Uncertainty-Aware Single-Cell Annotation

Architecture, selective prediction, and the limits of transfer

Bence Lukacsy

Stony Brook University | CSE 495/496 Honors Project
Adviser: Professor Scott Stoller

Fall 2026

https://cse496.com

The question is not just "Which cell type?"

1. **Which reference is applicable?** Lung, brain, fish, or unsupported input.
2. **How specific can the answer be?** Broad class through fine subtype.
3. **When should the system abstain?** Confidence and reference support must be tested.

Research question

Can one pipeline compare neural and neighbor-based annotation while exposing the limits of its predictions?

Evidence: CSE 495 proposal; CSE 496 progress report.

Earlier work established the foundation

Reusable training infrastructure

- AnnData processing and ordered gene alignment.
- Flat classifiers and a cascade of specialist models.
- Five lung comparisons; 414 mouse-brain models.

What changed the direction

- Hierarchy helped broad brain labels, not every fine level.
- Lung transfer degraded under dataset shift.
- Striatum predictions collapsed despite aligned inputs.

Follow-up

Test applicability and uncertainty explicitly instead of forcing every cell to a fine label.

Evidence: Historical Reports 1–3; the new conditional hierarchy is not the historical specialist cascade.

Seven primary references; 176,784 main test cells

Reference	Training cells	Test cells	Fine training paths
Lung controls	9,618	3,185	23
Lung 30-day	16,612	5,510	30
Lung sc78	6,430	2,116	28
Lung 60-day	19,192	6,376	27
Lung 90-day	12,486	4,137	26
Zebrafish	24,192	7,838	262
Mouse brain	444,446	147,622	1,320

Mouse brain means the **isocortex/striatum subset**, not the entire atlas.
Each model uses up to 3,500 training-selected features.

Evidence: Archived split manifests; training-role counts exclude tuning and calibration cells.

A fixed expanded study, not an open-ended search

Primary comparison

- Small flat: 256 → 128.
- Large flat: 1,024 → 512 → 256.
- Conditional hierarchy; gene-masked flat.
- Three seeds per variant; probability ensembles.
- PCA+kNN: up to 100,000 reference cells.

Generalization tests

- Six sample/library holdouts.
- Seven held-out-type experiments.
- Twenty directed lung transfers.
- Five leave-one-lung-file-out tests.
- Pooled lung, router, corruption and missing genes.

Completed: 108 neural fits, 27 configurations. Early stopping: 31–177 epochs.

Evidence: Frozen v2 protocol, PLAN_SUMMARY.json, TRAINING.csv. sc78 group test skipped for insufficient groups.

Separate fitting from evaluation

Disjoint roles

Training → tuning → temperature → support → thresholds → test

- Features, scaling and PCA use training cells.
- Tuning balanced accuracy chooses the candidate method.
- Test scores do not choose models or thresholds.
- Existing main split roles and the mouse external holdout are preserved.

This is a follow-up study informed by v1, not a pristine confirmatory benchmark.

Evidence: Frozen plan and independently checked file/row manifests. Seeds are optimization repeats, not biological replicates.

The annotation pipeline

Input contract

Cells $\times$ named genes; explicit identifier alignment; reference-native expression values.

Reference choice

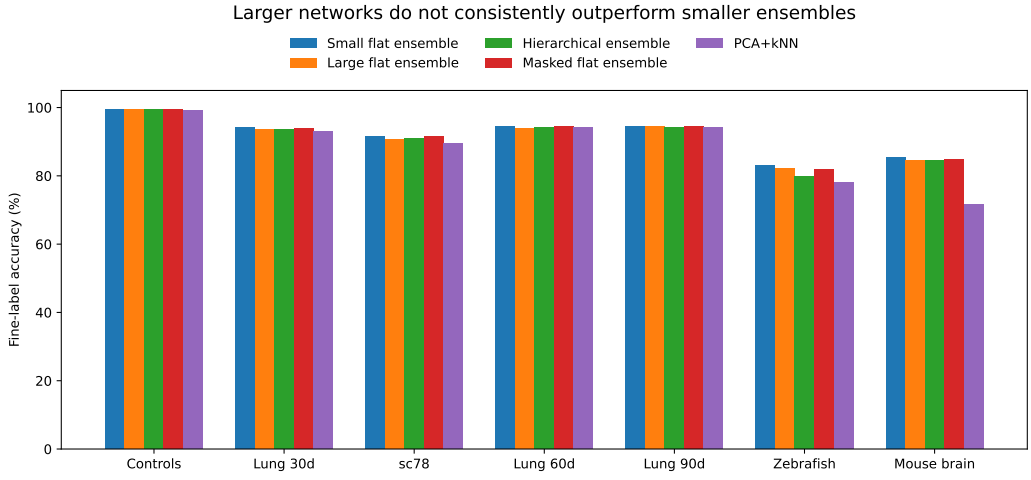
Mouse: shared-panel ranks $\rightarrow$ neural lung/brain router.
Fish: gene-panel compatibility filter, not a learned species probability.

Cell-type answer

Selected expert $\rightarrow$ calibrated probabilities and support checks
$\rightarrow$ accepted taxonomy prefix, or no annotation.

Evidence: Executed cross.py, score.py and data.py. Gene alignment does not make different expression scales equivalent.

More capacity is not a consistent accuracy gain



Evidence: Fine-label reference agreement on all seven main holdouts. Neural bars are three-member ensembles.

The mouse-brain improvement is real on matched cells

Method on the same 20,000 cells	Accuracy
Initial local single flat model	73.96%
Expanded small flat ensemble	**85.02%**
Expanded masked-flat ensemble	84.465%

Matched-cell gain

$+11.06$ percentage points

The combined pipeline changed training size, feature selection, optimization and ensembling. This is not the isolated effect of network width.

Evidence: Cell IDs and full annotation paths matched between v1 and v2; v1_v2_matched_cells.csv.

Accuracy and balanced accuracy tell different stories

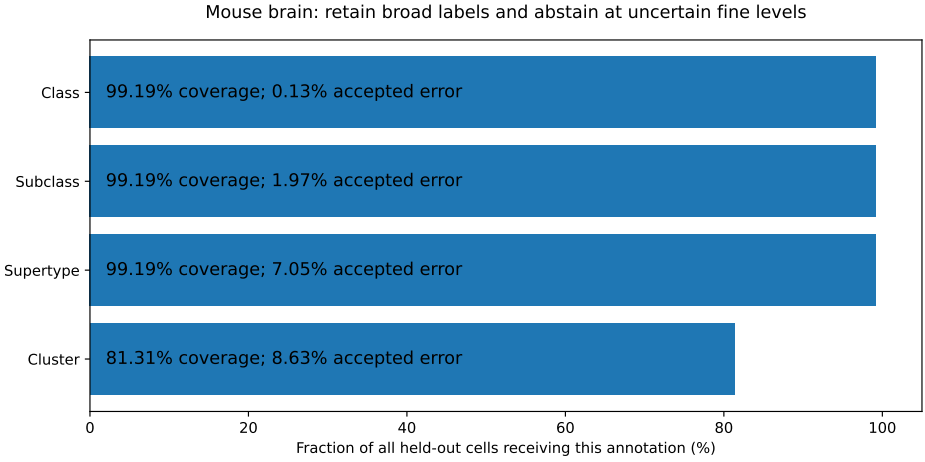
Mouse-brain method	Accuracy	Balanced accuracy
Small flat ensemble	**85.38%**	72.77%
Large flat ensemble	84.37%	73.80%
Hierarchical ensemble	84.51%	**75.06%**
Masked flat ensemble	84.82%	74.34%
PCA+kNN	71.59%	57.48%

Balanced accuracy gives every target type equal weight.

The hierarchy is more competitive when rare-type recall matters. The demo candidate remains masked flat, selected on tuning data, not replaced by the test winner.

Evidence: 147,622-cell holdout; archived selection.json and independently verified classification counts.

Abstain at the fine level, not necessarily everywhere



Evidence: Tuning-selected masked-flat mouse expert; greedy taxonomy-consistent decoder, combined policy.

Calibration is not an automatic improvement

Selected expert	Measure	Before	After temperature
Mouse masked flat	NLL	0.4936	0.4604
Mouse masked flat	ECE	2.45%	3.93%
Pooled-lung flat	NLL	0.2343	0.2248
Pooled-lung flat	ECE	2.34%	0.57%

Interpretation

Optimizing one calibration objective does not ensure improvement on every metric or under dataset shift.

Evidence: Recorded calibration summaries; post-temperature ECE independently reconstructed from compact predictions.

Shared-panel routing resolves the earlier identifier confound

96.76%

Mouse domain coverage

2

Reference-origin disagreements among 23,647 accepted routes

- 24,438 sampled mouse cells.
- 99.9915% agreement among accepted routes.
- Explicit common identifiers; no symbol-versus-Ensembl shortcut.
- End-to-end labels still have their own errors.

Lung-versus-brain provenance is not universal tissue or novelty recognition.

Evidence: Verified routing JSONL records. Evaluated domain model: flat neural ensemble; fish uses panel compatibility.

Holding out whole recorded groups is a harder test

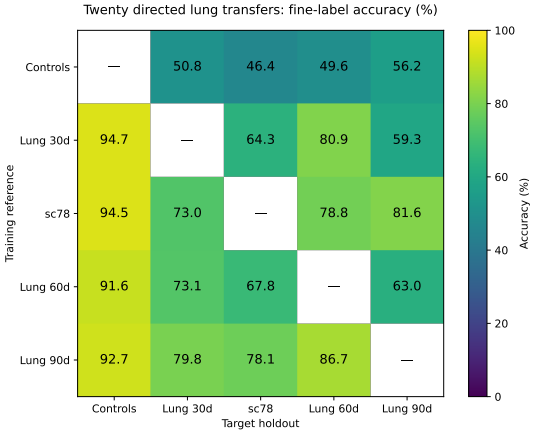
Group-held-out test	Large flat accuracy	Balanced accuracy
Zebrafish	79.11%	60.81%
Mouse brain	82.01%	67.94%
Lung controls	98.31%	96.35%
Lung 30-day	92.46%	91.32%
Lung 60-day	91.56%	87.40%
Lung 90-day	92.96%	88.03%

Each row uses one large flat fit, not a three-seed ensemble. Test populations and class mixtures differ from the primary tests.

Sample/library holdouts are not automatically donor holdouts.

Evidence: Group configurations; sc78 omitted by the frozen insufficient-group rule.

Transfer can be confidently wrong



Controls → 60-day lung

49.58% global-leaf accuracy

100% fine-label coverage

50.44% accepted path error

Source-unseen target types are retained. The two prediction decoders can differ.

Evidence: 20 directed transfers; large-flat source ensembles. Transfer is native-scale, not proven batch-harmonized.

Nearest neighbors can be stronger under dataset shift

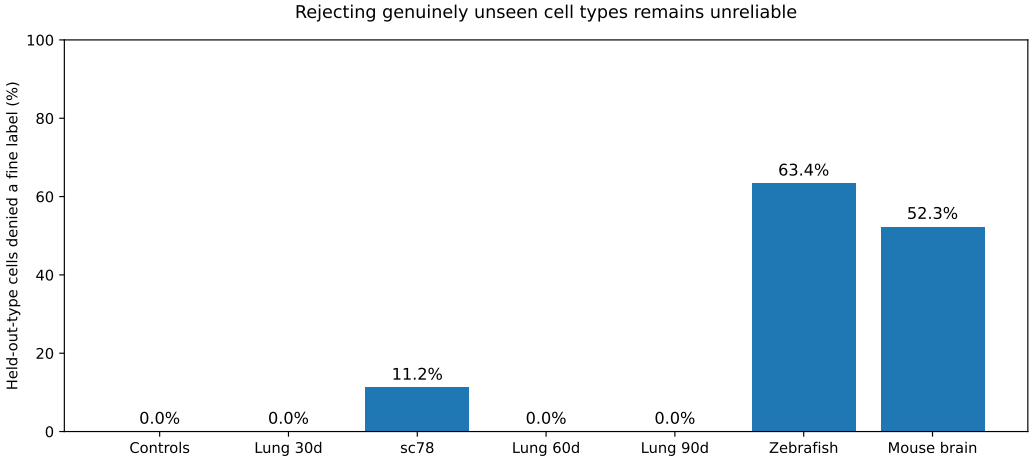
Entire held-out lung file	Single large flat	PCA+kNN
Controls	93.97%	94.01%
30-day	75.67%	77.30%
sc78	71.87%	76.65%
60-day	82.18%	82.22%
90-day	**61.05%**	**81.92%**

Train on four lung files; the fifth file supplies no fitting or calibration cells.

Even the stronger 90-day kNN result has **18.44% accepted error** at 96.18% coverage. It is not a universally safe fallback.

Evidence: Five pooled leave-file-out experiments; full target populations, not the directed-transfer samples.

Unknown-type detection remains an explicit limitation



Evidence: Entire fine types removed from every fitting role. Policies selected on tuning data. Four lung tests reject no unknown fine types.

Gene overlap is necessary, not sufficient

Random missing genes

- At 10% removal: masked mouse ensemble 84.30% accuracy; large flat 82.43%.
- Three masks on the same 1,000 cells.
- At 20/30%, the fixed overlap contract deliberately rejects.

Unlabelled striatum

- 3,469/3,500 genes match: 99.11%.
- All tested methods reject all 5,000 sampled cells.
- No verified target labels: no accuracy or novelty-success estimate.

Zero coverage is not robust classification. Aligned names do not establish aligned expression scales.

Evidence: Recorded fixed-mask tests and striatum diagnostics; new features differ from the historical gene-overlap calculation.

Live demonstration: cse496.com

1. Load held-out mouse cells; confirm the reference-native input convention.
2. Inspect the exact three-seed router, selected expert, and support signals.
3. Change from class to cluster: a broad label can remain while a fine label is withheld.
4. Export the decisions; show rejection of an unsupported gene panel.

All query inference runs on the visitor's device

Named CSV or single-cell JSON; no expression upload. The map is a training-reference projection, not evidence of annotation accuracy.

Evidence: Matched v2 weights and gates; 312 real-cell/method browser-engine checks. Complete the supplied local browser check before recording.

Conclusion: useful annotation, honest boundaries

1. **Architecture matters, but not monotonically.** Bigger networks do not always win; hierarchy and kNN are competitive in specific settings.
2. **Selective depth is useful.** Preserve supported broad labels rather than forcing fine certainty.
3. **Routing is not novelty detection.** Cross-dataset confidence and unseen-type rejection remain failure modes.

Contribution

A reproducible multi-reference annotation pipeline, supported by a broad comparison and explicit measurements of when it fails.

Evidence: Completed v2 study; successful and negative results retained. No matched commercial-tool comparison or universal novelty claim.

Appendix: what each metric means

Accuracy Fraction of cells whose global leaf argmax agrees with the reference label.

Balanced accuracy Mean recall over distinct target label paths; unseen target types remain distinct.

Coverage Fraction of all query cells that receive an annotation at a particular depth.

Selective error Error among the cells accepted at that depth, using the actual greedy path decoder.

Reference support Empirical reference-relative compatibility, not the posterior probability of a tissue.

Open-set AUROC Score ranking for held-out types, not the performance of a fixed rejection threshold.

Evidence: Frozen protocol and executed scoring implementation.

Appendix: verification and reproducibility

- 175 saved prediction sets independently recomputed; 18,249 checks passed.
- Recorded file/row roles are disjoint and agree with the frozen plan.
- Executed-code hashes and 108 training log traces checked.
- Selected experts rerun on 240 real cells; browser engine matches 312 real-cell/method checks, including the exact neural router.

Full raw matrices, all individual-seed predictions, and all full-probability calibration quantities were not rerun in the review. The release includes a local HTTP/Worker check; live deployment and visual rehearsal remain recording checks.

Evidence: Independent research audit; real-cell export checks; browser_engine_parity.json.

Appendix: evidence map and scope

Evidence	Location
Original question and work plan	CSE 495 proposal; CSE 496 progress report
Historical experiments	Reports 1–3; repository Project 1
Frozen follow-up design	THOROUGH_PROTOCOL.md; plan.json
Model comparisons and repeats	SUMMARY.csv; SEEDS.csv; TRAINING.csv
Transfer and reference routing	cross/; routing/; striatum/
Independent review	CSE496_V2_Review/

Limits: follow-up design, cell dependence, native-scale differences, unverified donor independence, and incomplete unknown detection. Model/data publication permissions remain a separate requirement.