HiLL: LLM-based Recursive Semantic Refinement for Reference-Free Cell Type Annotation

Anonymous authors
Paper under double-blind review

Abstract

Single-cell RNA sequencing and spatial transcriptomics provide high-resolution profiles of tissues, but translating these profiles into biological insights critically depends on accurate cell type annotation. Manual annotation is labor-intensive, difficult to scale, and dependent on expert knowledge, motivating automatic annotation methods. However, automatic cell type annotation often relies on external annotated atlases, which may be unavailable, mismatched to the target data, or insufficiently resolved to capture fine-grained cell subtypes. Reference-free annotation addresses these limitations by inferring cell identities directly from the data. Existing reference-free LLM-based methods usually operate on a fixed clustering partition and rely on conventional marker extraction, limiting adaptive subtype refinement and accurate annotation. To address this limitation, we propose HiLL, a hierarchical LLM-based framework for reference-free cell type annotation. HiLL formulates reference-free annotation as a hierarchical refinement process, using LLM-based split validation to determine when further refinement is warranted and comparative evidence to support accurate annotation. Across human and mouse scRNA-seq and single-cell spatial transcriptomics datasets, HiLL consistently outperforms reference-free LLM baselines, with especially strong gains on fine-grained subtype distinctions. It also remains competitive with reference-based methods and zero-shot foundation models, supporting recursive local refinement as an effective paradigm for reference-free annotation.

1 Introduction

Single-cell RNA sequencing (scRNA-seq) and spatial transcriptomics (ST) provide high-resolution measurements for characterizing cellular identity and tissue organization (Plass et al., 2018; Cao et al., 2017; Codeluppi et al., 2018; Moffitt et al., 2018). Reliable cell type annotation is therefore foundational for downstream analysis, yet manual annotation remains labor-intensive, difficult to scale, and sensitive to annotator expertise (Mikulová et al., 2022; Neves & Ševa, 2021), motivating automatic annotation methods.

Existing automatic annotation methods are mostly reference-reliant and fall into three broad categories: similarity-based label transfer (e.g., SingleR (Aran et al., 2019), scmap (Kiselev et al., 2018), CHETAH (de Kanter et al., 2019)), supervised deep learning (e.g., TOSICA (Chen et al., 2023), scAdapt (Zhou et al., 2021), mtANN (Xiong et al., 2023)), and pretrained foundation models (e.g., scGPT (Cui et al., 2024), CellPLM (Wen et al., 2023), scFoundation (Hao et al., 2024)). Although these approaches can perform well when reference and query data are well aligned, they depend on external label resources and may degrade under domain shift, incomplete training coverage, or mismatched annotation granularity.

These limitations motivate reference-free cell type annotation, especially when suitable references are unavailable or when the target data contains cell identities absent from existing atlases. In this setting, annotation becomes an open-vocabulary semantic inference problem rather than classification within a predefined label space. Large language models (LLMs) are therefore appealing because they can map marker-derived molecular evidence to flexible biological labels without being

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.

restricted to a fixed taxonomy. Recent methods such as GPTCelltype (Hou & Ji, 2024), scPilot (Gao et al., 2026), and LICT (Ye et al., 2025) have begun to explore this direction.

However, accurate reference-free annotation depends on two coupled decisions that are often overlooked in existing pipelines: deciding at what granularity a population should be interpreted, and determining what molecular evidence is most informative for assigning a semantic label at that granularity. In practice, clustering resolution should not be treated merely as a technical hyperparameter, but as a biological assumption about the level at which cell identity should be described. This granularity choice is coupled to the choice of evidence: a neuronal population, for example, may be meaningfully labeled at a coarse level as GABAergic neurons, while finer distinctions such as SST^+, $PVALB^+$, and VIP^+ interneurons become interpretable only after further local subdivision and with the appropriate marker evidence. At the same time, not every geometrically separable subcluster corresponds to a biologically meaningful subtype: over-splitting may fragment a coherent population into weakly supported groups whose marker evidence is unstable or difficult to interpret. Existing reference-free LLM-based methods typically operate on a fixed global partition and rely on conventional differential-expression summaries under a generic comparison setting, which limits both adaptive subtype refinement and context-appropriate semantic interpretation.

To address this limitation, we propose HiLL, a hierarchical LLM-based framework for reference-free cell type annotation. HiLL formulates annotation as a recursive semantic refinement process over a hierarchy of cell populations.F or each active population, it first proposes a candidate subdivision supported by the current expression representation, then retains the split only if branch-local comparative evidence indicates that the candidate child clusters are biologically distinguishable enough to warrant separate interpretation. For accepted splits, HiLL performs level-wide marker extraction, uses an LLM to generate open-vocabulary labels and natural-language descriptions, and feeds these descriptions back as semantic targets to regularize the shared expression representation used in later refinement rounds. This design allows annotation granularity to vary across branches while grounding semantic labeling in comparison contexts matched to the current refinement level.

Across human and mouse scRNA-seq and ST datasets spanning pancreas and brain tissues, HiLL consistently outperforms existing reference-free LLM baselines, especially on fine-grained subtype distinctions, while remaining competitive with several reference-based and pretrained methods. Our main contributions are:

- We formulate reference-free cell type annotation as a recursive semantic refinement problem, in which granularity selection and semantic labeling are distinct but coupled decisions.
- We introduce HiLL, a framework that combines branch-local partition proposal, LLM-based split validation using local comparative evidence, level-wide marker extraction for annotation, and iterative semantic regularization.
- We demonstrate across human and mouse scRNA-seq and ST datasets that HiLL consistently improves reference-free annotation performance, with gains driven mainly by recursive refinement and level-wide marker extraction.

2 Related Work

Cell type annotation with external supervision. Many automatic cell type annotation methods rely on external supervision, including similarity-based label transfer methods such as SingleR (Aran et al., 2019), scmap (Kiselev et al., 2018), and CHETAH (de Kanter et al., 2019), supervised models such as TOSICA (Chen et al., 2023), mtANN (Xiong et al., 2023), and scAdapt (Zhou et al., 2021), and large pretrained models such as scGPT (Cui et al., 2024), CellPLM (Wen et al., 2023), and scFoundation (Hao et al., 2024). These approaches can perform well when suitable references, labeled training data, or large-scale pretraining resources are available. However, they assume some form of external label information and therefore do not directly address the fully reference-free open-vocabulary setting considered here.

Reference-free LLM-based annotation. Recent LLM-based methods, including GPTCelltype (Hou & Ji, 2024), scPilot (Gao et al., 2026), LICT (Ye et al., 2025), and related work (Chen et al., 2025; Huang et al., 2025), reduce dependence on external atlases by mapping marker genes or cluster summaries to cell type labels. Most of these methods follow a common cluster-then-annotate pipeline: cells are first partitioned by a conventional clustering method, and each resulting cluster is

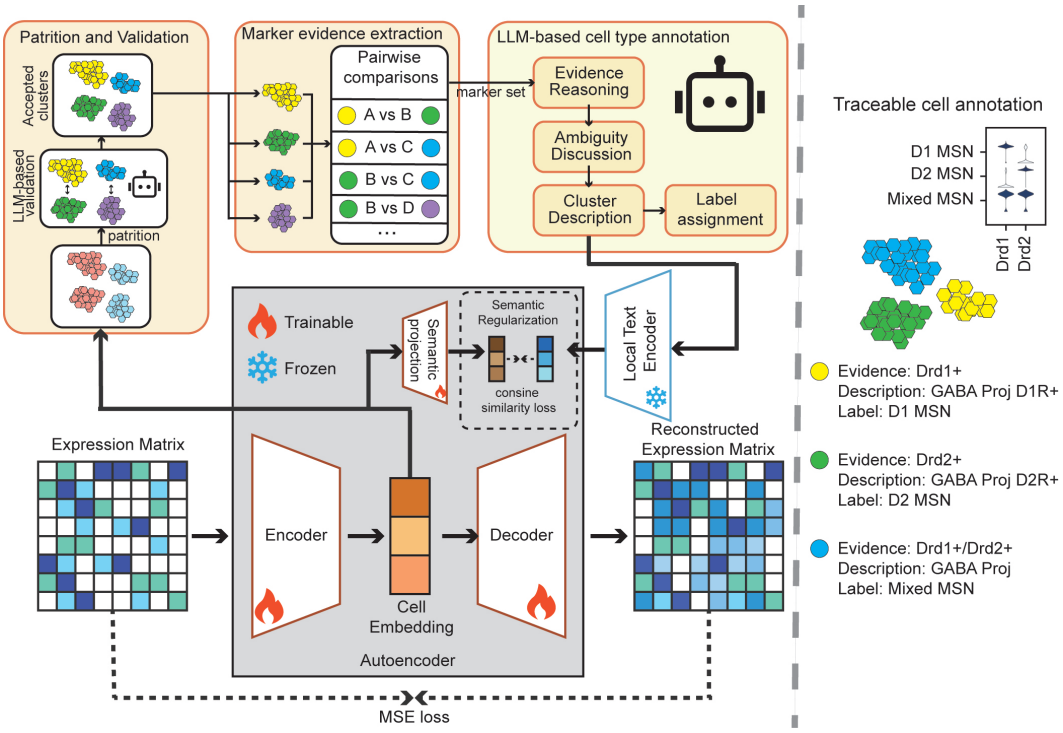

Figure 1: Overview of HiLL. At each refinement level, HiLL performs branch-local partition proposal, LLM-based split validation, level-wide marker extraction, LLM-based cell type annotation, and semantic regularization. The right panel shows the traceable annotation output for each cluster.

then summarized by differential expression gene analysis for LLM-based interpretation. This design enables flexible semantic labeling beyond a fixed reference taxonomy. As a result, existing methods are less suited to settings where different cell populations require different refinement depths and where semantic interpretation depends on comparison context tailored to the current branch of refinement.

Hierarchical annotation and multi-resolution cell identity. Prior work on hierarchical clustering, multi-resolution graph partitioning, and ontology-aware organization highlights that cell identity is inherently coarse-to-fine (Žurauskienė & Yau, 2016; Zou et al., 2021; Seth et al., 2022; Peng et al., 2021; Wang et al., 2021). In the reference-free setting, however, the challenge is not only to represent hierarchy but also to decide where refinement is needed, when a parent population should stop subdividing, and how marker evidence should be summarized at each refinement level for semantic annotation. Thus, hierarchical organization alone does not determine which candidate splits are biologically interpretable enough to retain, nor which comparison context is most appropriate for deriving marker evidence for LLM-based labeling.

3 Methods

3.1 Problem Formulation

Let $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ denote a preprocessed gene expression matrix, where n is the number of cells or spatial units and p is the number of genes. In the reference-free setting, the goal is to assign biologically meaningful labels to data-derived populations without access to an external atlas. This is inherently an open-vocabulary and multi-resolution problem: the method must infer both what a population represents and whether it should be further subdivided. We formulate annotation as a recursive decision process over a hierarchy of cell populations. Starting from the full dataset, each active population is either assigned a terminal semantic label or partitioned into child populations for further refinement.

3.2 HiLL: A FRAMEWORK FOR RECURSIVE SEMANTIC REFINEMENT

HiLL is a reference-free framework for recursive cell type annotation. At each refinement level, it performs branch-local partition proposal, LLM-based split validation, level-wide marker extraction, LLM-based annotation, and semantic regularization. For each active population, HiLL proposes a latent-representation-supported partition and retains it only if the resulting children are biologically distinguishable. Different branches may thus stop at different resolutions.

3.3 SHARED EXPRESSION REPRESENTATION

HiLL maintains a shared expression representation throughout recursive refinement. Given a cell expression profile $\mathbf{x}_i$, a Transformer-based encoder E (Zou et al., 2025) maps it to a latent representation $\mathbf{z}_i = E(\mathbf{x}_i) \in \mathbb{R}^d$. These embeddings define the representation space used for clustering and split proposal at each refinement level and are updated across refinement rounds.

To learn this representation, HiLL uses a decoder D_{rec} to reconstruct the input from $\mathbf{z}_i$, yielding

$$\mathcal{L}_{\text{recon}} = \frac{1}{b} \sum_{i=1}^{b} \|\mathbf{x}_i - D_{\text{rec}}(\mathbf{z}_i)\|_2^2 , \tag{1}$$

where b is the batch size. At the first refinement level, when semantic targets are not yet available, the representation is learned using only $\mathcal{L}_{\text{recon}}$. This ensures that early partition proposals are grounded in transcriptomic structure before language-derived semantic constraints are introduced.

3.4 PARTITION PROPOSAL AND LLM-BASED SPLIT VALIDATION

Because different cell populations may require different annotation granularities, HiLL determines subdivision for each active cluster through metric-based partition proposal followed by LLM-based split validation.

For each active cluster P, Leiden clustering is evaluated in the current latent embedding space over a candidate resolution set $\Gamma \subset [0.01, 1]$. For each $\gamma \in \Gamma$, we obtain a partition $\mathcal{C}_\gamma(P)$ and score it using the Silhouette Coefficient (SC) (Rousseeuw, 1987), Davies–Bouldin Index (DBI) (Davies & Bouldin, 1979), and Calinski–Harabasz Index (CHI). After transforming each metric so that larger values indicate better partitions and normalizing each metric to $[0, 1]$ within Γ, we define the representation-support score

$$Q_{\text{rep}}(\mathcal{C}_\gamma(P); P) = \sum_{m \in \mathcal{M}} \tilde{q}_m(\mathcal{C}_\gamma(P)) , \tag{2}$$

where $\mathcal{M}$ is the set of internal evaluation metrics and $\tilde{q}_m(\mathcal{C}_\gamma(P)) \in [0, 1]$ denotes the direction-adjusted normalized score of metric m for partition $\mathcal{C}_\gamma(P)$. We then select

$$\gamma^*(P) \in \arg\max_{\gamma \in \Gamma} Q_{\text{rep}}(\mathcal{C}_\gamma(P); P). \tag{3}$$

This stage identifies representation-supported candidate partitions, which are then subjected to semantic validation. The resulting partition is therefore treated only as a candidate. HiLL then validates the proposed split using branch-local comparative evidence derived from the candidate clusters. Only LLM-approved splits are adopted; otherwise, P remains unsplit.

3.4.1 BRANCH-LOCAL COMPARATIVE EVIDENCE

To assess whether a candidate partition reflects biologically meaningful heterogeneity, HiLL constructs branch-local comparative evidence within the proposed split of the current cluster. Let

$$\mathcal{C}_{\gamma^*}(P) = \{c_1, \ldots, c_m\}$$

denote the candidate child clusters obtained from the selected partition of parent cluster P. For every ordered pair (c_i, c_j) with $i \neq j$, we perform pairwise differential expression analysis and define

$$\mathcal{E}_{i \to j} = \{g : g \text{ is significantly upregulated in } c_i \text{ relative to } c_j\} , \tag{4}$$

where significance is determined by the same differential expression criteria used in Section 3.5. For each candidate child cluster c_i, we then collect its sibling-relative comparative profile

$$\mathcal{B}(c_i; P) = \{(j, \mathcal{E}_{i\to j}, \mathcal{E}_{j\to i}) : j \neq i\}. \tag{5}$$

This restricts validation to sibling comparisons: a split is accepted only if its children are distinguishable relative to the heterogeneity the split is meant to explain, avoiding justification by genes that merely separate the parent branch from distant populations. A compact summary of $\mathcal{B}(c_i; P)$ is provided to the LLM only for split validation, not for final labeling.

3.5 Level-wide marker extraction

After split validation at level t, HiLL performs level-wide marker extraction for the accepted clusters $\mathcal{C}_t = \{c_1, \ldots, c_m\}$ by comparing each cluster against all other clusters. This stage follows a level-wise contrastive principle: a marker for cluster c_i should significantly distinguish c_i from as many other clusters as possible in pairwise comparisons, rather than only in a one-vs-rest comparison.

For every ordered pair (c_i, c_j) with $i \neq j$, we perform pairwise differential expression analysis, computing $\log_2$ fold change, expression fraction, and a Wilcoxon rank-sum test with Benjamini–Hochberg correction. A gene g is treated as upregulated in c_i relative to c_j if

$$\log_2 \mathrm{FC}_g^{(c_i,c_j)} \geq 2, \qquad p_{g,\mathrm{adj}}^{(c_i,c_j)} < 0.05, \qquad \rho_g^{(c_i)} \geq 0.5, \tag{6}$$

where $\rho_g^{(c_i)}$ is the fraction of cells in cluster c_i with nonzero expression of gene g.

We summarize the discriminative strength of gene g for cluster c_i by

$$u_{g,c_i} = \sum_{\substack{j=1 \\ j\neq i}}^{m} \mathbf{1}\left[\log_2 \mathrm{FC}_g^{(c_i,c_j)} \geq 2 \ \wedge \ p_{g,\mathrm{adj}}^{(c_i,c_j)} < 0.05 \ \wedge \ \rho_g^{(c_i)} \geq 0.5\right]. \tag{7}$$

For each cluster c_i, the top-k genes ranked by u_{g,c_i} are used as markers for LLM-based annotation. Thus, genes ranked highly for c_i are repeatedly discriminative against multiple same-level alternatives, yielding markers that are more stable for semantic annotation than single-comparison DEG summaries.

3.6 LLM-based cell type annotation

Given the marker set of each accepted cluster, HiLL uses an external LLM to generate an open-vocabulary cell type label and a natural-language description. These descriptions serve both as interpretable summaries and as semantic targets for regularization. We use an external LLM for annotation and a frozen local text encoder for embedding computation. Implementation details are provided in Appendix A.1.

3.7 Semantic regularization

From the second refinement round onward, HiLL regularizes the shared expression representation using LLM-generated cluster descriptions. A semantic projection head D_{sem} maps $\mathbf{z}_i$ into the text embedding space, giving $\mathbf{s}_i = D_{\mathrm{sem}}(\mathbf{z}_i)$. Let $\mathbf{t}_c$ be the text embedding of the generated description for cluster c, and let $c(i)$ denote the cluster assignment of cell i. For each cell i, we align $\mathbf{s}_i$ with $\mathbf{t}_{c(i)}$ using cosine distance:

$$\mathcal{L}_{\mathrm{align}} = \frac{1}{b}\sum_{i=1}^{b}\left(1 - \frac{\mathbf{s}_i^\top \mathbf{t}_{c(i)}}{\|\mathbf{s}_i\|\,\|\mathbf{t}_{c(i)}\|}\right). \tag{8}$$

The full objective from the second refinement round onward is $\mathcal{L} = \mathcal{L}_{\mathrm{recon}} + \lambda\mathcal{L}_{\mathrm{align}}$, where λ balances reconstruction and semantic regularization. This loss function encourages cells from the same annotated cluster to align with a shared semantic prototype in text space while preserving transcriptomic reconstruction. As refinement proceeds, updated cluster descriptions provide progressively finer semantic targets for subsequent partition proposal. Architectural details are provided in Appendix A.1.4.

3.8 RECURSIVE EXECUTION AND STOPPING CONDITIONS

HiLL starts from the full dataset and proceeds level by level. For each active cluster, it proposes a candidate partition, validates it with branch-local evidence, and, if accepted, performs level-wide marker extraction, LLM-based annotation, and semantic regularization. Recursion stops when the maximum depth is reached, no non-trivial approved split remains, or a cluster is too small to subdivide.

4 EXPERIMENTS

We evaluate HiLL on human and mouse datasets and compare it with reference-free LLM-based methods, classical reference-mapping methods, supervised deep models, and zero-shot pretrained baselines. Dataset statistics are summarized in Appendix A.2. The benchmark comprises human pancreas, human brain, and mouse brain datasets spanning both scRNA-seq and single-cell-resolution ST platforms (Li et al., 2025; Siletti et al., 2023; Allen et al., 2023). Foundation-model baselines are evaluated in a zero-shot setting without task-specific fine-tuning to measure direct transferability.

Because predicted labels may differ from reference annotations in wording or granularity, we evaluate all methods under a Cell Ontology-grounded hierarchical matching protocol. Predicted and ground-truth labels are normalized and mapped to Cell Ontology terms (McLaughlin et al., 2025). Under **Descendant** matching, a prediction is considered correct if the mapped term is identical to the ground-truth term or corresponds to a descendant of it in the ontology, accommodating biologically compatible fine-grained predictions when the reference annotation is coarse. Under **Ancestor** matching, we additionally accept ancestor-level matches, allowing ontologically compatible predictions at a coarser resolution. Full details are provided in the supplementary material (A.3). All methods are benchmarked under the same ontology normalization and matching pipeline.

Table 1: Performance comparison on human pancreas, human brain, and mouse brain benchmarks under Cell Ontology-based hierarchical matching. Reference-free methods are reported under both ancestor and descendant matching; reference-based methods under ancestor matching; foundation models are evaluated zero-shot. Results are mean±std across datasets. Underlined values denote the best result within each method category, and bold values denote the overall best.

Method		Ref	Human Pancreas		Human Brain	Mouse Brain
			Best Ref	Median Ref	Acc./macro-F1	Acc./macro-F1
			Acc./macro-F1	Acc./macro-F1		
HiLL	Ancestor	×	**0.916$_{\pm 0.081}$**/0.583$_{\pm 0.151}$	**0.916$_{\pm 0.081}$**/0.583$_{\pm 0.151}$	0.945$_{\pm 0.074}$/0.667$_{\pm 0.098}$	**0.950$_{\pm 0.024}$/0.782$_{\pm 0.064}$**
GPTCellType (Hou & Ji, 2024)	Ancestor	×	0.586$_{\pm 0.171}$/0.592$_{\pm 0.142}$	0.586$_{\pm 0.171}$/0.592$_{\pm 0.142}$	0.654$_{\pm 0.145}$/0.489$_{\pm 0.163}$	0.882$_{\pm 0.117}$/0.707$_{\pm 0.057}$
LICT (Ye et al., 2025)	Ancestor	×	0.488$_{\pm 0.168}$/0.543$_{\pm 0.129}$	0.488$_{\pm 0.168}$/0.543$_{\pm 0.129}$	0.409$_{\pm 0.106}$/0.354$_{\pm 0.009}$	0.492$_{\pm 0.166}$/0.483$_{\pm 0.120}$
scPilot (Gao et al., 2026)	Ancestor	×	0.676$_{\pm 0.249}$/0.496$_{\pm 0.183}$	0.676$_{\pm 0.249}$/0.496$_{\pm 0.183}$	0.116$_{\pm 0.070}$/0.354$_{\pm 0.099}$	0.790$_{\pm 0.147}$/0.670$_{\pm 0.105}$
HiLL	Descendant	×	0.784$_{\pm 0.190}$/0.413$_{\pm 0.142}$	0.784$_{\pm 0.190}$/0.413$_{\pm 0.142}$	0.945$_{\pm 0.074}$/0.667$_{\pm 0.098}$	0.938$_{\pm 0.026}$/0.735$_{\pm 0.059}$
GPTCellType (Hou & Ji, 2024)	Descendant	×	0.407$_{\pm 0.197}$/0.261$_{\pm 0.089}$	0.407$_{\pm 0.197}$/0.261$_{\pm 0.089}$	0.654$_{\pm 0.145}$/0.489$_{\pm 0.163}$	0.704$_{\pm 0.110}$/0.598$_{\pm 0.067}$
LICT (Ye et al., 2025)	Descendant	×	0.344$_{\pm 0.199}$/0.233$_{\pm 0.095}$	0.344$_{\pm 0.199}$/0.233$_{\pm 0.095}$	0.407$_{\pm 0.105}$/0.326$_{\pm 0.092}$	0.340$_{\pm 0.147}$/0.351$_{\pm 0.089}$
scPilot (Gao et al., 2026)	Descendant	×	0.528$_{\pm 0.293}$/0.280$_{\pm 0.156}$	0.528$_{\pm 0.293}$/0.280$_{\pm 0.156}$	0.099$_{\pm 0.080}$/0.111$_{\pm 0.077}$	0.724$_{\pm 0.153}$/0.610$_{\pm 0.103}$
SingleR (Aran et al., 2019)		✓	0.855$_{\pm 0.175}$/**0.831$_{\pm 0.090}$**	0.742$_{\pm 0.257}$/**0.627$_{\pm 0.145}$**	0.990$_{\pm 0.004}$/0.920$_{\pm 0.065}$	0.867$_{\pm 0.031}$/0.821$_{\pm 0.021}$
CHETAH (de Kanter et al., 2019)		✓	0.789$_{\pm 0.147}$/0.673$_{\pm 0.169}$	0.420$_{\pm 0.165}$/0.441$_{\pm 0.118}$	0.762$_{\pm 0.060}$/0.847$_{\pm 0.084}$	0.539$_{\pm 0.082}$/0.643$_{\pm 0.061}$
scmap (Kiselev et al., 2018)		✓	0.273$_{\pm 0.098}$/0.345$_{\pm 0.115}$	0.084$_{\pm 0.064}$/0.167$_{\pm 0.078}$	0.519$_{\pm 0.023}$/0.551$_{\pm 0.145}$	0.444$_{\pm 0.070}$/0.430$_{\pm 0.045}$
TOSICA (Chen et al., 2023)		✓	0.877$_{\pm 0.125}$/0.623$_{\pm 0.106}$	0.547$_{\pm 0.271}$/0.370$_{\pm 0.095}$	0.984$_{\pm 0.004}$/0.638$_{\pm 0.105}$	0.465$_{\pm 0.107}$/0.452$_{\pm 0.065}$
scAdapt (Zhou et al., 2021)		✓	0.853$_{\pm 0.228}$/0.825$_{\pm 0.074}$	0.515$_{\pm 0.236}$/0.612$_{\pm 0.123}$	0.916$_{\pm 0.003}$/0.874$_{\pm 0.085}$	0.788$_{\pm 0.036}$/0.826$_{\pm 0.027}$
mtANN (Xiong et al., 2023)		✓	0.841$_{\pm 0.191}$/0.805$_{\pm 0.109}$	0.521$_{\pm 0.196}$/0.537$_{\pm 0.127}$	0.905$_{\pm 0.091}$/0.800$_{\pm 0.119}$	0.740$_{\pm 0.035}$/0.674$_{\pm 0.036}$
scFoundation (zero-shot) (Hao et al., 2024)		✓	0.906$_{\pm 0.082}$/0.785$_{\pm 0.090}$	0.702$_{\pm 0.230}$/0.527$_{\pm 0.102}$	**0.999$_{\pm 0.001}$/0.931$_{\pm 0.058}$**	0.571$_{\pm 0.066}$/0.639$_{\pm 0.055}$
CellPLM (zero-shot) (Wen et al., 2023)		✓	0.805$_{\pm 0.151}$/0.667$_{\pm 0.086}$	0.601$_{\pm 0.237}$/0.491$_{\pm 0.081}$	0.954$_{\pm 0.004}$/0.521$_{\pm 0.076}$	0.399$_{\pm 0.065}$/0.404$_{\pm 0.090}$
scGPT (zero-shot) (Cui et al., 2024)		✓	0.823$_{\pm 0.134}$/0.684$_{\pm 0.076}$	0.639$_{\pm 0.226}$/0.501$_{\pm 0.092}$	0.998$_{\pm 0.001}$/0.869$_{\pm 0.099}$	0.744$_{\pm 0.032}$/0.701$_{\pm 0.021}$

4.1 OVERALL ANNOTATION QUALITY ACROSS DIVERSE BENCHMARKS

Across datasets, HiLL outperforms existing reference-free LLM-based methods and remains competitive with several reference-based and pretrained baselines (Table 1). Under ancestor matching, HiLL achieves the best accuracy on human pancreas and the best accuracy and macro-F1 on mouse brain, while on human brain it remains competitive with the strongest zero-shot foundation model, scFoundation. For reference-based methods, only ancestor results are reported, as this setting more closely reflects their better performance, while descendant matching produces only marginal

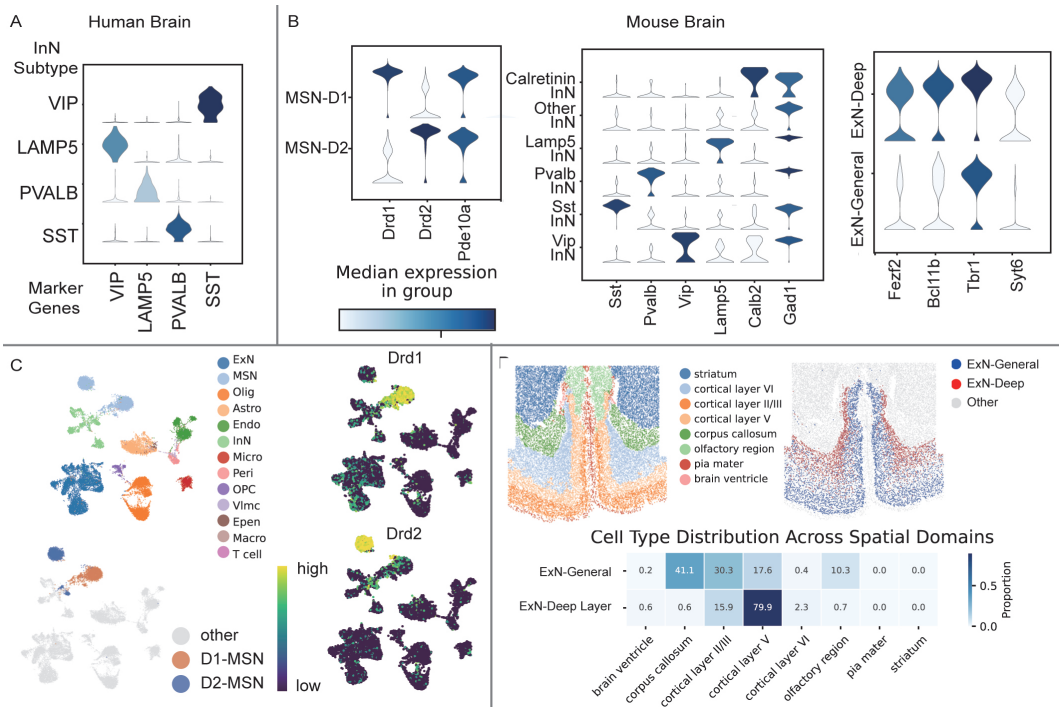

Figure 2: Examples of fine-grained subtype annotation in human and mouse brain datasets. (A) Marker expression of interneuron subtypes in human brain. (B) Marker expression of MSN, interneuron, and excitatory neuron subtypes in mouse brain. (C) Latent embedding of MSN subtypes with Drd1/Drd2 expression. (D) Spatial distribution of excitatory neuron subtypes in mouse brain.

changes. We evaluated statistical significance using paired Wilcoxon signed-rank tests, comparing HiLL against each baseline method across all datasets within the same tissue benchmark (Appendix A.11). These comparisons show that HiLL's improvements over reference-free LLM-based baselines are statistically significant in the majority of datasets.

4.2 Fine-grained subtype refinement

We next examine fine-grained subtype resolution in neuronal populations, where broad categories are easily identified, but the real difficulty lies in resolving transcriptionally similar subtypes. We illustrate this with two representative case studies.

Human brain (10X146_2). HiLL separates neurons into excitatory and interneuron populations, and then further resolves major interneuron subtypes, including *SST*+, *PVALB*+, *VIP*+, and *LAMP5*+ populations, consistent with marker expression patterns (Figure 2). By contrast, GPTCelltype recovers only a subset of these subtypes, as further illustrated by the Sankey diagrams in Appendix Figure 14.

Mouse brain spatial data (slice 12_0). HiLL resolves medium spiny neurons into D1 or D2, and also recovers canonical inhibitory and excitatory neuronal subtypes with spatially coherent distributions (Figure 2). GPTCelltype misannotates MSNs and misses several neuronal subtypes. The corresponding Sankey diagrams in Appendix Figure 15 further highlight this pattern.

4.3 Ablation study

We ablate four key components of HiLL: LLM-based split validation, level-wide marker extraction, semantic regularization, and recursive refinement. Table 2 compares the full model at different refinement level with three main variants: replacing level-wide marker extraction with DEG-based marker summary [HiLL (deg)], removing semantic regularization [HiLL (w/o SR)], and removing LLM-based split validation [HiLL (w/o SV)].

Table 2: Ablation results of HiLL variants on human pancreas (10 datasets) and mouse brain (18 datasets). Results are reported as ancestor-level means across datasets. Levels 0/1/2 denote increasing refinement depth.

Methods	Level	Human Pancreas		Mouse Brain	
		Acc	F1	Acc	F1
HiLL	Level 0	$0.849_{\pm 0.153}$	$0.386_{\pm 0.100}$	$0.943_{\pm 0.036}$	$0.677_{\pm 0.070}$
	Level 1	$0.925_{\pm 0.070}$	$0.545_{\pm 0.128}$	$0.955_{\pm 0.025}$	$0.747_{\pm 0.052}$
	Level 2	$0.916_{\pm 0.081}$	$0.583_{\pm 0.151}$	$0.950_{\pm 0.025}$	$0.782_{\pm 0.066}$
HiLL (deg)	Level 0	$0.647_{\pm 0.212}$	$0.355_{\pm 0.077}$	$0.824_{\pm 0.113}$	$0.634_{\pm 0.073}$
	Level 1	$0.688_{\pm 0.154}$	$0.483_{\pm 0.098}$	$0.770_{\pm 0.146}$	$0.669_{\pm 0.081}$
	Level 2	$0.707_{\pm 0.147}$	$0.514_{\pm 0.115}$	$0.727_{\pm 0.150}$	$0.682_{\pm 0.072}$
HiLL (w/o SR)	Level 0	$0.849_{\pm 0.153}$	$0.386_{\pm 0.100}$	$0.946_{\pm 0.024}$	$0.686_{\pm 0.048}$
	Level 1	$0.865_{\pm 0.147}$	$0.599_{\pm 0.159}$	$0.948_{\pm 0.032}$	$0.754_{\pm 0.056}$
	Level 2	$0.884_{\pm 0.135}$	$0.605_{\pm 0.154}$	$0.953_{\pm 0.025}$	$0.777_{\pm 0.048}$
HiLL (w/o SV)	Level 0	$0.850_{\pm 0.150}$	$0.382_{\pm 0.095}$	$0.935_{\pm 0.045}$	$0.679_{\pm 0.070}$
	Level 1	$0.898_{\pm 0.093}$	$0.519_{\pm 0.140}$	$0.956_{\pm 0.036}$	$0.748_{\pm 0.061}$
	Level 2	$0.889_{\pm 0.128}$	$0.583_{\pm 0.156}$	$0.952_{\pm 0.027}$	$0.784_{\pm 0.056}$

Effect of level-wide marker extraction. HiLL (deg) performs substantially worse than HiLL on human pancreas and mouse brain, indicating that level-wide marker extraction is more effective than the conventional DEG summaries used in existing LLM-driven methods. On the mouse brain datasets, its performance declines progressively as the number of clusters increases, suggesting that conventional DEG summaries may become less effective at capturing cluster-specific cell identity. This matches the intended role of level-wide marker extraction as context-stable evidence construction, favoring genes that remain discriminative across multiple same-level alternatives.

Effect of LLM-based split validation. HiLL differs only marginally from HiLL (w/o SV). Cluster-number statistics and deeper-round evaluations (Appendix Figure 10) confirm that LLM-based split validation curbs over-fragmentation and reduces redundant subclusters, while maintaining or even improving agreement with ground truth without sacrificing overall performance.

Effect of semantic regularization. Removing semantic regularization causes only minor changes in annotation accuracy, consistent with the pattern that it is not the primary driver of end-task performance. Its main effect is representation-level (Figure 3): improving embedding structure, cell-to-ontology retrieval, and cross-modal consistency between cell and text representations.

Effect of recursive refinement. Increasing refinement depth (Levels 0/1/2) consistently improves macro-F1 for HiLL on both benchmarks, with only marginal accuracy variation. This supports HiLL's central design principle that annotation granularity is better treated as a sequence of local refinement decisions than as a single global clustering choice.

Overall, recursive refinement and level-wide marker extraction are the key contributors to HiLL's gains, consistent with Table 2. LLM-based split validation mildly regularizes against over-fragmentation with little accuracy impact, while semantic regularization mainly affects embedding structure rather than annotation metrics (Figure 3). HiLL also shows robust stability across backends and random seeds (Appendix A.8). Furthermore, replacing the LLM-based labeling module with CellMarker-basedHu et al. (2023) lookup (using either HiLL markers or conventional DEG markers) generally lowers annotation quality, especially under descendant matching, indicating that LLM-based labeling is better able to resolve fine-grained subtype identities (Appendix Table 5).

4.4 Representation-level effects of semantic regularization

Table 2 shows that semantic regularization does not drive annotation performance, we examine its role in shaping the representation space, focusing on mouse brain data and reporting pancreas results in supplementary analyses.

Embedding geometry. Semantic regularization improves embedding geometry: as shown in Figure 3(a), increasing λ from 0.01 to 100 gradually raises the Silhouette Coefficient, with only minor changes in annotation accuracy and macro-F1.

Cell-to-ontology retrieval. To assess ontology-level semantic organization, we projected cell embeddings into the text space and retrieved Cell Ontology terms by cosine similarity, using candidate terms constructed from each dataset's annotated cell types and encoded with the same frozen local text encoder (Appendix A.4). Retrieval quality, measured by mean reciprocal rank (MRR), was higher for HiLL than for HiLL (w/o SR), indicating that semantic regularization improves alignment between cell representations and ontology-level semantic concepts.

Cross-modal structural consistency. We further examined whether semantic regularization induces a meaningful correspondence between the cell and text representation spaces. For each dataset, we computed cluster-level similarity matrices in both the learned cell embedding space and the text embedding space of LLM-generated cluster descriptions. As illustrated by the mouse brain heatmaps in Figure 3(d), the relative similarity structure among clusters is broadly preserved across modalities, with consistent positive correlations across datasets (most exceeding 0.5; the vast majority significant under Mantel permutation testing, $p = 0.0001$; Appendix A.5). An additional representative heatmap is shown in Appendix Figure 5.

Overall, semantic regularization mainly improves embedding organization, cell-to-ontology retrieval, and cross-modal consistency, with limited impact on annotation accuracy. These effects are stronger in mouse brain data; in human pancreas, geometry-level effects are weaker, although cell-to-ontology retrieval still improves significantly under semantic regularization (Appendix Figure 4).

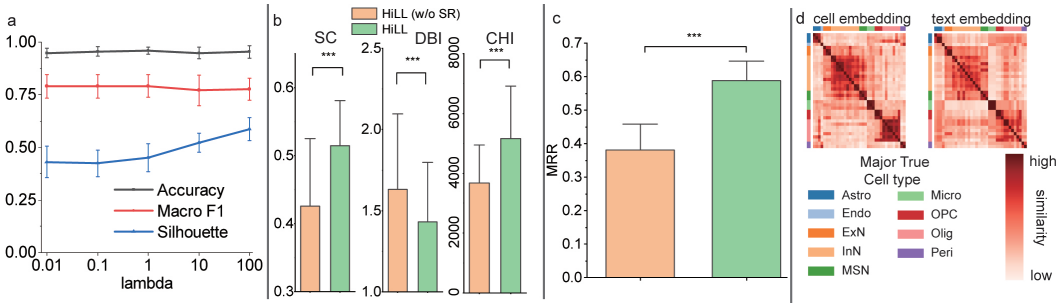

Figure 3: Representation-level effects of semantic regularization in mouse brain datasets. (a) Accuracy, macro-F1, and Silhouette Coefficient of HiLL at different λ. (b) Embedding geometry without and with semantic regularization, measured by Silhouette Coefficient, DBI, and CHI. (c) Cell-to-ontology retrieval performance. (d) Cross-modal similarity structure in the cell and text embedding spaces.

5 Conclusion and Limitation

We presented HiLL, a hierarchical LLM framework for reference-free cell type annotation that enables adaptive subtype refinement and semantic labeling. Across human and mouse scRNA-seq and ST datasets, HiLL consistently outperforms prior reference-free LLM baselines, particularly on fine-grained subtype distinctions, while remaining competitive with several reference-based and zero-shot pretrained methods. Ablation results show that the main gains come from recursive refinement and level-wide marker extraction, whereas semantic regularization mainly improves representation quality. Overall, HiLL provides an effective and interpretable framework for open-vocabulary cell type annotation without external references, supporting the view that reference-free annotation is better treated as a sequence of local refinement decisions than as a one-shot cluster-then-label pipeline. However, several limitations remain. As a cluster-level method, HiLL may degrade when candidate subpopulations are subtle, weakly separated, or highly imbalanced; its recursive refinement is path-dependent, so early split or stop decisions can shape subsequent hierarchy levels; and because it is designed for the fully reference-free setting, it does not currently incorporate optional reference information, limiting its ability to exploit high-quality labeled datasets when available.

AI USE STATEMENT

In this work, we use generative AI tools only for auxiliary purposes during research and manuscript editing. The core algorithm, model architecture, and evaluation metrics were designed by the authors and were not generated by AI. At a later stage of development, generative AI was used to assist in code performance optimization, without changing the core algorithmic logic or methodological design. Generative AI was also used to help produce scripts for figure generation and statistical data processing, as well as to improve the readability and clarity of the manuscript. In addition, the proposed framework itself is LLM-based: an LLM is used as a cell type annotation engine as part of the method.

REPRODUCIBILITY STATEMENT

To support reproducibility, the supplementary materials include the HiLL codebase, together with scripts for preprocessing, training, recursive refinement, and evaluation. All datasets used in this work are publicly available, and their sources are documented in the main text and Appendix. The paper and supplementary materials provide key implementation details, including model architecture, hyperparameter settings, prompt design, and evaluation protocol. We additionally include representative experimental outputs and partial result files generated by the released code to help verify the reported findings.

REFERENCES

William E Allen, Timothy R Blosser, Zuri A Sullivan, Catherine Dulac, and Xiaowei Zhuang. Molecular and spatial signatures of mouse brain aging at single-cell resolution. *Cell*, 186(1): 194–208, 2023.

Dvir Aran, Agnieszka P. Looney, Leqian Liu, Esther Wu, Valerie Fong, Austin Hsu, Suzanna Chak, Ram P. Naikawadi, Paul J. Wolters, Adam R. Abate, Atul J. Butte, and Mallar Bhattacharya. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature Immunology*, 20(2):163–172, 2019. ISSN 1529-2916. doi: 10.1038/s41590-018-0276-y. URL https://doi.org/10.1038/s41590-018-0276-y.

Junyue Cao, Jonathan S Packer, Vijay Ramani, Darren A Cusanovich, Chau Huynh, Riza Daza, Xiaojie Qiu, Choli Lee, Scott N Furlan, Frank J Steemers, et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, 357(6352):661–667, 2017.

Jiawei Chen, Hao Xu, Wanyu Tao, Zhaoxiong Chen, Yuxuan Zhao, and Jing-Dong J. Han. Transformer for one stop interpretable cell type annotation. *Nature Communications*, 14(1):223, 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-35923-4. URL https://doi.org/10.1038/s41467-023-35923-4.

Jiawen Chen, Jianghao Zhang, Huaxiu Yao, and Yun Li. Celltypeagent: Trustworthy cell type annotation with large language models. *ArXiv*, pp. arXiv–2505, 2025.

Simone Codeluppi, Lars E. Borm, Amit Zeisel, Gioele La Manno, Josina A. van Lunteren, Camilla I. Svensson, and Sten Linnarsson. Spatial organization of the somatosensory cortex revealed by osmfish. *Nature Methods*, 15(11):932–935, 2018. ISSN 1548-7105. doi: 10.1038/s41592-018-0175-z. URL https://doi.org/10.1038/s41592-018-0175-z.

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*, 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02201-0. URL https://doi.org/10.1038/s41592-024-02201-0.

D. L. Davies and D. W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, 1979. ISSN 1939-3539. doi: 10.1109/TPAMI.1979.4766909.

J. K. de Kanter, P. Lijnzaad, T. Candelli, T. Margaritis, and F. C. P. Holstege. Chetah: a selective, hierarchical cell type identification method for single-cell rna sequencing. *Nucleic Acids Res*, 47 (16):e95, 2019. ISSN 0305-1048 (Print) 0305-1048. doi: 10.1093/nar/gkz543.

Yiming Gao, Zhen Wang, Jefferson Chen, Mark Antkowiak, Mengzhou Hu, JungHo Kong, Dexter Pratt, Jieyuan Liu, Enze Ma, and Zhiting Hu. scpilot: Large language model reasoning toward automated single-cell analysis and discovery. *arXiv preprint arXiv:.11609*, 2026.

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 21(8):1481–1491, 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02305-7. URL https://doi.org/10.1038/s41592-024-02305-7.

Wenpin Hou and Zhicheng Ji. Assessing gpt-4 for cell type annotation in single-cell rna-seq analysis. *Nature Methods*, 21(8):1462–1465, 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02235-4. URL https://doi.org/10.1038/s41592-024-02235-4.

Congxue Hu, Tengyue Li, Yingqi Xu, Xinxin Zhang, Feng Li, Jing Bai, Jing Chen, Wenqi Jiang, Kaiyue Yang, Qi Ou, et al. Cellmarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scrna-seq data. *Nucleic acids research*, 51(D1):D870–D876, 2023.

Yidi Huang, Ivan Cohen, Van Quynh-Thi Truong, Pedram B Bayat, Sameer A Bhatti, Luca Paruzzo, Mark M Painter, Shirong Zheng, Derek Alan Oldridge, Joost Wagenaar, et al. Reference-free cell-type annotation with llm agents. In *ICLR 2025 Workshop on Machine Learning for Genomics Explorations*, 2025.

Vladimir Yu Kiselev, Andrew Yiu, and Martin Hemberg. scmap: projection of single-cell rna-seq data across data sets. *Nature Methods*, 15(5):359–362, 2018. ISSN 1548-7105. doi: 10.1038/nmeth.4644. URL https://doi.org/10.1038/nmeth.4644.

Mengwei Li, Kok Siong Ang, Brian Teo, Uddamvathanak Rom, Minh N Nguyen, Sebastian Maurer-Stroh, and Jinmiao Chen. Rediscovering publicly available single-cell data with the disco platform. *Nucleic Acids Research*, 53(D1):D932–D938, 01 2025. ISSN 1362-4962. doi: 10.1093/nar/gkae1108. URL https://doi.org/10.1093/nar/gkae1108.

James McLaughlin, Josh Lagrimas, Haider Iqbal, Helen Parkinson, and Henriette Harmse. Ols4: a new ontology lookup service for a growing interdisciplinary knowledge ecosystem. *Bioinformatics*, 41(5):btaf279, 2025. ISSN 1367-4811. doi: 10.1093/bioinformatics/btaf279. URL https://doi.org/10.1093/bioinformatics/btaf279.

Marie Mikulová, Milan Straka, Jan Štěpánek, Barbora Štěpánková, and Jan Hajic. Quality and efficiency of manual annotation: Pre-annotation bias. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 2909–2918, 2022.

J. R. Moffitt, D. Bambah-Mukku, S. W. Eichhorn, E. Vaughn, K. Shekhar, J. D. Perez, N. D. Rubinstein, J. Hao, A. Regev, C. Dulac, and X. Zhuang. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science*, 362(6416), 2018. ISSN 0036-8075 (Print) 0036-8075. doi: 10.1126/science.aau5324.

Mariana Neves and Jurica Ševa. An extensive review of tools for manual annotation of documents. *Briefings in bioinformatics*, 22(1):146–163, 2021.

Minshi Peng, Brie Wamsley, Andrew G Elkins, Daniel H Geschwind, Yuting Wei, and Kathryn Roeder. Cell type hierarchy reconstruction via reconciliation of multi-resolution cluster tree. *Nucleic acids research*, 49(16):e91–e91, 2021.

Mireya Plass, Jordi Solana, F Alexander Wolf, Salah Ayoub, Aristotelis Misios, Petar Glažar, Benedikt Obermayer, Fabian J Theis, Christine Kocks, and Nikolaus Rajewsky. Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science*, 360(6391): eaaq1723, 2018.

Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. ISSN 0377-0427. doi: https://doi.org/10.1016/0377-0427(87)90125-7. URL https://www.sciencedirect.com/science/article/pii/0377042787901257.

Soumita Seth, Saurav Mallik, Tapas Bhadra, and Zhongming Zhao. Dimensionality reduction and louvain agglomerative hierarchical clustering for cluster-specified frequent biomarker discovery in single-cell sequencing data. *Frontiers in genetics*, 13:828479, 2022.

Kimberly Siletti, Rebecca Hodge, Alejandro Mossi Albiach, Ka Wai Lee, Song-Lin Ding, Lijuan Hu, Peter Lönnerberg, Trygve Bakken, Tamara Casper, Michael Clark, Nick Dee, Jessica Gloe, Daniel Hirschstein, Nadiya V. Shapovalova, C. Dirk Keene, Julie Nyhus, Herman Tung, Anna Marie Yanny, Ernest Arenas, Ed S. Lein, and Sten Linnarsson. Transcriptomic diversity of cell types across the adult human brain. *Science*, 382(6667):eadd7046, 2023. doi: 10.1126/science.add7046. URL https://www.science.org/doi/abs/10.1126/science.add7046.

Sheng Wang, Angela Oliveira Pisco, Aaron McGeever, Maria Brbic, Marinka Zitnik, Spyros Darmanis, Jure Leskovec, Jim Karkanias, and Russ B Altman. Leveraging the cell ontology to classify unseen cell types. *Nature communications*, 12(1):5556, 2021.

Hongzhi Wen, Wenzhuo Tang, Xinnan Dai, Jiayuan Ding, Wei Jin, Yuying Xie, and Jiliang Tang. Cellplm: pre-training of cell language model beyond single cells. *bioRxiv*, pp. 2023.10.03.560734, 2023.

Yi-Xuan Xiong, Meng-Guo Wang, Luonan Chen, and Xiao-Fei Zhang. Cell-type annotation with accurate unseen cell-type identification using multiple references. *PLOS Computational Biology*, 19(6):e1011261, 2023. ISSN 1553-734X.

Hao Xu, Shuyan Wang, Minghao Fang, Songwen Luo, Chunpeng Chen, Siyuan Wan, Rirui Wang, Meifang Tang, Tian Xue, Bin Li, et al. Spacel: deep learning-based characterization of spatial transcriptome architectures. *Nature Communications*, 14(1):7603, 2023.

Wenjin Ye, Yuanchen Ma, Junkai Xiang, Hongjie Liang, Jintian Luo, Yuantao Li, Tao Wang, Qiuling Xiang, Wu Song, and Weiqiang Li. Evaluation of cell type annotation reliability using a large language model-based identifier. *Communications Biology*, 8(1):1360, 2025. ISSN 2399-3642.

Xiang Zhou, Hua Chai, Yuansong Zeng, Huiying Zhao, and Yuedong Yang. scadapt: virtual adversarial domain adaptation network for single cell rna-seq data classification across platforms and species. *Briefings in Bioinformatics*, 22(6):bbab281, 2021. ISSN 1467-5463.

Zhiyi Zou, Ying Liu, Yuting Bai, Jiawei Luo, and Zhaolei Zhang. sctrans: Sparse attention powers fast and accurate cell type annotation in single-cell rna-seq data. *PLOS Computational Biology*, 21(4):e1012904, 2025.

Ziheng Zou, Kui Hua, and Xuegong Zhang. Hgc: fast hierarchical clustering for large-scale single-cell data. *Bioinformatics*, 37(21):3964–3965, 2021.

Justina Žurauskienė and Christopher Yau. pcareduce: hierarchical clustering of single cell transcriptional profiles. *BMC bioinformatics*, 17(1):140, 2016.

A Appendix

A.1 Implementation details

A.1.1 Roles of branch-local and level-wide evidence.

HiLL uses two different comparison contexts for two different purposes. Branch-local comparative evidence is constructed within a candidate split and is used only for LLM-based split validation, i.e., to assess whether the proposed subdivision is biologically meaningful enough to refine further. After all accepted clusters at a refinement level have been determined, HiLL then performs level-wide

marker extraction for semantic annotation. In this stage, each accepted cluster is compared against the other accepted clusters at the same level so that the resulting marker set reflects its distinctiveness under the final refinement context. Thus, branch-local evidence is used for split decision, whereas level-wide evidence is used for marker summarization and labeling.

In contrast, the conventional approach extracts markers for each cluster in a one-vs-rest manner, i.e., by comparing that cluster against all remaining clusters pooled together into a single background. Such a comparison collapses heterogeneous alternatives into one aggregate, so a marker can appear significant simply by differing from the pooled average, even if it fails to distinguish the target cluster from any individual neighboring cluster. Consequently, the resulting markers may reflect a coarse separation from the global background rather than a genuine, cluster-specific distinction.

Thus, branch-local evidence is used for split decision, whereas level-wide evidence is used for marker summarization and labeling.

A.1.2 Roles of the external LLM backend and local text encoder

HiLL relies on two language-related components with clearly separated roles.

Roles of the language components. HiLL uses two language components with distinct roles. An external LLM backend is used only for cluster-level semantic generation, producing labels and descriptions from marker evidence. A separate frozen local text encoder is used only for embedding computation, including encoding generated descriptions for expression–text regularization and initializing gene embeddings from textual gene summaries.

External LLM backend. The external LLM backend is used only during the LLM-based annotation stage. Given the marker set of a candidate child cluster, it generates two textual outputs: (i) an open-vocabulary cell type label, and (ii) a natural-language cluster description. These outputs are used as annotation results and as semantic supervision targets for later stages. The external LLM backend is not used to compute embeddings and is not updated during training.

Local text encoder. A separate frozen local text encoder is used only for embedding computation. First, it encodes the LLM-generated cluster descriptions into text embeddings for expression–text alignment. Second, it is used to initialize gene embeddings from textual gene summaries collected from NCBI. The local text encoder does not generate labels or descriptions.

This separation is intentional: semantic generation is delegated to a powerful external LLM, while all embedding computation is handled by a fixed local encoder to ensure stable and reproducible text representations throughout recursive refinement.

A.1.3 Local Text encoder

Unless otherwise specified, we use the open-source Qwen3-0.6B model as the local text encoder. The text encoder serves two purposes in HiLL: encoding LLM-generated cluster descriptions for semantic regularization, and initializing gene embeddings from textual gene summaries. To assess the sensitivity of HiLL to the encoder choice, we replace Qwen3-0.6B with two alternative models in an ablation study: sciBert and BioBERT. BioBERT is a biomedical-domain encoder further pretrained on large-scale biomedical text. The corresponding results are provided in the supplementary ablation section.

For all text encoders, we use the `[CLS]` token representation as the text embedding. The encoder is frozen throughout training.

A.1.4 Autoencoder architecture

The semantically guided autoencoder comprises a sparse-attention encoder, a reconstruction decoder, and a semantic decoder.

Encoder. We use a Transformer-style encoder with sparse attention to model transcriptomic profiles. Each gene is first mapped to a dense embedding vector, and gene-level information is then

aggregated through self-attention. To better match the sparse nature of transcriptomic data and improve computational efficiency, attention is computed only over genes with nonzero expression values in each cell. To improve encoder robustness, we apply lightweight stochastic perturbations to the input during training.

Gene embeddings are initialized from textual gene descriptions. Specifically, we collect the summary text of each gene from NCBI and encode it with the text encoder to obtain the corresponding gene embedding. If no textual description exists for a gene, that gene is excluded from the entire representation training process.

The encoder uses a gene embedding dimension of 128, 2 attention heads, 1 Transformer layer, and a feedforward hidden dimension of 384.

Decoders. Both the reconstruction decoder and the semantic decoder are implemented as 3-layer multilayer perceptrons (MLPs) with ReLU activations. The input to each decoder is the latent embedding produced by the encoder. The reconstruction decoder outputs the reconstructed gene expression vector, whereas the semantic decoder projects the latent representation into the text embedding space. For both decoders, the hidden-layer dimension is set to $(inputdimension + outputdimension)/2$.

Input perturbation during autoencoder training. To improve robustness of the expression encoder, we apply lightweight stochastic perturbations to the input during training. For each mini-batch, we randomly sample $0 \sim 2$ perturbation strategies from the following set: (1) random masking of observed gene tokens, with retention probabilities sampled uniformly from 0.7 to 0.95; (2) additive Gaussian noise on nonzero expression values, with the noise scale sampled uniformly from 0.05 to 0.15; and (3) partial shuffling of a subset of observed gene-expression features within each cell, with shuffle rates sampled uniformly from 0.1 to 0.3. These perturbations are used during training and serve as generic robustness enhancement for representation learning.

Training hyperparameters. At each level, the autoencoder is trained for 40 epochs using AdamW, with a batch size of 128, a learning rate of 5×10^{-4} and a weight decay of 0.01.

A.2 Datasets

To characterize benchmark difficulty more objectively, we summarize each dataset using quantitative statistics of label structure rather than assigning subjective categories such as coarse or fine-grained. For each datasets, we report the total number of cells/spots, unique annotated labels, the median number of cells per label, the minimum and maximum class sizes.

Table 3 reports these statistics for all datasets. The benchmark spans substantial variation in dataset size, label cardinality, and class balance. In particular, the human pancreas collection exhibits the largest variation in label cardinality and datasets contain extremely small classes, indicating the presence of rare cell populations. All datasets used in this work are publicly available. The corresponding download portals for the three benchmark groups are: Human pancreas (Li et al., 2025), Human brain (ventrolateral prefrontal cortex) (Siletti et al., 2023), and Mouse brain (Allen et al., 2023).

For reference-based methods, the human brain reference dataset was taken from the public SPACEL(human_MTG_snrna) (Xu et al., 2023). The mouse brain reference dataset was obtained from the GEO study GSE113576 (`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE113576`), and the corresponding cell type annotations were taken from the associated publication (Moffitt et al., 2018).

A.3 Cell Ontology Mapping for Open-Vocabulary Evaluation

Because reference-free annotation methods may produce open-vocabulary labels, direct string matching between predicted labels and dataset annotations can be unreliable. To standardize evaluation, we map both predicted labels and ground-truth labels to **Cell Ontology (CL)** terms before applying the hierarchical matching criteria used in the main text.

Table 3: Quantitative dataset-complexity statistics for the benchmark datasets. Median/Min/Max Count denote the median/minimum/maximum number of cells per label.

dataset		median count	min count	max count	num labels	total count
human pancreas	CRA001160	23.5	1	24726	28	48304
	GSE148073	68.5	1	3314	18	9686
	GSE150724	261	1	3865	16	14441
	GSE153834	14.5	1	3992	20	7783
	GSE197177	2	1	342	18	492
	GSE202494	29	1	6874	21	9118
	GSE205049	33	1	1071	31	5333
	GSE212966	219	5	2690	35	20168
	GSE217837	186.5	1	3820	18	11102
	GSE221156	210.5	1	22374	24	61942
human brain	10X146_2	17	1	6983	9	7297
	10X146_3	30	1	7024	8	7341
	10X269_1	20.5	1	5884	10	6361
	10X269_2	11.5	1	6309	10	6768
	10X276_5	44	1	4671	9	5175
	10X276_6	94	1	4823	6	5317
mouse brain	12_0	791	3	6521	13	21727
	12_1	831	4	6616	13	24452
	13_0	622	8	4829	13	17778
	13_1	451	1	3349	13	10937
	13_2	487	13	3118	13	12481
	14_0	889	23	5368	13	22517
	14_1	1175	8	7990	13	26678
	15_0	505	6	5540	13	15358
	15_1	868	19	5904	13	24021
	16_0	912	7	6680	13	23922
	16_1	771	7	6542	13	21492
	17_0	1187	20	7467	13	25676
	17_1	885	20	4916	13	20151
	18_0	999	9	6699	13	22573
	18_1	850	40	5018	13	17734
	19_0	421	4	3184	13	10758
	19_1	695	11	5079	13	18631
	19_2	281	4	2394	13	9048

A.3.1 Ontology resource

We use the Cell Ontology distributed in OBO Graph JSON format from From this resource, we extract all ontology nodes with type `CLASS` that have a valid identifier and label and are not marked as deprecated. For each retained ontology term, we collect (i) the CL identifier, (ii) the canonical ontology label, and (iii) the available synonym strings. Ontology identifiers are normalized to a unified `CL:xxxxxxx` format.

A.3.2 Automatic candidate generation

Each predict label is compared against Cell Ontology terms using a fixed sequential procedure to generate candidate mappings for review:

(1) Exact label match. The normalized query is first matched against normalized canonical ontology labels.

(2) Exact synonym match. If no exact label match is found, the normalized query is matched against normalized ontology synonyms.

(3) Fuzzy candidate retrieval. If exact matching fails, we perform approximate matching against the union of normalized ontology labels and normalized ontology synonyms using token-sort string similarity. The highest-ranking candidate(s) are retained as references for subsequent review.

(4) No automatic candidate. If none of the above steps yields a reasonable candidate, the label is marked for direct manual inspection.

A.3.3 MANUAL CURATION OF ONTOLOGY MAPPING

Automatic matching is only used to assist candidate identification and is not treated as a final mapping procedure. All automatically generated matches are manually reviewed, and final CL term assignments are determined through manual curation. This step is particularly important for labels containing compositional modifiers, dataset-specific terminology, ambiguous abbreviations, or uncommon lexical variants, for which automated matching alone may be insufficient.

During manual curation, inappropriate automatic matches are corrected, ambiguous cases are resolved, and labels without reliable automatic candidates are assigned to the most appropriate CL term when supported by the label expression and annotation context. The same manual review procedure is applied uniformly to predictions from all methods and to the corresponding ground-truth labels when ontology mapping is required.

A.3.4 HIERARCHICAL MATCHING ON THE ONTOLOGY GRAPH

After final ontology curation, evaluation is performed on ontology terms rather than raw strings. Let y denote the ground-truth CL term and $\hat{y}$ denote the predicted CL term.

- **Ancestor matching.** A prediction is considered correct if $\hat{y} = y$, or if $\hat{y}$ is the direct parent or direct child of y in the Cell Ontology graph.
- **Descendant matching.** A prediction is considered correct if $\hat{y} = y$, or if $\hat{y}$ is a direct child of y.

A.4 CELL-TO-ONTOLOGY RETRIEVAL

To assess ontology-level semantic regularization, we performed cell-to-ontology retrieval in the shared text embedding space. For each dataset, we collected the unique cell types in the reference annotations, mapped them to Cell Ontology terms, and encoded the corresponding term names and textual descriptions with the same frozen text encoder used in HiLL to form a dataset-specific candidate set. Ontology term names and descriptions were obtained from Cell Ontology through OLS4.

For each cell i, we used its semantic projection $s_i = D_{\text{sem}}(z_i)$ as its text-space representation and ranked candidate ontology terms by cosine similarity. The correct target was the ontology term corresponding to the cell's ground-truth annotation. We report mean reciprocal rank (MRR), averaged over cells within each dataset.

This metric evaluates whether the learned representation is organized according to ontology-level semantic structure, independently of the open-vocabulary labels generated by the LLM. Ontology term names and descriptions were obtained from Cell Ontology through OLS4(McLaughlin et al., 2025).

A.5 CROSS-MODAL SIMILARITY STRUCTURE BETWEEN CELL AND TEXT EMBEDDINGS

To further examine the basis of semantic regularization, we asked whether the semantic embedding space defined by LLM-generated cluster descriptions preserves similarity structure that is consistent with the learned cell representation space.

For each dataset, we first obtained a cluster-level representation in the cell embedding space by averaging normalized cell embeddings within each predicted cluster to form a cluster centroid. We then embedded the corresponding LLM-generated natural-language description of each cluster using the same frozen text encoder employed for semantic regularization. Based on these cluster-level representations, we constructed two cosine-similarity matrices: a cell-space similarity matrix among cluster centroids and a text-space similarity matrix among cluster descriptions.

To quantify the agreement between the two modalities, we extracted the upper-triangular entries of the two similarity matrices and computed their Pearson correlation. Because pairwise similarity entries are not independent, we additionally assessed significance using a permutation-based Mantel test with 9,999 permutations. In each permutation, cluster identities in one matrix were jointly shuffled across rows and columns, and the matrix-level correlation was recomputed to form the null distribution.

Across datasets, we observed consistent positive agreement between the two similarity structures. Most datasets showed Pearson correlations greater than 0.5, and were significant under the Mantel permutation test ($p < 0.05$ at the resolution of 9,999 permutations). This pattern was observed in both human pancreas and mouse brain benchmarks, indicating that the semantic space induced by generated cluster descriptions captures a substantial part of the relational organization present in the learned cell embedding space.

Representative examples of human pancreas are shown in Figure 5, where the block structure of the cluster-level similarity matrices is visibly consistent across the two modalities. Complete dataset-level statistics are provided in Table 4. Overall, these results suggest that semantic regularization is not merely injecting arbitrary textual information, but is instead coupling the learned cell representation space to a text space with partially aligned cluster-level geometry.

A.6 Representation-level effects of semantic regularization on human pancreas

To assess whether the representation-level effects of semantic regularization extend beyond mouse brain data, we repeated the embedding-geometry and ontology-retrieval analyses on the human pancreas benchmarks, comparing HiLL with HiLL (w/o SR).

HiLL showed favorable average trends in embedding geometry, with higher Silhouette Coefficient and Calinski–Harabasz index and lower Davies–Bouldin index than the control variants, although these differences were not consistently statistically significant across datasets. By contrast, cell-to-ontology retrieval performance, measured by mean reciprocal rank (MRR), was significantly improved relative to both HiLL (w/o SR) (Figure 4). Overall, these results suggest that semantic regularization yields clearer ontology-level benefits than geometry-level gains in the human pancreas benchmarks.

A.7 Replacing LLM-based labeling with CellMarker

To assess the contribution of the labeling module itself, we replaced the LLM-based cell type annotation step with the official CellMarker APIHu et al. (2023), which infers cell types directly from an input marker gene list. For a controlled comparison, CellMarker was applied to the same final clusters produced by HiLL. Specifically, cellmarker (hill) uses the marker genes that HiLL supplies to the LLM, whereas cellmarker (deg) uses conventional DEG-based marker lists.

As shown in Table 5, replacing the LLM with direct database lookup generally reduces annotation quality, especially under the descendant-level evaluation. The drop is most pronounced in human pancreas, where both CellMarker-based variants perform substantially worse than HiLL under ancestor and descendant matching. In human brain and mouse brain datasets, CellMarker can achieve competitive or even higher accuracy in some ancestor-level settings, but this advantage does not consistently translate to macro-F1 or descendant performance. This pattern suggests that direct marker-to-database matching may favor frequent or coarse labels that improve top-level agreement, while being less effective at recovering balanced and fine-grained subtype identities.

These results highlight an important advantage of LLM-based labeling in the reference-free setting. Rather than performing direct lookup against a fixed marker database, the LLM can integrate

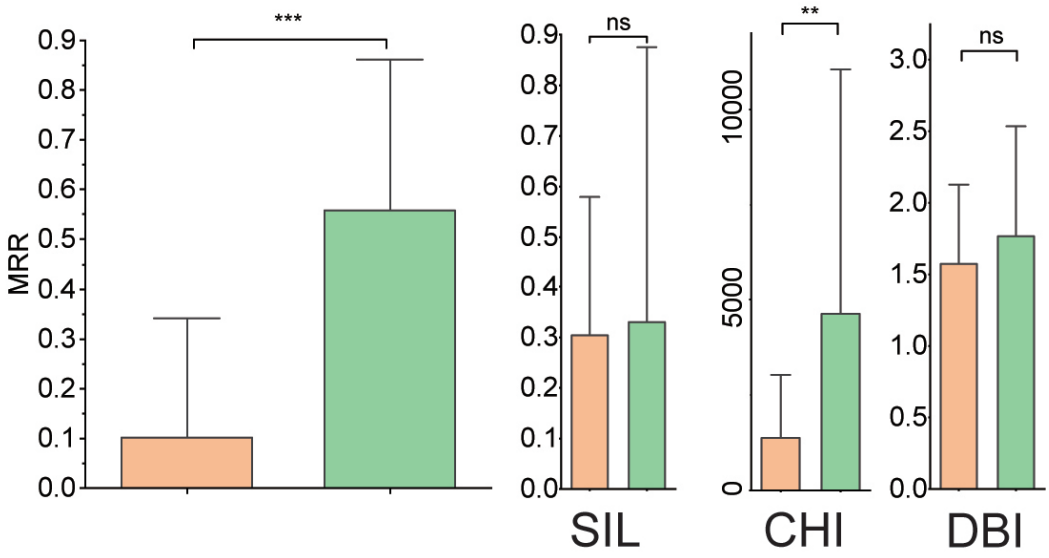

Figure 4: Effects of semantic regularization on human pancreas datasets. (A) Embedding geometry for HiLL and HiLL (w/o SR), measured by SC, DBI, and CHI. (B) Cell-to-ontology retrieval for HiLL and HiLL (w/o SR), measured by mean reciprocal rank (MRR). HiLL shows favorable but not consistently significant trends in embedding geometry, while achieving significant improvement in cell-to-ontology retrieval.

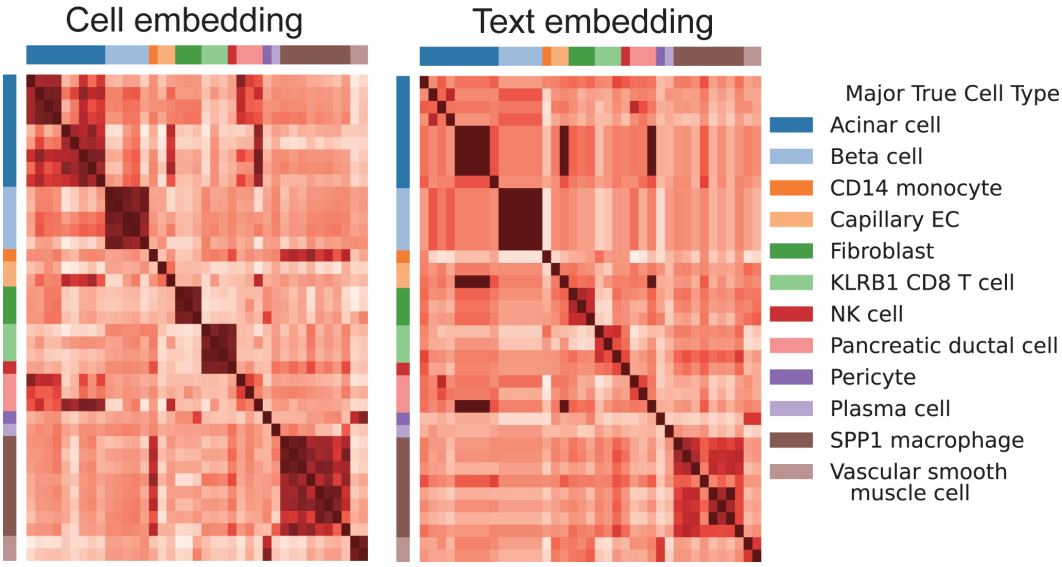

Figure 5: Representative examples of cluster-level similarity structure in the cell and text embedding spaces. Clusters are ordered by their major reference cell type for visualization only. Left: cosine-similarity matrix among cluster centroids in the learned cell embedding space. Right: cosine-similarity matrix among text embeddings of the corresponding LLM-generated cluster descriptions. Examples from a human pancreas dataset show similar blockwise structure across the two modalities, consistent with positive cross-modal structural correspondence.

Table 4: Dataset-level agreement between cluster-level similarity structure in the cell and text embedding spaces. For each dataset, we report the Pearson correlation between the upper-triangular entries of the two cosine-similarity matrices, together with the Mantel permutation-test p-value (9,999 permutations). Numerical values are rounded to two decimal places.

	r	p	datasets
mouse brain	0.77	0.0001	12_0
	0.76	0.0001	12_1
	0.77	0.0001	13_0
	0.76	0.0001	13_1
	0.72	0.0001	13_2
	0.75	0.0001	14_0
	0.78	0.0001	14_1
	0.64	0.0023	15_0
	0.70	0.0001	15_1
	0.69	0.0001	16_0
	0.75	0.0001	16_1
	0.75	0.0001	17_0
	0.79	0.0001	17_1
	0.82	0.0001	18_0
	0.74	0.0005	18_1
	0.83	0.0001	19_0
	0.66	0.0001	19_1
	0.80	0.0001	19_2
human pancreas	0.61	0.0001	CRA001160
	0.46	0.0001	GSE148073
	0.71	0.0001	GSE150724
	0.68	0.0001	GSE153834
	0.87	0.1663	GSE197177
	0.60	0.0001	GSE202494
	0.64	0.0001	GSE205049
	0.45	0.0001	GSE212966
	0.49	0.0001	GSE217837
	0.47	0.0001	GSE221156

Table 5: Performance comparison of HiLL and CellMarker-based variants under ancestor and descendant matching on the human pancreas, human brain, and mouse brain benchmark groups. Results are reported as means across datasets within each group.

Setting	**Method**	**Human Pancreas**		**Mouse Brain**		**Human Brain**	
		Accuracy	**Macro F1**	**Accuracy**	**Macro F1**	**Accuracy**	**Macro F1**
Ancestor	HiLL	0.916	0.583	0.950	0.782	0.945	0.667
	cellmarker (hill)	0.881	0.603	0.839	0.702	0.997	0.672
	cellmarker (deg)	0.843	0.572	0.882	0.723	0.993	0.672
Descendant	HiLL	0.784	0.413	0.938	0.735	0.945	0.667
	cellmarker (hill)	0.618	0.277	0.665	0.488	0.992	0.579
	cellmarker (deg)	0.537	0.279	0.699	0.504	0.990	0.598

Table 6: Stability of HiLL across five random seeds on 18 mouse brain datasets. For each evaluation metric, we report the Friedman test statistic and p-value across seeds, together with the mean standard deviation (SD), mean range, and mean coefficient of variation (CV) across datasets.

Metric	Friedman Stat	Friedman p	Mean SD	Mean Range	Mean CV
Ancestor Accuracy	0.747	0.9455	0.0136	0.0387	0.0145
Ancestor Macro F1	1.467	0.8325	0.0316	0.0840	0.0408
Descendant Accuracy	4.489	0.3439	0.0162	0.0453	0.0173
Descendant Macro F1	3.333	0.5037	0.0297	0.0757	0.0395

the full marker context more flexibly, resolve ambiguous or partially overlapping signatures, and generate labels that better match the target granularity of the inferred clusters. This advantage is particularly relevant for descendant evaluation, where successful annotation requires not only broad category recognition but also semantically appropriate fine-grained subtype assignment. Overall, the comparison indicates that curated marker databases remain useful sources of prior knowledge, but replacing LLM-based interpretation with direct database querying leads to a less adaptive and less fine-grained annotation module.

A.8 Stability of HiLL

A.8.1 Stability across random seeds on mouse brain datasets

We further assessed the stability of HiLL by repeating experiments on the 18 mouse brain datasets with five different random seeds. Performance was evaluated under both ancestor and descendant matching, and variation across seeds was summarized by the mean standard deviation (mean SD), mean range, and mean coefficient of variation (mean CV).

As shown in Table 6, no metric shows a significant seed effect under the Friedman test (all $p > 0.3$). Variation across runs is also small, with mean SD values between 0.0136 and 0.0316, mean ranges between 0.0387 and 0.0840, and mean CV values between 0.0145 and 0.0408. The per-dataset bar plots in Figure 6 and Figure 7 further show consistent performance patterns across the five seeds under ancestor and descendant matching, respectively, with only minor fluctuations on individual datasets. Overall, these results indicate that HiLL is stable across random initializations on mouse brain benchmarks.

A.8.2 Sensitivity to local text encoders and LLM backends

We further evaluated the sensitivity of HiLL to the choice of local text encoder and external LLM backend on the mouse brain benchmarks, under both ancestor and descendant matching.

For local text encoders, the results show no significant performance differences across the tested alternatives. As shown in Figure 8, performance remains highly consistent under both ancestor and descendant evaluation, suggesting that HiLL is relatively insensitive to the specific frozen local text encoder used for semantic embedding computation and that the semantic regularization module does not rely on a narrowly tuned text representation choice.

For external LLM backends, we compared five models: gpt5.6luna, gpt5.6terra, gpt5.6sol, deepseek-v4-pro, and deepseek-v4-flash, where luna, terra, and sol form an increasing-capacity model series. As shown in Figure 9, the default backend, deepseek-v4-flash, shows no significant performance difference relative to most alternative backends, indicating that HiLL remains effective under a relatively lightweight default configuration. At the same time, stronger backends generally yield better annotation performance. In particular, within the gpt5.6 series, sol consistently outperforms the smaller luna and terra models, suggesting that backend capability can further improve fine-grained semantic interpretation.

Overall, these results indicate that HiLL is robust to the choice of local text encoder and reasonably stable across LLM backends, while still being able to benefit from stronger external language models when available.

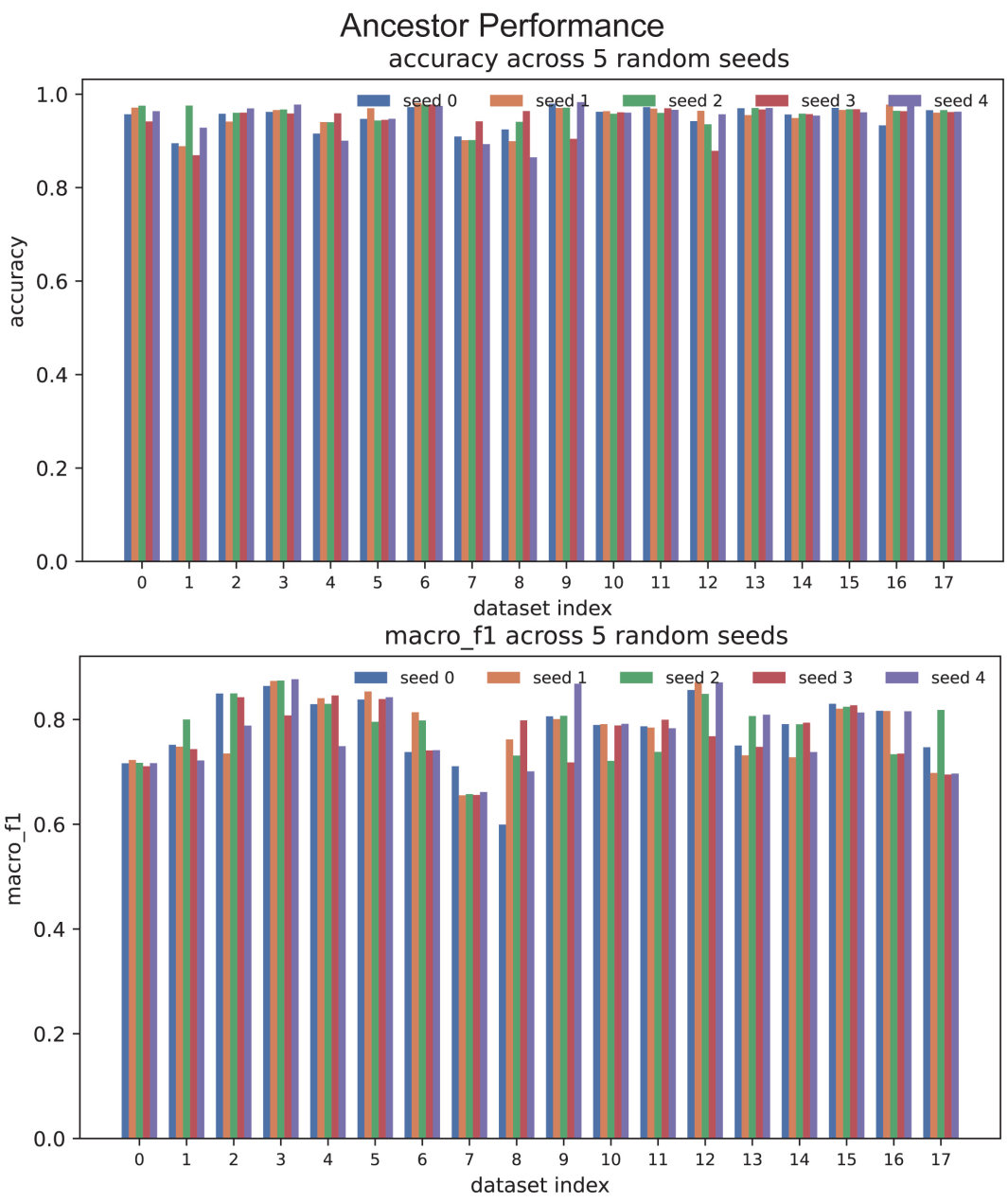

Figure 6: Per-dataset ancestor-level performance of HiLL across five random seeds on 18 mouse brain datasets. Bars show results for each seed on each dataset.

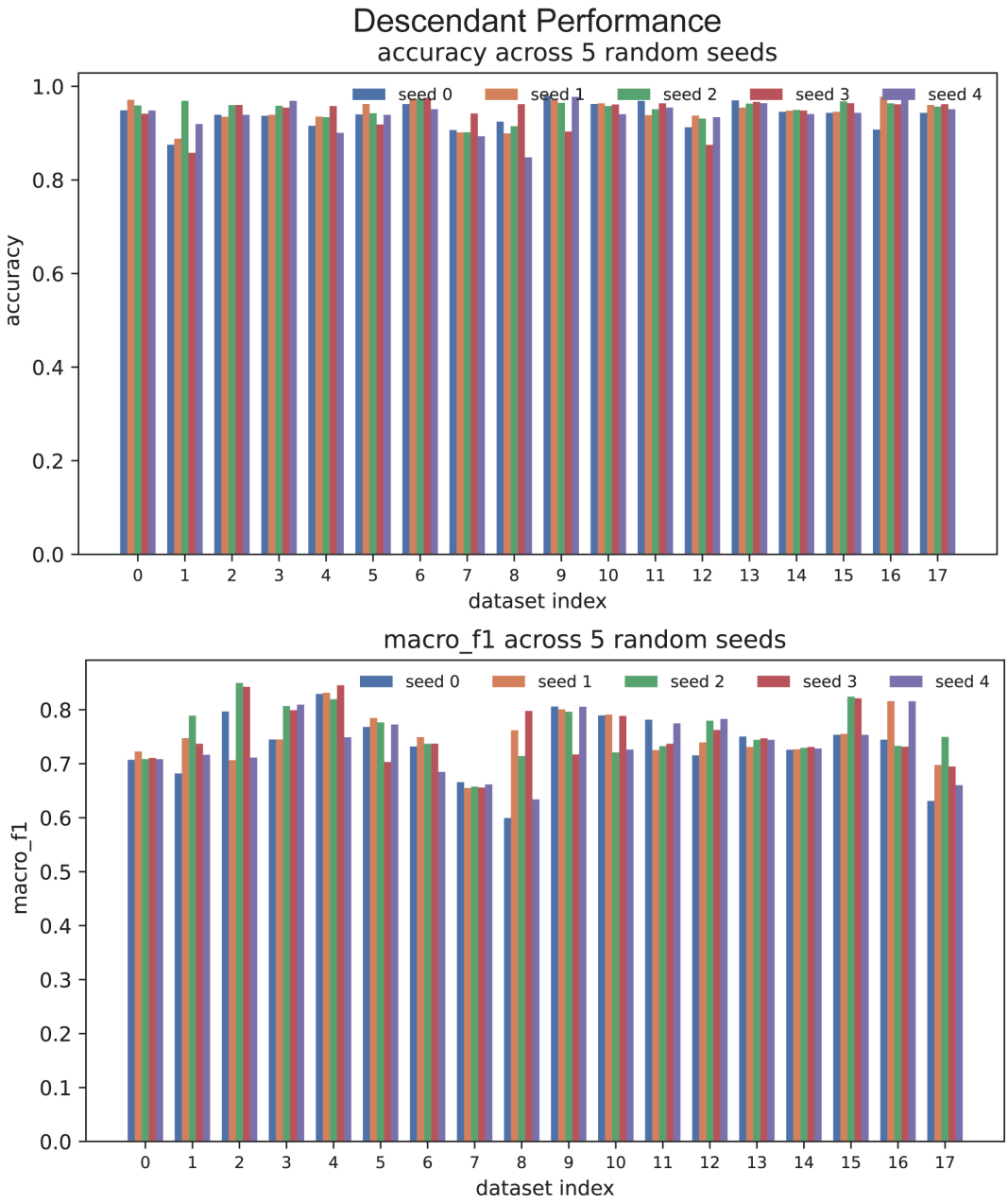

Figure 7: Per-dataset descendant-level performance of HiLL across five random seeds on 18 mouse brain datasets. Bars show results for each seed on each dataset.

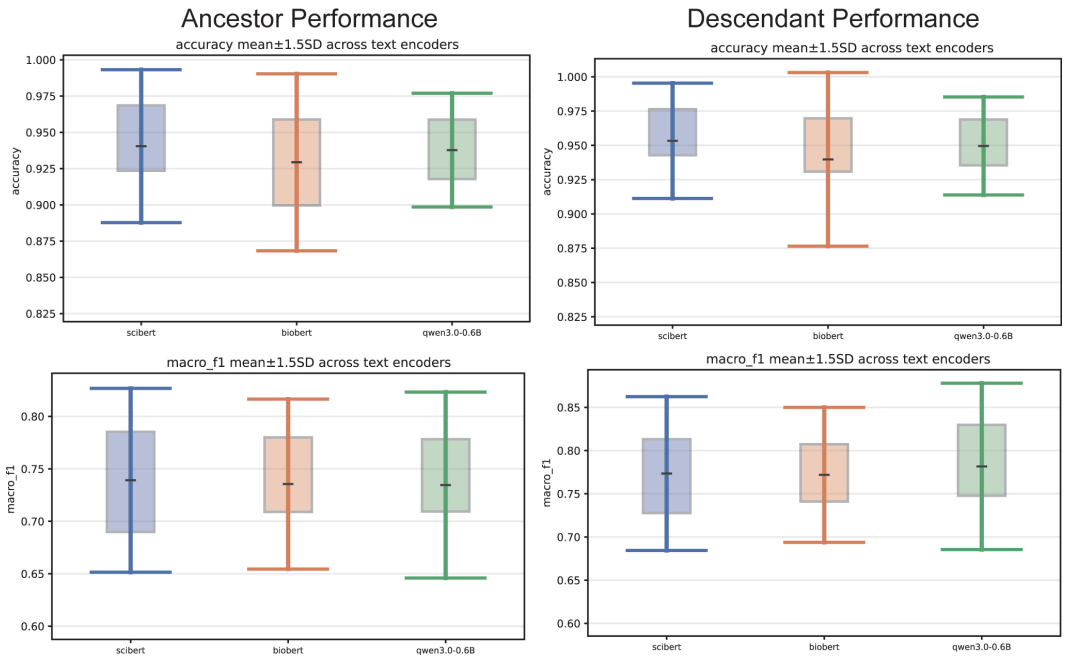

Figure 8: Performance of HiLL with different frozen local text encoders on mouse brain datasets under ancestor and descendant matching. Differences across encoder choices are small and not statistically significant.

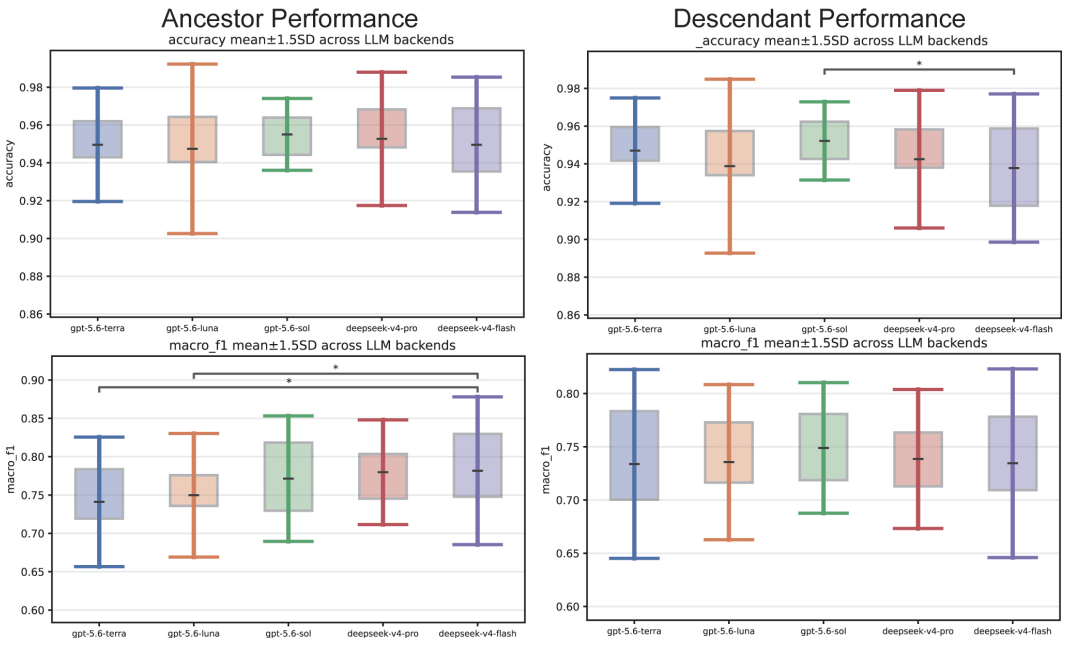

Figure 9: Performance of HiLL with different external LLM backends on mouse brain datasets under ancestor and descendant matching. The default backend performs comparably to most alternatives, while stronger backends generally achieve better results.

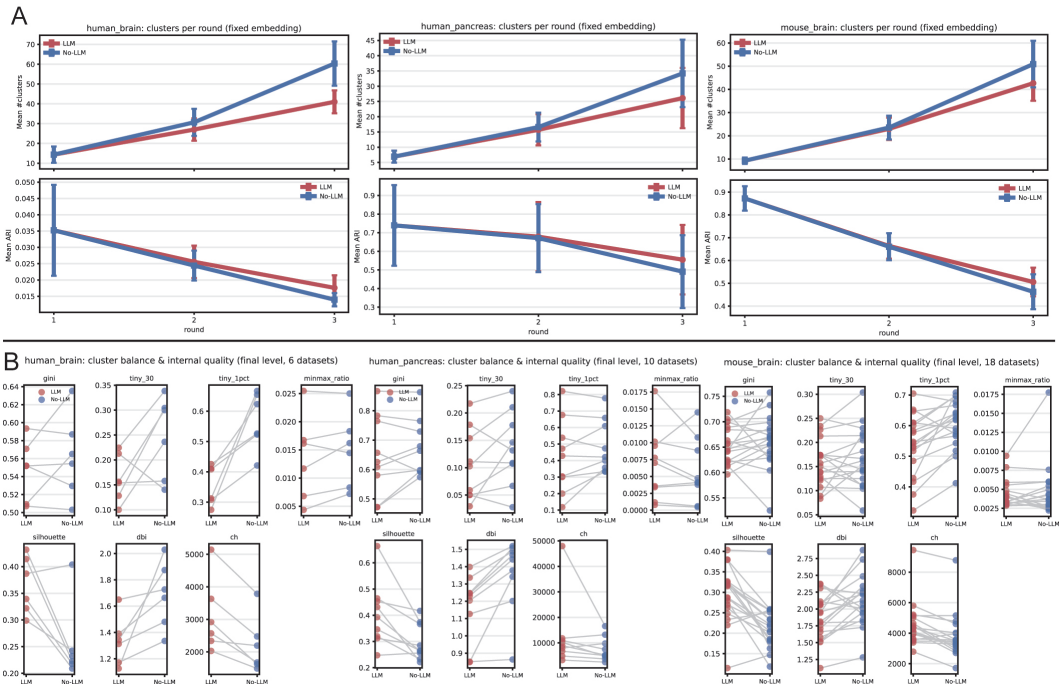

Figure 10: Effect of LLM-based split validation in HiLL. (A) Fixed-embedding ablation across three clustering rounds, showing the mean number of inferred clusters and adjusted Rand index (ARI) against ground truth for human brain, mouse brain, and human pancreas datasets. (B) Final-level comparison of cluster-size structure and internal clustering quality between the LLM-controlled and no-LLM variants. Each point represents one dataset, with paired lines connecting the two variants.

A.9 Analysis of LLM-based split validation

To assess whether LLM-based split validation provides a real benefit beyond added complexity, we compared HiLL with a no-LLM variant that always applies the best-scoring partition under the same fixed embedding. Across human brain, mouse brain, and human pancreas datasets, LLM control consistently produced fewer clusters at deeper rounds while maintaining or improving agreement with the ground truth. As shown in Figure 10A, at round 3 the mean number of clusters was reduced by 19.33, 8.22, and 8.10, respectively, with paired Wilcoxon $p < 0.05$ in all three tissues, while ARI was higher by 0.0036, 0.0431, and 0.0642.

As further shown in Figure 10B, internal clustering metrics followed the same trend: LLM-controlled partitions generally had higher silhouette, lower DBI, and larger CH scores, together with fewer extremely small clusters and slightly better cluster-size balance. In addition, five-seed experiments on mouse brain showed no significant seed effect ($p > 0.05$), indicating good stability. Overall, these results suggest that LLM-based split validation effectively suppresses biologically uninformative over-partitioning while preserving or improving clustering quality.

A.10 Difference between level-wide marker extraction and conventional DEG summarization.

A conventional DEG pipeline typically compares one cluster against all remaining clusters pooled together and returns genes that are globally upregulated relative to this aggregated background. While effective for broad cell type discrimination, this strategy may miss genes that are informative only with respect to specific subsets of clusters, because their signal can be diluted when the remaining clusters are merged. In contrast, HiLL performs pairwise comparisons between each accepted cluster and the other clusters at the same refinement level, and then aggregates how often each gene is significantly upregulated across these comparisons. This level-wide strategy can better preserve markers that distinguish a cluster from particular neighboring populations and is therefore

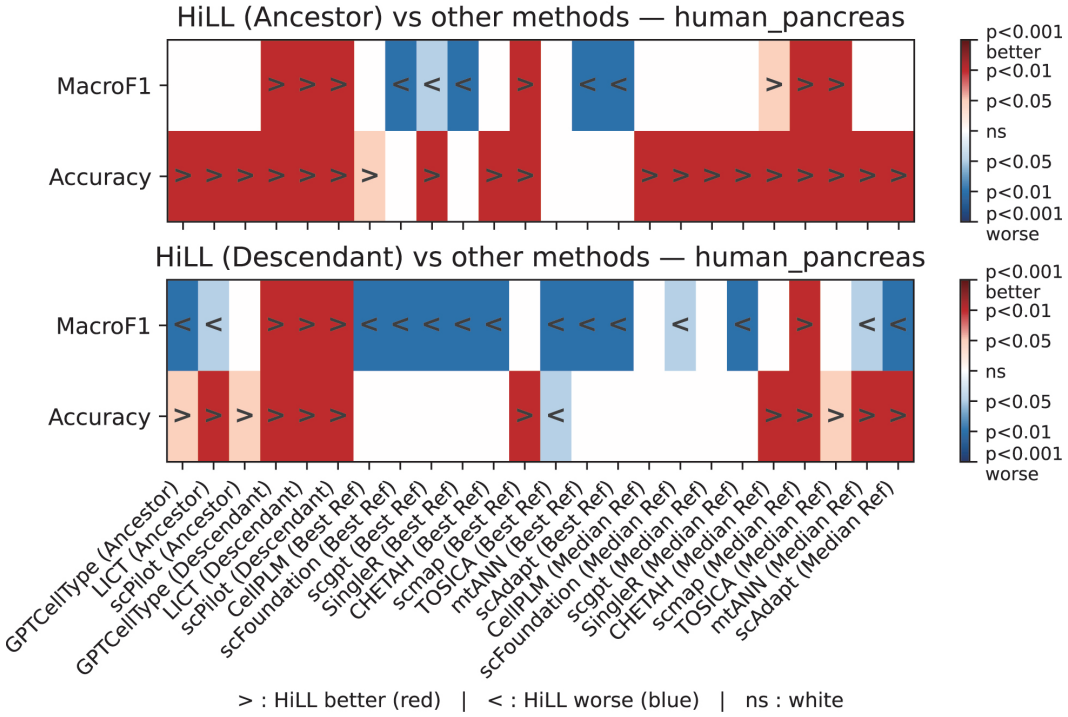

Figure 11: Pairwise statistical comparisons between HiLL and competing methods under ancestor and descendant evaluation for the human pancreas benchmarks, shown for accuracy and macro-F1. Significance is computed using the Wilcoxon signed-rank test across datasets. Red indicates significantly better performance of HiLL, blue indicates significantly worse performance, and color intensity reflects significance level.

better suited to fine-grained semantic annotation. Accordingly, the HiLL (deg) ablation replaces these level-wide markers with markers obtained from a conventional DEG summary, allowing the contribution of the marker extraction strategy itself to be isolated.

A.11 Statistical methods & Results

To assess whether the performance differences reported in the main paper are statistically significant, we perform paired Wilcoxon signed-rank tests between HiLL and each baseline method on the human pancreas, human brain, and mouse brain benchmarks. The tests are conducted separately for each dataset group and evaluation metric.

For each pair of methods, we compare their paired performance values across datasets within the same benchmark collection. Specifically, for a given metric, we compute the paired differences between the two methods on each dataset and apply a two-sided Wilcoxon signed-rank test. When the resulting p-value is below the significance threshold ($\alpha = 0.05$), we determine the direction of the difference according to the median of the paired performance differences. A positive median difference indicates that the first method performs better, whereas a negative median difference indicates the opposite. Pairs with fewer than three valid matched observations are excluded from the analysis.

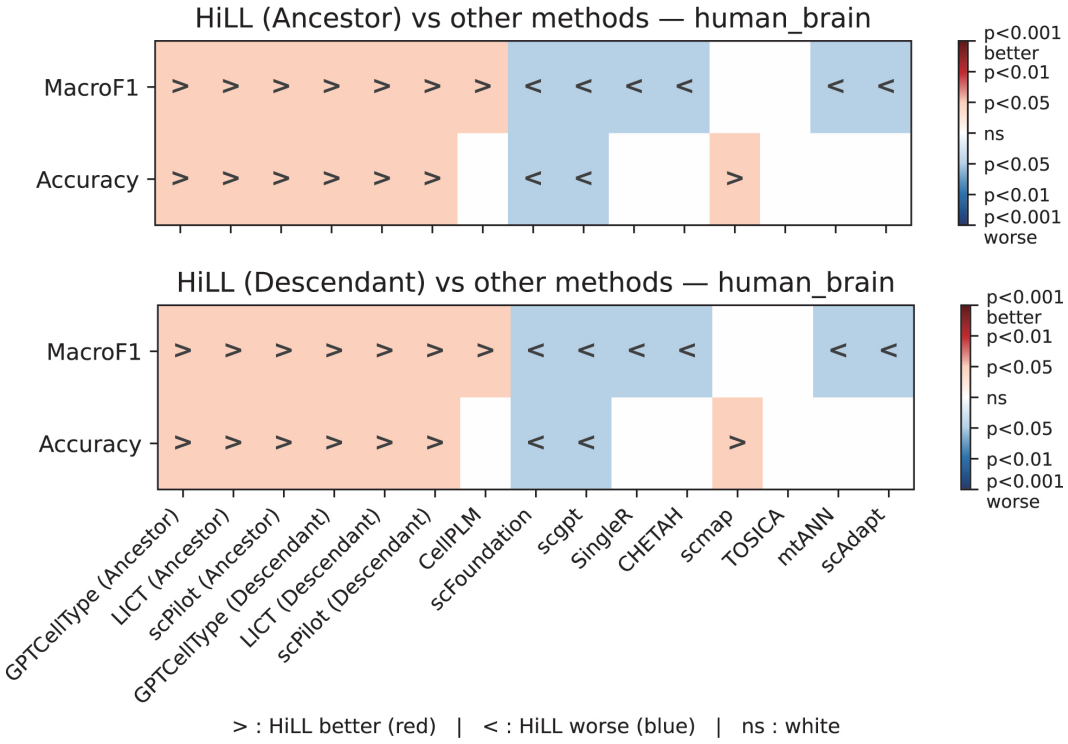

Figure 12: Pairwise statistical comparisons between HiLL and competing methods under ancestor and descendant evaluation for the human brain benchmarks, shown for accuracy and macro-F1. Significance is computed using the Wilcoxon signed-rank test across datasets. Red indicates significantly better performance of HiLL, blue indicates significantly worse performance, and color intensity reflects significance level.

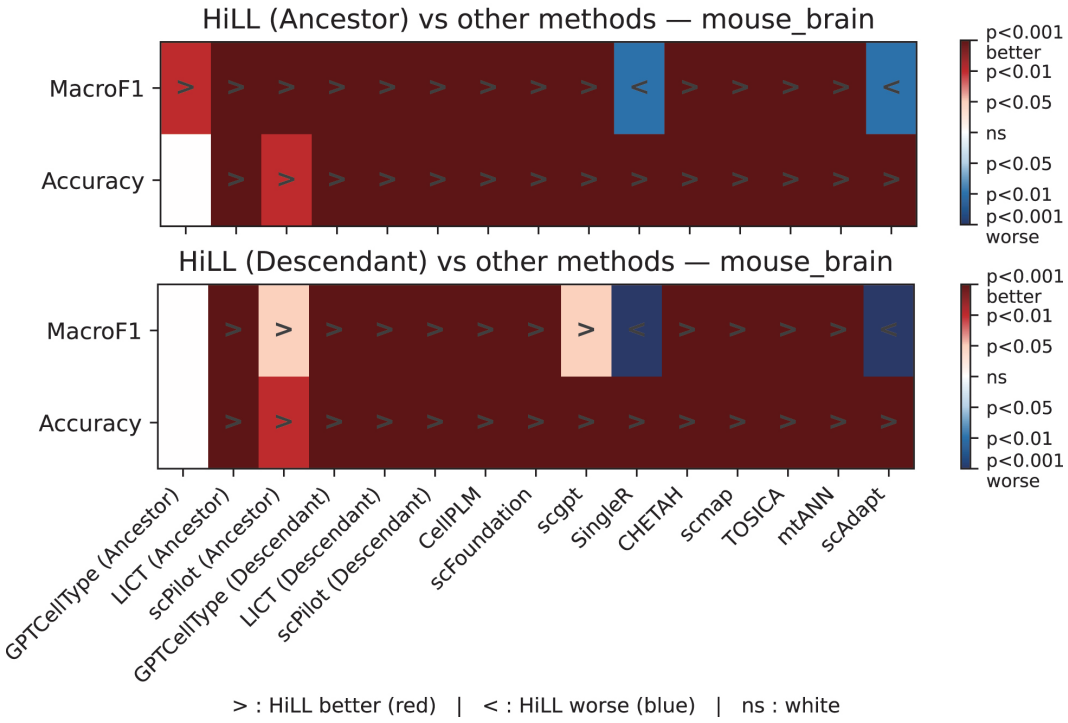

Figure 13: Pairwise statistical comparisons between HiLL and competing methods under ancestor and descendant evaluation for the mouse brain benchmarks, shown for accuracy and macro-F1. Significance is computed using the Wilcoxon signed-rank test across datasets. Red indicates significantly better performance of HiLL, blue indicates significantly worse performance, and color intensity reflects significance level.

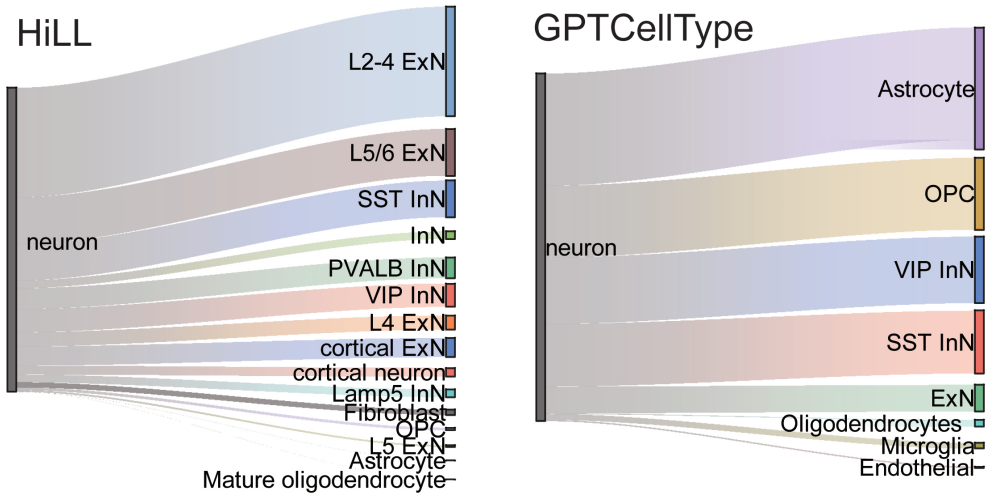

Figure 14: Sankey diagrams of human brain datasets (10X146_2) for HiLL and GPTCelltype on the fine-grained annotation task. Ground-truth cell types are shown on the left, and predicted subtypes on the right. Flow widths are proportional to cell counts and visualize how each method preserves or collapses fine-grained subtype structure.

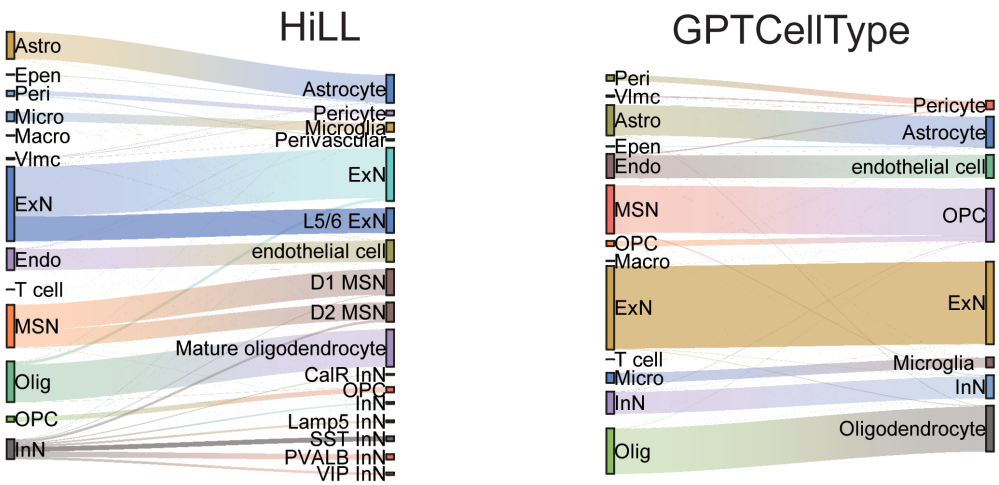

Figure 15: Sankey diagrams of mouse brain datasets (slice 12_0) for HiLL and GPTCelltype on the fine-grained annotation task. Ground-truth cell types are shown on the left, and predicted subtypes on the right. Flow widths are proportional to cell counts and visualize how each method preserves or collapses fine-grained subtype structure.