Cross-Species Cell-Type Recovery with Hypothalamic Atlases

Sofía Prado Arenzana Shantell Luna Greta VanZetten

Abstract

We evaluate cross-species cell-type recovery in the hypothalamus by mapping mouse and human atlases to one-to-one orthologs and comparing KNN, Seurat v4, and Harmony+MLP for predicting human cell types from mouse references. By resampling mouse cell-type compositions, we quantify how distributional mismatch affects performance: Seurat v4 is strong on average (mean weighted F1 ≈ 0.94) but can lose precision under imbalance, while Harmony+MLP remains more stable across donors (mean weighted F1 ≈ 0.96). These results emphasize class imbalance as a key limitation in cross-species label transfer.

1. Introduction

Cross-species cell-type recovery refers to the process of taking an annotated cell atlas from one species and using it to identify or predict cell types in another (Song et al., 2023). Drug transfer research relies heavily on cross-species cell type recovery. If a drug works on certain cells in a house mouse, *mus musculus*, it is important that humans also have those cells, and they perform similar biological functions. Since many clinical trials use mice as test subjects, accurately and reliably transferring cell-type information from mice to humans becomes crucial for improving preclinical experiment interpretations. With this motivating information, we asked three main questions: i) What modeling approaches are best suited for cross-species cell-type recovery in the hypothalamus? ii) How do data-processing choices impact model performance? iii) Are certain human cell types more easily recovered than others and what might this tell us evolutionarily?

To explore these issues, we used two atlases, a mouse hypomap (Steuernagel et al., 2022) and a human one (Tadross et al., 2025). Since both atlases had different gene identifiers, we made them compatible by mapping the mouse gene symbols and human stable IDs with an Ensembl BioMart ortholog table and restricting both datasets to the 17,853 ortholog genes present in both species. This created a unified feature space and allowed us to directly train models on the mouse data to predict on the human.

We compared three model approaches: a distance-based baseline (KNN), an anchor-based integration method (Seurat v4), and a latent-space neural classifier combining Harmony with a multilayer perceptron (MLP). To each model, we input our labeled mouse reference atlas and an unlabeled human atlas with matching features, aiming to learn a mapping from gene expression to cell type in the mouse hypothalamus and use it to predict cell labels for the human atlas.

2. Related Work

While different species may be evolutionarily related, their gene expression profiles can substantially differ (Shafer, 2019). Thus, cells must be placed into a shared expression space before performing cross-species gene mapping. Early work in this field typically analyzed each species, performed clustering in each atlas, and then matched clusters between species by hand using marker genes or other correlation measures. For example, Shafer shares a method that computes gene-specificity indices for each cluster and then correlates them across species (Shafer, 2019). The "separate analysis" strategies preserve species-specific heterogeneity, but they do not account for systematic expression shifts between species.

Recent methods perform algorithmic integration to embed species cells into a joint low-dimensional space. Seurat, fastMNN, Harmony, LIGER, Scanorama, scVI, and scANVI were originally developed for batch correction and cross-condition integration, but have also been found useful in cross-species settings, as species differences (or species effects) can be viewed as a stronger version of traditional batch effects (Shafer, 2019; Song et al., 2023). In particular, species effects occur because cells from the same species share more gene expression patterns between their own cells rather than with their cross-species counterparts. In Seurat v4, PCA with mutual nearest-neighbor "anchors" defines non-linear correlation vectors that are used to transfer cell labels across atlases (Shafer, 2019; Song et al., 2023). Harmony uses an iterative clustering-based approach to correct for batch effects in a lower-dimensional embedding, leading it to become a valuable tool for harmonizing heterogeneous single-cell data. These integration methods are mainly automated and greatly reduce species effects (Song et al., 2023). These methods can also over-correct and merge distinct but related cell types.

Song et al. benchmarked 28 combinations of gene-homology mapping and integration algorithms across multiple species using their BENGAL pipeline (Song et al., 2023). They showed that combining otholog mapping with ENSEMBL and scVI, scANVI, or Seurat v4 can balance species mixing and conserve biological structure. They also noted that SAMap, which uses self-assembling manifolds and gene homology graphs to capture evolutionary correspondence, performed exceptionally well for evolutionarily

distant species. We initially planned to include SAMap in our comparison to span the full spectrum of complexity in cross-species cell-type recovery strategies. However, due to dependency and implementation issues, we were unable to successfully apply SAMap to our datasets.

These studies informed our decision to create an ortholog-based mapping strategy to create a shared human-mouse feature space and compare a simple distance-based classifier (KNN) with an integration and label-transfer approach (Seurat v4) and SAMap to span the spectrum of cross-species cell-type recovery strategies.

3. Data

The HypoMap project conducted by Steuernagel et al. provided a fully integrated reference atlas of single-nucleus RNA sequencing data comprising 384,925 murine hypothalamic cells profiled across 57,362 genes. This dataset was used to train our predictive models for a corresponding human hypothalamus dataset generated by Tadross et al.. The human HypoMap includes 433,369 single-cell transcriptomes, measured across 36,924 genes using scRNA-seq (Tadross et al., 2025). Both the murine and human datasets were supplied as log-normalized expression matrices. However, for the murine dataset, raw count matrices were not released, which limited our downstream normalization and modeling choices, such as the scVI/scANVI pipeline (Xu et al., 2021).

To facilitate classification between species, we restricted the feature space to known one-to-one orthologous genes shared between mice and humans, which are genes that evolved from a common ancestral gene and usually retain similar biological functions. We obtained a Human-House Mouse ortholog table from Ensembl BioMart, and used human ortholog gene identifiers to label both ortholog-restricted datasets. This filtering step reduced each gene expression matrix to 17,853 orthologous genes.

Eight cell types were annotated in the human dataset (oligodendrocytes, neurons, astrocytes, microglia, mural cells, endothelial cells, ependymal cells, and fibroblasts), whereas twelve were identified in the murine dataset (the same eight human cell types, plus dividing, parstuber, hypendymal, and erythroid-like cells). We recognized that our original aim to infer evolutionary relationships could be confounded by the presence of highly dominant cell-type groups in each dataset. Such class imbalances could bias predictions toward dominant populations, particularly in models that rely on majority-based decision rules like KNN, and hide true relationships between orthologous marker genes and their corresponding cell types.

In particular, the murine dataset was overrepresented by neuronal cells, which accounted for approximately 57% of all profiled nuclei. The human dataset similarly showed skewed composition, with oligodendrocytes and neurons each comprising roughly 18% of cells. This imbalance motivated a curiosity for strategies that investigate dominance effects prior to cross-species analysis. Accordingly, we evaluated several data-ingestion strategies by subsampling the murine hypothalamus to generate three distinct training datasets (Mouse-1, Mouse-2, and Mouse-3) with varying cell-type distributions. Notably, each of these mouse datasets contains cells aggregated from multiple original biological samples to retain sample diversity. We then used these training sets to examine how different cell-type distributions influenced cell-type recovery in five individual human donor samples (Human-1 to Human-5). This restricted testing scope was necessary due to our computational limits, but the natural extension would be to measure performance against the complete human atlas.

4. Methods

Cross-species transfer learning literature strongly emphasizes the need to address species effects. To contextualize our approach within this landscape, we incorporated a baseline k-nearest neighbors (KNN) classifier alongside established cross-species integration methods that explicitly correct for batch effects, namely Seurat v4 and Harmony, allowing us to evaluate the degree to which these approaches improve model performance over a naïve classifier, although we would also encourage the use of the highly specialized SAMap algorithm as another extension as it is designed for manifold alignment in distant species.

We sought to compare a range of modeling strategies to evaluate how different methods for integration affect performance. We selected a k-nearest neighbors (KNN) classifier as a simple, distance-based baseline. Next, we incorporated two methods designed to explicitly mitigate species effects. Seurat v4 employs a Mutual Nearest Neighbors (MNN) framework to identify cross-dataset anchors for integration, which are then used to probabilistically project and transfer cell labels. In contrast, our third approach combined Harmony, an iterative batch correction algorithm used to harmonize species data into a joint latent space, with a Multilayer Perceptron (MLP) classifier trained on this corrected space for non-linear cell-type prediction.

For evaluation, all models were trained using the three subsampled mouse reference atlases (Mouse-1, Mouse-2, and Mouse-3) and tested against five individual human donor samples. Cell-type composition for each subsample and donor may be viewed in Table 2.

4.1. KNN as a baseline model

Given an unlabeled human cell x, KNN identifies the k closest mouse reference cells and assigns the cell type label through a weighted vote among those neighbors. In our implementation, we used $k = 5$ neighbors with the standard Euclidean distance:

$$d(x_i, x_j) = \sqrt{\sum_{l=1}^{d} (x_{il} - x_{jl})^2}$$

For prediction, each neighbor contributes a weight inversely proportional to its distance:

$$w_i = \frac{1}{d(x, x_i)}$$

and the predicted label is

$$\hat{y}(x) = \arg\max_{c} \sum_{i \in N_k(x)} w_i \mathbf{1}\{y_i = c\}$$

This approach captures the local similarity in expression, where nearby mouse cells have more influence and distant cells contribute less. We implemented a pipeline where a common set of orthologous genes is first used to define the feature space, followed by PCA for dimensionality reduction (from 17,853 genes to 100 components) before the KNN classification is performed.

4.2. Seurat v4

Seurat v4 uses an anchor-based label transfer framework to map the annotated mouse hypothalamic cells onto the human dataset. Since both atlases in our analysis were provided as log-normalized expression matrices, we constructed Seurat objects directly from them without additional normalization before scaling. (Shafer, 2019).

4.2.1. Preprocessing and Feature Selection

For each dataset, we identified highly variable genes (HVGs) using Seurat's default FindVariableFeatures() function applied to the ortholog-restricted expression matrices. We then scaled the log-normalized data using ScaleData(), which centers each gene and adjusts for gene-level variance within species. We did not include covariates (such as mitochondrial content or library size).

Both datasets were independently projected into a low-dimensional space using PCA. To ensure consistency in anchor detection, we selected 1,200 shared integration features using SelectIntegrationFeatures() applied jointly to the mouse and human objects.

4.2.2. Anchor Identification and Label Transfer

Using the mouse atlas as reference and the human dataset as query, we computed cross-species anchors using Seurat's FindTransferAnchors() function with PCA as the reference reduction, 30 principal components, and log-normalization as the normalization method:

$$A = \{(i, j) : i \in \text{mouse},\ j \in \text{human},\ i \in \text{NN}(j),\ j \in \text{NN}(i)\}$$

Each anchor corresponds to a pair of mutual nearest neighbors, forming the basis for mapping transcriptomic similarity across species. For each anchor pair (i, j), Seurat computes a similarity weight α_{ij}, which is normalized across all anchors associated with a human cell j:

$$\tilde{\alpha}_{ij} = \frac{\alpha_{ij}}{\sum_{k \in A_j} \alpha_{kj}}$$

During label transfer, mouse cell-type annotations were assigned to human cells using TransferData(), which combines anchor-derived similarity weights to generate a probabilistic distribution for each human cell over the possible mouse cell-types:

$$P(y_j = c) = \sum_{i \in A_j} \tilde{\alpha}_{ij} \cdot \mathbf{1}(y_i = c)$$

The predicted label for each human cell is the class with the highest probability.

4.3. Harmony and Multilayer Perceptron (MLP)

To ensure the mitigation of species effects, we combined a neural network classifier with Harmony batch correction. Harmony constructs a shared latent space that mitigates species-driven differences and a multilayer perceptron (MLP) trains on labeled mouse data to predict cell-types.

We limited features to the highly variable genes independently identified in the mouse and human datasets by Scanpy's HVG selection with the Seurat flavor. We concatenated the datasets into a single AnnData object with a categorical species label. Then, we performed PCA to obtain 50 principal components that served as the initialization for Harmony's integration.

Harmony operates on the PCA embedding matrix $Z \in \mathbb{R}^{n \times d}$, and models each embedding Z_i as the sum of a global mean μ, a batch (here, species) effect β_{b_i}, and residual biological variation: $Z_i = \mu + \beta_{b_i} + \epsilon_i$

It then alternates between clustering and estimating species correction terms. It assigns cells into clusters using soft assignments:

$$r_{ik} = \frac{\exp\left(-\frac{1}{2\sigma^2}\|Z_i - \theta_k\|^2\right)}{\sum_{k'} \exp\left(-\frac{1}{2\sigma^2}\|Z_i - \theta_{k'}\|^2\right)}$$

where θ is the centroid of each cluster k. Within a cluster, Harmony calculates species specific differences

$$\beta_{kb} = \frac{\sum_{i:b_i=b} r_{ik}\,(Z_i - \theta_k)}{\sum_{i:b_i=b} r_{ik}}$$

Then, the corrected Harmony embedding is updated as: $Z_i^{\text{new}} = Z_i - \sum_k r_{ik}\,\beta_{kb_i}$

The updates approximately minimize:

$$\mathcal{L} = \sum_{i,k} r_{ik}\|Z_i - \theta_k - \beta_{kb_i}\|^2 + \lambda \sum_{k,b} \|\beta_{kb}\|^2$$

This approach, in principle, shrinks the species effects while still keeping biological structure in a lower dimensional space.

4.3.1. Cross-Species Prediction

After computing the Harmony embeddings, we used them as the input to train a multilayer perceptron (MLP) classifier. With $x \in \mathbb{R}^d$ as the Harmony corrected latent representation, the MLP has the functional form:

$$f(x) = W_3\,\sigma(W_2\,\sigma(W_1 x + b_1) + b_2) + b_3$$

where $\sigma() = \text{ReLU}()$ is applied elementwise. The softmax transformation computes predicted cell-type probabilities from logits:

$$p(y = c \mid x) = \frac{\exp(f_c(x))}{\sum_{c'=1}^{C} \exp(f_{c'}(x))}$$

To account for cell-type imbalance in the mouse data, we optimized the class-weighted cross-entropy loss:

$$\mathcal{L} = -\sum_{i=1}^{N} w_{y_i} \log p(y_i \mid x_i)$$

where w_{y_i} is the weight assigned to the true class of cell i.

The rationale for including this method was an interest in exploring nonlinear cross-species label transfer in a simple way. While we added Harmony to ensure that batch effects were explicitly controlled, we recognize that more specialized deep learning methods may have better predictive power. This is especially true for models that explicitly focus on batch effects, such as scVI and scANVI, which unfortunately fell outside the scope of this study due to the lack of raw counts in the mouse dataset.

4.4. Evaluation Methods

We evaluated our three models based on homogeneity, ARI, NMI, and the weighted average F1 score by cell type.

Homogeneity measures whether predicted clusters contain cells that share the same true cell type.

$$h = 1 - \frac{H(C|K)}{H(C)}$$

Adjusted Rand Index (ARI) quantifies the similarities between two clusters by calculating how often true and predicted labels agree on data points, after subtracting the expected agreement that would have naturally occurred by chance.

$$\text{ARI}(P^*, P) = \frac{\sum_{i,j}\binom{N_{ij}}{2} - \frac{\sum_i \binom{N_i}{2}\sum_j \binom{N_j}{2}}{\binom{N}{2}}}{\frac{1}{2}\left[\sum_i \binom{N_i}{2} + \sum_j \binom{N_j}{2}\right] - \frac{\sum_i \binom{N_i}{2}\sum_j \binom{N_j}{2}}{\binom{N}{2}}}$$

Normalized Mutual Information Score (NMI) measures how much the predicted labels reduce uncertainty about true labels.

$$\text{NMI}(X, Y) = \frac{2 \cdot I(U; V)}{H(U) + H(V)}$$

Weighted Average F1 Score measures the harmonic mean of precision and recall for each cell type, then averages these scores using the proportion of each cell type as the weight. Precision measures how correct the model's predictions are, namely the fraction of cells correctly predicted as class c out of all cells predicted as c. Recall refers to the fraction of cells correctly predicted as class c out of all cells that truly belong to class c. The F1 score for each class c is the harmonic mean of these two metrics, which means that it harshly penalizes models that are either precise or accurate but not both, such as a model that achieves perfect recall for a cell type by simply predicting all cells as that cell type which is imprecise. In particular, we include the weighted average across all cell types:

$$\text{F1}_{\text{Weighted}} = \sum_{c=1}^{C} \frac{N_c}{N} \cdot \left(2 \cdot \frac{\text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}\right)$$

where N_c is the number of true cells that belong to cell-type c, and N is the total number of cells.

4.5. Compositional Metrics

Given our expectation that dominant classes in our training and testing data may have an impact in the performance of our models, we also incorporate metrics to analyze cell-type distributions for the data we use. In particular, we quantify the diversity of cell types in our training dataset and include a comparison measure of diversity against the testing dataset.

Normalized Shannon Entropy: We quantify the cell-type diversity of each human test sample using Normalized Shannon Entropy (NSE). For a set of cell-type proportions in

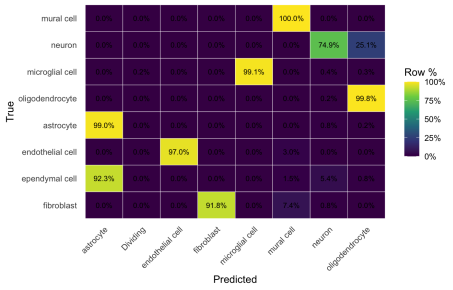

Figure 1. Row-normalized confusion matrix for Seurat v4 Human-3-Mouse-2 pairing

the human data, $P = \{p_1, p_2, \ldots, p_8\}$, the raw Shannon Entropy $H(P)$ is divided by the maximum possible entropy ($\log_2 8$).

$$H_{norm}(P) = \frac{H(P)}{\log_2(8)} = \frac{-\sum_{c=1}^{8} p_c \log_2(p_c)}{\log_2(8)}$$

The resulting value ranges between 0 and 1, where 1 indicates a perfectly uniform distribution of cell types (maximum diversity) and 0 indicates the sample contains only one cell type (minimum diversity). We believe that models may struggle more on samples with low diversity, as it may lead to cells being misspecified as the dominant class.

Jensen-Shannon Divergence (JSD): We measure the distributional shift between the cell-type composition of each mouse training dataset P and human test dataset Q using the Jensen-Shannon Divergence (JSD). JSD is a symmetric and smoothed measure of the difference between two probability distributions, calculated as the average of the Kullback-Leibler (KL) divergence to the average distribution M:

$$\text{JSD}(P||Q) = \frac{1}{2} D_{\text{KL}}(P||M) + \frac{1}{2} D_{\text{KL}}(Q||M)$$

where $M = \frac{1}{2}(P + Q)$.

JSD ranges from 0 to 1, with 0 indicating that the mouse and human cell-type distributions are identical and 1 the opposite, that they have a maxmimum divergence in cell-type composition. Including this metric allows us to evaluate whether poor model performance may stem from differences in the cell-type composition of the training and testing samples.

5. Discussion of Results

Interestingly, we obtained very high performance across all three models throughout the different mouse-human subsample pairs. Across the 15 unique mouse training and human test combinations, no single method was universally superior. As expected, both integration-based methods, Seurat and MLP, consistently outperformed the KNN baseline, but only marginally.

Seurat demonstrated the best overall performance, emerging as the clear winner (achieving the 3 or 4 highest metrics) in 6 of the 15 pairings, and ranking highest in at least one metric for the remaining combinations. MLP was the winner in 4 combinations. No clear winner was found for the rest.

MLP produced higher Homogeneity scores than Seurat in every mouse training sample tested against Human-5, which had one of the highest internal cell-type diversities of the samples included (NSE). It also outranked Seurat in Human-4 which had the highest diversity of all human samples, especially when trained on the mouse datasets most different in cell-type distribution from the testing data. In particular, the Human-4-Mouse-2 combination had the highest divergence out of all pairs, 0.5683, and while this worsened KNN's and Seurat's performance compared to other pairs, Harmony and MLP retained extremely high predictive power, with a 0.9741 weighted average F1 score. This could suggest that the Harmony batch correction and subsequent non-linear MLP classifier generate a more tightly defined latent space as this is beneficial for separating clusters in more diverse query samples.

In fact, while Seurat's performance may have been "better" across a range of mouse-human dataset pairs, it was more variable than that of Harmony + MLP. Beside Human-3-Mouse-2, Harmony + MLP scored consistently high scores across all metrics, whereas Seurat v4 had slightly more variable performance. Concretely, the Mouse-3-Human-2 pairing was surprising, because Seurat v4 experienced a catastrophic failure, recording the lowest scores across all 15 combinations (NMI $= 0.4689$, $\text{F1}_{\text{Weighted}} = 0.5645$).

Human-2 is highly dominated by neurons (75.6% of total cells), but Mouse-3 is the training reference with the lowest neuronal proportion (34%). Seurat's poor metrics are owed to a 74.88% recall for this relatively large neuron class, which sunk the F1 score. This suggests that misclassification was owed to the dominant class in the query data.

The KNN model achieved its best performance in the Mouse-2-Human-2 pairing, which had the lowest observed cell-type distributional divergence between the training and testing data (JSD $= 0.0287$). It's likely that this result stems from the use of Euclidean distance, because smaller shifts between training and test distributions would naturally provide more reliable neighborhood definitions.

6. Conclusion and Future Work

Nevertheless, we did not find a clear, monotonic correlation between the compositional metrics of the training and testing datasets with model performance. The

Mouse-2/Human-3 pairing exhibited the highest overall divergence (JSD $= 0.5683$) and high intrinsic diversity. Under these extreme conditions, both KNN and Seurat v4 maintained strong predictive performance, yet Harmony + MLP recorded some of its lowest results. And still, Harmony + MLP was the clear winner for Human-4 (NSE $= 0.4711$, highest overall diversity) and Human-5 (high diversity/divergence). The lack of a clear, consistent pattern suggests that the nature of the compositional difference, not merely its magnitude, influences which model is optimal, which suggest that a more detailed analysis into the recovery of specific cell types and identified HVGs per class could be necessary.

Notwithstanding, the fact that there is no singular model that wins across the 15 train-test combinations implies that there is no universally optimal pipeline out of our three model candidates. In itself, this notion demonstrates that prediction success is context-dependent, driven by the specific distributional relationship between the reference and query datasets, which was the only modification made to the models in all 15 combinations. As such, it strongly indicates that cell-type and genetic composition influence cross-species prediction, and that the ingested data directly dictates which integration and classification strategy is the most effective.

Beyond our analysis of data ingestion, our models performed exceptionally well. One of the reasons for this was our feature selection approach. Starting with over 17,000 orthologous genes, our initial baseline struggled with a space too large and noisy for accurate cross-species transfer given the evolutionary distance. The use of PCA in all three models reduced the feature space by orders of magnitude and effectively concentrated the classification signal within the most informative features. This simple pre-processing step was so effective that it transformed KNN into a competitor against the advanced integration methods of our other models. In addition, we obtained each model's ultimate parametrization via trial and error on systematic grids.

6.1. Future Work

Our analysis of the compositional metrics (JSD and NSE) suggests that performance is not necessarily a linear function of divergence. The non-linear classification step (MLP) requires the latent space generated by Harmony to be well-structured. High divergence in the cell-type proportion alone may not be the issue; instead, high divergence paired with complex or subtle genetic differences in the misaligned cells may cause the Harmony step to misinterpret local similarities, leading to a degraded latent space where the MLP cannot effectively discriminate cell types. Therefore, our current analysis should be expanded into the genetic composition of misaligned cells.

Furthermore, given the preliminary nature of our analysis, which exhibited a limited range of within-sample cell-type diversity (0.3295 to 0.4711) and a concentration of JSD values in the $0.1-0.3$ range, no clear trends can be observed. Further work must, therefore, expand the experimental space to explore a greater number of dataset combinations with deliberately wide variations in JSD and NSE to be able to assess if this relationship is true, not solely with our current mouse and human datasets, but perhaps with other species' region-focused atlases as well.

The failure case of Seurat v4 on Mouse-3-Human-2 suggests a need to go beyond aggregate metrics. The significant drop in Seurat's F1 score was not due to small, scattered errors but to a single, critical misclassification of the largest cell type in the testing data (neurons). This encourages us to investigate prediction by cell-type, instead of sample, to identify which cell types (e.g., rare or dominant ones) are most sensitive to train-test composition mismatches for each model.

Additionally, our current analysis focused on cross-species cell type recovery methods that worked directly on log-normalized expression matrices that enabled a controlled comparison across models. An important next step would be to include raw count data into our ingestion pipeline. Unfortunately, this was not available to us for the murine hypothalamus complete dataset.

Similarly, while we discussed SAMap in the context of prior benchmarking work, a full application of SAMap would be an interesting addition to our model universe, as well as scVI and scANVI.

6.2. The Unanswered Question

We set out this analysis with the particular goal of identifying evolutionary relationships between the house mouse and human hypothalamus cell types. In the experimental process, we realized issues relating to the performance of our classifiers, which left evolutionary biology as a second-order concern. Thus, it remains a large unanswered question that could guide consequent analysis, or reframe the classification process entirely. As the present approach stands, it is possible to conduct focused genetic analysis of the misclassified cells to understand the underlying genetic and evolutionary distance contributing to the model's inability to correctly transfer certain labels, not just in genetic expression, but in potentially disaparate roles that recurring orthologous genes in misidentified classes have developed since their common ancestor.

Table 1. Cross-species cell-type recovery metrics in 5 human donors for multiple differently sampled mouse training datasets. Diversity is computed per human sample (normalized Shannon entropy of human cell-type proportions), and JSD is the Jensen–Shannon divergence between each mouse dataset and each human sample (lower indicates more similar compositions).

Testing data	Human sample cell-type diversity	Training data	JSD human-mouse	Model	Homogeneity	ARI	NMI	Weighted F1
Human-1	0.3297	Mouse-1	0.1598	KNN	0.8871	0.9646	0.8976	0.9731
				Seurat v4	0.9743	**0.9944**	**0.9663**	**0.9950**
				Harmony + MLP	**0.9780**	0.9847	0.9350	0.9840
		Mouse-2	0.0334	KNN	**0.9681**	0.9952	**0.9595**	0.9848
				Seurat v4	0.9643	**0.9871**	0.9540	**0.9930**
				Harmony + MLP	0.9407	0.8538	0.7958	0.9350
		Mouse-3	0.1634	KNN	0.8067	0.9038	0.8573	0.9654
				Seurat v4	0.9743	**0.9959**	0.9581	0.9937
				Harmony + MLP	**0.9958**	0.9919	**0.9726**	**0.9960**
Human-2	0.3295	Mouse-1	0.1547	KNN	0.8682	0.9470	0.8928	0.9798
				Seurat v4	**0.9915**	0.9770	**0.9934**	0.9552
				Harmony + MLP	0.9876	**0.9896**	0.9516	**0.9926**
		Mouse-2	0.0287	KNN	0.9755	0.9966	0.9695	0.9917
				Seurat v4	**0.9763**	**0.9943**	**0.9724**	**0.9960**
				Harmony + MLP	0.9645	0.9294	0.8774	0.9690
		Mouse-3	0.1587	KNN	0.7456	0.8665	0.8172	0.9506
				Seurat v4 [1]	0.8055	0.6904	0.4689	0.5645
				Harmony + MLP	**0.9823**	**0.9647**	**0.9320**	**0.9841**
Human-3	0.4702	Mouse-1	0.1394	KNN	0.8595	0.9420	0.8760	0.9637
				Seurat v4	**0.9261**	**0.9662**	**0.9261**	**0.9849**
				Harmony + MLP	0.9069	0.8875	0.8423	0.9441
		Mouse-2	0.2776	KNN	0.9322	**0.9668**	0.9198	**0.9673**
				Seurat v4	**0.9855**	0.9349	**0.9727**	0.9349
				Harmony + MLP	0.8591	0.8170	0.7090	0.8269
		Mouse-3	0.1118	KNN	0.6874	0.7304	0.7146	0.8803
				Seurat v4	0.9234	**0.9678**	**0.9678**	0.9234
				Harmony + MLP	**0.9652**	0.9267	0.9067	**0.9763**
Human-4	0.4711	Mouse-1	0.2541	KNN	0.7773	0.9192	0.7761	0.9269
				Seurat v4	0.9121	**0.9773**	0.8912	0.9740
				Harmony + MLP	**0.9690**	0.9708	**0.8966**	**0.9749**
		Mouse-2	0.5683	KNN	0.8764	0.9695	0.8709	0.9457
				Seurat v4	0.8912	0.9681	0.8656	0.9659
				Harmony + MLP	**0.9716**	**0.9688**	**0.8926**	**0.9741**
		Mouse-3	0.2223	KNN	0.7565	**0.8945**	0.7659	0.8745
				Seurat v4	0.8396	0.8423	**0.9360**	0.8423
				Harmony + MLP	**0.9258**	0.8660	0.8368	**0.9530**
Human-5	0.4235	Mouse-1	0.1565	KNN	0.8676	0.9572	0.8896	0.9708
				Seurat v4	0.9403	**0.9780**	**0.9403**	**0.9885**
				Harmony + MLP	**0.9492**	0.9091	0.8630	0.9694
		Mouse-2	0.2914	KNN	0.9401	0.9790	0.9340	0.9819
				Seurat v4	0.9398	**0.9792**	**0.9399**	**0.9895**
				Harmony + MLP	**0.9506**	0.9099	0.8519	0.9648
		Mouse-3	0.1241	KNN	0.7020	0.7743	0.7319	0.8951
				Seurat v4	0.9483	**0.9812**	**0.9461**	**0.9884**
				Harmony + MLP	**0.9846**	0.9787	0.9424	0.9879

Note: The Seurat v4 integration for the human donor 3/Mouse-2 pair showed inconsistent results compared to the other tested pairs. While the model achieved high prediction precision across the non-neuronal cell types, its overall testing accuracy was reduced due to a significant misclassification of neurons. Specifically, the recall for the neuron class was only 74.88%.

Table 2. Cell Type Composition (Proportions %) Across Human and Mouse Datasets

CellType	Human-1	Human-2	Human-3	Human-4	Human-5	Mouse-1	Mouse-2	Mouse-3
Oligodendrocyte	14.62	14.51	**55.47**	**63.41**	**61.60**	23.54	9.64	**28.97**
Neuron	**75.98**	**75.60**	26.24	2.37	24.60	**36.31**	**77.82**	**34.51**
Mural cell	0.48	0.12	1.27	2.57	0.28	1.87	0.47	1.87
Microglial cell	4.05	2.88	7.71	13.77	4.88	6.94	1.40	5.93
Fibroblast	0.34	0.30	0.15	0.84	0.27	0.78	0.32	0.56
Ependymal cell	0.09	0.32	0.02	0.99	0.19	9.78	4.48	7.91
Endothelial cell	0.27	0.16	0.15	1.36	0.29	6.04	0.64	6.80
Astrocyte	4.17	6.11	9.00	14.70	7.89	13.49	4.32	12.49
ParsTuber	0.00	0.00	0.00	0.00	0.00	0.60	0.32	0.33
Hypendymal	0.00	0.00	0.00	0.00	0.00	0.06	0.02	0.08
Erythroid-like	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.05
Dividing	0.00	0.00	0.00	0.00	0.00	0.52	0.58	0.49

7. Acknowledgments

We would like to thank our course instructor, Dr. Bianca Dumitrascu, for her guidance and feedback throughout STAT W4243. This report was produced collaboratively by Sofía Prado Arenzana (M.S. student in Statistics, Columbia University), Shantell Luna (undergraduate student in Data Science, Columbia University), and Greta VanZetten (undergraduate student in Data Science, Columbia University). Any remaining errors are our own.

References

Shafer, M. E. R. Cross-species analysis of single-cell transcriptomic data. *Frontiers in Cell and Developmental Biology*, 7:175, 2019. ISSN 2296-634X. doi: 10.3389/fcell.2019.00175.

Song, Y., Miao, Z., Brazma, A., et al. Benchmarking strategies for cross-species integration of single-cell rna sequencing data. *Nature Communications*, 14:6495, 2023. doi: 10.1038/s41467-023-41855-w.

Steuernagel, L., Lam, B. Y. H., Klemm, P., et al. Hypomap—a unified single-cell gene expression atlas of the murine hypothalamus. *Nature Metabolism*, 4:1402–1419, 2022. doi: 10.1038/s42255-022-00657-y.

Tadross, J. A., Steuernagel, L., Dowsett, G. K. C., et al. A comprehensive spatio-cellular map of the human hypothalamus. *Nature*, 639:708–716, 2025. doi: 10.1038/s41586-024-08504-8.

Xu, C., Lopez, R., Mehlman, E., Regier, J., Jordan, M., and Yosef, N. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Molecular Systems Biology*, 17(1):e9620, 2021. doi: 10.15252/msb.20209620.