL AMI	0.6607 ± 0.0028 [4/4]	0.1885 ± 0.2667 [2/4]	0.5822 ± 0.1654 [2/4]
All-PREDNOVEL NMI	0.6622 ± 0.0028 [4/4]	0.1892 ± 0.2673 [2/4]	0.5825 ± 0.1652 [2/4]
TP-only ARI	0.3610 ± 0.0032 [4/4]	0.0852 ± NA [1/4]	0.7897 ± NA [1/4]
TP-only AMI	0.6043 ± 0.0020 [4/4]	0.3772 ± NA [1/4]	0.4659 ± NA [1/4]
TP-only NMI	0.6053 ± 0.0020 [4/4]	0.3781 ± NA [1/4]	0.4658 ± NA [1/4]

AUROC/AP are reported descriptively, without cross-system deltas, because they derive from different native scoring functions. Undefined values remain missing and sample SD is NA when only one run is valid. Paired incomplete-case clustering deltas need not equal differences of marginal means. TP-only scores reuse the identical all-PREDNOVEL Leiden partition and apply the ground-truth-novel mask only after clustering.

Table 8: Equal-dataset end-to-end Known summary (four runs; mean ± sample SD; $n = 4$).

Method	Known retention	False-novel rate	Retained-known conditional accuracy	Joint correct-known rate
scOLAR	0.7411 ± 0.0145	0.2589 ± 0.0145	0.9926 ± 0.0003	0.7359 ± 0.0141
strict scBOL	1.0000 ± 0.0000	0.0000 ± 0.0000	0.9584 ± 0.0064	0.9584 ± 0.0064

Joint correct-known is cellwise, not the product of aggregate retention and conditional accuracy. Strict scBOL receives the true target class count during training.

For the comparator, supplementary AUROC/AP use $-$`source_max_logit` as a ranking diagnostic, while routing metrics follow the method's native prediction procedure.

D.2 PUBLISHED-PROTOCOL AND COMMON-EVALUATOR EVALUATION

Table 9 retains all five datasets, eight methods, and the three published metrics with all supplied scOLAR SDs. Published Known, Novel, and Overall preserve the method-specific settings of Zhai et al. (2024) and are distinct from the native metrics. Table 9 reproduces the established literature comparison; Table 10 separately reports the closest direct comparison under a common downstream evaluator. Figure 6 summarizes the two evaluation views.

Table 9: Complete published-protocol comparison (percentages). K: Known; N: Novel; O: Overall; SS2: Smart-seq2. Comparator entries are published point estimates. The scOLAR mean row is followed by its four-run sample SDs; bold marks the highest mean per column. These metrics retain method-specific benchmark settings, not native selective-prediction definitions.

	Cao			Quake_10x			Quake_SS2			Wagner			Zeisel_2018		
Method	K	N	O	K	N	O	K	N	O	K	N	O	K	N	O
scCNC	55.8	42.8	28.7	84.4	65.3	68.5	58.0	35.3	38.5	83.1	57.4	59.8	64.9	79.1	70.1
scDECL	51.4	41.6	26.3	31.7	45.9	26.1	22.0	30.3	25.9	32.5	48.8	35.0	55.8	69.4	49.5
MARS	92.7	58.2	63.1	96.4	49.8	67.5	88.1	78.9	78.3	78.0	53.0	54.8	89.8	87.0	83.9
ItClust	3.4	45.0	48.0	54.3	43.3	53.2	10.9	62.0	65.9	29.4	30.8	36.0	32.1	69.3	63.7
scNym	98.5	63.1	61.2	98.5	48.1	53.7	95.3	69.6	65.9	93.9	44.6	43.5	99.3	62.2	62.4
scArches	78.0	45.4	57.7	90.3	57.3	70.1	64.0	56.1	58.1	65.2	41.0	46.8	73.1	63.2	63.6
scBOL	96.5	74.6	77.4	98.0	65.8	77.5	96.2	**82.1**	82.5	94.9	54.7	62.5	96.5	91.7	89.1
scOLAR	**99.2**	**78.1**	**89.5**	**99.3**	**68.6**	**87.2**	**97.8**	78.3	**88.3**	**99.9**	**58.0**	**72.1**	**99.6**	**99.3**	**99.4**
SD	±0.2	±1.7	±0.9	±0.3	±2.0	±1.0	±0.1	±1.2	±0.7	±0.0	±2.5	±1.6	±0.1	±0.1	±0.1

The common evaluator first fixes representations and predictions, then extracts the ground-truth-novel subset and row-normalizes its embeddings. Both systems receive identical k-means settings: true novel K, `k-means++`, a fixed random state, ten restarts selected by inertia, and one numerical thread. A single Hungarian matching maximizes correct novel cells (Kuhn, 1955). Novel weighted is correct matched cells divided by all novel cells; Novel macro averages per-type recall under the same mapping. Known is classification accuracy on all ground-truth-known cells, distinct from conditional retained accuracy. ARI measures the novel benchmark partition. OverallJ uses joint benchmark outputs and the evaluator's global matching.

Table 10: **Common-evaluator benchmark over four matched runs** (mean ± sample SD). Paired differences are reported as scOLAR minus the comparator.

Metric	scOLAR	strict scBOL	Paired delta
Known	0.9682 ± 0.0006	0.9584 ± 0.0064	0.0098 ± 0.0065
Novel weighted	0.6389 ± 0.0225	0.6656 ± 0.0319	-0.0267 ± 0.0531
Novel macro	0.6586 ± 0.0187	0.5867 ± 0.0197	0.0720 ± 0.0367
ARI	0.5093 ± 0.0277	0.5280 ± 0.0382	-0.0188 ± 0.0623
OverallJ	0.7549 ± 0.0143	0.7673 ± 0.0232	-0.0124 ± 0.0368

Ground-truth-novel cells and true K enter benchmark clustering after outputs are fixed. This comparison evaluates complete systems, not native deployment or individual ontology components. Strict scBOL receives the true target class count during training.

The common evaluator standardizes downstream benchmark scoring rather than model training. Representations, native predictions, and saved outputs are fixed before ground-truth-novel membership and true K are introduced. Both systems then use the same novel subset, k-means initialization and restart policy, Hungarian matching, and numerical-thread setting. Thus, the comparison equalizes clustering and metric computation while preserving each method's training information and native rejection behavior. All four matched runs are retained irrespective of downstream score.

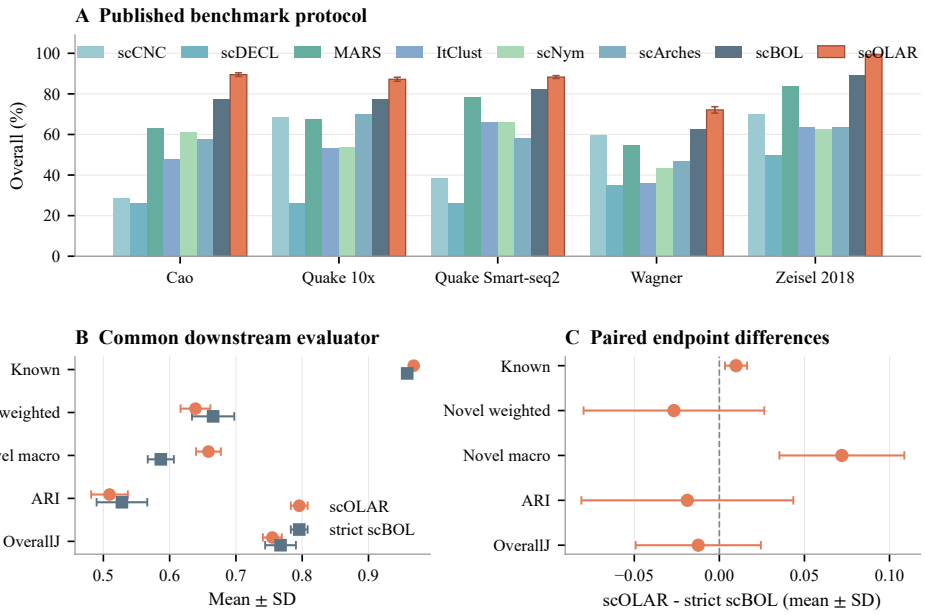

Figure 6: Benchmark comparison protocols. (A) Published benchmark with method-specific settings. (B) Common downstream evaluator for scOLAR and strict scBOL. (C) Paired endpoint differences across four matched runs; signs indicate the higher system per endpoint, not an overall ranking.

D.3 Machine-readable results

The primary native and common-evaluator results are reported above as aggregate summaries. The supplementary package provides the corresponding per-dataset and per-run records, paired summaries, and aggregate statistics across five datasets and four matched computational runs. For the ontology study, the accompanying files additionally provide all 60 run records, dataset-by-condition summaries, and paired-effect records. Appendix E describes these files and their inferential roles; the machine-readable tables preserve full numerical precision and undefined values.

D.4 Scope and limitations

scOLAR is transductive. Target expression participates in joint preprocessing and representation learning, although target labels, the target-private class count, and evaluation metrics are withheld until outputs are fixed. The reported experiments therefore evaluate adaptation to a specified unlabeled target dataset and do not establish inductive performance on future targets that were absent during training. The five benchmark datasets are the inferential units in the ontology study, while the four computational runs quantify algorithmic variability rather than independent biological replication. Dataset-level inferential power is consequently limited.

The common evaluator standardizes downstream scoring but does not equalize training information, so Table 10 should be interpreted as a system-level comparison rather than a matched-information ablation. The ontology interventions identify pipeline-level dependence rather than isolating HCL, DBR, or LCC. The depth-preserving shuffle breaks the correct class-to-ontology correspondence while retaining term depth, but uses only one fixed mapping per dataset and does not preserve degree, ancestor count, or pairwise relations. The generic-star control jointly changes HCL availability, DBR coarse affinity, and native LCC behavior, so it does not isolate a single pathway.

Finally, PredNovel can mix target-private and falsely rejected represented cells, so its Leiden communities are unresolved units rather than validated novel types. The post-hoc Zeisel_2018 example uses a known benchmark identity, same-dataset markers, and a separate computational run. It illustrates post-rejection context rather than final matched-protocol reproduction.

E ONTOLOGY CONTROLS AND SUPPORTING RESULTS

The intervention contains five datasets, four matched computational runs, and three conditions, for 60 completed runs. Model dimension, source-label rows, partitions, training schedule, calibration rule, nSep, Leiden settings, and evaluator are fixed across conditions; training is repeated in each condition.

Effects are paired within dataset and run, averaged over four runs within each dataset, and then averaged equally across five datasets. Datasets are inferential units. Hierarchical-bootstrap intervals use 10,000 resamples of datasets and matched runs. Two-sided exact sign-flip tests operate on five dataset-level paired means. Holm correction covers six primary tests: two control contrasts for joint correct-known, AP, and full-PREDNOVEL AMI.

Table 11: Ontology interventions and retained consequences.

Condition	Definition
Real CL	Original source-class-to-CL mapping; structure enters source HCL, source DBR, and LCC.
Depth shuffled	Circular derangement within exact-depth source strata; singleton strata use an unused non-source term at equal depth. One mapping is fixed per dataset. Depth is preserved, but degree, ancestor count, and pairwise relations need not be.
Generic star	A synthetic root with source leaves at depth 1 and inactive padding rows. Source identity remains available, but no HCL triplets satisfy the shared-ancestor rule, coarse DBR affinity is zero under the retained coefficient, and the fixed LCC depth rule yields no native descriptions.

For Real minus Depth-shuffled, the equal-dataset effects are +1.34 percentage points for joint correct-known (95% CI [−0.78, +3.78]), +0.00005 for AP, and −0.00564 for AMI; the latter two intervals include zero. The corresponding Real-minus-Star primary means are non-positive. None of the six primary contrasts passes multiplicity-adjusted exact testing.

Secondary LCC outputs have distinct denominators. Assignment yield is the proportion of eligible communities receiving a term. Compatibility is conditional on evaluable assigned cells and is undefined when none are assigned. Utility U includes coverage, compatibility, and depth specificity over all eligible evaluable cells, with unassigned cells contributing zero. Real CL increases mean U over shuffling by 0.0546 (bootstrap 95% CI [0.0109, 0.1044]), while the dataset-level exact test is $P = 0.25$. Synthetic labels have no real-CL semantics, so native star compatibility and utility are not applicable. These endpoints measure agreement with ontology-mapped benchmark labels, not independent biological lineage validation.

E.1 COMMUNITY-CONTEXT DENOMINATORS

An eligible community contains at least five cells in a run with at least ten PREDNOVEL cells. Assigned-cell coverage divides assigned eligible cells by all eligible cells, whereas assignment yield weights eligible communities equally and therefore gives each community the same contribution regardless of its size. Evaluation maps benchmark types through a unique valid CL identifier when available, otherwise through a unique exact term name or synonym, or the fixed Wagner alias. All supplied runs have complete benchmark-label mapping coverage under this procedure; this guarantees evaluability of the reported lineage comparisons but does not imply that the reference structurally covers all relevant ontology branches present in the target data.

An assigned term is compatible if it equals or is an ancestor of the mapped benchmark type under real CL. Conditional cell-weighted compatibility divides compatible assigned evaluable cells by all assigned evaluable cells, and is NA when there are no assignments. For eligible evaluable cells $\mathcal{E}$, let a_i denote assignment, c_i compatibility, $\hat{t}_i$ the predicted term, t_i the benchmark term, and d the real-CL depth. The cell-weighted lineage utility is

$$U = \frac{1}{|\mathcal{E}|} \sum_{i \in \mathcal{E}} a_i c_i \min\left\{1, \frac{d(\hat{t}_i)}{d(t_i)}\right\}. \tag{16}$$

Unassigned cells contribute zero. Both compatibility and utility include falsely rejected known cells. Star nodes lack real-CL semantics, so their native biological compatibility and utility are NA, not zero; zero native yield follows from the retained depth rule.

E.2 Absolute dataset-by-condition results

Tables 12–14 retain all dataset-by-condition means and sample SDs from the matched study. These fresh runs are separate from the primary deployment summaries in Appendix D.

Table 12: Known-cell routing and novelty ranking in the matched ontology experiment.

Dataset	Condition	Retention	Cond. Acc.	Joint	AUROC	AP
Cao	Real	0.7016 ± 0.0291	0.9974 ± 0.0014	0.6998 ± 0.0293	0.9509 ± 0.0038	0.9758 ± 0.0015
	Shuffled	0.6578 ± 0.0573	0.9976 ± 0.0012	0.6563 ± 0.0577	0.9514 ± 0.0038	0.9760 ± 0.0015
	Star	0.7270 ± 0.0193	0.9971 ± 0.0015	0.7249 ± 0.0186	0.9595 ± 0.0035	0.9792 ± 0.0015
Quake 10x	Real	0.8782 ± 0.0058	0.9976 ± 0.0003	0.8761 ± 0.0056	0.9975 ± 0.0000	0.9982 ± 0.0000
	Shuffled	0.8772 ± 0.0058	0.9972 ± 0.0004	0.8747 ± 0.0057	0.9975 ± 0.0001	0.9982 ± 0.0000
	Star	0.8770 ± 0.0043	0.9977 ± 0.0003	0.8750 ± 0.0042	0.9974 ± 0.0001	0.9982 ± 0.0000
Quake SS2	Real	0.7737 ± 0.0044	0.9892 ± 0.0014	0.7653 ± 0.0045	0.9574 ± 0.0021	0.9784 ± 0.0011
	Shuffled	0.7726 ± 0.0057	0.9897 ± 0.0016	0.7646 ± 0.0056	0.9543 ± 0.0024	0.9769 ± 0.0012
	Star	0.7725 ± 0.0038	0.9893 ± 0.0014	0.7642 ± 0.0038	0.9558 ± 0.0022	0.9779 ± 0.0010
Wagner	Real	0.6224 ± 0.0386	0.9821 ± 0.0008	0.6112 ± 0.0379	0.9569 ± 0.0038	0.9828 ± 0.0022
	Shuffled	0.6407 ± 0.0340	0.9819 ± 0.0009	0.6291 ± 0.0337	0.9584 ± 0.0039	0.9838 ± 0.0021
	Star	0.6362 ± 0.0334	0.9839 ± 0.0010	0.6259 ± 0.0324	0.9618 ± 0.0039	0.9849 ± 0.0021
Zeisel	Real	0.7332 ± 0.0051	0.9967 ± 0.0007	0.7308 ± 0.0050	0.9998 ± 0.0001	0.9998 ± 0.0001
	Shuffled	0.6929 ± 0.0320	0.9976 ± 0.0006	0.6912 ± 0.0317	0.9997 ± 0.0001	0.9998 ± 0.0001
	Star	0.7107 ± 0.0901	0.9979 ± 0.0007	0.7092 ± 0.0895	0.9998 ± 0.0000	0.9999 ± 0.0000

Values are mean ± sample SD over four runs; datasets are not pooled by cell count.

The absolute results show that ontology effects are not monotone across datasets or endpoints. Relative to the depth-preserving shuffle, Real CL increases joint correct-known most clearly in Cao and Zeisel, changes little in the two Quake datasets, and decreases it in Wagner. Novelty ranking and complete-PredNovel grouping remain relatively stable across conditions within most datasets, although individual grouping metrics move in either direction. This agrees with the primary paired analysis, which is more sensitive in routing than ranking or complete-population grouping. Tables 12–14 give the absolute dataset-by-condition values underlying the paired contrasts, not additional primary endpoints.

Table 13: Threshold-based novelty detection and clustering of all PredNovel cells.

Dataset	Condition	Precision	Recall	ARI	AMI	NMI
Cao	Real	0.8449 ± 0.0120	0.9576 ± 0.0040	0.2733 ± 0.0150	0.6057 ± 0.0112	0.6080 ± 0.0113
	Shuffled	0.8266 ± 0.0240	0.9589 ± 0.0077	0.3094 ± 0.0200	0.6251 ± 0.0167	0.6272 ± 0.0165
	Star	0.8571 ± 0.0070	0.9651 ± 0.0042	0.2706 ± 0.0193	0.6050 ± 0.0099	0.6071 ± 0.0098
Quake 10x	Real	0.9150 ± 0.0033	0.9975 ± 0.0001	0.3436 ± 0.0316	0.7152 ± 0.0061	0.7175 ± 0.0060
	Shuffled	0.9143 ± 0.0032	0.9969 ± 0.0002	0.3546 ± 0.0191	0.7223 ± 0.0078	0.7245 ± 0.0077
	Star	0.9142 ± 0.0022	0.9970 ± 0.0003	0.3667 ± 0.0301	0.7261 ± 0.0105	0.7283 ± 0.0104
Quake SS2	Real	0.8897 ± 0.0024	0.9593 ± 0.0023	0.8655 ± 0.0051	0.8399 ± 0.0033	0.8415 ± 0.0032
	Shuffled	0.8890 ± 0.0029	0.9576 ± 0.0029	0.8658 ± 0.0050	0.8410 ± 0.0022	0.8426 ± 0.0022
	Star	0.8890 ± 0.0022	0.9575 ± 0.0026	0.8675 ± 0.0024	0.8420 ± 0.0017	0.8437 ± 0.0017
Wagner	Real	0.8671 ± 0.0115	0.9759 ± 0.0030	0.1376 ± 0.0191	0.4329 ± 0.0155	0.4342 ± 0.0154
	Shuffled	0.8725 ± 0.0100	0.9737 ± 0.0034	0.1332 ± 0.0082	0.4255 ± 0.0097	0.4268 ± 0.0097
	Star	0.8717 ± 0.0097	0.9793 ± 0.0035	0.1403 ± 0.0138	0.4331 ± 0.0146	0.4344 ± 0.0145
Zeisel	Real	0.9033 ± 0.0014	1.0000 ± 0.0000	0.3001 ± 0.0304	0.7162 ± 0.0135	0.7165 ± 0.0135
	Shuffled	0.8904 ± 0.0097	1.0000 ± 0.0000	0.3086 ± 0.0342	0.7242 ± 0.0159	0.7245 ± 0.0158
	Star	0.8967 ± 0.0282	1.0000 ± 0.0000	0.3165 ± 0.0330	0.7265 ± 0.0097	0.7269 ± 0.0097

Values are mean ± sample SD over four runs; datasets are not pooled by cell count.

Table 14: Absolute community assignment and lineage outputs.

Dataset	Condition	Yield	Coverage	Compatibility	U
Cao	Real	0.2988 ± 0.0369	0.3160 ± 0.0306	0.9268 ± 0.0275	0.2752 ± 0.0376
Cao	Shuffled	0.1508 ± 0.0445	0.1779 ± 0.0239	0.7344 ± 0.0905	0.1282 ± 0.0297
Cao	Star	0.0000 ± 0.0000	0.0000 ± 0.0000	NA	NA
Quake 10x	Real	0.0000 ± 0.0000	0.0000 ± 0.0000	NA	0.0000 ± 0.0000
Quake 10x	Shuffled	0.0074 ± 0.0147	0.0002 ± 0.0004	0.0000 (1)	0.0000 ± 0.0000
Quake 10x	Star	0.0000 ± 0.0000	0.0000 ± 0.0000	NA	NA
Quake Smart-seq2	Real	0.0000 ± 0.0000	0.0000 ± 0.0000	NA	0.0000 ± 0.0000
Quake Smart-seq2	Shuffled	0.0000 ± 0.0000	0.0000 ± 0.0000	NA	0.0000 ± 0.0000
Quake Smart-seq2	Star	0.0000 ± 0.0000	0.0000 ± 0.0000	NA	NA
Wagner	Real	0.4409 ± 0.0461	0.3659 ± 0.0525	0.2132 ± 0.0452	0.0765 ± 0.0080
Wagner	Shuffled	0.4199 ± 0.0843	0.3162 ± 0.0807	0.0172 ± 0.0067	0.0052 ± 0.0017
Wagner	Star	0.0000 ± 0.0000	0.0000 ± 0.0000	NA	NA
Zeisel 2018	Real	0.1881 ± 0.0335	0.0761 ± 0.0022	0.9753 ± 0.0253	0.0611 ± 0.0007
Zeisel 2018	Shuffled	0.0978 ± 0.0217	0.0501 ± 0.0010	0.2224 ± 0.0102	0.0067 ± 0.0003
Zeisel 2018	Star	0.0000 ± 0.0000	0.0000 ± 0.0000	NA	NA

Values are mean $\pm$ sample SD over four computational runs; datasets are not pooled by cell count. Compatibility is conditional on evaluable assigned cells and is NA without a description. (1) denotes one evaluable run and no estimable between-run SD. Star compatibility and utility are inapplicable; zero yield follows directly from the fixed minimum-depth rule used by LCC.

E.3 All primary contrasts and uncertainty

The five datasets, not individual cells or computational repeats, are the inferential units. The hierarchical bootstrap resamples datasets and matched runs within them. The exact two-sided sign-flip test retains zero effects; with five datasets its minimum attainable probability is $2/32 = 0.0625$. Bootstrap intervals and exact tests therefore answer related but non-identical questions. Holm correction covers all six primary contrasts; utility belongs to a separate two-contrast secondary family, with the native star contrast not evaluable. No failed runs are omitted and no undefined measurement is imputed.

Table 15: Prespecified primary ontology contrasts.

Contrast	Endpoint	Effect	95% interval
Real − shuffled	Joint correct-known	0.01343	[-0.00783, 0.03780]
Real − shuffled	Novelty AP	0.00005	[-0.00058, 0.00081]
Real − shuffled	Full PREDNOVEL AMI	-0.00564	[-0.01510, 0.00324]
Real − star	Joint correct-known	-0.00322	[-0.02076, 0.02166]
Real − star	Novelty AP	-0.00100	[-0.00242, 0.00022]
Real − star	Full PREDNOVEL AMI	-0.00459	[-0.01051, 0.00011]

Contrast	Endpoint	Exact P	Holm P	Matched run pairs
Real − shuffled	Joint correct-known	0.3125	1.0000	20
Real − shuffled	Novelty AP	0.8750	1.0000	20
Real − shuffled	Full PREDNOVEL AMI	0.2500	1.0000	20
Real − star	Joint correct-known	0.7500	1.0000	20
Real − star	Novelty AP	0.3750	1.0000	20
Real − star	Full PREDNOVEL AMI	0.1875	1.0000	20

Matched run pairs are descriptive computational repeats; the five datasets remain the equal-weight inferential units. Intervals are hierarchical-bootstrap estimates, and exact tests use dataset-level sign flips. Blank or inapplicable adjustments are shown as NA. A synthetic native-lineage contrast has no evaluable pairs because the synthetic labels do not carry corresponding real-CL semantics.

Table 16: Secondary assignment and lineage utility contrasts.

Contrast	Endpoint	Effect	95% interval
Real − shuffled	Assignment yield	0.05038	[0.00039, 0.10833]
Real − shuffled	Assigned-cell coverage	0.04271	[0.00512, 0.09271]
Real − shuffled	Lineage utility U	0.05456	[0.01088, 0.10445]
Real − star	Assignment yield	0.18556	[0.03763, 0.33582]
Real − star	Assigned-cell coverage	0.15159	[0.01522, 0.29058]
Real − star	Lineage utility U	NA	[NA, NA]

Contrast	Endpoint	Exact P	Holm P	Matched run pairs
Real − shuffled	Assignment yield	0.2500	NA	20
Real − shuffled	Assigned-cell coverage	0.2500	NA	20
Real − shuffled	Lineage utility U	0.2500	0.5000	20
Real − star	Assignment yield	0.2500	NA	20
Real − star	Assigned-cell coverage	0.2500	NA	20
Real − star	Lineage utility U	NA	NA	0

Matched run pairs are descriptive computational repeats; the five datasets remain the equal-weight inferential units. Intervals are hierarchical-bootstrap estimates, and exact tests use dataset-level sign flips. Blank or inapplicable adjustments are shown as NA. A synthetic native-lineage contrast has no evaluable pairs because the synthetic labels do not carry corresponding real-CL semantics.

E.4 Machine-readable run-level results

Complete run-level records are provided in the accompanying machine-readable CSV files. `results_run_level.csv` contains all 60 runs, including calibration thresholds, community counts, mapping coverage, and LCC numerators and denominators. `aggregate_by_dataset.csv` gives dataset-by-condition means and SDs; `paired_effects_by_dataset.csv` gives paired dataset effects; and `paired_effects_and_CI.csv` gives equal-dataset contrasts and uncertainty. These files retain full numerical precision and undefined values.

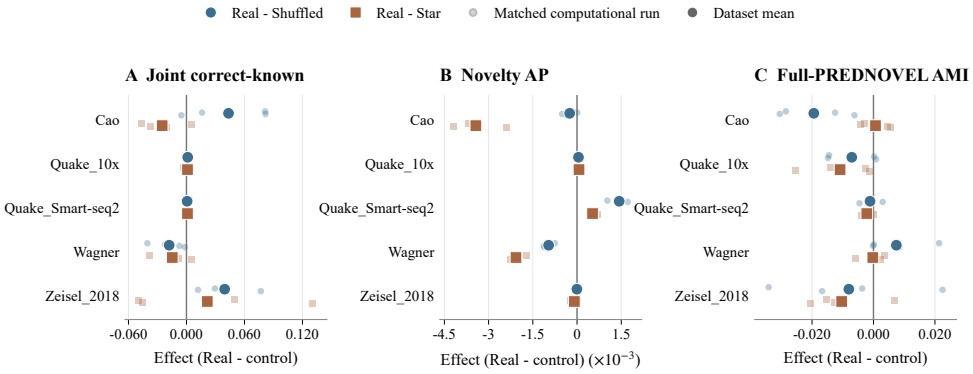

Figure 7: Run-level variation underlying Figure 3. Small points are paired effects from the four matched computational runs, and larger points are dataset means. Effects are Real CL minus the stated control. The runs are computational repeats, not biological replicates. This visualization exposes the run-level variation summarized in the main figure; dataset-level inference remains the analysis reported in Appendix E.

Run-level points expose optimization variability but remain nested within datasets and do not increase the number of inferential units. Dataset-level paired means, rather than the 20 individual run pairs, enter the exact sign-flip tests and equal-dataset contrasts. Figure 7 is therefore descriptive of computational variability, while the formal inference remains at the dataset level.

F Quake Smart-seq2 Post-hoc Diagnostic

This diagnostic followed the disagreement between cell-weighted and class-balanced recovery in the common-evaluator results. On Quake_Smart-seq2, scOLAR Novel weighted is 0.6389 ± 0.0511

versus strict scBOL's 0.7172 ± 0.0374, whereas Novel macro is 0.6854 ± 0.0649 versus 0.5272 ± 0.0395. Local purity, 15-nearest-neighbor agreement, and same-label edge agreement remain high for scOLAR but do not determine globally matched recovery.

Class 2 contains 4,394 cells (30.33% of ground-truth-novel cells) and contributes -0.1114 to the four-run weighted difference. Its abundance-weighted recall gap can reverse the summary despite higher macro recovery. This decomposition is descriptive, not biological or causal evidence.

The diagnostic is derived from the stored per-dataset, per-run common-evaluator outputs and the frozen class-level decomposition underlying Figure 4. Reported means and sample SDs summarize four matched computational runs. The supplementary records preserve run identifiers, benchmark class labels, metric denominators, and the signed per-class contributions at full precision.

G Glial Community Illustration

G.1 Selection and analysis

The Zeisel_2018 application includes 78,700 target cells, with 67,066 routed to PredNovel. A community was selected post hoc by preference for an LCC description containing "glia," and expression markers were compared against the remaining cells in the same dataset. The highlighted community contains 9,156 cells, all annotated as astrocytes in the exported benchmark labels. Astrocytes belong to the represented reference set, so these cells are false rejections of a known type.

The model description is *glial cell* (CL:0000125). Its Top-3 coverage score is approximately 0.4911 under the fixed LCC rule. Expression of *Aqp4*, *Slc1a2*, *Aldoc*, and *Gja1* supports the finer astrocyte-like reading within this broader glial context. The source export contains cell coordinates, labels, marker values, and metadata. This same-dataset analysis is post hoc and is not independent validation.

G.2 Relation to the quantitative experiments

Figure 5 is an illustrative post-hoc application analysis from a separate computational run and is kept distinct from the final matched ontology-control experiment. The community was selected for qualitative inspection and is not part of the prespecified primary ontology contrasts; no quantitative ontology-control claim is derived from it. Its purpose is to illustrate how a rejected community can retain coherent expression structure and receive broader ontology context even when its represented fine label is not accepted.

H Reproducibility and Training Algorithm

H.1 Information access and computational replication

The training interface exposes labeled reference expression and unlabeled target expression, raw counts, and size factors. Target annotations are unavailable to optimization, checkpoint selection, source calibration, native routing, Leiden, and LCC. Evaluation consumes fixed saved outputs. The common-evaluator benchmark introduces oracle novel membership and true K only after saved outputs are fixed.

The common-evaluator study and matched ontology-control study each use four frozen computational runs under their respective locked protocols. These experimental collections are analyzed separately and are not pooled. Historical benchmark records and the illustrative Zeisel application are likewise kept separate from the final matched analyses. All repeats quantify computational variability rather than biological replication. Machine-readable results preserve run identifiers, protocol hashes, calibrated thresholds, counts, and metric denominators.

Algorithm 3 scOLAR training and final prediction

Require: Labeled reference, unlabeled target expression, CL graph, fixed configuration
Jointly preprocess expression; retain raw counts and size factors
Compile ontology, source mapping, ancestry, and coarse memberships
Initialize encoder, decoder, prototype head, and logit scale
Freeze twenty source-only exclusion draws for Algorithm 1
for $t = 1, \ldots, 100$ **do**
    Set learning rate, loss weights, and margin from Appendix B
    if $t \in \{60, 70, 80, 90, 100\}$ **then**
        Refresh training threshold using the fixed source plan
    end if
    for paired reference and target mini-batches **do**
        Compute embeddings, logits, reconstruction, and source cross-entropy
        Add source HCL when $t \geq 20$ and source DBR when $t \geq 50$
        Add target nSep when $t \geq 60$ with detached novelty weights
        Backpropagate Eq. (6); clip gradients and update with AdamW
    end for
end for
Save final epoch-100 checkpoint; reload it
Recalibrate and freeze α using Algorithm 1
Run native prediction and context (Algorithm 2); save outputs
Evaluate saved outputs separately with target annotations

H.2 Environment and data access

Public accessions are GSE98561 (Cao), GSE109774 (both Tabula Muris arms), GSE112294 (Wagner), and SRP135960 (Zeisel_2018). Appendix A specifies the CL release and preprocessing. The anonymous supplement provides the executable implementation, locked protocol, split metadata, reproduction scripts, and machine-readable results. Public expression datasets and the ontology are obtained separately; identifying repository and archival links are withheld during double-blind review.

Table 17: Environment (one GPU per run).

Item	Configuration
OS	Debian GNU/Linux 12
Python	3.9.25
PyTorch	2.8.0; CUDA 12.8
Scanpy	1.10.3; AnnData 0.10.8
CL parsing	obonet 1.1.1; pronto 2.7.3
GPU	RTX 3090 or RTX 4090
CPU / RAM	Xeon Platinum 8168; 96 cores; 251 GiB