




Deep learning-based cell type profiles reveal signatures of Alzheimer's disease resilience and resistance

Eloise Berson,^{1,2,3} Amalia Perna,¹ Syed Bukhari,¹ Yeasul Kim,^{2,3,4} Lei Xue,^{2,3,4} David Seong,^{2,5} Samson Mataraso,^{2,3,4} Marc Ghanem,² Alan L. Chang,^{2,3,4} Kathleen S. Montine,¹ C. Dirk Keene,⁶ Maya Kasowski,¹  Nima Aghaeepour^{2,3,4,†} and Thomas J. Montine^{1,†}

[†]These authors contributed equally to this work.

See Dammer (<https://doi.org/10.1093/brain/awaf323>) for a scientific commentary on this article.

Neurological disorders result from the complex and poorly understood contributions of many cell types. It is therefore essential to uncover mechanisms behind these disorders and identify specific therapeutic targets. Single-nucleus technologies have advanced brain disease research, but remain limited by their low nuclear transcriptional coverage, high cost and technical complexity.

To address this, we applied a transformer-based deep learning model that restores cell type-specific investigation transcriptional programs from bulk RNA sequencing, significantly outperforming previous methods. This enables large-scale and cost-effective investigation of cell type-specific transcriptomes in complex and heterogeneous phenotypes such as cognitive resilience or brain resistance to Alzheimer's disease.

Our analysis identified astrocytes as the major cell mediator of Alzheimer's disease resilience across cerebral cortex regions, while excitatory neurons and oligodendrocyte progenitor cells emerged as the major cell mediators of resistance, maintaining synaptic function and preserving neuron health.

Finally, we show that our approach could restore the whole tissue transcriptome, offering an unbiased framework for exploring cell-specific functions beyond single-nucleus data.

1 Department of Pathology, Stanford University, Stanford, CA 94305, USA

2 Department of Anesthesiology, Perioperative and Pain Medicine, Stanford University, Stanford, CA 94305, USA

3 Department of Biomedical Data Science, Stanford University, Stanford, CA 94305, USA

4 Department of Pediatrics—Neonatal and Developmental Medicine, Stanford University, Stanford, CA 94305, USA

5 Immunology Program, Stanford University School of Medicine, Stanford, CA 94305, USA

6 Department of Laboratory Medicine and Pathology, University of Washington School of Medicine, Seattle, WA 98195, USA

Correspondence to: Eloise Berson

Department of Pathology, 300 Pasteur Drive, Stanford University, Stanford, CA 94305, USA

E-mail: eloiseb@stanford.edu

Keywords: machine learning; cell type deconvolution; Alzheimer's disease resilience; Alzheimer's disease resistance

Received February 12, 2025. Revised May 29, 2025. Accepted June 29, 2025. Advance access publication August 5, 2025

© The Author(s) 2025. Published by Oxford University Press on behalf of the Guarantors of Brain.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

Introduction

The human brain has more than 100 billion interconnected, diverse cells that continuously synergize across regions to deliver neural activity.¹ Given the brain's inherent dynamic nature and the relatively small cohort sizes in human neuroscience studies, understanding cell type-specific roles in neurological disorders and identifying robust cell type-specific targets remain challenging.

Single-cell RNA sequencing (RNA-seq) offers a powerful approach to cell-specific transcriptional analyses for those tissues from which single cells can be isolated reliably. While this constraint can be met for many tissue types, the brain presents a unique structure where interacting cells are enmeshed in a dense network of neuronal and glial processes, termed neuropil. Such an intricate structure limits the application of traditional cell dissociation protocols. A widely employed workaround is single-nucleus (sn)RNA-seq, which has enabled the detection of cell type-specific nuclear transcriptomes from brain regions. However, snRNA has major limitations. First, it remains a skilled and cost-demanding technique, often limiting cohort size, which not only hinders the generalization of biological findings but also restricts the exploration of highly heterogeneous phenotypes, such as Alzheimer's disease (AD) and its subtypes.² Second, snRNA-seq only recapitulates ~20%–50% of the cellular transcriptome, excluding comprehensive extranuclear content, including synaptic and transcriptional profiling.³ Finally, snRNA-seq is more prone to technical challenges such as dropout and cell type-specific sensitivity to nuclear dissociation, potentially confounding data analysis and increasing the likelihood of false discoveries.^{3,4}

Disentangling cell type-specific information from bulk tissue sequencing offers an attractive framework for large, robust and low-cost investigation of cell type-specific molecular features in brain disorders.^{5,6} However, current approaches to disentangle bulk RNA-seq data have multiple limitations: they rely on accurate reference data and accurate detection of a limited number of marker genes to predict the underlying cell type-specific expression from bulk tissue sequencing and often lack cross-validation to ensure the method generalizes to unseen samples.^{5,7–9} Transformer-based deep learning models can mitigate several of these limitations with their strong ability to learn prior knowledge from long-sequence data using vast amounts of data.^{10,11} For instance, we previously developed a transformer-based model, Cellformer, that deconvolutes bulk brain epigenetic Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) data into cell type-specific data.¹²

Here, we show that Cellformer can also accurately deconvolute bulk brain transcriptional data (RNA-seq) into cell type-specific RNA profiles, preserving biological variations and phenotypic transcriptional signatures. Trained on more than 3.3 million brain nuclei from different brain regions and different population studies, our model enabled accurate and reference-free cell type-specific transcriptomic profiles of the main cell types in the brain. Notably, this model generalizes across species, unseen data, or unseen phenotype and significantly outperforms previous methods^{5,6} while preserving cell type-specific biological variations. We demonstrate that pre-trained Cellformer can reveal novel potential cell type-specific transcriptional features of complex phenotypes such as resistance and resilience to AD. When applied on RNA-seq from mice brain tissue, we showed that Cellformer retrieved additional cell-type RNA information from tissue compared with nuclei-based approaches. Together, our results recommend Cellformer as a novel computational approach for comprehensively and efficiently surveying cell type-specific

transcriptomes in the brain. It is freely available and can be applied to any bulk RNA-sequence data through a straightforward command-line interface.

Materials and methods

Single-nucleus RNA sequencing

Human cohort

Primary brain samples were obtained at the University of Washington, from post-mortem tissue following informed consent and Institutional Review Boards approval. Clinical diagnosis of cognitive decline and neuropathological assessment of AD neuropathologic change and other neuropathologic comorbidities were determined using current consensus guidelines. Alzheimer's disease dementia (ADD) cases were characterized by cognitive impairment and an overall AD neuropathological change superior to three. Controls were defined as individuals with an overall AD neuropathological change equal to zero and no dementia.

Mouse cohort

All mouse procedures were conducted in accordance with the guidelines of the Institutional Animal Care and Use Committee (IACUCs) at Stanford University. Three-month-old C57BL/6 male mice were injected intraperitoneally (i.p.) with 3 mg/kg of body weight of kainic acid (KA) ($n = 3$). Control mice ($n = 7$) were injected with an equivalent volume of vehicle (0.9% saline solution; NaCl). Mice were observed for 1 h after treatment and sacrificed at 12 h after injection. Mice were transcardially perfused with saline solution, and their brains were rapidly removed. The hippocampi were collected and stored at -80°C prior to nuclei isolation.

Tissue preparation, quality control and data preprocessing for snRNA and bulk RNA-seq are described in the [Supplementary material](#).

Cellformer RNA

Synthetic dataset creation

To create a synthetic pseudo-bulk with known composition, we leveraged annotated total count and log-normalized snRNA-seq samples. For each individual, N synthetic samples were created by: (i) aggregating a random number of cells ranging from 100 to 800 cells per cell type, leading to cell type-specific pseudo-bulk RNA samples which are Cellformer's ground-truths; (ii) aggregating the created cell type-specific pseudo-bulk sample to derive pseudo-bulk samples, mimicking real bulk RNA-seq. The synthetic dataset enables the creation of a large number of paired bulk cell type-specific pseudo-bulk samples to train a deep learning model. Only the common genes, expressed across all the snRNA-seq and bulk RNA-seq human brain datasets were used, resulting in a model input size of 14 719 genes for the pre-trained model and 18 265 genes for the human middle temporal gyrus (MTG) model. Similarly, only the common genes expressed across all the mouse samples were used, resulting in 19 299 genes for the mouse model. To train the pre-trained model, we generated the synthetic dataset, leveraging the total count and log normalized snRNA from publicly available datasets.^{13–18} The model requires a fixed set of genes as input. When tested on bulk RNA-seq data from a novel external dataset, genes required by the model but

not found in the external dataset are set to zero to meet this requirement. The raw count matrix was used. No further batch effect correction was applied.

Model training

Cellformer was evaluated using a 5-fold cross-validation strategy, ensuring that the test set samples came from individuals not included in the training set. The training set was further randomly split into a training (80%) and validation (20%) set to train and optimize the deep learning model. Cellformer is a transformer-based model that leverages the dual-path strategy to compute both local and global attentions. All the hyperparameters for the different models are available (<https://github.com/elo-nsrb/CellformerRNA/>). The models were trained using the Adam optimizer, initialized with a learning rate of 0.001 to minimize the mean square error (MSE) loss. Best iteration and optimal weights were selected using an early-stop algorithm or a maximum of 80 epochs. The training stability was ensured by using gradient clipping to limit the MSE error to five. To evaluate the model, we reported both the mean cross-validated Spearman and Pearson correlations between predicted and ground-truth samples across genes and between predicted and ground-truth gene expression across the whole dataset. A bimodal distribution was observed in the gene Spearman correlation distribution across cell types for the human MTG model (Supplementary Fig. 1A). A threshold of 0.3 was chosen to exclude genes that were non-predictable for downstream analysis, corresponding to the first mode centred around zero. This threshold enables the predictable genes that lie within 99% confidence across cell types to be kept. The number of genes for the human MTG model ranges from 10 344 to 12 800 genes per cell type for this threshold and from 3920 to 8945 for a very conservative threshold of 0.7, while for the human brain model more than 12 500 genes can be predicted across all the cell types with a Spearman correlation of 0.5 (Supplementary Fig. 1B). The list of predictable genes >0.5 for the universal model is available at <https://github.com/elo-nsrb/CellformerRNA>. Once evaluated, a model trained on the whole cohort was derived for downstream analysis.

Model comparison

We compared Cellformer against several methods: a linear regression model, the unsupervised non-negative matrix factorization (NMF) approach commonly used for mixture deconvolution, and two state-of-the-art techniques, CIBERSORTx (high-resolution)⁵ and BayesPrism,⁶ using identical cross-validation splits. To generate fold-specific signature matrices for CIBERSORTx and BayesPrism, we randomly sampled 50 cells per cell type from the training set. During each fold iteration, all these methods were evaluated on 20 individual-specific pseudo-bulk samples randomly selected from the fold's test set. Genes that were not imputed by CIBERSORTx or BayesPrism were removed for testing.

Multivariate analysis

Conventional machine learning (ML) models were employed for multivariate analysis, including separating AD and Control (CTRL) phenotypes, AD and resilience or AD and resistance. The input to the models was either a deconvoluted cell type-specific RNA profile or the cell type-specific snRNA-derived pseudo-bulk profile derived by summing all the cells from the same cell type. A five-repeated 5-fold cross-validation framework, with area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC) criteria, was exploited to reduce

the estimated error bias. While AUROC measures the model's ability to discriminate between two classes, AUPRC evaluates how well the model can identify the positive cases. It is reported here to account for imbalanced classes when classifying ADD, resilient to AD and resistant to AD. The performance of the model with the highest AUROC on the test samples from unseen individuals is reported. All the tested models (least absolute shrinkage and selection operator or LASSO, Ridge, Random Forest and XGboost) were run on Python using *scikit-learn* packages (v.1.4.1) with default hyperparameters.

Univariate and statistical analysis

Pythonic implementation of Deseq2 was leveraged to perform differential expression analysis^{19,20} with default parameters. Other statistical tests were performed using Scipy (v.1.14.0). Figures were drawn using either Seaborn (v.12.0), Matplotlib (v.3.8.3) or Scanpy (v.1.9.8) in Python or Biorender. Gene Ontology (GO) enrichment was performed using GSEAPY (v.1.1.3) using GO ontology databases released in 2023. Only significantly enriched terms ($P < 0.05$) with more than one gene were considered.

Results

Accurate cell-type profiles from human brain bulk RNA-seq

Bulk tissue deconvolution offers a promising solution to overcome the limitations of snRNA-seq and to gain deeper insight into the whole brain cellular transcriptome. We adapted and optimized Cellformer,¹² a transformer-based model that previously showed high performance at accurately deconvoluting bulk ATAC-seq data, for application with bulk RNA-seq. Cellformer relies on synthetic bulk samples with corresponding synthetic ground-truth cell type-specific expression, created by aggregating a random number of single-nucleus data samples for its training and evaluation. While this approach accurately simulates realistic bulk DNA data given that DNA is confined within the cell nucleus, it remains an approximation of bulk tissue RNA, as transcripts are not restricted to the nucleus but are instead distributed throughout brain tissue. To evaluate Cellformer's ability to accurately deconvolute bulk RNA-seq and characterize transcriptional information included in deconvoluted bulk tissue data, we developed a multi-assay experiment illustrated in Fig. 1A.

Tissue samples from 16 middle and temporal gyri from AD ($n = 8$) and healthy control ($n = 8$) individuals were collected and divided into two portions (see 'Materials and methods' section). One half of each bulk tissue sample underwent RNA-seq to assay the bulk tissue RNA transcriptome. The other half of each tissue sample was used for nuclei isolation and sequenced either as snRNA or as bulk nuclei RNA to infer the total RNA content contained in the nuclei only (Fig. 1A). From snRNA data, synthetic bulk RNA-seq samples with known composition and corresponding ground-truth cell type-specific expression were generated, allowing effective training and testing of the model to resolve cell type-specific profiles of the seven main cell types in the brain: astrocytes (AST), excitatory neurons (EXC), endothelial-mural cells (Endo-Mural), inhibitory neurons (INH), microglia (MIC), oligodendrocyte (OLD) and oligodendrocyte progenitor cells (OPCs). A cross-validation scheme was employed to ensure the model's robustness on unseen individual samples (see 'Materials and methods' section). To enhance the correlation between synthetic bulk RNA from snRNA

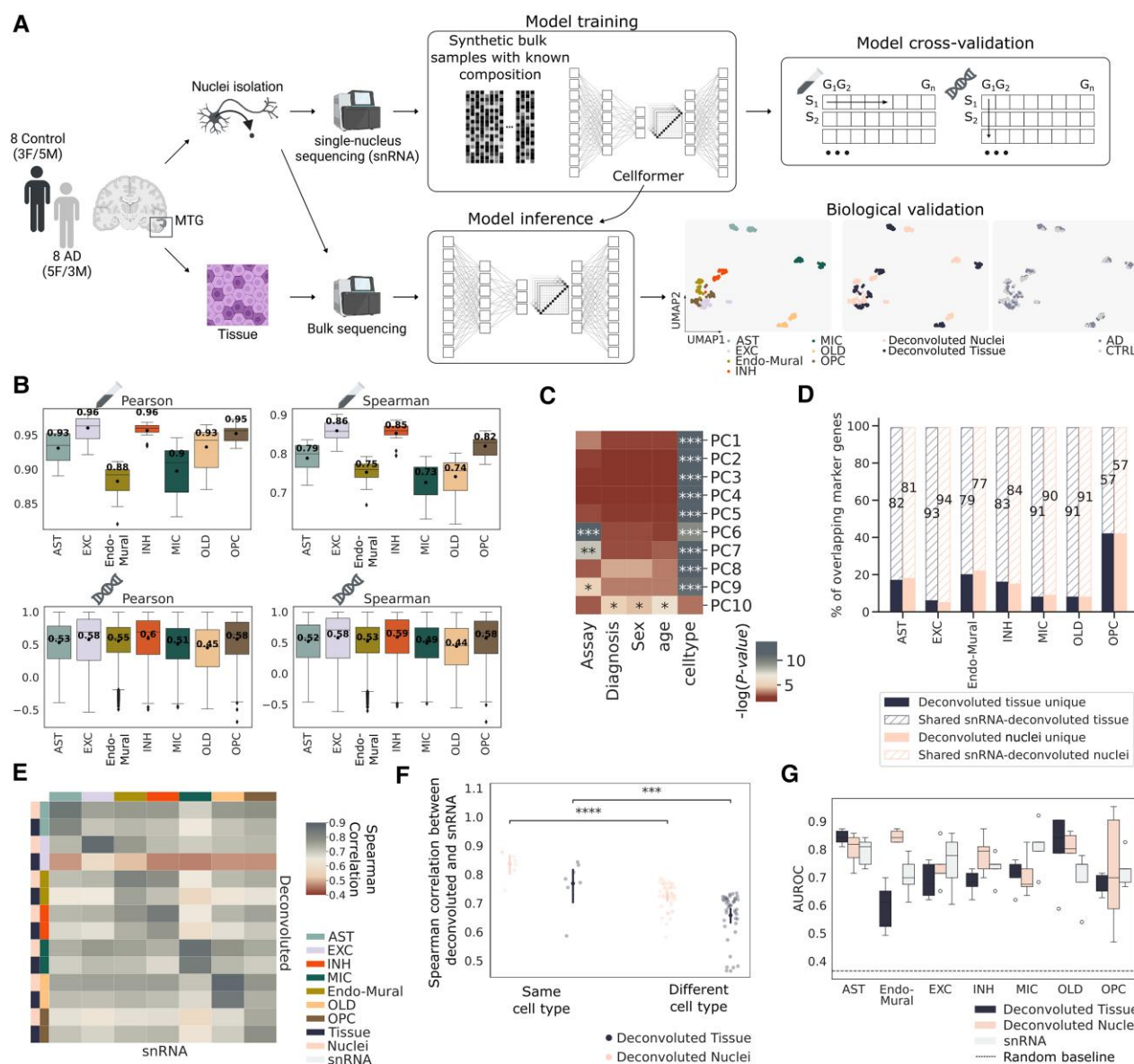


Figure 1 Cellformer accurately recapitulates cell type-specific gene expression in bulk RNA-seq from human brains. (A) Study overview. Middle temporal gyrus (MTG) tissue samples were collected from individuals with Alzheimer's disease (AD; $n = 8$) and healthy control ($n = 8$) subjects (CTRL). One portion of the tissue was processed to extract nuclei, which were then used to run either single-nucleus RNA-seq (snRNA) or bulk (nuclei) RNA-seq. The other portion of the tissue underwent bulk (tissue) RNA-seq. From the snRNA, synthetic bulk samples with known composition and realistic transcriptional profiles were created to train and cross-validate the transformer-based Cellformer model. Once the model was trained, it was used to infer cell type-specific gene expression from both bulk tissue and bulk nuclei RNA-seq. Created in BioRender. Berson, E. (2025) <https://BioRender.com/28q32fr> (B) Cellformer cross-validation performances. Pearson and Spearman correlations were computed between the predicted sample expression profile and ground-truth ($n = 16$) (top) and between predicted ground-truth gene profiles across samples ($n = 18\,265 \times 5$ -fold) (bottom). Created in BioRender. Berson, E. (2025) <https://BioRender.com/28q32fr> (C) Associations between deconvoluted data principal components and biological and technical covariates. P-values are derived using Benjamini-Hochberg adjusted ANOVA test. *** $P < 0.0001$, ** $P < 0.001$, * $P < 0.05$. For visualization purposes, the adjusted P-values, represented in this panel, were capped at 10^{-6} [corresponding to $-\log(P\text{-value})$ cap of 13.82]. (D) Cell type marker genes overlap between deconvoluted data and snRNA. (E) Distribution of the Spearman correlation computed between deconvoluted data and snRNA cell type-specific expression from the same or different cell types. Significantly higher correlations were found between deconvoluted data and snRNA from the same cell type than from different cell types using the Mann-Whitney test. *** $P < 0.001$, **** $P < 0.0001$. (F) Spearman correlation between deconvoluted tissue and nuclei and snRNA. (G) Model classification performance at distinguishing AD versus CTRL using either snRNA, deconvoluted tissue or deconvoluted nuclei data. No significant differences were found between the cross-validated AUROC from the different data types using the Benjamini-Hochberg adjusted Mann-Whitney test. The box plots show the median (centre), 25th and 75th percentiles (box bounds). AST = astrocytes; AUROC = area under the receiver operating characteristic curve; EXC = excitatory neurons; Endo-Mural = endothelial-mural cells; F = female; INH = inhibitory neurons; M = male; MIC = microglia; OLD = oligodendrocytes; OPC = oligodendrocyte progenitor cells; RNA-seq = RNA sequencing.

samples and the corresponding actual bulk RNA and to improve the model's performance, snRNA samples were normalized by total count and log transformed (Supplementary Fig. 1C). Given the relationship profile observed between snRNA-seq pseudo-

bulk and bulk nuclei or tissue profiles, both Pearson and Spearman correlations are reported. Spearman correlations were higher than Pearson correlations in the presence of outliers (Supplementary Fig. 1D).

The model evaluation quantifies the accuracy of predicting a realistic cell type-specific profile using both Pearson and Spearman correlations, using a cross-validation strategy. In addition, we quantified the model's ability to accurately predict cell type-specific gene expression variation by computing the correlation between the predicted and ground-truth gene profile across all samples. For both evaluations, the model accurately predicted the cell type-specific expression for all cross-validation iterations with a mean Pearson correlation superior to 0.88 and a mean Spearman correlation superior to 0.73 across cell types (Fig. 1B). Cross-validation reveals that Cellformer accurately predicted most of the genes across seven main brain cell types with a mean Pearson correlation ranging from 0.45 (in OLD) to 0.6 (in INH) and a mean Spearman correlation ranging from 0.44 to 0.6. Notably, more than 10 000 transcripts per cell type were found highly predictable by Cellformer per cell type, with an average cross-validation Spearman correlation exceeding 0.3. Cellformer enables accurate deconvolution of more than 20% genes per cell type than the previous state-of-the-art method⁶ at different thresholds (Supplementary Fig. 1E). Especially for highly similar cell types such as OLD and OPCs, BayesPrism accurately predicts less than 5000 genes for a threshold of 0.3, which is two times less than Cellformer and less than 50 genes at a more stringent threshold (Supplementary Fig. 1E).

The trained model was then used to infer cell type-specific RNA-seq profiles from both bulk tissue and bulk nuclei RNA-seq data. As expected, cell type was the dominant source of variation in the deconvoluted data from both bulk tissue and bulk nuclei RNA-seq (adjusted ANOVA $P < 0.0001$) (Fig. 1C and Supplementary Fig. 1F). Additional variability observed in the deconvoluted data was significantly associated with covariates such as the assay (bulk tissue or bulk nuclei), sex, diagnosis and age (Fig. 1C), suggesting the model's ability to preserve biological and technical variations inherent in bulk data.

To assess the model's effectiveness in preserving biological information and effectively deconvoluting cell type-specific expression from bulk data, we intersected marker genes identified in the deconvoluted data (bulk tissue and bulk nuclei) with marker genes found using snRNA. The overlap between deconvoluted data markers and snRNA markers ranged from 55% to 94% across cell types (Fig. 1D), demonstrating the model's ability to preserve cell type identities. Furthermore, comparisons between deconvoluted cell type-specific data and snRNA cell type-specific pseudo-bulk data (derived by summing all the cells within the same cell type), revealed significantly higher correlations within the same cell types compared with different cell types (Mann–Whitney $P < 0.001$) (Fig. 1E and F). Notably, deconvoluted cell type-specific data from bulk nuclei samples exhibited a significantly stronger correlation with snRNA pseudo-bulk data than data from bulk tissue RNA-seq ($P < 0.001$; Fig. 1F), suggesting that the model could preserve the differences in RNA content between nuclei and whole tissue.

We then assessed the model's ability to preserve phenotype information. For this, we trained assay-specific and cell type-specific classifiers to distinguish between healthy controls and AD and compared performance across assays. Overall, the classifiers accurately separated AD from healthy controls with a mean cross-validated AUROC ranging from 0.60 to 0.85 and AUPRC ranging from 0.72 to 0.89 across experiments and cell type, which are significantly superior to a random baseline across all the cell types (Fig. 1G and Supplementary Fig. 1G). These results align well with previous studies, implicating the different main cell types in AD.^{13,14,21} No significant differences were found between cross-

validated AUROC derived from a model trained using deconvoluted bulk tissue data (mean AUROC = 0.71 ± 0.11 and mean AUPRC = 0.78 ± 0.09), deconvoluted bulk nuclei data (mean AUROC = 0.79 ± 0.8 and mean AUPRC = 0.81 ± 0.06) or snRNA-derived pseudo-bulk data (mean AUROC = 0.75 ± 0.8 and mean AUPRC = 0.81 ± 0.08) using multi-testing the corrected Mann–Whitney test ($P > 0.05$) across cell types. These results underscore the ability of Cellformer to unveil cell type-specific transcriptional profiles associated with disease or biological function using fast and cost-effective bulk tissue RNA-seq data.

We noticed better performance with nuclei-based RNA-seq data than with deconvoluted tissue data. We therefore sought to disentangle the sources of variability in model predictions to determine the relative contributions to observed differences between deconvoluted nuclei and deconvoluted tissue data from inherent differences in RNA information between the two modalities compared with that from other sources such as computational errors or technical factors. To address this, we evaluated the model's robustness to: (i) dropout noise, simulating technical variability such as nuclei preparation or RNA integrity (RIN); and (ii) Gaussian noise, simulating random computational noise. Our results show that potential computational errors or technical variability have a relatively small impact on model performance. Additionally, we observed that RNA quality correlates with model performance, while the model remains robust to random computational noise (Supplementary Fig. 1H).

For validation in another tissue type, we applied Cellformer to peripheral blood mononuclear cells (PBMC) using data from different snRNA probing methods (Supplementary Fig. 2A and B). Similar results were observed for PBMC data sequenced using 10x Genomics, with a mean Pearson correlation of 0.93 and a mean Spearman correlation of 0.69. Slightly lower performance was noted with other sequencing technologies—particularly Seq-Well—which may partly explain the overall decrease in correlation observed in PBMC samples compared with brain samples.

A generalizable model for accurate cell type-specific analysis of the human brain

The current deconvolution method performance largely depends on the availability of accurate reference data. Advances in artificial intelligence have shown the power of large transformer models to generalize to unseen contexts. Combined with the exponential amount of available snRNA data, they offer an unprecedented opportunity for overcoming the shortcomings of current deconvolution methods.

After confirming and validating our model, we next built a generalizable Cellformer model, trained on more than 3.3 million nuclei collected from different human brain regions and therefore coming from patients from different studies (Fig. 2A). Cross-validation performances demonstrate the model's ability to generalize across studies and samples, outperforming the single-dataset model (Fig. 2B and see Fig. 1B). Using the same cross-validation scheme, the generalizable Cellformer model significantly outperforms previous state-of-the-art deconvolution methods and other machine learning-based approaches used for RNA-seq deconvolution (Mann–Whitney test $P < 0.001$) (Fig. 2C).^{5,6,22,23} Notably, Cellformer more accurately preserves gene variation across samples, crucial for real-world applications.

We observed that gene-wise correlations across samples were consistently lower than sample-wise correlations across genes. To assess Cellformer's robustness to variability in cell type

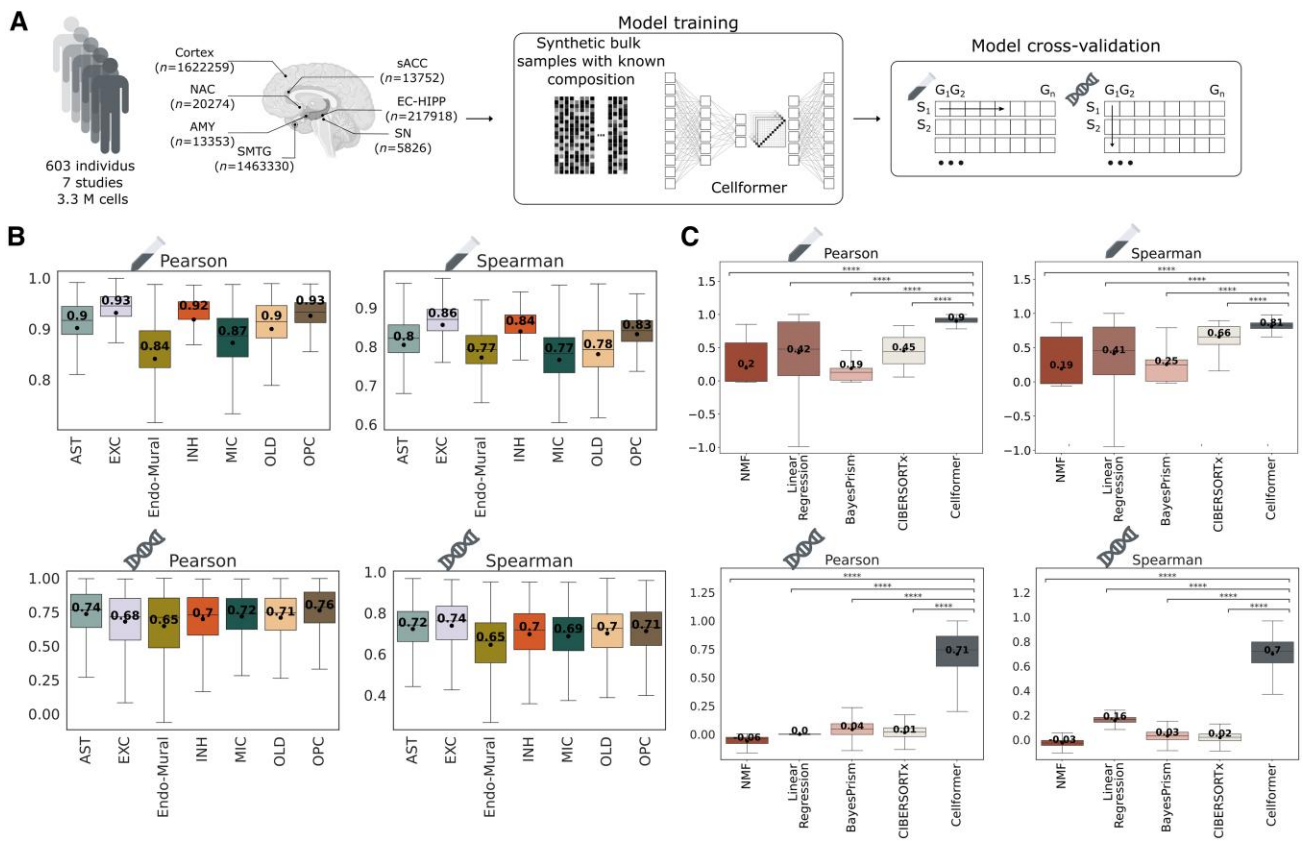


Figure 2 The generalizable human brain deconvolution model accurately resolves cell type-specific expression profiles, outperforming previous baselines. (A) Cellformer was trained on synthetic bulk expression generated from more than 3.3 million cells, from 603 individuals collected in seven studies and validated using a cross-validation strategy to ensure model generalization to unseen individuals' samples. Created in BioRender. Berson, E. (2025) <https://BioRender.com/28q32fr>. (B) Model cross-validation performances: sample-wise correlations per cell type ($n = 566$) (top) and gene-wise correlations per cell type ($n = 14\,712 \times 5$ -fold) (bottom). (C) Cellformer significantly outperforms other baseline models, including BayesPrism⁶ and CIBERSORTx,⁵ using a cross-validation strategy using Mann-Whitney test. *** $P < 0.001$, **** $P < 0.0001$. Created in BioRender. Berson, E. (2025) <https://BioRender.com/28q32fr>. AST = astrocytes; EXC = excitatory neurons; Endo-Mural = endothelial-mural cells; INH = inhibitory neurons; M = million; MIC = microglia; NMF = non-negative matrix factorization; OLD = oligodendrocytes; OPC = oligodendrocyte progenitor cells.

proportions, a common source of sample-to-sample variation, we compared the cell type-specific variation in Spearman correlations within deconvoluted pseudo-bulk samples (generated from the same individuals but differing in cell type proportions) to the cell type-specific total variation observed across cell types. We observed that cell type proportion effect ranges from none to 0.2 on the Spearman correlation between predicted and ground-truth cell type-specific expression. However, the within-sample variation due to cell type proportion change was significantly lower than the overall variation across cell types (Wilcoxon test, $P < 0.05$). Similarly, we compared the cell type-specific variation in gene-wise Spearman correlations across pseudo-bulk samples generated from different individuals with varying cell type proportions to the total cell type-specific gene-wise variation. Again, we observed that the within-individual gene-wise variation was significantly lower than the total across cell types (Wilcoxon test, $P < 0.05$) (Supplementary Fig. 2C). These results indicate that while changes in cell type proportion contribute to sample-to-sample variation, the majority of variability in correlation likely arises from intrinsic technical noise in single-cell data rather than compositional differences alone. We then investigated the effect of low expressed genes on gene-wise correlation and found that gene-wise correlation using only highly expressed genes leads to similar results as sample-wise correlation with mean Pearson

correlation superior to 0.83 and a mean Spearman correlation superior to 0.78 (Supplementary Fig. 2D).

Next, we assessed the model's ability to generalize to real-world scenarios by applying it to predict cell type-specific expressions from unseen phenotypes and previously unused datasets. First, we evaluated the model's ability to accurately predict realistic cell type expression and gene profile in samples with a phenotype unseen by the model. For this, a model trained on samples from control only and tested on samples from individuals with AD dementia was compared with a cross-validated model trained on samples from people both control and AD. Importantly, the model tested on samples from an unseen phenotype performs as well as a model trained using data from both phenotypes (Fig. 3A). Similarly, we compared a model trained using data from five datasets and tested on data from a sixth dataset to a cross-validated model trained with all six datasets. We observed that the model tested on an unseen dataset also performs as well as a model trained using all the datasets (Fig. 3B). These results suggest a high capacity to generalize the model to real-world scenarios.

Beyond its validation using synthetic data, we also evaluated the model's ability to deconvolute bulk RNA-seq from five independent studies across various brain regions.^{24–28} Unsupervised dimension reduction of deconvoluted data using uniform manifold approximation and projection (UMAP) provides visual evidence that

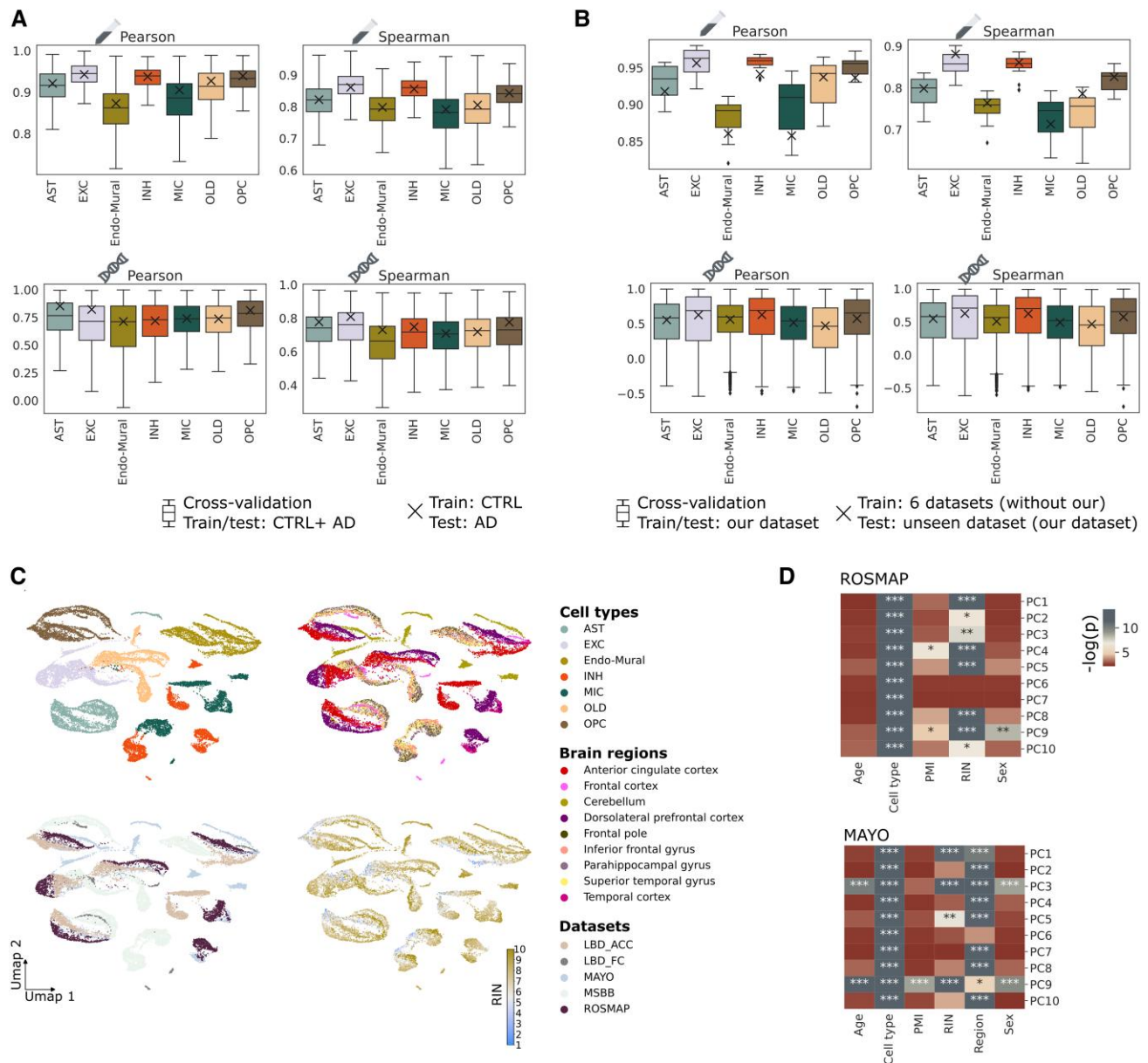


Figure 3 Cellformer model effectively generalizes on bulk RNA from unseen patients and unseen phenotypes. (A) Zero-shot model testing trained with transcriptomic data from control individuals (CTRL) and tested on transcriptomic data from individuals with Alzheimer's disease (AD). Both Spearman and Pearson correlation performances fall within the cross-validated performances of a model trained on both CTRL and AD samples. (B) Zero-shot model testing trained with transcriptomic data from 6/7 datasets and tested the left-out dataset transcriptomic data. Both Spearman and Pearson correlation performances fall within the cross-validated performances of a model trained on all seven datasets. (C) The UMAP represents the deconvoluted data by the generalizable model fed with bulk RNA-seq from five datasets from various brain regions coloured by cell type, brain region, study and RNA integrity number (RIN) score. Five datasets include bulk RNA-seq from individuals with AD and Lewy body disease (LBD): from anterior cingulate cortex (LBD_ACC),²⁴ from frontal cortex (LBD_FC),²⁵ from the Religious Order Study and Memory and Aging Project (ROSMAP),²⁶ the Mount Sinai Brain Bank (MSBB)²⁷ and the Mayo Clinic Alzheimer's Disease Genetics Studies (MAYO).²⁸ (D) Association between deconvoluted data principal components from ROSMAP (top) and MAYO (bottom) and known confounders. P-values were derived using the Benjamini–Hochberg adjusted ANOVA test. * $P < 0.05$, ** $P < 0.001$, *** $P < 0.0001$. For visualization purposes, P-values were capped at 10^6 . AST = astrocytes; EXC = excitatory neurons; Endo-Mural = endothelial-mural cells; INH = inhibitory neurons; M = million; MIC = microglia; NMF = non-negative matrix factorization; OLD = oligodendrocytes; OPC = oligodendrocyte progenitor cells; PMI = post-mortem interval; RNA-seq = RNA sequencing; UMAP = Uniform Manifold Approximation and Projection.

deconvoluted cell type-specific data retained the inherent biological and technical variations present in the original bulk RNA-seq while separating the cell type information (Fig. 3C). Interestingly, the model appears sensitive to data quality and batch effect (Fig. 3C and D and Supplementary Fig. 3E). Variance analysis of two datasets reveals that the main source of variance in deconvoluted data is not only associated with cell type and brain region, as expected, but also RNA

integrity number (RIN) ($P < 0.001$). Smaller portions of variance are significantly associated to sex and age ($P < 0.001$) (Fig. 3D). A reduced portion of variance is associated with the post-mortem interval (PMI) ($P < 0.05$). These results highlight the high fidelity of the model in deconvoluting independent bulk tissue RNA-seq data. They also demonstrate the importance of high RNA-seq quality to ensure the preservation of cell type-specific transcriptomic signatures.

Table 1 Neuropathological and clinical criteria used to identify Alzheimer’s disease-resilient, -resistant and -matched Alzheimer’s disease dementia individuals across three publicly available studies

	Age	NFT-Braak	CERAD	Thal	LBD	Cognitive imp. (AD only)
Resilience	>80	>4	Moderate-high	Moderate-high	No	No
Resistance	>80	<3	Low-sparse	Low-sparse	No	No
ADD	>80	>4	Moderate-high	Moderate-high	No	Yes

AD = Alzheimer’s disease; ADD = Alzheimer’s disease dementia; CERAD = Consortium to Establish a Registry for Alzheimer’s Disease; LBD = Lewy body disease; NFT = neurofibrillary tangles.

Table 2 Cohort information

	Resilience (F/M)	Resistance (F/M)	ADD (F/M)	Brain region	Assay
ROSMAP	9 (4/5)	27 (15/12)	52 (34/18)	Dorsolateral prefrontal cortex	Bulk
MAYO	–	8 (3/5)	39 (24/15)	Temporal Cortex	Bulk
LBD_ACC	–	11 (5/6)	25 (13/12)	Anterior cingulate cortex	Bulk
SEA-AD	15 (9/4)	3 (0/3)	26 (16/10)	Dorsolateral prefrontal cortex	snRNA

ACC = anterior cingulate cortex; ADD = Alzheimer’s disease dementia; F = female; LBD = Lewy body disease; M = male; MAYO = Mayo Clinic Alzheimer’s Disease Genetics Studies; NFT = neurofibrillary tangles; SEA-AD = Seattle Alzheimer’s disease Brain Cell Atlas; ROSMAP = Religious Orfer Study and Memory and Aging Project; snRNA = single-nucleus RNA.

Cellformer unveils robust cell type-specific role in AD resistance and resilience

Cell type specificity in complex and heterogeneous phenotypes such as resistance to AD and resilience to ADD, two well-described but uncommon clinico-pathologic states that underscore the potential to prevent and treat AD,^{29,30} requires large, meticulously annotated datasets to obtain an accurate molecular profile of the underlying biological processes, which may not be achievable when analysing small datasets independently.

To overcome these challenges, we merged bulk RNA data from three datasets after filtering cases using stringent established age, neuropathological and clinical criteria^{2,31} (Tables 1 and 2). Merging datasets helps extract more robust and generalizable targets, leading to higher classification accuracies,³² increases statistical power, allowing for the detection of subtle biological differences and limits systematic technical bias.^{33,34} Leveraging Cellformer-derived cell type-specific samples from 46 individuals resistant to ADD (normal cognitive function and no/minimal AD neuropathological hallmarks) and 116 ADD cases (severe cognitive impairment and high AD neuropathological hallmarks), we applied multivariate and univariate analysis to identify cell type-specific signatures of these different groups; multivariate analysis suggests that we can detect differences between the AD resistance and ADD in all the cell types with AUROC > 0.7 and AUPRC > 0.5, outperforming a random baseline (Fig. 4A). Differential expression analysis reveals that the main cell types contributing to AD resistance are AST and then OLD (Fig. 4B). Gene ontology analysis exhibits enrichment in ADD of processes associated with cellular response to cytokine stimulus in AST, lipid-mediated signalling in MIC, histone acetylation in INH and tau binding in OPC (Fig. 4C). This analysis reveals enrichment in AD resistance of processes associated with mRNA binding in AST, Endosome in EXC, cytoplasmic translation in Endo-Mural, mitochondrial function in INH, Exosome and SMC5-SMC6 Complex in OLD (Fig. 4C).

Using only cases in ROSMAP, we next compared AD resilience (high cognitive function) with ADD, with both groups matched for high AD neuropathological hallmarks. To capture the strongest

difference between ADD and AD resilience, we applied stringent neuropathological and clinical criteria to define AD resilience as previously described,^{2,31} therefore discarding cases with intermediate neuropathologic changes or mild cognitive impairment. Aligning with a recent study, most of the differences between ADD and AD resilience were found in cerebral cortical AST with 16 genes (Fig. 4D and E).²¹ Among the 16 AST-specific targets, 63% (10/16) were also found differentially upregulated in AST in AD resilience in an external snRNA dataset, including TMSB4X, MPHOSPH8, TUBB4A, BASP1, RUNDC3A, DSTN, NUDC, RPL7, ENC1 and RPL13A (Fig. 4F). Interestingly, protein phosphatase 3 catalytic subunit alpha (PPP3CA) encoding the protein calcineurin A, which plays a crucial role in synaptic plasticity by regulating the activity of N-methyl-D-aspartate receptor channel gating,^{19,20} was found upregulated in AD resilience in EXC. PPP3CA was also found significantly differentially expressed in most of the cerebral cortical excitatory neuron subtypes in the external validation snRNA SEA-AD dataset (Supplementary Fig. 3A). Univariate does not account for the interaction between genes, we therefore also quantified differences between the two groups using multivariate analysis. Cross-validated multivariate analysis suggests that not only AST but also EXC, Endo-Mural, MIC and INH might contribute to AD resilience (Supplementary Fig. 3B).

Finally, we compared AD resistant (low/minimal AD neuropathologic hallmarks) and AD resilient (high AD neuropathologic hallmarks), with both groups matched for high cognitive function. We found that most univariate differences were in OPC and EXC (Fig. 4G). Retinoic acid receptor responder 2, RARRES2, encoding chemerin, an adipokine known to regulate adipogenesis and adipocyte lipid metabolism³⁵ and potential tumour development regulator in numerous cancers^{36,37} shows significant upregulation in MIC in AD resistance. A correlation network of the 56 AD resistance-upregulated genes in EXC is shown (Fig. 4H). Among them, NRGN (neurogranin), CAMK2B (calcium/calmodulin-dependent protein kinase II beta), GRIA1 (glutamate ionotropic receptor AMPA type subunit 1), PFN1 (profilin 1) could contribute to maintaining synaptic function and MC4R (melanocortin 4 receptor), IPO13 (importin 13), STARD7 (star

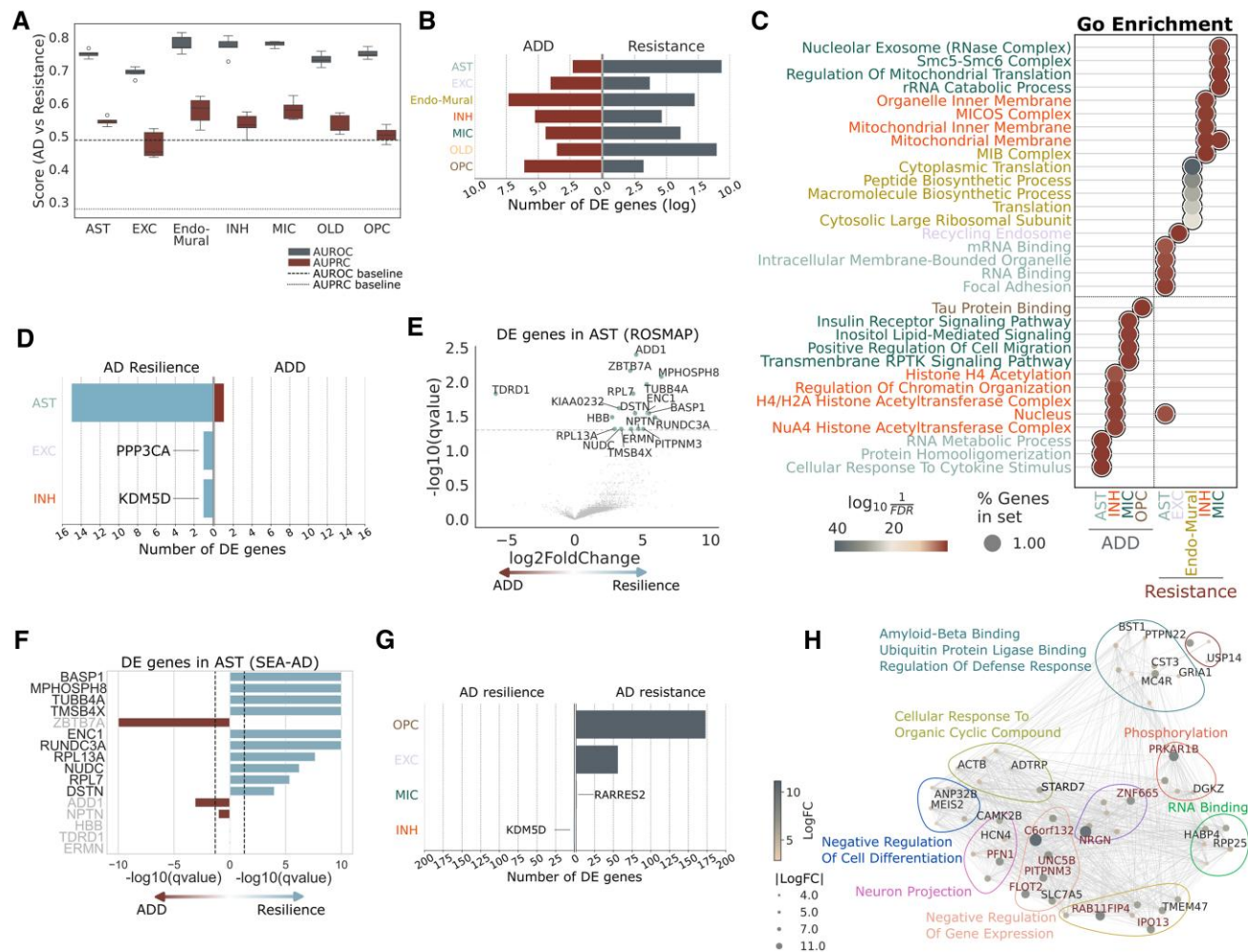


Figure 4 The generalist Cellformer model identified putative glial-specific candidates of Alzheimer's disease resilience and resistance. (A) Cross-validation model performance using three datasets to separate Alzheimer's disease dementia (ADD) and Alzheimer's disease (AD)-resistant individuals. The box plots show the median (centre), 25th and 75th percentiles (box bounds). (B) Number of reproducible differentially expressed (DE) genes between ADD and AD-resistant individuals across three datasets using sex and study-adjusted Deseq2 algorithm. Created in BioRender. Berson, E. (2025) <https://BioRender.com/28q32fr>. (C) Gene ontology (GO) enrichment applied on reproducible upregulated and downregulated DE genes between AD-resistant and AD individuals. Only the top 500 genes were used per cell type with a minimum of 0.25 absolute fold-change. (D) Number of reproducible DE genes between ADD and AD resilient individuals found in the Religious Order Study and Memory and Aging Project (ROSMAP) using sex-adjusted Deseq2 algorithm. (E) Volcano plot representing astrocyte-specific DE genes between ADD and AD-resilient individuals in deconvoluted data from the ROSMAP dataset. Sixteen targets were found upregulated in AD-resilient individuals compared with AD individuals. (F) Benjamini-Hochberg adjusted q-values of 16 putative resilient-specific targets in astrocytes using multi-testing corrected Wilcoxon's test applied on all genes detected in single-nucleus RNA (snRNA) from astrocytes found in the Seattle Alzheimer's Disease Brain Cell Atlas (SEA-AD) dorsolateral prefrontal cortex dataset. Ten targets were also found upregulated in astrocytes in AD-resilient individuals compared with ADD individuals in snRNA from the SEA-AD study. (G) Number of reproducible DE genes between AD-resilient and AD-resistant individuals found in ROSMAP using sex-adjusted Deseq2 algorithm. (H) Correlation networks of the 56 genes upregulated in AD-resistant compared with AD-resilient individuals in excitatory neurons (EXC). Node size and node colours were determined using the log fold-change in resistant and resilient comparison. Edges correspond to absolute Spearman correlation between nodes. Top 10 genes are annotated in red. Leiden clustering was used to cluster the nodes. Clusters were annotated using GO Enrichment. Other genes involved in GO processes are annotated in black. AST = astrocytes; Endo-Mural = endothelial-mural cells; INH = inhibitory neurons; MIC = microglia; OLD = oligodendrocytes; OPC = oligodendrocyte progenitor cells.

related lipid transfer domain containing 7), *SLC7A5* (solute carrier family 7 member 5), *USP14* (ubiquitin specific peptidase 14), *TMEM47* (transmembrane protein 47) could help to preserve neuronal health and function. GO enrichment identifies groups of genes involved in neuron projection, amyloid-beta binding, phosphorylation, regulation of defence response, gene expression regulation and RNA binding processes (Fig. 4H). Cross-validated multivariate analysis suggests the most significant transcriptional signals to separate AD resilience from AD resistance are in AST and EXC (Supplementary Fig. 3C).

Interestingly, lysine demethylase 5D (*KDM5D*), located on the Y chromosome was found to be significantly upregulated in AD resilience in comparison to both AD resistance and ADD, suggesting potential sex differences in AD resilience.³⁸

While Cellformer does not output cell proportions, we investigated potential changes in cell type proportions between ADD, AD resilience and AD resistance by comparing the relative content proportion differences among the different groups (Supplementary Fig. 4). Aligning with previous work, we found a significant relative increase in MIC (corrected Mann-Whitney $P < 10^{-5}$), a significant

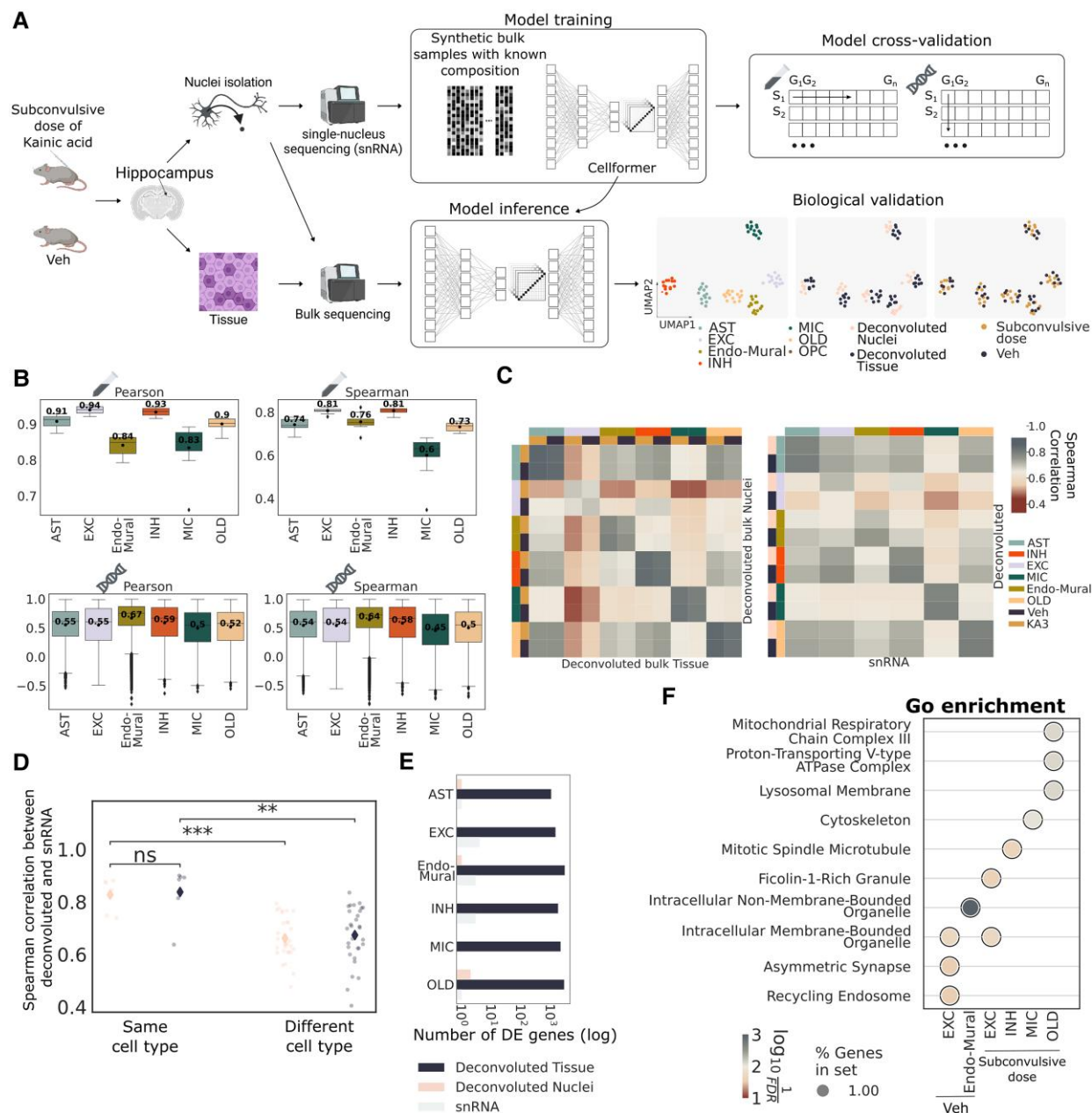


Figure 5 Cellformer could resolve extranuclear transcriptional information involved in low-induced neurotoxicity. (A) Study overview. Hippocampal tissue was collected from healthy ($n = 7$) and mice injected with a subconvulsive dose of kainic acid ($n = 4$). One tissue portion was processed to extract nuclei, which were then used to run either single-nucleus RNA-seq (snRNA) or bulk (nuclei) RNA-seq. The other portion underwent bulk (tissue) RNA-seq. From the snRNA, synthetic bulk samples with known composition and realistic transcriptional profiles were created to train and cross-validate the transformer-based Cellformer model. Once the model was trained, it was used to infer cell type-specific gene expression from both bulk tissue and bulk nuclei RNA-seq. Created in BioRender. Berson, E. (2025) <https://BioRender.com/28q32fr>. (B) Cellformer cross-validation performances applied to mouse bulk RNA-seq data: sample-wise correlations per cell type ($n = 10$) (top) and gene-wise correlations per cell type ($n = 17\,265 \times 5$ -fold). Created in BioRender. Berson, E. (2025) <https://BioRender.com/28q32fr>. (C) Correlation matrix between deconvoluted bulk tissue and nuclei (left) and deconvoluted and snRNA pseudo-bulk data. (D) Spearman correlation distribution computed between deconvoluted data and snRNA cell type-specific expression from the same or different cell types. Significantly higher correlations were found between deconvoluted data and snRNA from the same cell type than from different cell types using the Mann-Whitney test. $***: P < 0.001$. (E) Number of differentially expressed (DE) genes from healthy ($n = 4$) and low-dose injected ($n = 4$) mice across different modalities with an absolute fold change superior to one and multi-testing corrected $\text{Deseq2 } P < 0.05$. (F) Top three cell-type specific Gene Ontology Enrichment applied to upregulated DE genes in subconvulsively dosed mice and upregulated DE genes in vehicle mice with an absolute fold change superior to one. AST = astrocytes; EXC = excitatory neurons; Endo-Mural = endothelial-mural cells; INH = inhibitory neurons; MIC = microglia; OLD = oligodendrocytes; OPC = oligodendrocyte progenitor cells.

relative decrease in other glial cells (corrected Mann-Whitney $P < 10^{-5}$), a predicted relative increase in neuronal cells, and a predicted relative decrease in endothelial cells in ADD compared with AD resistance.^{39,40} A similar trend was observed for AD resilience;

however, our analysis reveals no significant difference in the relative proportion of MIC between AD resilience and ADD, along with a significant decrease in endothelial cells in AD resilience compared with ADD (corrected Mann-Whitney $P < 10^{-5}$).

Resolving cell type-specific extranuclear transcriptional profiles using Cellformer

To better investigate the ability of our model to accurately recapitulate cell type-specific gene expression in tissue, we performed a similar experiment using RNA data from ideally preserved (PMI = 0) tissue from healthy mice and mice with neuronal stress from kainic acid exposure (Fig. 5A). These conditions ensure a maximal preservation of RNA integrity. Cellformer accurately predicted the six main brain cell type-specific RNA profiles in mice, achieving a mean Pearson correlation greater than 0.88 and a mean Spearman correlation greater than 0.77 (Fig. 5B). The mean Pearson correlation between predicted and ground-truth gene profiles across samples ranged from 0.44 to 0.60 (Fig. 5B), similar to the model trained on human data. Correlation matrices between bulk deconvoluted nuclei and tissue, as well as between deconvoluted and snRNA-derived pseudo-bulk data, displayed a similar profile (Fig. 5C). A significantly higher correlation coefficient was observed when comparing the same cell type versus different cell types between deconvoluted and snRNA-derived pseudo-bulk data ($P < 0.001$), suggesting that the model's output accurately restores the inherent cell type-specific signature in its cell type-specific profile (Fig. 5D).

We performed differential expression analysis by comparing cell type-specific expression from healthy mice and mice exposed to a sub-convulsive dose of kainate sufficient to produce excitotoxic stress with minimal neuron death,⁴¹ and compared three different modalities: deconvoluted tissue, deconvoluted nuclei and snRNA-derived pseudo-bulk data. Despite the high similarity in cell type-specific profiles and marker genes between deconvoluted tissue data and both deconvoluted nuclei and snRNA data (Fig. 5C, Fig. S5), we observed 100 times more differentially expressed (DE) genes in the deconvoluted tissue compared with nuclei-based RNA data (Fig. 5E). Gene ontology enrichment analysis was applied on significantly DE genes in kainate-exposed mice ($P < 0.05$) in deconvoluted bulk tissue, but not in nucleus probing assays. Excitatory neurons in kainate-exposed mice showed downregulation of 20 genes associated with asymmetric synapses and 19 genes associated with recycling endosome processes. In MIC, this analysis also revealed 71 genes associated with cytoskeleton processes upregulated in kainate-exposed mice, notably including the immune-related genes *SPPL2B* found to be involved in AD pathology⁴² and *BIN1* a widely replicated AD genetic risk candidate and key regulator of brain inflammatory response.⁴³ OLD showed upregulation of 62 genes associated with lysosomal membrane processes (Fig. 5F). These analyses suggest that deconvoluting bulk tissue captures additional transcriptomic signals, potentially extranuclear information, transparent to nucleus-based RNA-seq data.

Discussion

Determining accurate and robust cell type-specific molecular signatures for diseases of the brain is crucial for understanding disease mechanisms and improving the specificity of potential therapeutic targets. However, the cost and complexity of current technologies prevent large-scale analysis of intricate or rare phenotypes. Hence, we established a thoroughly validated framework that enables low-cost, more comprehensive, and robust identification of cell type-specific transcriptional changes in human AD subtypes and mouse models of neurodegeneration. Our approach leverages an advanced transformer-based model and the large amount of available snRNA-seq from the brain to derive a generalizable model, suitable for real-world scenarios without any reference data. Our approach enables rapid inference of cell type-specific transcriptomic

profiles. It outperforms previous methods,^{5,6} preserves cell type-specific biological variations and increases the transcriptional signal by up to two orders of magnitude over snRNA-seq in ideally acquired samples. Applied to large available datasets, Cellformer unveiled novel potential cell type-specific targets discriminative of complex phenotypes such as AD resilience and resistance. Finally, we demonstrated that Cellformer effectively generalizes to different species and different tissues.

The intricate association between ageing, development of neuropathologic hallmarks and clinical symptoms of AD remains poorly understood. Although strong correlations exist for most individuals on the AD continuum, there are well-described subsets of individuals who exhibit no or minimal neuropathological hallmarks despite elevated risk of disease (resistant to AD) or maintain high cognitive function despite high levels of pathological hallmarks (resilient to AD dementia or ADD).^{2,44,45} Understanding the mechanisms that suppress neuropathological hallmark development despite elevated risk or preserve cognitive function in the presence of extensive neuropathologic hallmarks is crucial to gaining therapeutic insights into preventing or treating AD.

We, therefore, applied our approach to infer cell type-specific transcriptomic features of AD resistance and AD resilience. The definitions used to classify AD resistance and AD resilience vary,⁴⁶ limiting the comparison to other studies that also sought their molecular signatures; we use stringent criteria,^{12,47} which has the advantage of enriching for more extreme phenotypes but the disadvantage of limiting sample size by excluding more marginal cases. With this in mind, in previous work using a single cohort, we determined epigenetic and proteomic features of AD resilience, defined as we have done here, in multiple brain regions;^{12,47} there were too few resistant cases in this single cohort for robust investigation. Our proteomic work on bulk tissue highlighted changes in synaptic and axonal proteins, especially in the hippocampus, and glial injury response in AD resilience.⁴⁷ Application of Cellformer to ATAC-seq bulk tissue data from the same samples inferred AD resilience cell type-specific changes in open chromatin regions that were most common in EXC, again related strongly to genes that encode or regulate components of excitatory neurotransmission, and in MIC; epigenetic features of AD resilience again were strongest in the hippocampus.¹² It is important to note that the multi-site data used in the current study included only regions of the cerebral cortex and not hippocampus. Here, application of Cellformer to transcriptomic data reinforced that the predominant molecular features of AD resilience are expressed by EXC synaptic apparatus and glia, now highlighting a larger contribution of cerebral cortical AST transcripts with a smaller signal from MIC. Our larger multi-site data enables our first assessment of the molecular features of AD resistance. Cellformer inferred transcriptomic differences in cerebral cortical regions between AD resistance and ADD or AD resilience in a multivariate analysis that again focused on EXC synaptic apparatus and AST response. In aggregate, these multi-modal, multi-site, multi-region data indicate that molecular features of AD resistance and AD resilience share a focus on the regulation of excitatory neurotransmission in the hippocampus and cerebral cortex, which is intriguing since one of the three approved treatments for symptomatic AD targets one component of excitatory neurotransmission.⁴⁸ The picture for glia is more complex, with AD resilience predominated by hippocampal MIC and AD resistance by cerebral cortical AST responses.

We showed that Cellformer can generalize to other tissues, including PBMC. However, the model was built on a small number of samples. As with the brain, we expect to improve performance

on PBMC cell types by incorporating additional publicly available scRNA-seq data. Future work will focus on developing a more generalizable model for blood cells.

While Cellformer does not directly predict bulk cell type proportions,⁴⁹ we investigated the changes in relative cell type-specific RNA content between ADD, AD resilience and AD resistance. Previous studies investigating brain cell type deconvolution in Alzheimer's disease compared with controls reported a relative decrease in endothelial cells and a relative increase in neuronal cells and MIC in ADD.^{39,40} Similar trends were observed here between ADD and AD resistance, underscoring the value of Cellformer in investigating cell type-specific content in the human brain. Interestingly, the major shifts in cell type proportions between AD resilience and AD resistance were found in MIC and endothelial cells. Notably, no significant difference was found in the relative proportion of MIC between AD resilience and ADD. A significant decrease in endothelial cells was observed in AD resilience compared with both ADD and AD resistance. Previous studies across different brain regions had reported no significant differences in MIC and endothelial cell proportions between AD resilience and ADD.^{38,50} The discrepancies in endothelial cells may stem from the more stringent criteria used here to define AD resilience. These results underscore the differences in changes to cellular composition and potential alterations in molecular function within the human brain, reinforcing the importance of this approach.

Our approach has limitations, the main one being the trade-off between cell type resolution and computation cost. A more fine-grained brain cell typing, resolving RNA profiles of different subtypes of neurons, incurs a higher computational cost using our current neural network architecture. Tensor parallelism⁵¹ or advanced transformer layers⁵² will be investigated to preserve the model performance while reducing the memory cost. Another limitation of our generalizable model is the lack of several brain regions in the training set. Large consortia such as the Seattle Alzheimer's Disease Brain Cell Atlas (SEA-AD) and the Religious Orders Study and Memory and Aging Project (ROSMAP) will help to fill this gap. Additionally, Cellformer outputs cell type-specific whole tissue pseudo-bulk data. While pseudo-bulk prevents a fine-grained cell state analysis, it is the recommended approach for single-cell analysis to reduce false discovery rate in cell type-specific molecular feature identification or response to perturbations.⁴ The deconvoluted pseudo-bulk data, including information from the whole tissue, were compared with a ground-truth derived from snRNA-seq but not from single cells, as recommended for brain tissue.³ Unlike deconvoluted data, snRNA-seq does not capture cytoplasmic RNA content, potentially introducing a bias in both model training and evaluation. This discrepancy could explain the decreased performance of the classifier trained using deconvoluted data of the neuronal cell types, as expected for the most affected cell type. The interaction between assays and cell types was preserved or even increased after removing Endo and neuronal cell types in Fig. 1C—the ones most affected by discrepancies between bulk tissue and nucleus-based assays in the human MTG model. This may be explained by the fact that the remaining cells are glial, which are more transcriptionally similar to each other, and thus less variability is attributed to cell type differences. This discrepancy may also account for the lower correlation and reduced shared information content observed between bulk tissue and nuclei-based assays in these cell types (see Fig. 1C and G and Fig. 5C). Interestingly, this discrepancy between RNA content in the nucleus and in the cytoplasm or dendroplasm^{53,54} accounts for more variability than computational errors or technical dropout (see Supplementary Fig. 1H).

This suggests that Cellformer could handle zero imputation on novel bulk samples. More advanced imputation strategies⁵⁵ will be explored in the future to better recover missing gene expression.

Cellformer shows very high sample-level correlation across genes, suggesting that it accurately predicts realistic cell type profiles. In contrast, we observed lower gene-wise correlation across samples. When focusing only on highly expressed genes, gene-wise correlation improves, indicating that low expressed genes—less reliably detected across individuals—tend to be less predictable and contribute to the overall decrease in gene-wise correlation. For future applications, we recommend focusing on genes that can be predicted with high confidence for downstream analyses (Supplementary Fig. 1A).

Cellformer performance can vary up to 0.2 Spearman correlation with extreme cell type proportion change (Supplementary Fig. 2C). Notably, we observed a reduction in downstream prediction performance for rare cell types, such as Endo-mural cells, which account for only 2%–4% of brain cells⁵⁶ (Fig. 1C). Similarly, we observed reduced marker gene overlap with single-cell data for OPCs, representing less than 10% of brain cells⁵⁷ (see Fig. 1D). The higher abundance on neuronal cell type in grey matter potentially leads to more robust expression estimates of the deconvoluted cell type expression. Tissue heterogeneity thus plays a critical role, e.g. if more white matter had been sampled, we would expect improved performance in glial cell types and potentially reduced performance in neurons. We anticipate this sensitivity to decrease with larger training datasets, as both gene-wise and sample-wise prediction performance increase with more data (see Fig. 2B and Supplementary Fig. 1A).

Finally, our analysis of AD resistance and resilience is limited by the low number of samples from stringently defined cases, reducing statistical power in univariate analysis, and perhaps explaining the differences observed between univariate and multivariate analyses. Further investigation leveraging large, well-annotated cohorts is therefore mandatory.

Emerging evidence points to significant heterogeneity in ageing and AD trajectories, underscoring the need to scale biological studies to reflect more accurately population diversity. While initiatives like SEA-AD and ROSMAP offer unique opportunities for deep profiling across the disease spectrum, these efforts are resource intensive. Through Cellformer, we showcase the potential of deep learning to amplify diversity in population-level, high-density molecular research, making such needed studies more accessible and cost-effective.

Data availability

The code for processing the data, training, evaluating and using the model is available at <https://github.com/elo-nsrb/CellformerRNA/>.

Data used in this study are available on Dryad at 10.5061/dryad.6hdx7sr9d. We leveraged snRNA data from¹⁶ available at (https://github.com/LieberInstitute/10xPilot_snRNAseq-human), snRNA-seq data from¹⁵ downloaded from GEO (GSE186538), snRNA-seq data from¹⁷ downloaded from GEO (GSE140231), snRNA from¹³ downloaded from Synapse (syn18485175), snRNA from¹⁸ downloaded from synapse (syn52293417) and snRNA from SEA-AD Atlas¹⁴ downloaded from (<https://portal.brain-map.org/explore/seattle-alzheimers-disease>), bulk RNA-seq from²⁴ available at Short Read Archive (SRA) (BioProject PRJNA1023207), bulk RNA-seq from²⁵ downloaded from GEO (GSE216281), bulk RNA-seq from the Religious Orders Study and the Memory and Aging Project (ROSMAP),²⁶ MAYO²⁸ and Mount Sinai Brain Bank (MSBB)²⁷ cohort available on synapse (syn26403544).

Acknowledgements

The thumbnail image for the online table of contents was created in BioRender. Berson, E. (2025) <https://BioRender.com/28q32fr>.

Funding

This work was supported by the National Institutes of Health: AG077443 (T.J.M.), AG072573 (T.J.M.) and R35GM138353 (N.A.) and by the Phil & Penny Knight Initiative for Brain Resilience at the Wu Tsai Neurosciences Institute, Stanford University (E.B.) and the Stanford Bio-X Interdisciplinary Initiatives Seed Grants Program (I.I.P.) (R11-74) (M.K.).

Competing interests

The authors report no competing interests.

Supplementary material

Supplementary material is available at [Brain](#) online.

References

- Herculano-Houzel S. The human brain in numbers: A linearly scaled-up primate brain. *Front Hum Neurosci*. 2009;3:31.
- Montine TJ, Cholerton BA, Corrada MM, et al. Concepts for brain aging: Resistance, resilience, reserve, and compensation. *Alzheimers Res Ther*. 2019;11:22.
- Bakken TE, Hodge RD, Miller JA, et al. Single-nucleus and single-cell transcriptomes compared in matched cortical cell types. *PLoS One*. 2018;13:e0209648.
- Squair JW, Gautier M, Kathe C, et al. Confronting false discoveries in single-cell differential expression. *Nat Commun*. 2021;12:5692.
- Newman AM, Steen CB, Liu CL, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat Biotechnol*. 2019;37:773–782.
- Chu T, Wang Z, Pe'er D, Danko CG. Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. *Nat Cancer*. 2022;3:505–517.
- Avila Cobos F, Alquicira-Hernandez J, Powell JE, Mestdagh P, De Preter K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat Commun*. 2020;11:5650.
- Jin H, Liu Z. A benchmark for RNA-Seq deconvolution analysis under dynamic testing environments. *Genome Biol*. 2021;22:102.
- Sutton GJ, Poppe D, Simmons RK, et al. Comprehensive evaluation of deconvolution methods for human brain gene expression. *Nat Commun*. 2022;13:1358.
- Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In Guyon I, Von Luxburg U, Bengio S et al. eds. 31st Conference on Neural Information Processing Systems (NIPS). Advances in Neural Information Processing Systems. Vol. 30. Curran Associates; 2017. doi:10.48550/arXiv.1706.03762.
- Bommasani R, Hudson DA, Adeli E, et al. On the opportunities and risks of foundation models. *arXiv:210807258*. 2021.
- Berson E, Sreenivas A, Phongpreecha T, et al. Whole genome deconvolution unveils Alzheimer's resilient epigenetic signature. *Nat Commun*. 2023;14:4947.
- Mathys H, Davila-Velderrain J, Peng Z, et al. Single-cell transcriptomic analysis of Alzheimer's disease. *Nature*. 2019;570:332–337.
- Gabitto M, Travaglini K, Ariza J, et al. Integrated multimodal cell atlas of Alzheimer's disease. *Nat Neurosci*. 2024;27:2366–2383.
- Franjic D, Skarica M, Ma S, et al. Transcriptomic taxonomy and neurogenic trajectories of adult human, macaque, and pig hippocampal and entorhinal cells. *Neuron*. 2021;110:452–469.e14.
- Tran MN, Maynard KR, Spangler A, et al. Single-nucleus transcriptome analysis reveals cell-type-specific molecular signatures across reward circuitry in the human brain. *Neuron*. 2021;109:3088–3103.
- Agarwal D, Sandor C, Volpato V, et al. A single-cell atlas of the human substantia nigra reveals cell-specific pathways associated with neurological disorders. *Nat Commun*. 2020;11:4183.
- Mathys H, Peng Z, Boix CA, et al. Single-cell atlas reveals correlates of high cognitive function, dementia, and resilience to Alzheimer's disease pathology. *Cell*. 2023;186:4365–4385.
- Lieberman DN, Mody I. Regulation of NMDA channel function by endogenous Ca^{2+} -dependent phosphatase. *Nature*. 1994;369:235–239.
- Arendt KL, Zhang Z, Ganesan S, et al. Calcineurin mediates homeostatic synaptic plasticity by regulating retinoic acid synthesis. *Proc Natl Acad Sci USA*. 2015;112:E5744–E5752.
- Mathys H, Boix CA, Akay LA, et al. Single-cell multiregion dissection of Alzheimer's disease. *Nature*. 2024;632:858–868.
- Newman AM, Liu CL, Green MR, et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods*. 2015;12:453–457.
- Gong T, Szustakowski JD. DeconRNASeq: A statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. *Bioinformatics*. 2013;29:1083–1085.
- Olney KC, Rabichow BE, Wojtas AM, et al. Distinct transcriptional alterations distinguish Lewy body disease from Alzheimer's disease. *Brain*. 2024;148:69–88.
- Cappelletti C, Henriksen SP, Geut H, et al. Transcriptomic profiling of Parkinson's disease brains reveals disease stage specific gene expression changes. *Acta Neuropathol*. 2023;146:227–244.
- Mostafavi S, Gaiteri C, Sullivan SE, et al. A molecular network of the aging human brain provides insights into the pathology and cognitive decline of Alzheimer's disease. *Nat Neurosci*. 2018;21:811–819.
- Wang M, Beckmann ND, Roussos P, et al. The Mount Sinai cohort of large-scale genomic, transcriptomic and proteomic data in Alzheimer's disease. *Sci Data*. 2018;5:180185.
- Allen M, Carrasquillo MM, Funk C, et al. Human whole genome genotype and transcriptome data for Alzheimer's and other neurodegenerative diseases. *Sci Data*. 2016;3:160089.
- Montine TJ, Corrada MM, Kawas C, et al. Association of cognition and dementia with neuropathologic changes of Alzheimer disease and other conditions in the oldest old. *Neurology*. 2022;99:e1067–e1078.
- Montine TJ, Bukhari SA, White LR. Cognitive impairment in older adults and therapeutic strategies. *Pharmacol Rev*. 2021;73:152–162.
- Montine TJ, Phelps CH, Beach TG, et al. National institute on aging-Alzheimer's association guidelines for the neuropathologic assessment of Alzheimer's disease: A practical approach. *Acta Neuropathol*. 2012;123:1–11.
- Krepel J, Kircher M, Kohls M, Jung K. Comparison of merging strategies for building machine learning models on multiple independent gene expression data sets. *Stat Anal Data Min*. 2022;15:112–124.
- Ryu Y, Han GH, Jung E, Hwang D. Integration of single-cell RNA-Seq datasets: A review of computational methods. *Mol Cells*. 2023;46:106–119.
- Glorigijević V, Pržulj N. Methods for biological data integration: Perspectives and challenges. *J R Soc Interface*. 2015;12:20150571.
- Goralski KB, McCarthy TC, Hanniman EA, et al. Chemerin, a novel adipokine that regulates adipogenesis and adipocyte metabolism. *J Biol Chem*. 2007;282:28175–28188.

36. Li YQ, Sun FZ, Li CX, et al. RARRES2 regulates lipid metabolic reprogramming to mediate the development of brain metastasis in triple negative breast cancer. *Mil Med Res.* 2023;10:34.
37. Shin WJ, Zabel BA, Pachynski RK. Mechanisms and functions of chemerin in cancer: Potential roles in therapeutic intervention. *Front Immunol.* 2018;9:2772.
38. de Vries LE, Jongejan A, Monteiro Fortes J, et al. Gene-expression profiling of individuals resilient to Alzheimer's disease reveals higher expression of genes related to metallothionein and mitochondrial processes and no changes in the unfolded protein response. *Acta Neuropathol Commun.* 2024;12:68.
39. Yap CX, Vo DD, Heffel MG, et al. Brain cell-type shifts in Alzheimer's disease, autism, and schizophrenia interrogated using methylomics and genetics. *Sci Adv.* 2024;10:eadn7655.
40. Johnson TS, Xiang S, Dong T, et al. Combinatorial analyses reveal cellular composition changes have different impacts on transcriptomic changes of cell type specific genes in Alzheimer's disease. *Sci Rep.* 2021;11:353.
41. Walls AB, Eyjolfsson EM, Schousboe A, Sonnewald U, Waagepetersen HS. A subconvulsive dose of kainate selectively compromises astrocytic metabolism in the mouse brain in vivo. *J Cereb Blood Flow Metab.* 2014;34:1340-1346.
42. Maccioni R, Travisan C, Badman J, et al. Signal peptide peptidase-like 2b modulates the amyloidogenic pathway and exhibits an A β -dependent expression in Alzheimer's disease. *Prog Neurobiol.* 2024;235:102585.
43. Sudwarts A, Ramesha S, Gao T, et al. BIN1 is a key regulator of proinflammatory and neurodegeneration-related activation in microglia. *Mol Neurodegener.* 2022;17:33.
44. Arenaza-Urquijo EM, Vemuri P. Resistance vs resilience to Alzheimer disease. *Neurology.* 2018;90:695-703.
45. Montine KS, Berson E, Phongpreecha T, et al. Understanding the molecular basis of resilience to Alzheimer's disease. *Front Neurosci.* 2023;17:1311157.
46. Stern Y, Albert M, Barnes CA, Cabeza R, Pascual-Leone A, Rapp PR. A framework for concepts of reserve and resilience in aging. *Neurobiol Aging.* 2023;124:100-103.
47. Huang Z, Merrihew GE, Larson EB, et al. Brain proteomic analysis implicates actin filament processes and injury response in resilience to Alzheimer's disease. *Nat Commun.* 2023;14:2747.
48. Reisberg B, Doody R, Stöffler A, Schmitt F, Ferris S, Möbius HJ. Memantine in moderate-to-severe Alzheimer's disease. *N Engl J Med.* 2003;348:1333-1341.
49. Huang P, Cai M, Lu X, McKennan C, Wang J. Accurate estimation of rare cell-type fractions from tissue omics data via hierarchical deconvolution. *Ann Appl Stat.* 2024;18:1178-1194.
50. O'Neill N, Stein TD, Olayinka OA, et al. Cognitive resilience to Alzheimer's disease characterized by cell-type abundance. *Alzheimers Dementia.* 2024;20:6910-6921.
51. Korthikanti VA, Casper J, Lym S, et al. Reducing activation re-computation in large transformer models. *Proc Mach Learn Syst.* 2023;5:341-353.
52. Ding J, Ma S, Dong L, et al. Longnet: Scaling transformers to 1,000,000,000 tokens. *arXiv.* [Preprint] arXiv:230702486.
53. Buckley PT, Khaladkar M, Kim J, Eberwine J. Cytoplasmic intron retention, function, splicing, and the sentinel RNA hypothesis. *Wiley Interdiscip Rev RNA.* 2014;5:223-230.
54. Glanzner J, Miyashiro KY, Sul JY, et al. RNA splicing capability of live neuronal dendrites. *Proc Natl Acad Sci USA.* 2005;102:16859-16864.
55. Linderman GC, Zhao J, Roulis M, et al. Zero-preserving imputation of single-cell RNA-Seq data. *Nat Commun.* 2022;13:192.
56. Kimble AL, Silva J, Omar OM, et al. A method for rapid flow-cytometric isolation of endothelial nuclei and RNA from archived frozen brain tissue. *Lab Invest.* 2022;102:204-211.
57. Raff MC, Miller RH, Noble M. A glial progenitor cell that develops in vitro into an astrocyte or an oligodendrocyte depending on culture medium. *Nature.* 1983;303:390-396.