Uncertainty-Aware Single-Cell Annotation

Architecture, Selective Prediction, and the Limits of Transfer

Bence Lukacsy | Adviser: Professor Scott Stoller
Stony Brook University • CSE 495/496 Honors Project • Fall 2026

Problem and contribution

Single-cell annotation assigns a cell type to a gene-expression vector. This project investigates when neural classifiers and nearest-neighbor label transfer provide useful annotations, and when uncertainty should limit the specificity of an answer. I developed a reproducible multi-reference pipeline with explicit gene alignment, classifier comparisons, calibration, coarse-to-fine selective prediction, and a reference-family router. The central finding is a distinction: **strong performance within a reference and accurate reference selection do not establish reliable recognition of unseen cell types or correctness under dataset shift**.

Research progression and data

Earlier Project 1 experiments compared flat models with a specialist cascade. Hierarchy improved broad mouse-brain labels but lost at finer resolutions; lung and striatum transfer failures showed that gene alignment alone was insufficient. These findings motivated the uncertainty-aware follow-up.

The completed follow-up uses seven references: five lung datasets (controls, granulomas, sc78, sc92 and sc93), zebrafish, and an isocortex/striatum subset of the mouse-brain atlas. The main mouse experiment uses 444,446 training-role cells, 3,500 selected genes, 22 classes and 1,320 fine clusters, evaluated on the preserved 147,622-cell external holdout. Zebrafish has 39,712 cells with usable annotations after filtering, with 88 broad and 262 fine training paths. The five lung main holdouts total 21,324 cells. Reference annotations supply the evaluation labels; agreement is not independent biological validation.

Methods and evaluation

The final batch comprises **108 neural fits across 27 configurations**. Each primary reference compares small flat, large flat, conditional hierarchical, and gene-masked flat networks, each with three optimization seeds, against a training-only PCA+kNN baseline. Neural ensemble probabilities are averaged across seeds. The follow-up hierarchy shares a trunk and conditional heads; it is not the earlier specialist cascade. Actual fits ended after 31–177 epochs under a 250-epoch cap with early stopping.

Training, tuning, temperature calibration, support calibration, threshold calibration and testing use disjoint recorded file/row assignments. Feature selection and scaling use training cells. Methods are chosen on tuning balanced accuracy, not test performance. Six sample/library holdouts, seven held-out-cell-type tests, 20 directed lung transfers, five pooled leave-one-lung-file-out tests, and gene-removal/corruption diagnostics evaluate generalization. The sc78 group test is explicitly skipped for insufficient groups. Native expression conventions are retained and are not declared equivalent merely because gene identifiers align.

Primary results

Reference	Small flat	Large flat	Hierarchy	Masked flat	PCA+kNN
Lung controls	99.43	99.56	99.56	99.56	99.15
Lung 30-day	94.07	93.68	93.59	93.83	93.07
Lung sc78	91.54	90.69	90.88	91.64	89.51
Lung 60-day	94.53	93.84	94.07	94.43	94.21
Lung 90-day	94.34	94.51	94.08	94.34	94.20
Zebrafish	83.08	82.28	79.71	81.73	78.09
Mouse brain	85.38	84.37	84.51	84.82	71.59

Fine-label accuracy (%). All four neural columns are three-seed ensembles.

Larger networks do not consistently win. On mouse-brain balanced accuracy, however, hierarchy scores 75.06%, above the small flat ensemble's 72.77%. On the same 20,000 cells used in the initial local run, small-flat ensemble accuracy improves from the initial single model's 73.96% to 85.02%. Multiple pipeline components changed, so this gain cannot be attributed to architecture size alone.

Selective annotation and calibration

The tuning-selected masked-flat mouse model accepts class-level labels for 99.19% of test cells with 0.13% error among accepted assignments. At the finest cluster level, it accepts 81.31% with 8.63% accepted error, compared with 15.28% error when the same greedy path decoder labels every cell. The selected zebrafish model retains 74.60% at 8.47% accepted error; pooled lung retains 98.20% at 7.86%. Thus an uncertain fine label need not discard a well-supported broad annotation. Global leaf-argmax accuracy and greedy accepted-path error are distinct, explicitly identified metrics.

Temperature scaling has mixed effects. Pooled-lung expected calibration error (ECE) improves from 2.34% to 0.57%. For the selected mouse model, negative log-likelihood improves from 0.4936 to 0.4604, but ECE worsens from 2.45% to 3.93% and Brier score also worsens. Calibration is therefore assessed with multiple measures rather than presented as an automatic improvement.

Reference routing versus biological novelty

The evaluated mouse router is a three-member neural ensemble using within-cell ranks on an explicitly shared mouse gene panel. It accepts domain decisions for 23,647 of 24,438 sampled mouse cells (96.76% coverage), with two disagreements against lung/brain reference provenance (99.9915% agreement among accepted decisions). This removes the earlier symbol-versus-Ensembl routing confound. The fish branch remains a gene-panel compatibility check, not a learned species probability. Reference origin is only a proxy for applicability; shared cell types need not uniquely identify tissue origin.

Novelty detection is much weaker. In held-out-fine-type tests, the selected policy rejects 52.31% of unknown mouse cells and 63.37% of unknown zebrafish cells at the fine level, but rejects **none** in four of five lung experiments. Fine rejection differs from rejection at every depth: a known ancestor can still be useful. Ranking uncertainty scores can separate unknowns without the fixed operating threshold rejecting them. The system must not be described as a universal novel-cell detector.

Transfer, feature mismatch and failure analysis

Directed transfer retains target types absent from the source. The large flat controls model reaches only 49.58% accuracy on the 60-day target while accepting fine labels for every sampled cell, with 50.44% accepted greedy-path error. Holding out the entire 90-day lung file yields 61.05% neural accuracy versus 81.92% for PCA+kNN trained on the other four files; kNN's accepted error is still 18.44%. High within-reference scores do not prevent confident transfer errors. Historical transfers used different methods and excluded source-unseen target labels.

At 10% random gene removal, the mouse masked-flat ensemble reaches 84.30% unconditional accuracy versus 82.43% for the large flat ensemble, averaged across three fixed masks on 1,000 cells. This is evidence for a bounded masking test, not arbitrary feature loss. Larger removals violate the fixed 90% overlap contract and deliberately produce rejection; zero coverage is not classification robustness. In the unlabelled striatum diagnostic, 3,469 of 3,500 selected genes match, yet all tested methods reject all 5,000 sampled cells. With no verified target labels, no accuracy or novelty-detection success rate is assigned to that result.

Limitations and conclusion

The expanded design was informed by earlier results and is a follow-up study, not a pristine confirmatory benchmark. Seed repeats measure optimization variability, not biological replication. Sample/library holdouts are not assumed to be donor holdouts. Raw preprocessing equivalence, independent donor validation, and a matched commercial-tool comparison remain unestablished. The nominal threshold target and label-set calibration do not guarantee error control under distribution shift.

The completed project provides a reproducible annotation-and-abstention pipeline and a quantitative account of both its usefulness and its limits. Neural ensembles and selective depth control provide useful reference annotations; nearest neighbors remain competitive and can be stronger in transfer. Reliable unfamiliar-cell detection remains a limitation rather than a solved feature. The browser prototype for cse496.com reproduces the evaluated router and selected experts, exposing accepted depth and diagnostics without uploading query expression values.

Evidence. Reports 1–3; Project 1 code; frozen v2 protocol and results. Review checked 175 archived prediction sets and recorded splits. JavaScript matched original-Python decisions in 312 real-cell/method checks. Raw matrices were not rerun.