Cell-Type-Aware Pooling for Robust Sample Classification in Single-Cell RNA-seq Data

Soorin Yim [*1] **Kyungwook Lee** [*1] **Dongyun Kim** [*†1 2] **Sungjoon Park** [1] **Doyeong Hwang** [1] **Soonyoung Lee** [1] **Amy Dunn** [3] **Daniel Gatti** [3] **Elissa Chesler** [3] **Kristen O'Connell** [3] **Kiyoung Kim** [1]

Abstract

Single-cell RNA sequencing (scRNA-seq) enables high-resolution profiling of cellular heterogeneity, offering a promising foundation for predicting phenotypes such as disease status. We propose a pooling strategy that utilizes cell type annotations by first aggregating cell representations within each cell type, followed by integration of cell type representations into a sample-level representation. Evaluated across three scRNA-seq datasets of varying sizes and biological contexts, our model consistently outperforms baseline models in phenotype classification. Our model is particularly effective in datasets with missing or sparsely represented cell types. These results underscore the importance of carefully incorporating cell type information for robust phenotype prediction from scRNA-seq data.

1. Introduction

Single-cell RNA sequencing (scRNA-seq) enables high-resolution characterization of the transcriptomic landscape across heterogeneous cellular populations. Recent advances have demonstrated the potential of scRNA-seq data for predicting sample-level phenotypes such as disease status and treatment response (Xiong et al., 2023; Verlaan et al., 2025; Do & Lähdesmäki, 2025). These predictions offer critical insights for understanding disease mechanisms and advancing personalized medicine (He et al., 2021; Litinetskaya et al., 2024).

Only a limited number of studies have explored cell-type-aware modeling for phenotype prediction from scRNA-seq data (Xiong et al., 2023; Do & Lähdesmäki, 2025). ProtoCell4P (Xiong et al., 2023) uses cell type information by encouraging cells of the same type to cluster in the latent space. On the other hand, (Do & Lähdesmäki, 2025) introduced a hierarchical pooling framework based on cell types, but relies on attention mechanisms at cell and cell type levels. While effective in some settings, attention-based pooling can overfit in noisy, low-sample regimes typical of scRNA-seq data (Lin et al., 2022; Zhong et al., 2025).

To address these limitations, we propose Cell Type Mean (CTMean), a simple yet effective two-step pooling framework for sample-level phenotype prediction. CTMean first aggregates cell representations within each annotated cell type via mean pooling and then integrates the resulting cell type representations via another mean pooling step. This strategy enforces shared global weights to cell types across samples, improving robustness, particularly in settings with high noise or limited sample sizes. By explicitly leveraging the hierarchical structure of biological data, CTMean provides a scalable and interpretable solution for robust phenotype prediction.

We evaluated CTMean on three scRNA-seq datasets, showing consistent improvements over baselines that either ignore cell type information or rely on attention-based pooling. These results highlight that careful design of the pooling mechanism is critical for accurate and robust phenotype prediction.

Our contributions are summarized as follows:

- We propose CTMean, a novel, cell-type-aware pooling strategy for phenotype prediction from scRNA-seq data.
- We demonstrate that CTMean is particularly robust in settings with incomplete cell type compositions.
- We show that incorporating auxiliary prediction at the cell type level can improve model performance.

[*]Equal contribution . [†]Work done during an internship at LG AI Research. [1]LG AI Research, Seoul, South Korea [2]Department of Chemistry, Seoul National University, Seoul, South Korea [3]The Jackson Laboratory, Bar Harbor ME, USA. Correspondence to: Kiyoung Kim <elgee.kim@lgresearch.ai>.

Proceedings of the ICML 2025 Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Vancouver, Canada. PMLR 267, 2025.

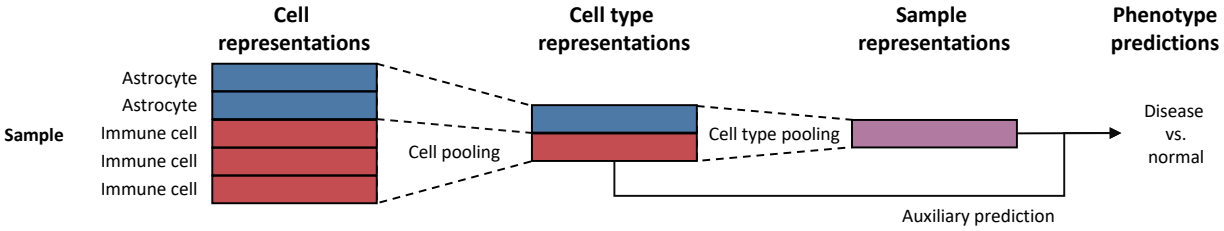

Figure 1. Model architecture. A sample has multiple cells, which is embedded into cell representations. Cell pooling aggregates cell representations into cell type representations, which in turn is aggregated into sample representations by cell type pooling. The sample representation is used for phenotype predictions. Additionally, each cell type representation passes through the same phenotype predictor for auxiliary prediction.

2. Method

Figure 1 illustrates the overall architecture of our proposed model, CTMean. Each sample consists of multiple cells, where each cell is annotated with a cell type label (e.g., astrocyte, immune cell). Each cell is represented by its gene expression profile, which is embedded into a latent representation via a multi-layer perceptron (MLP).

The model employs a hierarchical two-step pooling strategy. First, cell-level pooling aggregates cell representations within each cell type to form a cell-type-specific representation. Next, cell-type-level pooling aggregates these representations to form a final sample-level representation, which is used for phenotype classification (e.g., disease vs. control).

While simple mean pooling is effective, it does not explicitly encourage phenotype-relevant signals to emerge in the intermediate representations. To address this, we introduced auxiliary prediction tasks at the cell type level, allowing the cell type representations to capture phenotype information directly. Each cell type representation is passed through the same MLP classifier used for sample-level prediction, encouraging the intermediate features to retain discriminative information relevant to the target phenotype.

The total loss function is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{sample}} + \frac{1}{N_{\text{ct}}} \sum_{i=1}^{N_{\text{ct}}} \mathcal{L}_i \tag{1}$$

where $\mathcal{L}_{sample}$ is the cross-entropy loss for sample-level prediction, $\mathcal{L}_i$ is the cross-entropy loss for auxiliary classification using ith cell type, and N_{ct} denotes the number of distinct cell types.

3. Experiments

3.1. Datasets

We evaluated our model on three publicly available scRNA-seq datasets spanning a range of sample sizes, cell counts, and numbers of cell types, as summarized in Table 1. Each dataset was preprocessed to produce normalized, log-transformed gene expression values using up to 1,000 genes. We also removed unannotated cells without cell type information. This ensures robust and fair comparison across diverse biological contexts. We briefly describe each dataset below; detailed preprocessing steps are provided in the Appendix A.

The Immune Checkpoint Blockade (ICB) dataset consists of single-cell transcriptomic profiles from 57 patients who underwent immune checkpoint blockade therapy (Gondal et al., 2025). The task involves binary classification of treatment outcome (38 responders vs. 19 non-responders). Cell type labels were provided by the authors of (Do & Lähdesmäki, 2025), using SingleR (Aran et al., 2019) for automatic annotation, resulting in 23 distinct cell types.

COVID dataset includes scRNA-seq profiles from 56 individuals spanning three clinical conditions: 15 healthy controls, 35 patients with COVID-19, and 6 patients with respiratory failure (Ziegler et al., 2021). Samples labeled as "long COVID" were excluded due to the small number of such cases (n=2). The dataset includes 18 cell types, as annotated in the original study.

The Religious Orders Study and Memory and Aging Project (ROSMAP) dataset contains single-nucleus RNA-seq (snRNA-seq) data from 139 postmortem human brain samples (Mathys et al., 2023). The binary classification task is to distinguish between 78 Alzheimer's Disease (AD) and 61 cognitively healthy individuals. The dataset includes 7 cell types from the original study (Mathys et al., 2023).

3.2. Performance Comparison

We evaluated phenotype classification performance on three datasets using Area Under the receiver operating characteristic Curve (AUC) as the primary evaluation metric, as shown in Table 2. We compare CTMean against several

Table 1. Summary statistics of the datasets used for phenotype classification from scRNA-seq.

DATASET	SAMPLES (CLASS DISTRIBUTION)	CELLS	AVG. CELLS PER SAMPLE	CELL TYPES	AVG. CELL TYPES PER SAMPLE
ICB	57 (38 + 19)	9,292	163	23	8.3 (36%)
COVID	56 (15 + 35 + 6)	30,282	541	18	11.4 (63%)
ROSMAP	139 (78 + 61)	890,314	6,405	7	7 (100%)

Table 2. Performance comparison with baseline models. "O" indicates cell type annotation used; "X" indicates it was not used. AUC ± standard deviation on test dataset across 10 repetitions. **Bold** values indicate the best performance, while underlined values indicate the second-best.

MODEL	CELL TYPE ANNOTATION	CELL POOLING	CELL TYPE POOLING	ICB	COVID	ROSMAP
PROTOCELL4P	O	X	X	0.63±0.05	0.82±0.02	0.77±0.01
CELL ATTENTION (CA)	X	ATTENTION	X	0.67±0.03	0.81±0.05	0.71±0.04
MEAN-POOLING	X	MEAN	X	0.73±0.05	0.79±0.05	0.74±0.03
HIERARCHICAL ATTENTION (HA)	O	ATTENTION	ATTENTION	0.76±0.04	0.75±0.05	0.73±0.04
CELL TYPE ATTENTION (CTA)	O	MEAN	ATTENTION	0.71±0.04	0.76±0.03	0.77±0.05
CTMEAN	O	MEAN	MEAN	**0.79±0.01**	0.84±0.04	**0.80±0.03**
CTMEAN + AUXILIARY PREDICTION	O	MEAN	MEAN	0.79±0.03	**0.86±0.03**	0.79±0.02

baselines, including ProtoCell4P (Xiong et al., 2023), and pooling-based models from (Do & Lähdesmäki, 2025). All models were evaluated according to their original paper using repeated, nested cross-validation as done in (Do & Lähdesmäki, 2025). For pooling-based models from (Do & Lähdesmäki, 2025), we fix the dimensionality of all cell representations to the same dimensions, 32, to ensure that all models have the same expressive power. Details of training procedures and hyperparameter optimization are provided in the Appendix B.

Mean-pooling and Cell Attention (CA) aggregate cell representations directly into a sample-level representation without incorporating cell type information. Mean-pooling simply averages all cell representations, being equivalent to a weighted average of cell type representation based on their proportions. CA instead applies attention over individual cells, following a standard Multiple Instance Learning (MIL) approach.

Hierarchical Attention (HA) and Cell Type Attention (CTA) follow a two-step pooling strategy: they first aggregate cells within each cell type, then combine cell type representations using attention-based pooling. In contrast, CTMean also adopts a two-step pooling approach but replaces attention with mean pooling at both levels. This simplified design consistently outperforms attention-based models, particularly in datasets with limited or missing cell types, such as ICB and COVID. This advantage may arise from the tendency of attention mechanisms to overfit under high noise and small sample sizes, which are common in scRNA-seq data (Lin et al., 2022; Zhong et al., 2025).

Lastly, unlike ICB and COVID, the ROSMAP dataset provides complete cell type coverage across all samples and contains over ten times more cells per sample. This richer structure enables models that incorporate cell type information to achieve improved performance. In particular, we observed that CTA achieved relatively strong performance on ROSMAP, likely due to its ability to exploit the abundant and consistent cell type signals. For example, cell types such as immune cells - suggested to be implicated in AD (Green et al., 2024; Verlaan et al., 2025)—consistently received higher attention weights (Appendix Figure 4). However, these scores remained sensitive to random initialization, suggesting limited robustness. In contrast, the use of shared cell type weights in CTMean encourages more stable and generalizable representations, leading to improved performance even in large and complete datasets.

3.3. Model Analysis

To assess the effectiveness of auxiliary prediction, we evaluated the performance of CTMean both with and without it. As shown in Table 2, auxiliary prediction led to an improvement on the COVID dataset, while its effect was marginal on ICB and ROSMAP. This suggests that auxiliary prediction is especially beneficial in datasets with limited signal or higher variability. Furthermore, UMAP visualizations in Appendix Figure 3 confirm that auxiliary prediction helps enforce phenotype-relevant structure in the cell type representations, making them more discriminative.

To understand why mean pooling exhibits greater robustness than attention mechanisms when some cell types are missing, we analyzed the ICB dataset, which has the most diverse cell type distribution. We trained the CTA using 10 random seeds and computed the attention scores assigned to each cell type within individual samples. Then, for each cell type, we averaged these scores across the 10 random seeds. To quantify variability of attention scores, we calculated the relative attention deviation by subtracting the uniform attention score and normalizing the result by the uniform score. As shown in Figure 2, most samples exhibit deviations close to zero, indicating near-uniform attention. Additionally, 95% of samples show a relatively low coefficient of variation which is below 0.3, confirming that attention is nearly uniform across present cell types.

We hypothesize that this near-uniform attention arises from the high variability and sparsity of cell types across samples, which hinders the model's ability to learn meaningful and sample-specific attention patterns. Consequently, the model defaults to a mean pooling-like behavior that treats all present cell types more equally. While slight deviations from uniformity do occur, these may act as a noise and lead to overfitting. This may explain why CTMean outperforms CTA in scenarios involving missing cell types.

3.4. The Use of Single-Cell Foundation Models

We also tested whether pretrained single-cell foundation models could improve performance in low-sample settings. Specifically, we replaced normalized gene expression values with embeddings from scGPT (Cui et al., 2024), evaluating both general-purpose (whole-human) and organ-specific variants in a zero-shot setting. As shown in Appendix Table 5, using scGPT embeddings consistently degraded performance across all datasets. This aligns with recent findings that zero-shot application of single-cell foundation models may underperform compared to task-specific models (Kedzierska et al., 2025). These results suggest that fine-tuning may be necessary to fully leverage foundation models for phenotype prediction.

4. Conclusion

We introduced CTMean, a cell-type-aware framework for phenotype prediction from scRNA-seq data. CTMean employs a two-step mean pooling strategy: aggregating cell representations within each cell type, followed by combining cell type representations into a sample-level representation. Evaluated across three datasets with varying numbers of cells and cell types, CTMean consistently outperforms existing methods. Notably, its simple and robust design makes it particularly effective in datasets with missing cell types or a limited number of cells.

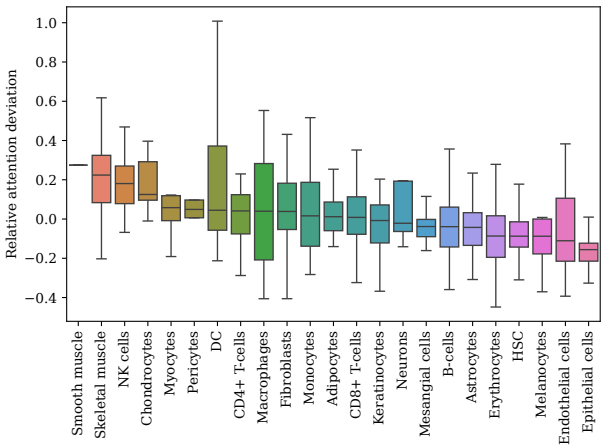

Figure 2. Distribution of relative attention deviation for each cell type in ICB dataset, computed as (attention - uniform attention) / uniform attention.

Despite its strong performance, CTMean has room for improvement. It does not explicitly incorporate cell type proportions, which have been shown to be informative for phenotype prediction (Litinetskaya et al., 2024; Verlaan et al., 2025); incorporating this information may further enhance accuracy. In addition, CTMean depends on predefined cell type annotations, making it sensitive to annotation quality. Future work could explore end-to-end models that jointly learn cell type groupings and phenotype predictions, potentially uncovering novel cell populations. Lastly, CTMean offers a promising foundation for extension to multimodal single-cell omics data, such as paired scRNA-seq and scATAC-seq (Litinetskaya et al., 2024).

Acknowledgements

The results published here are in part based on data obtained from the AD Knowledge Portal (https://adknowledgeportal.org). Study data were provided by the Rush Alzheimer's Disease Center, Chicago. This work was supported by LG AI Research and in part by the National Institutes of Health (NIH) grant RF1AG059778, and the Alzheimer's Association Research Fellowship AARF-18-565506.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Alvarez-Breckenridge, C., Markson, S. C., Stocking, J. H., Nayyar, N., Lastrapes, M., Strickland, M. R., Kim, A. E., De Sauvage, M., Dahal, A., Larson, J. M., et al. Microenvironmental landscape of human melanoma brain metastases in response to immune checkpoint inhibition. *Cancer immunology research*, 10(8):996–1012, 2022.

Aran, D., Looney, A. P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R. P., Wolters, P. J., Abate, A. R., et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature immunology*, 20(2):163–172, 2019.

Bassez, A., Vos, H., Van Dyck, L., Floris, G., Arijs, I., Desmedt, C., Boeckx, B., Vanden Bempt, M., Nevelsteen, I., Lambein, K., et al. A single-cell map of intratumoral changes during anti-pd1 treatment of patients with breast cancer. *Nature medicine*, 27(5):820–832, 2021.

Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., and Wang, B. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 21(8):1470–1480, 2024.

Do, C. and Lähdesmäki, H. Incorporating hierarchical information into multiple instance learning for patient phenotype prediction with scrna-seq data. *bioRxiv*, pp. 2025–02, 2025.

Gondal, M. N., Cieslik, M., and Chinnaiyan, A. M. Integrated cancer cell-specific single-cell rna-seq datasets of immune checkpoint blockade-treated patients. *Scientific Data*, 12(1):139, 2025.

Green, G. S., Fujita, M., Yang, H.-S., Taga, M., Cain, A., McCabe, C., Comandante-Lou, N., White, C. C., Schmidtner, A. K., Zeng, L., et al. Cellular communities reveal trajectories of brain ageing and alzheimer's disease. *Nature*, 633(8030):634–645, 2024.

He, B., Thomson, M., Subramaniam, M., Perez, R., Ye, C. J., and Zou, J. Cloudpred: Predicting patient phenotypes from single-cell rna-seq. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2022*, pp. 337–348. World Scientific, 2021.

Kedzierska, K. Z., Crawford, L., Amini, A. P., and Lu, A. X. Zero-shot evaluation reveals limitations of single-cell foundation models. *Genome Biology*, 26(1):101, 2025.

Lin, T., Wang, Y., Liu, X., and Qiu, X. A survey of transformers. *AI open*, 3:111–132, 2022.

Litinetskaya, A., Shulman, M., Hediyeh-zadeh, S., Moinfar, A. A., Curion, F., Szałata, A., Omidi, A., Lotfollahi, M., and Theis, F. J. Multimodal weakly supervised learning to identify disease-specific changes in single-cell atlases. *bioRxiv*, pp. 2024–07, 2024.

Mathys, H., Peng, Z., Boix, C. A., Victor, M. B., Leary, N., Babu, S., Abdelhady, G., Jiang, X., Ng, A. P., Ghafari, K., et al. Single-cell atlas reveals correlates of high cognitive function, dementia, and resilience to alzheimer's disease pathology. *Cell*, 186(20):4365–4385, 2023.

Pozniak, J., Pedri, D., Landeloos, E., Van Herck, Y., Antoranz, A., Vanwynsberghe, L., Nowosad, A., Roda, N., Makhzami, S., Bervoets, G., et al. A tcf4-dependent gene regulatory network confers resistance to immunotherapy in melanoma. *Cell*, 187(1):166–183, 2024.

Verlaan, T., Bouland, G., Mahfouz, A., and Reinders, M. scagg: Sample-level embedding and classification of alzheimer's disease from single-nucleus data. *bioRxiv*, pp. 2025–01, 2025.

Wan, Y.-W., Al-Ouran, R., Mangleburg, C. G., Perumal, T. M., Lee, T. V., Allison, K., Swarup, V., Funk, C. C., Gaiteri, C., Allen, M., et al. Meta-analysis of the alzheimer's disease human brain transcriptome and functional dissection in mouse models. *Cell reports*, 32(2), 2020.

Xiong, G., Bekiranov, S., and Zhang, A. Protocell4p: an explainable prototype-based neural network for patient classification using single-cell rna-seq. *Bioinformatics*, 39(8):btad493, 2023.

Zhong, J., Tian, W., Xie, Y., Liu, Z., Ou, J., Tian, T., and Zhang, L. Pmfsnet: Polarized multi-scale feature self-attention network for lightweight medical image segmentation. *Computer Methods and Programs in Biomedicine*, pp. 108611, 2025.

Ziegler, C. G., Miao, V. N., Owings, A. H., Navia, A. W., Tang, Y., Bromley, J. D., Lotfy, P., Sloan, M., Laird, H., Williams, H. B., et al. Impaired local intrinsic immunity to sars-cov-2 infection in severe covid-19. *Cell*, 184(18): 4713–4733, 2021.

A. Dataset Preprocessing

ICB. We used a preprocessed version of the Immune Checkpoint Blockade (ICB) dataset, publicly available on Zenodo (accession ID: 10407126) (Gondal et al., 2025). Following the data selection protocol of (Do & Lähdesmäki, 2025), we included only pre-treatment samples from three of the original studies (Alvarez-Breckenridge et al., 2022; Pozniak et al., 2024; Bassez et al., 2021). For each sample, cells were downsampled to a maximum of 200 in (Gondal et al., 2025).

COVID. The COVID-19 dataset was downloaded from the Single Cell Portal (accession ID: SCP1289) (Ziegler et al., 2021). We applied the following preprocessing steps: (1) removed genes expressed in fewer than five cells; (2) normalized total counts to 10,000 per cell; (3) applied a $\log(x+1)$ transformation; and (4) selected the top 1,000 most highly expressed genes. The dataset includes 18 cell types as annotated in the original study.

ROSMAP. We used the ROSMAP dataset, downloaded from the authors' website at compbio.mit.edu/ad_aging_brain. Following prior work (Wan et al., 2020; Verlaan et al., 2025), individuals were labeled as AD only if they met both clinical and neuropathological criteria. We removed genes expressed in fewer than 200 cells and cells expressing fewer than 200 genes. Total counts were normalized to 10,000 per cell and transformed using $\log(x+1)$. We then selected the top 1,000 most highly expressed genes. The dataset includes 7 cell types, as defined in the original study (Mathys et al., 2023).

B. Hyperparameter Optimization

All models are evaluated using a repeated, nested cross-validation (CV) procedure, following (Do & Lähdesmäki, 2025). For each repeat, we perform 5-fold CV in the outer loop to evaluate model performance. Within each outer training fold, we perform 3-fold CV in the inner loop to optimize hyperparameters using 30 Optuna trials. The combination of hyperparameters that yields the highest AUC across the inner folds is selected and used to train the model on the entire outer training fold. Predictions from all outer test folds are aggregated to compute the final AUC per repeat.

This entire procedure is repeated 10 times using different random seeds for CV splitting, resulting in 10 independent AUC scores per model. We report the mean and standard deviation of these 10 AUCs. The hyperparameter search space is summarized in Table 3.

Table 3. Hyperparameter search space used during optimization.

HYPERPARAMETER	SEARCH SPACE
NUMBER OF EPOCHS	{100, 500, 1000}
DROPOUT RATE	{0, 0.3, 0.5, 0.7}
WEIGHT DECAY	{1E-4, 1E-3, 1E-2}
NUMBER OF LAYERS	{1, 2}
LEARNING RATE	{1E-3, 5E-3}
ACTIVATION FUNCTION	{RELU, ELU}

C. Effect of Embedding Dimension

Table 4. AUC comparison across three datasets (ICB, COVID, ROSMAP) at embedding dimensions 64 and 128. Each value is reported as mean ± standard deviation.

MODEL	**ICB**		**COVID**		**ROSMAP**	
EMBEDDING DIMENSION	64	128	64	128	64	128
CA	0.68 ± 0.05	0.67 ± 0.04	0.81 ± 0.03	0.81 ± 0.05	0.71 ± 0.03	0.70 ± 0.04
MEAN-POOLING	0.74 ± 0.04	0.73 ± 0.04	0.79 ± 0.04	0.80 ± 0.04	0.75 ± 0.02	0.74 ± 0.03
HA	0.72 ± 0.05	0.74 ± 0.05	0.73 ± 0.07	0.72 ± 0.05	0.74 ± 0.03	0.72 ± 0.05
CTA	0.73 ± 0.04	0.73 ± 0.06	0.74 ± 0.04	0.73 ± 0.04	0.77 ± 0.03	0.77 ± 0.02
CTMEAN	0.78 ± 0.04	0.76 ± 0.04	0.81 ± 0.03	0.82 ± 0.03	0.79 ± 0.02	0.80 ± 0.01

In Table 2, we reported model performance using a fixed embedding dimension of 32. To assess the robustness of our results with respect to embedding size, we additionally evaluated all models with embedding dimensions of 64 and 128. As

shown in Table 4, CTMean consistently outperforms CA, Mean-pooling, HA, and CTA across all embedding dimensions, demonstrating that its superior performance is not sensitive to embedding size.

D. Auxiliary Prediction

To enhance the representation learning of individual cell types, we introduce an auxiliary loss that provides additional supervision during training. This loss encourages the model to learn more discriminative features specific to each cell type, thereby improving sample-level classification performance. Concretely, for each cell type, we obtain a cell type representation by aggregating the representations of all corresponding cells within a sample, and then apply an auxiliary classifier to predict the sample label from this cell type representation. Empirically, we find that incorporating this auxiliary loss leads to more well-separated latent space as shown in Figure 3, which may contribute to improved classification accuracy.

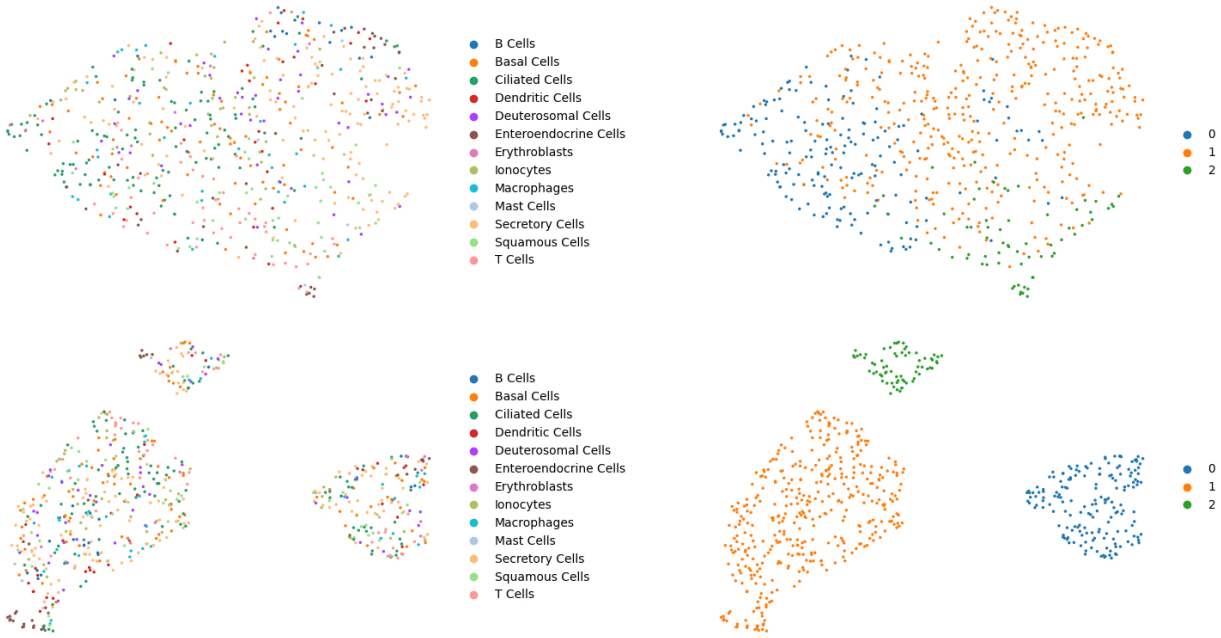

Figure 3. UMAP visualization of cell type representations from the COVID dataset. Representations were extracted after cell pooling but before cell type pooling to examine the effect of the auxiliary loss. Each point corresponds to a unique sample–cell-type pair. The top two panels show representations from the CTMean model trained without auxiliary loss, while the bottom two panels show representations from the model trained with auxiliary loss. As our model is a classifier, representations naturally cluster by patient label (right panels) with various cell types in the cluster (left panels). The introduction of the auxiliary loss leads to better label separation in the latent space, corresponding to improved sample-level classification performance.

E. Attention Distribution Across Cell Types

To understand why CTMean outperforms CTA, we analyzed how cell type attention scores vary in the CTA model using the ROSMAP dataset. Specifically, we trained the CTA model with 10 different random seeds, and for each sample, we computed the Jensen–Shannon divergence (JSD) between the attention score distributions across all cell types for every pair of seeds. We then averaged these pairwise JSD values to quantify the variability in attention scores caused by random initialization. Figure 4 shows that, among cell types implicated in Alzheimer's disease, immune cells consistently receive higher attention scores, indicating that the model relies on neuroinflammatory signals when making predictions. At the same time, although samples with the highest mean JSD naturally display the greatest seed-to-seed variability, even those with median or low mean JSD exhibit substantial fluctuations in which cell types are deemed most important. In other words, random initialization alone can shift the model's attention ranking, underscoring that attention outputs from a single seed should be interpreted with caution.

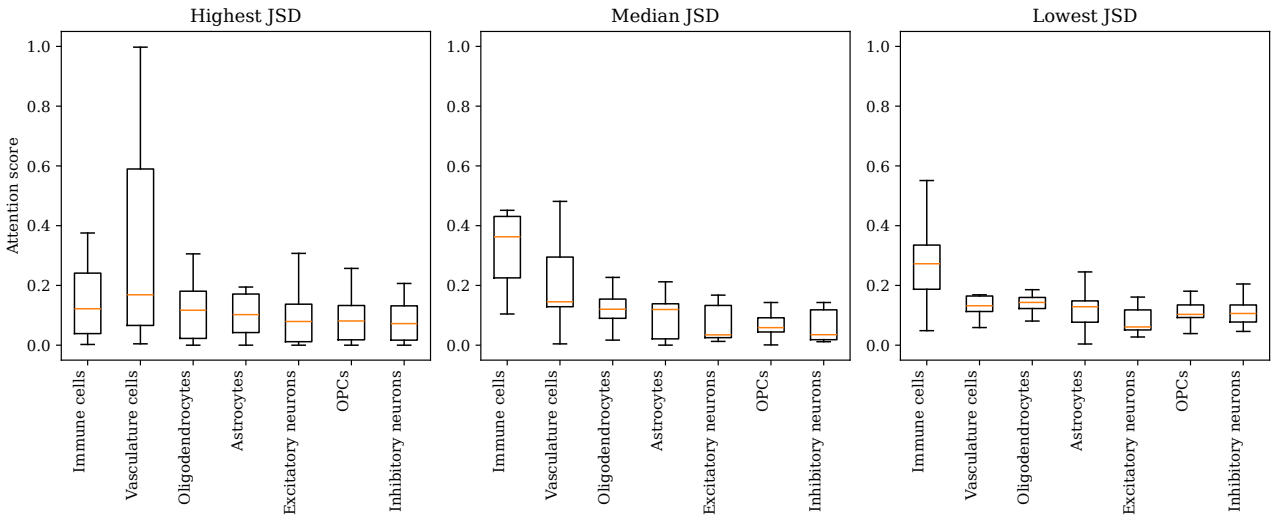

Figure 4. Seed-wise distribution of cell type attention scores for representative ROSMAP samples. For each sample, we computed the Jensen–Shannon divergence (JSD) between attention-score distributions across all cell types for every pair of 10 random seeds, then averaged those JSD values. Shown are the attention score distribution of samples with the highest mean JSD (left), the median mean JSD (center), and the lowest mean JSD (right).

F. The Use of Single-Cell Foundation Models

We investigated whether leveraging a single-cell foundation model could enhance phenotype prediction performance. Specifically, we utilized scGPT (Cui et al., 2024) to obtain pretrained embeddings for each cell, replacing normalized gene expression values as input to our model. To examine the effect of context-specific pretraining, we evaluated two variants of scGPT for each dataset: a general-purpose whole-human model and an organ-specific model. The organ-specific models were selected to match the biological context of each dataset—using the pan-cancer model for ICB, the lung model for COVID, and the brain model for ROSMAP. All experiments were conducted using the CTMean architecture without auxiliary prediction, with the scGPT encoder frozen, reflecting a zero-shot evaluation setting.

Table 5. Performance comparison of normalized counts and scGPT across datasets. Values are AUC ± standard deviation.

INPUT	ICB	COVID	ROSMAP
NORMALIZED COUNTS	0.79±0.01	0.84±0.04	0.80±0.03
EMBEDDINGS FROM SCGPT (WHOLE-HUMAN)	0.76±0.04	0.81±0.04	0.71±0.03
EMBEDDINGS FROM SCGPT (ORGAN)	0.73±0.04	0.80±0.04	0.72±0.03

As shown in Table 5, using scGPT embeddings resulted in decreased performance across all datasets. This aligns with prior findings that zero-shot applications of single-cell foundation models can underperform compared to simpler, task-specific models (Kedzierska et al., 2025). Additionally, the whole-human model outperformed the organ-specific models, which may be attributed to the broader training data available for the whole-human variant. These results suggest that fine-tuning foundation models on downstream tasks is likely necessary to fully realize their potential in phenotype prediction.