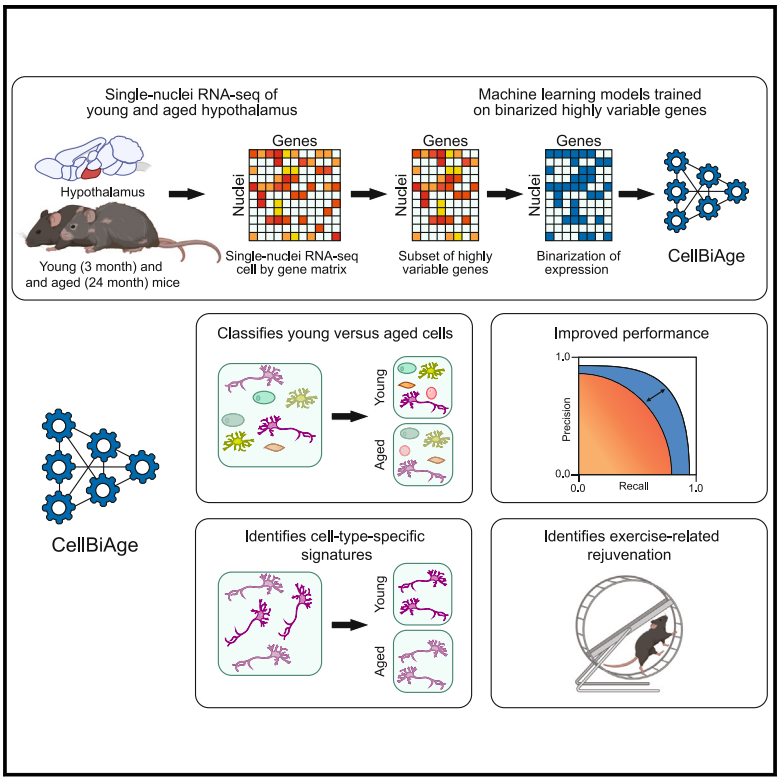
Cell Reports

Resource

CellBiAge: Improved single-cell age classification using data binarization

Graphical abstract



Highlights

- CellBiAge classifies organismal age groups of single cells through binarization
- CellBiAge model interpretation reveals cell-type-specific aging signatures
- Generalizable to ML models, sc/snRNA-seq techniques, sexes, and mouse brain regions
- The model captures exercise-induced rejuvenation in proliferating neural stem cells

Authors

Doudou Yu, Manlin Li, Guanjie Linghu, ..., An Wang, Ritambhara Singh, Ashley E. Webb

Correspondence

ritambhara_singh@brown.edu (R.S.), awebb@buckinstitute.org (A.E.W.)

In brief

Incorporating data binarization, Yu et al. develop the CellBiAge pipeline for accurate organismal age classification at the single-cell level in the mouse brain. CellBiAge demonstrates generalizability across techniques, sexes, and brain regions. Additionally, cell-type-specific models reveal distinct signatures and capture exercise-induced rejuvenation in proliferating neural stem cells.

Yu et al., 2023, Cell Reports *42*, 113500
December 26, 2023
https://doi.org/10.1016/j.celrep.2023.113500

CellPress

Cell Reports

CellPress
OPEN ACCESS

Resource

CellBiAge: Improved single-cell age classification using data binarization

Doudou Yu,[1,2,11] Manlin Li,[2,12] Guanjie Linghu,[2,12] Yihuan Hu,[2,12] Kaitlyn H. Hajdarovic,[3,11] An Wang,[4] Ritambhara Singh,[5,6,*] and Ashley E. Webb[7,8,9,10,11,13,*]

[1]Molecular Biology, Cell Biology, and Biochemistry Graduate Program, Brown University, Providence, RI 02912, USA
[2]Data Science Institute, Brown University, Providence, RI 02912, USA
[3]Neuroscience Graduate Program, Brown University, Providence, RI 02912, USA
[4]Department of Applied Mathematics & Statistics, Johns Hopkins University, Baltimore, MD 21218, USA
[5]Department of Computer Science, Brown University, Providence, RI 02912, USA
[6]Center for Computational Molecular Biology, Brown University, Providence, RI 02912, USA
[7]Department of Molecular Biology, Cell Biology, and Biochemistry, Brown University, Providence, RI 02912, USA
[8]Center on the Biology of Aging, Brown University, Providence, RI 02912, USA
[9]Carney Institute for Brain Science, Brown University, Providence, RI 02912, USA
[10]Center for Translational Neuroscience, Brown University, Providence, RI 02912, USA
[11]Present address: The Buck Institute for Research on Aging, Novato, CA 94945, USA
[12]These authors contributed equally
[13]Lead contact
*Correspondence: ritambhara_singh@brown.edu (R.S.), awebb@buckinstitute.org (A.E.W.)
https://doi.org/10.1016/j.celrep.2023.113500

SUMMARY

Aging is a major risk factor for many diseases. Accurate methods for predicting age in specific cell types are essential to understand the heterogeneity of aging and to assess rejuvenation strategies. However, classifying organismal age at single-cell resolution using transcriptomics is challenging due to sparsity and noise. Here, we developed CellBiAge, a robust and easy-to-implement machine learning pipeline, to classify the age of single cells in the mouse brain using single-cell transcriptomics. We show that binarization of gene expression values for the top highly variable genes significantly improved test performance across different models, techniques, sexes, and brain regions, with potential age-related genes identified for model prediction. Additionally, we demonstrate CellBiAge's ability to capture exercise-induced rejuvenation in neural stem cells. This study provides a broadly applicable approach for robust classification of organismal age of single cells in the mouse brain, which may aid in understanding the aging process and evaluating rejuvenation methods.

INTRODUCTION

Aging is a major risk factor for many diseases including cancer and neurodegeneration.[1,2] Thus, methods that characterize and quantify aging may have the power to predict age-associated diseases and evaluate rejuvenation methods. To this end, biomarkers have been developed to measure a variety of aging features, from molecular and cellular markers to organismal phenotypes.[3] For example, a number of hallmarks of aging, such as changes in transcriptional[4–8] and epigenetic networks,[9–12] loss of proteostasis,[13–15] stem cell dysfunction,[16–18] and frailty,[19] have all been used as metrics of aging.

Advances in high-throughput sequencing and 'omics techniques have enabled the development of aging biomarkers at large scales. To identify aging biomarkers with high-dimensional data, machine learning (ML) models that integrate a large number of features have been applied to learn to discriminate between ages. Aging clocks using biomarkers have been built on high-dimensional data from various modalities. For example, epigenetic biomarkers (clocks) built on DNA methylation profiles,[20–24] and incorporating prior biological knowledge or disease-associated markers,[25–27] predict age with high accuracy. These methods perform well despite a lack of understanding of the mechanisms involved. In contrast, transcriptome-based methods, which are more intuitive, have compromised performance due to increased noise of gene expression with age.[28] Methods to minimize noise, such as data binarization and relative age scaling, have been implemented in *Caenorhabditis elegans* and human fibroblasts, known as BiT age.[29] This approach improved performance in bulk RNA-seq datasets, but it has not been tested in the context of single-cell RNA-seq (scRNA-seq). Biomarkers identified using proteomics[30,31] and metabolomics[32,33] datasets, which harbor multi-tissue information across organs, reveal important biological pathways but suffer from scale and interpretability, respectively.

Most high-dimensional biomarkers and clocks identified to date have used bulk-tissue profiles, which average out tissue-specific or cell-type-specific aging signatures within an individual.[8,23,34–36] Recent developments in single-cell 'omics technologies enable the age prediction or classification of single

cells using transcriptomics[8,37–39] and methylation profiles.[23] For example, a cell-type-specific transcriptomic clock generated from the mouse subventricular zone (SVZ), a neurogenic region in the adult mammalian brain, revealed that each cell-type-specific clock selects unique gene sets for age prediction, revealing distinct aging trajectories in different cell types.[39]

Further identification of biomarkers to classify age in brain tissue has the potential to advance our understanding of the mechanisms of brain aging and how they impact vulnerability and resilience to neurodegeneration. The hypothalamus is a brain region that has been implicated in healthy aging as a top-down regulator of a variety of functions, including nutrient sensing, circadian rhythms, energy expenditure, and other homeostatic processes.[8,40] Longevity interventions such as caloric restriction target hypothalamic circuits involved in food intake[41] and circadian rhythms.[42] Transcriptomic profiles of cell types in the hypothalamus have revealed cell-type-specific aging signatures. For example, enrichment of inflammatory signatures in microglia, under-enrichment of cholesterol homeostasis gene sets in astrocytes, and under-enrichment of MYC targets gene sets in neurons.[8] In previous work, our laboratory demonstrated that single-nucleus transcriptomic (snRNA-seq) profiles of X chromosome genes could predict neurons as young or aged,[8] suggesting that this tissue could be leveraged to develop a cell-type-specific brain aging clock. However, predicting the age group of single cells across all cell types in the hypothalamus without prior knowledge is challenging, mostly due to the inherent sparsity and noise of single-cell RNA-seq datasets.

Here, we developed CellBiAge, a robust and easy-to-implement ML pipeline, to predict the age group (young or aged) of single cells in the mouse hypothalamus using snRNA-seq data. We showed that binarization of expression values for the top highly variable genes (HVGs) significantly improved the test performance across different ML models and brain regions. The interpretation of the all-cell model reveals potential age-related genes for model prediction. Cell-type-specific models for the most abundant cell types performed consistently well, with only a small subset of genes being shared as important features across cell types, indicating the cell-type-specific transcriptional signatures in age group prediction. In addition, the model captures the rejuvenating effect of exercise in proliferating neural stem cells (NSCs) in the mouse SVZ, which may aid in understanding the aging process and evaluating rejuvenation methods.

RESULTS

A robust ML pipeline to classify organismal age at single-cell resolution in the mouse hypothalamus

To test whether snRNA-seq profiles can predict organismal age group (young or aged), we used a publicly available dataset for the aging female mouse hypothalamus, previously generated by our laboratory (Figure 1A).[8] The data were generated in two independent batches using different library preparation kit versions and sequencing platforms. Each batch has four animals, two young (3 months old) and two aged (24 months old) (Figure 1B). After quality control, the dataset includes 40,064 nuclei in total. The nuclei number for each sample is shown in Figure 1B. Using these snRNA-seq data, we developed an age classification task to predict the organismal age group. In this task, the input gene expression matrix was the transcriptomic profiles of young and aged female hypothalamic nuclei (Figure 1C). Ultimately, the output was the probability of a nucleus belonging to the aged category.

To optimize task performance, we implemented a series of preprocessing and feature selection methods and used the area under the precision-recall curve (AUPRC) score as the evaluation metric. AUPRC score was preferred over other metrics due to imbalance of the two classes in the test set.[43] Specifically, for data preprocessing, three commonly used methods were implemented: log normalization, batch integration (batch correction using canonical correlation analysis, CCA), and scaling (standardization), all in Seurat.[44,45] Log normalization removes biases in sequencing coverage between nuclei, batch integration reduces the impact of potential confounding batch variables on the task, and scaling standardizes the range of features by adjusting the mean expression value to 0 and standard deviation to 1. For feature selection, we tested performance in classifying age using the top 2,000 (2k) HVGs, which have been shown to highlight biological signals in single-cell datasets,[45,46] or highly expressed genes (HEGs), ranked by gene expression (Figure S1). The benefit of these feature selection methods is that they do not require prior knowledge and are widely used in single-cell data analysis.[45,47] Furthermore, transcriptomic profiles of the HVGs could represent most nuclei in the dataset, and HEGs were the most abundant genes, which may harbor rich information. Lastly, we performed data binarization after preprocessing, which converted scaled values larger than 0 as 1 and remaining values as 0. To some extent, this binarization method resembles single-cell DNA methylation profiling. In methylation profiling, binarized single-cell methylomes, when coupled with bulk methylomes as references, have demonstrated the ability to predict cellular age with high accuracy.[23] Furthermore, binarization can preserve biological heterogeneity while alleviating noise embedded in RNA-seq data at the bulk[48,49] and single-cell resolutions,[50,51] yet its potential for age group prediction has not been explored. In our workflow, binarization significantly enhanced performance in the classification task when combined with HVG feature selection. The AUPRC score in this case was highly improved relative to the other preprocessing methods (Figure 1D). Thus, we leveraged the binarized matrix of HVGs for further ML modeling for organismal age classification and model interpretation (Figures 1E, S2, and S3).

Data binarization significantly improves performance across different ML models

We next wanted to determine if the data binarization approach could be useful across classification models. We applied linear models, tree-based models, support vector machines, as well as a fully connected neural network model before and after binarization (Figures 2A and 2B). After selecting the optimal hyperparameters, we retrained the model with the nuclei from animals 1–4 in the training data and tested on the previously held-out nuclei from animals 5–8 in the test data (Figure S4A). This test dataset was generated and preprocessed independently, and we report the performance of the final model over 10 random seeds

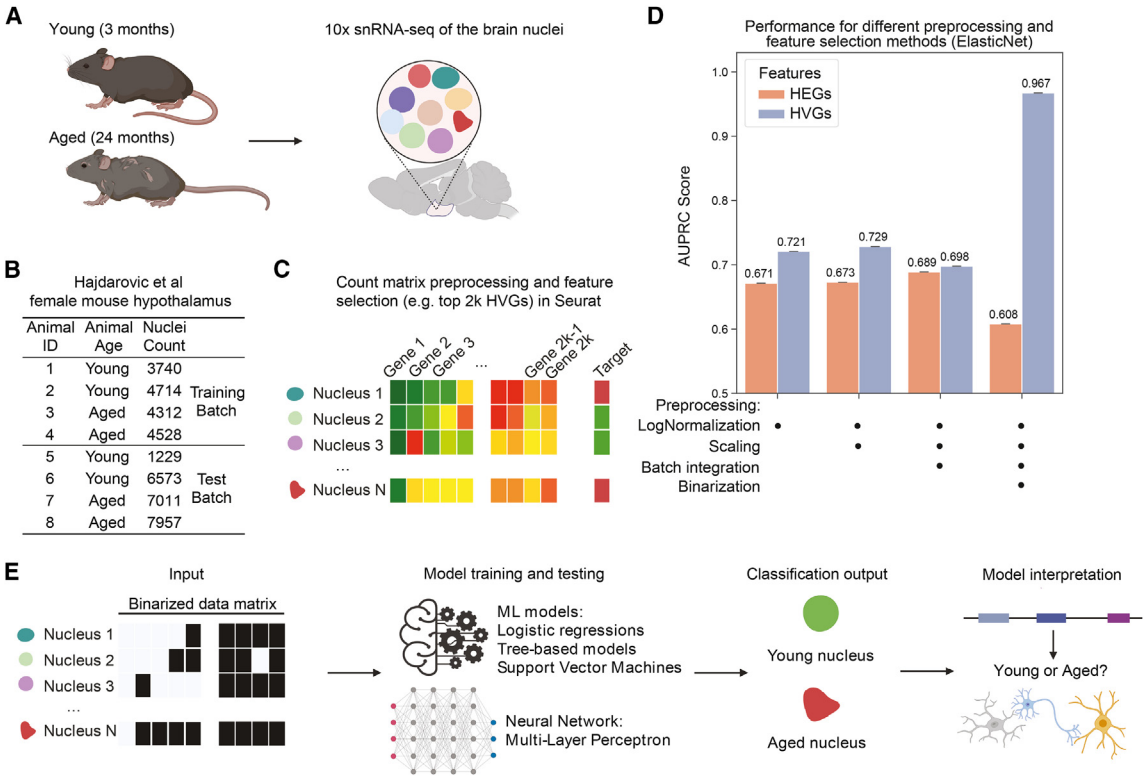


Figure 1. A robust ML pipeline to classify organismal age at single-cell resolution in the mouse hypothalamus

(A) Dataset description: single nuclei were isolated from the whole hypothalamus of young and aged mice (n = 4 per group), processed with the 10x Genomics snRNA-seq workflow.

(B) Nuclei counts for each sample in the training and test batches. The training batch was generated with 10x Chromium kit version 3 and Illumina NovaSeq, while the test batch was generated with 10x Chromium kit version 2 and Illumina HiSeq.

(C) The gene expression matrices of the training and test batches were preprocessed separately using the Seurat package in R. The target variable for this classification task is whether the nucleus is from the young or aged category.

(D) Bar plot showing ELN test performance with different feature selection and preprocessing methods. Orange: HEGs test performance; blue: HVGs test performance. All the preprocessing methods were cumulative (i.e., the binarized data was preprocessed with log normalization, batch integration, and scaling). Hyperparameters were chosen using the GridSearchCV on the training set. Data are represented as mean ± SD.

(E) The binarized gene expression matrix (rule: if entry value > 0, assign 1; else, assign 0) as input for models (logistic regressions with regularization, tree-based models, support vector machines, XGBoost classifier, and neural network). The model outputs the probability of a nucleus originating from the aged category. Models were evaluated using the AUPRC scores and interpreted for all cells and in a cell-type-specific manner. See also Figures S1–S3.

(Figure 2B). Models implemented before data binarization performed well in training but not testing. Interestingly, the gene expression matrix after binarization outperformed the non-binarized matrix across all models, with similar performance (AUPRC around 0.96) (Figures 2B and 2C). In parallel, we checked our experimental setup to ensure that the read depth of training data did not boost the performance on the test data. To do so, we swapped the training and test data, using the lower read-depth data to train the model, followed by testing on the higher read-depth data. Without finer hyperparameter tuning, the ElasticNet (ELN) model improved test performance from 0.60 (before binarization) to 0.94 (after binarization). Additionally, the binarization threshold implemented in CellBiAge (mean) outperformed the threshold in BiT age (median) on the test data, indicating the importance of threshold selection (Figure S4B). One concern with batch integration using Seurat CCA is that it may remove true biological age-related variation.[52] To confirm that our method is robust to batch correction, we selected animals from the two batches in the training and test (Figures S4C–S4G). The performance remained around 0.96 in the presence of batch effects within the training or test set, suggesting that data integration does not impair the model from detecting biological differences. Taken together, the data binarization method drastically improved the preprocessed HVG matrix performance regardless of the data or model applied.

ELN model interpretation reveals potential age-related genes for model prediction

We then focused on the ELN model for model interpretation because of its overall comparable performance, interpretability, and short run time (Figure S5A). We first ranked the 1,413 genes, the intersection of the top 2k genes in the training and test batches, in model training and testing by the absolute values of coefficients in the ELN model (Figure 3A, Table S1). The absolute values of the coefficients indicate the magnitude of their impact on the model prediction.

A

Model	Hyperparameters before binarization	Hyperparameters after binarization
L1 (Lasso) - L1	C=0.001	C=19
L2 (Ridge) - L2	C=0.001	C=0.077
ElasticNet - ELN	C=0.001 l1_ratio=0.35	C=0.046 l1_ratio=0.01
Random Forest - RFC	max_features=50 max_depth=10 min_sample_split=5	max_features=10 max_depth=20 min_sample_split=10
XGBoost - XGBC	max_depth=3	max_depth=5
Support Vector Machines - SVC	gamma=0.1 C=10	gamma=0.01 C=5.62
Multilayer Perceptron - MLP	n_hidden=1 n_neurons=320 learning_rate=0.004	n_hidden=1 n_neurons=256 learning_rate=3.11e-05

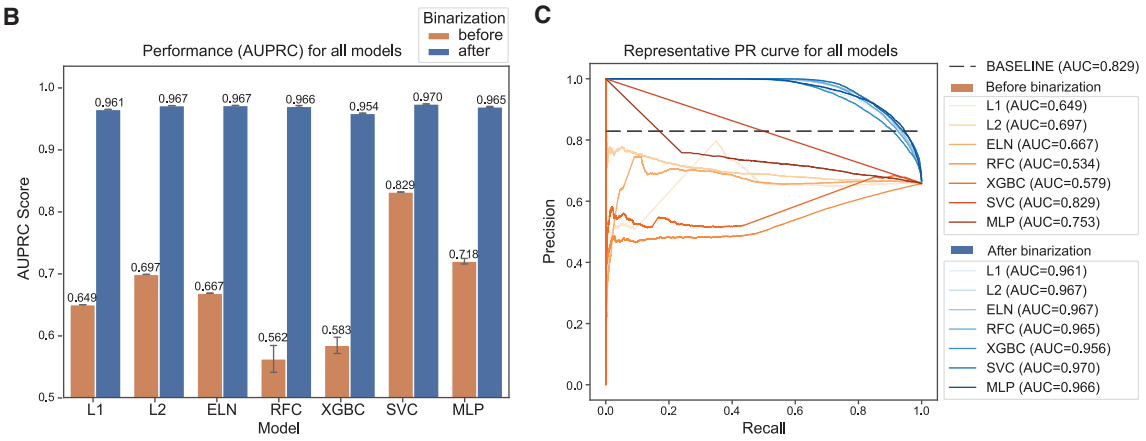


Figure 2. Data binarization significantly improves performance across different ML models

(A) Table reporting optimized hyperparameters for each model trained using the HVGs expression matrix before and after binarization. The hyperparameters were selected using GridSearchCV and KerasTuner in the group-based cross-validation method.

(B) Bar plot comparing test performance across models before (orange) and after (blue) binarization using 10 different random seeds. Data are represented as mean ± SD.

(C) The representative precision-recall (PR) curves for the models built with gene expression matrices before (orange lines) and after (blue lines) binarization for a single seed. The dashed line represents the baseline AUPRC score when the model predicts all cells to be in the same major category in the test set. See also Figure S4.

To better understand how many genes were essential for the model prediction, we shuffled the expression values for all ranked genes one by one in the binarized gene expression matrix and plotted the cumulative AUPRC scores. We chose shuffling as our perturbation strategy to preserve the original distribution. Interestingly, the performance dropped sharply upon perturbation of the last 200 genes (Figures 3B and S5B). For example, starting from gene #1,250, the derivatives changed from approximately 0 to negative values, indicating a rapid drop in model performance. Given that the top 200 genes were required for model performance, we then asked whether they were sufficient to restore the test performance. The top 80 genes with the highest absolute coefficients were recovered from a fully shuffled binarized gene expression matrix by replacing the shuffled values with the original ones sequentially. Interestingly, the top 40 genes restored the model performance to 0.90, and the top 80 restored the model performance to near the original 0.95 (Figure 3C). We also performed the perturbation in the opposite direction, from the most important genes to the least ones, which showed an earlier and gentler drop in performance (Figures S5C and S5D). Together, these results suggest that a relatively small number of genes are critical for model prediction.

Out of the top 80 genes we restored, we focused on the biological meaning of the top 20 genes for simplicity (Figure 3D). The binarized expressions of *Slc5a4b* and *Slc13a4*, which have the largest coefficients and encode predicted glucose and sulfate transporters, respectively, are important in the model prediction. Although the functional implications of these two genes in the context of aging remain unexplored, a recent study showed that knockout of genes encoding glucose import, such as *Slc2a4* (encodes the GLUT4 glucose transporter), rejuvenated old NSCs.[53] Another top gene, *Cd34*, which encodes endothelial cell glycoprotein, has been implicated in degeneration.[54] Thus, interpretation of the ELN model identifies potential genes related to the aging process.

In addition to the ELN coefficient-based perturbation, we performed permutations for individual genes without prior

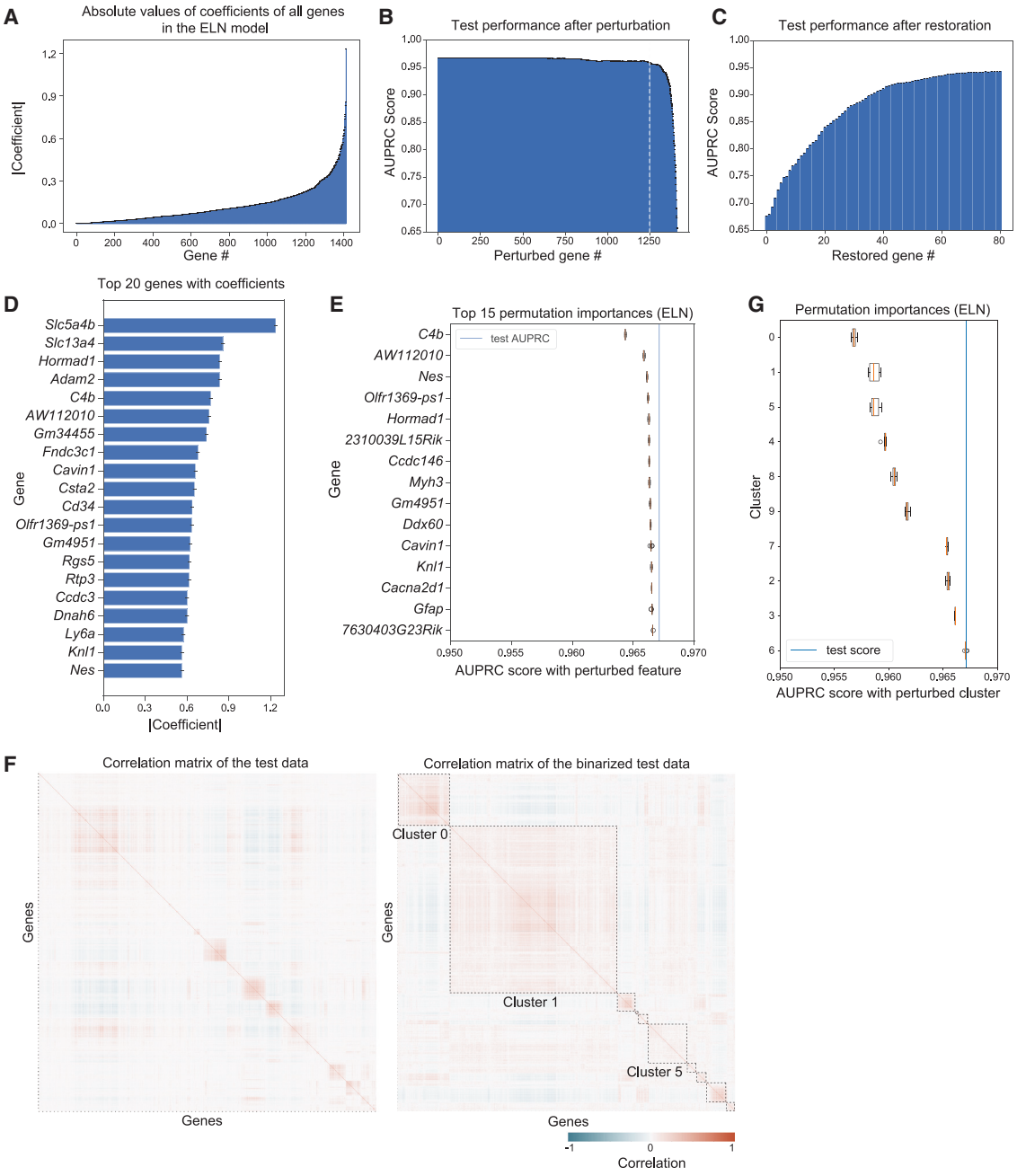


Figure 3. ELN model interpretation reveals potential age-related genes for model prediction

(A) Waterfall plot showing the absolute values of coefficients of all 1,413 genes (ranked). Data are represented as mean ± SD for training and test results over 10 random seeds.

(B) Cumulative test performance after feature perturbation (shuffling from the least important genes to the most important genes). Data are represented as mean ± SD over 10 random seeds. The white dashed line indicates a place with a sharp drop.

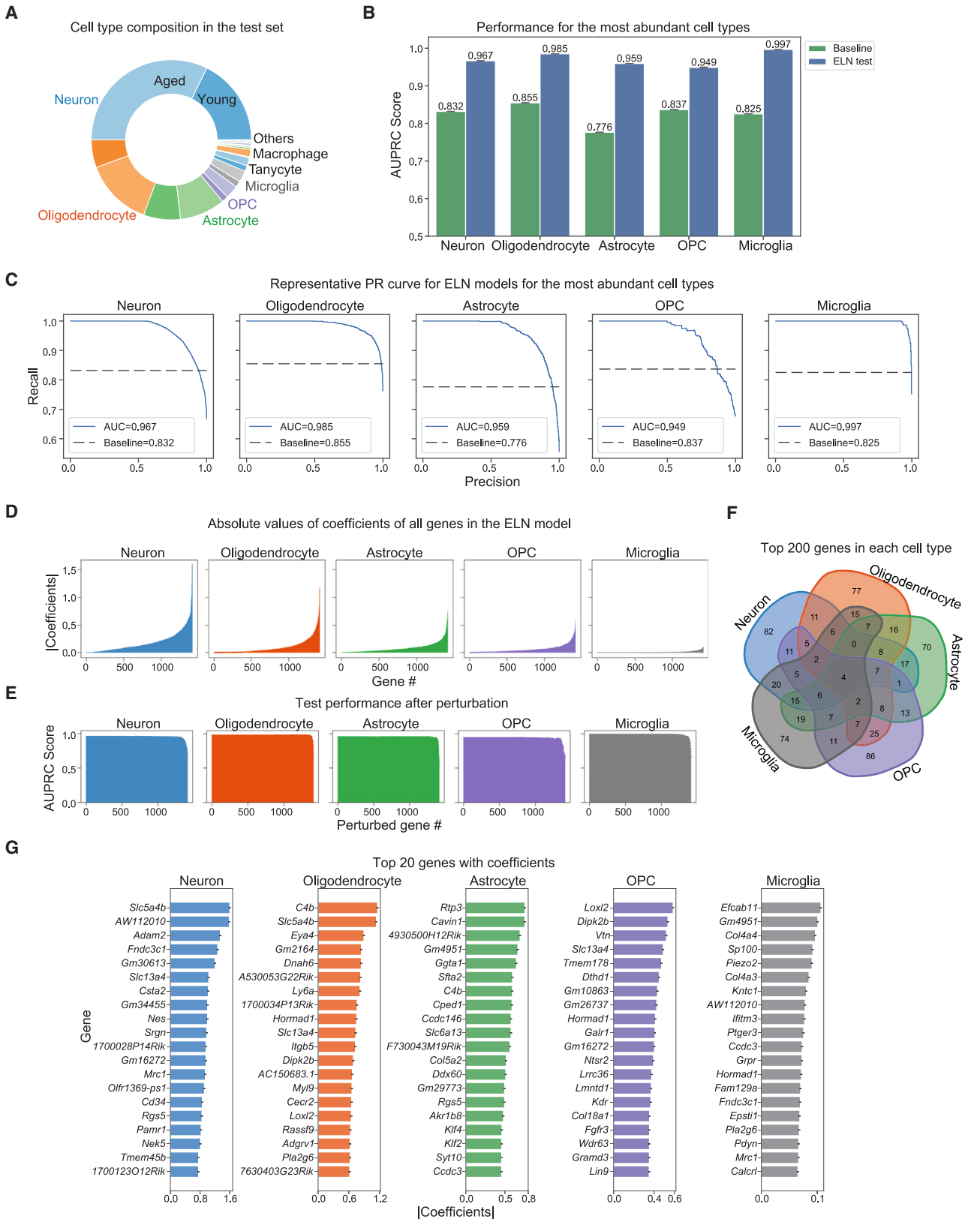
(C) Cumulative test performance after restoration of the top 80 genes in the fully shuffled test set.

(D) Bar plot showing the individual genes with top 20 coefficients. Data are represented as mean ± SD over 10 random seeds.

(E) Boxplot showing the top 15 most important genes in the feature permutation test. Data are represented as mean ± SD over 10 random seeds.

(F) Correlation matrix of the test set gene expression matrix before (left) and after (right) binarization. 10 clusters were manually subset from the binarized correlation matrix for the cluster perturbation test.

(G) Boxplot showing the test performance after perturbation of genes in clusters identified in (F). Data are presented as mean ± SD over 10 random seeds. See also Figure S5 and Table S1.



(legend on next page)

knowledge of the model. However, shuffling individual genes did not significantly affect the model performance (Figure 3E). Neither did the shuffling of genes on individual chromosomes (Figure S5E). Thus, we hypothesized that the correlation between genes contributed to the stable performance in the permutation test. Indeed, the correlation matrix of the cell-gene expression matrix showed correlated gene clusters regardless of binarization, with generally larger clusters after binarization (Figures 3F and S5F). We then permuted the genes within individual clusters manually identified by the pattern in the correlation matrix (Figure S5G). Perturbations on individual clusters affected the model performance to some extent, implicating the effect of gene-gene correlations (Figure 3G). Furthermore, the correlated expression matrix highlights the advantage of using the ELN model, as it evenly distributes weights across correlated features while retaining important features.[55] Together, these model interpretation analyses revealed an enhancement of gene-gene correlations within the binarized matrix.

Unique features underlie model prediction in a cell-type-specific manner

Single-cell 'omics approaches enable the discovery of cell-type-specific signatures and changes across different conditions. To understand the top cell-type-specific genes contributing to the age group prediction, we used our established cell type annotations[8] and retrained the ELN model for the top five most abundant cell types individually (neurons, oligodendrocytes, astrocytes, oligodendrocyte progenitor cells [OPCS], and microglia) (Figures 4A and S6A). Similar to the model using all cells (all-cell model), we built cell-type-specific models using binarized expression matrices of 1,413 HVGs, which were individually tuned, and repeated over ten random seeds. The clustering of major cell types was preserved after binarization (Figure S6B). The test performances were consistently improved compared to baseline AUPRC scores when the model predicted all cells to be in the same major category in the test set, across five different cell types, with microglia and oligodendrocytes performing the best (Figures 4B and 4C). Interestingly, microglia had the most zero coefficients, while neurons had the least among the five cell types (Figure 4D).

We next compared the cell-type-specific performance of CellBiAge with a bootstrap-based cell-type-specific aging clock (bootstrap_clock),[39] developed for major cell types in the mouse SVZ. Specifically, we evaluated test performance in the three shared major cell types: oligodendrocytes, astrocytes, and microglia. The bootstrap_clock generates 100 bootstrapped pseudo-cells for each cell type and animal combination, with each pseudo-cell generated from 15 cells in the population. The output of the bootstrap_clock is an estimated age of the pseudo-cell. Our results indicated that both CellBiAge and the bootstrap_clock demonstrated strong test performance in the three cell types, with CellBiAge slightly outperforming bootstrap_clock (Figure S6C).

We observed that, as with the all-cell models, the top 200 genes had the most impact on the sequential perturbation (Figure 4E). To investigate the shared and cell-type-specific signatures with age, we further analyzed the top 200 genes for each cell type. Interestingly, only four genes (*Nes, Ly6a, Slc5a4b, B230312C02Rik*) were shared by the top 200 genes in the five cell types, underlying the cell-type specificity for the top genes in the model prediction (Figure 4F). *Nes*, expressed in NSCs, neuroepithelial precursor cells, and reactive astrocytes, encodes the protein Nestin and plays a crucial role in neurogenesis.[56] *Ly6a* is involved in T cell activation and increased inflammation with age.[57] Additionally, out of the top 200 genes, the top 20 genes were mostly cell type specific. For example, *Adam2*, a gene that declines with age and is implicated in regulating neurogenesis,[58] was the top gene in neurons. *C4b*, involved in inflammation and age-associated neurodegenerative diseases,[59] was the top gene specifically for oligodendrocytes and astrocytes (Figure 4G).

In addition to the cell-type-specific model interpretations, we also broke down the test performance of the all-cell model by cell type. Across major cell types, the test performances exceeded 0.88 regardless of baseline or models (Figures S6D and S6E). In general, non-neuronal cells had greater test improvements in performance than neuronal cells (Figures S6F and S6G). Individual gene perturbations for each cell type revealed cell-type-specific genes. Interestingly, endothelial cells and ependymocytes were sensitive to perturbations, while neurons were resilient, indicating more gene-gene correlations in neurons (Figure S7). Together, data binarization significantly improved the test performance across different cell types.

Incorporating data binarization enhances model performance in an independent dataset

We then checked the generalizability of the data binarization to the scRNA-seq method. Specifically, we analyzed a publicly available scRNA-seq dataset profiling 79,123 cells from the aging male mouse SVZ, with animals assigned to either 5 weeks of

Figure 4. Unique features underlie model prediction in a cell-type-specific manner

(A) Doughnut chart showing the identity of the major cell types in the test set. The top seven most abundant cell types are labeled. Lighter shades represent aged nuclei.

(B) Bar plot showing the baseline (green) and test (blue) AUPRC scores in the cell-type-specific ELN models over 10 random seeds. Data are represented as mean ± SD.

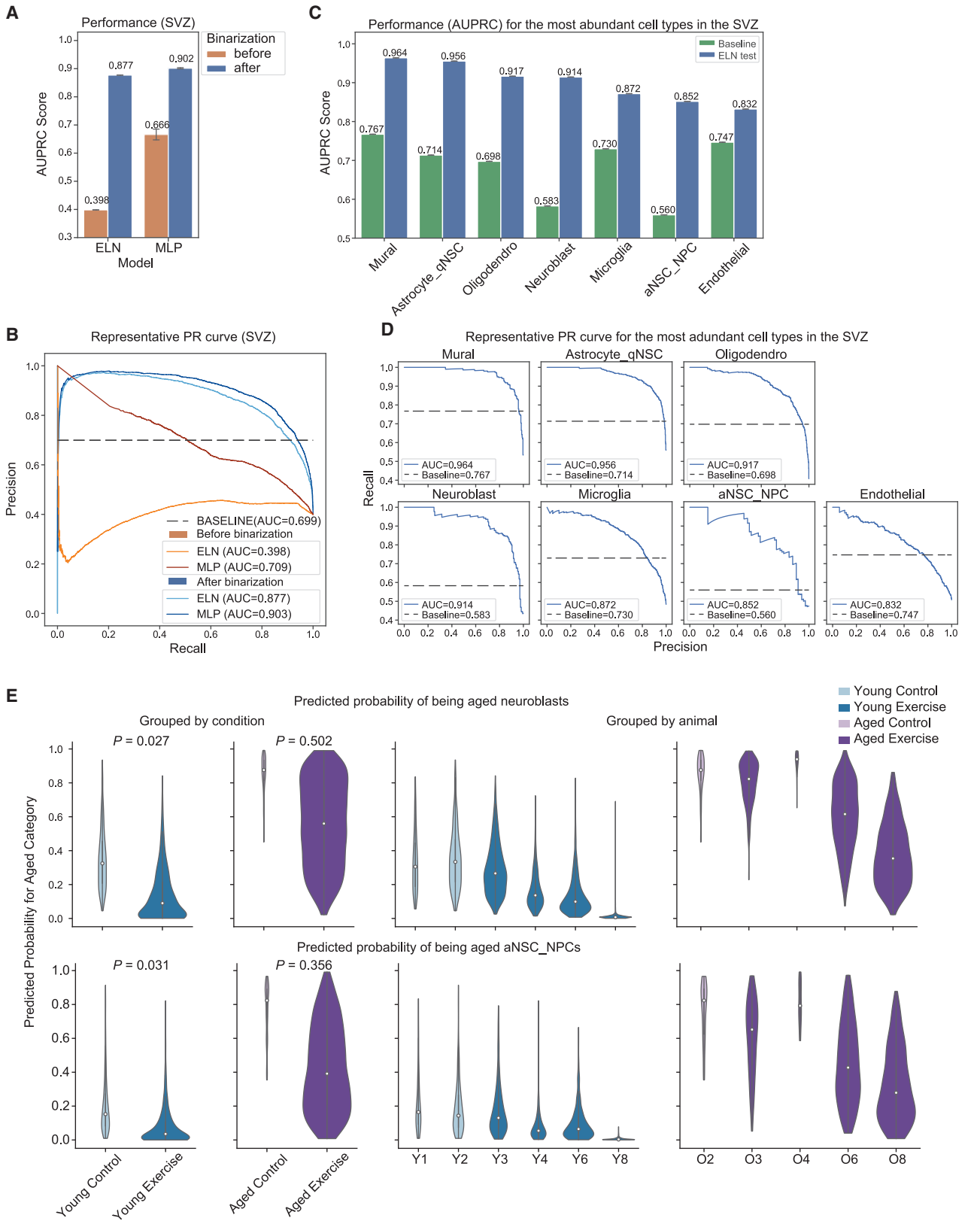
(C) The representative PR curves for models built with the binarized gene expression matrix. The baseline (dashed line) represents the AUPRC score when the model predicts all cells to be in the same major category in the test set for each cell type.

(D) Waterfall plots showing the absolute values of coefficients of all 1,413 HVGs (ranked) in each cell type. Data are represented as mean ± SD for training and test results over 10 random seeds.

(E) Cumulative test performance after feature perturbation (shuffling from the least important genes to the most important genes) in each cell type. Data are represented as mean ± SD over 10 random seeds.

(F) Venn diagrams showing the relationships between the top 200 genes in each cell type.

(G) Bar plot showing the genes with the top 20 coefficients in each cell type. Data are represented as mean ± SD over 10 random seeds. See also Figures S6 and S7.



(legend on next page)

sedentary (control) or voluntary running (exercise).[39,60] This high-quality dataset comprises 15 animals, including four young (6-month-old) and three aged (23-month-old) controls and four animals in each age-matched exercise condition, collected on two different days (Figure S8A). As single-nucleus and single-cell techniques capture different cell types, we selected the 1,617 shared top HVGs between the SVZ training and test controls and trained a new model (Figure S8A). We performed batch integration using the CCA in Seurat (Figure S8A and S8B, STAR Methods).[45] For the control animals, the model performance improved from 0.67 to 0.90 in both MLP (multilayer perceptron) and ELN (Figures 5A and 5B). We then constructed ELN models for the most abundant cell types to understand cell-type-specific features (Figures 5C and 5D). The cell-type-specific ELN models performed well especially for astrocytes and quiescent NSCs (qNSCs), mural cells, oligodendrocytes, and neuroblasts, suggesting that the binarization method is generalizable across brain regions (hypothalamus and SVZ), technique (nuclei and whole cells), and sexes.

CellBiAge captures exercise-induced rejuvenation in proliferating cells in the mouse SVZ

We next tested the performance of the model built on control mice in capturing the rejuvenating effects of exercise on all cells and the most abundant cell types. For all cells, after 5 weeks of exercise, we plotted the distributions of predicted probabilities of being classified as aged and grouped them by age and condition. Our results revealed decreasing probabilities in both the young and aged animals after exercise, although this was not statistically significant with the mixed-effect linear model accounting for the group structure (Figure S9A). We observed higher variability in the aged exercised animals compared to the young exercised mice when we separated the results by animals (Figure S9A). We further investigated the effect of exercise on the most abundant cell types (Figures 5E and S9B) and found that exercise significantly decreased the predicted probability of being aged in the neuroblast population ($p = 0.027$, Figure 5E) and activated NSCs and neural-progenitor cells (aNSCs_NPCs, $p = 0.031$, Figure 5E) in young animals, consistent with the finding that voluntary exercise increases neural stem cell proliferation.[61,62] This is in line with the finding that regular exercise benefits cognition and slows aging[63] measured by DNAmFitAge, a biological age predictor that incorporates physical fitness.[64] Together, these results demonstrate that the model built on control mice successfully captured the rejuvenating effects of exercise in the mouse SVZ, particularly in young neuroblasts and aNSCs_NPCs.

DISCUSSION

In this study, we developed an ML pipeline, CellBiAge, to classify the age of single cells in the mouse brain using sc/snRNA-seq data. Excitingly, an easy-to-implement denoising method—data binarization of the top HVGs—significantly improved test performance across different ML models and brain regions. We first trained and tested models using two independently generated snRNA-seq profiles of aging female mouse hypothalamus to ensure robustness. The model interpretation analysis revealed potential age-associated genes and cell-type-specific aging signatures. We next successfully implemented our pipeline on a scRNA-seq dataset from male mouse SVZ, demonstrating the generalizability of binarization across different models, techniques, sexes, and brain regions. The model built on control animals also captured the rejuvenating effects of exercise in proliferating NSCs in the SVZ, suggesting potential applications for assessing cellular rejuvenation.

The binarization of sc/snRNA-seq data was first introduced to infer gene regulatory networks[65] and then implemented in cell clustering,[66] trajectory inference,[67] and differential expression analysis.[68] Binary discretization has several advantages: it helps denoise inherently sparse data while preserving biological heterogeneity[50]; it takes fewer computational resources compared to the raw data[50]; and, from a biological perspective, the binarized single-cell gene expression data resemble the intrinsically sparse and binarized single-cell DNA methylation (DNAm) profiles. To date, bulk DNAm-based clocks have been useful estimators of age,[22] although the underlying mechanisms are still largely unknown. Interestingly, in *C. elegans*, a species lacking DNA methylation, binarization of bulk RNA-seq data improved the performance of an aging clock.[29] For single-cell DNAm profiles, a recently developed statistical framework accurately tracks the aging trajectory of different cell types.[23] In the future, single-cell multi-omics (RNA-seq and DNAm) studies profiling both modalities[69–71] in the context of aging will help reveal the relationship between the binarized RNA-seq and DNAm profiles in the same cell types. Such advances will inform the mechanisms underlying single-cell DNAm clocks and guide clock-derived perturbations at the expression level. For example, if CpG sites in the DNAm clock correspond to essential genes in the binarized RNA-seq clock, we can potentially perturb their expression by CRISPR or shRNA tools to interrogate the aging process.

Figure 5. Data binarization improves model performance in an independent mouse SVZ dataset and captures exercise-induced rejuvenation in proliferating NSCs

(A) Bar plot comparing ELN and MLP test performance before (orange) and after (blue) binarization over 10 random seeds. Data are represented as mean ± SD.

(B) The representative PR curves for baseline (the AUPRC score when the model predicts all cells to be in the same major category in the test set; dashed line) and models built with a binarized gene expression matrix.

(C) Bar plot showing the baseline (the AUPRC score when the model predicts all cells to be in the same major category in the test set; green) and test (blue) performance in the cell-type-specific ELN models over 10 random seeds. Data are represented as mean ± SD.

(D) The representative PR curves for baseline (the AUPRC score when the model predicts all cells to be in the same major category in the test set; dashed line) and models built with a binarized gene expression matrix.

(E) Violin plots showing the predicted probability of being aged (neuroblasts, top; aNSC_NPCs, bottom) grouped by condition (left) and animal (right) separately. The white dot represents the median. The p values above the violin plots are derived from a mixed-effect linear model test accounting for group structure. See also Figures S8 and S9.

Feature selection is important in our binarization pipeline. After binarization, the HVGs outperformed the HEGs expression matrix. In other words, the HVGs are more responsive to binarization than HEGs in this classification task. A simulation study has shown that marker genes (i.e., HVGs between cell types) in the raw single-cell count matrix are zero inflated. The zeros are driven by biological differences, which are preserved after binarization.[50] In contrast, HEGs are more stably expressed and not zero inflated, so their change with biological conditions may not be captured by binarization. Our findings indicate that many potential aging features are embedded in the top HVGs. These features are preserved and even enhanced after binarization and then captured by the ML models. On the other hand, HEGs are likely to be housekeeping genes whose fluctuation with age may not be useful in ML models after binarization. In addition to HEGs and HVGs, other feature sets, including differentially expressed genes and phenotypic features, may benefit the model's performance.

One advantage of the all-cell CellBiAge model is that it can predict the age group of single cells without relying on cell type annotation. This versatility becomes particularly beneficial for datasets that lack comprehensive cell type information, especially in the context of rejuvenating methods like cellular reprogramming, which lead to the emergence of highly heterogeneous cell populations.[72]

Our implementation of cell-type-specific models revealed the unique top features in the CellBiAge prediction (Figures 4E and 4G). For example, top features in microglia have genes that regulate the inflammatory response, whereas top features in neurons regulate neurogenesis (Figure 4E). Such unique features in different cell types highlight the heterogeneity of the aging process within individuals, which is in line with the findings in other aging clocks (models) at single-cell resolution.[23,39] Thus, assessing normal aging and age-associated diseases at the single-cell level will help elucidate vulnerable and resilient cell types or tissues with age, which will guide the development of precision medicine. For example, in the mouse SVZ dataset, we found a decreased number of proliferating NSCs during normal aging, consistent with findings demonstrating decreased neurogenesis with age.[16,73–75] Interestingly, our model showed that exercise led to a significant reduction in the predicted aged probability of these proliferating NSCs (Figure 5E), indicating the potential of single-cell level evaluation to improve our understanding of the mechanisms underlying anti-aging interventions. Moreover, the *in silico* perturbation analysis offers insights into important genes in the model prediction process, unveiling cell-type-specific signatures that may not be discernible through conventional bioinformatics analysis. This additional layer of analysis enhances the conceptual understanding of the underlying biology of aging and provides a comprehensive view of the predictive model's behavior.

In sum, CellBiAge is an easy-to-implement, generalizable, and interpretable ML pipeline that enables robust prediction of organismal age groups in all cells and specific cell types using sc/snRNA-seq data. Successful capture of the rejuvenating effects of exercise in proliferating NSCs highlights the potential of this pipeline to quickly evaluate pro-longevity interventions at single-cell resolution. In future work, CellBiAge may be further implemented to classify other conditions such as sex, genotypes, and disease stages. Thus, the interoperable linear models used in CellBiAge offer a unique perspective for identifying previously unknown genes related to various conditions, including aging.

Limitations of the study

One limitation of the study is that the classification model does not capture the continuous aging trajectory, in part due to the cost of single-cell experiments across multiple time points at large scales. A recent study profiled the transcriptome of 21,458 single nuclei for mouse SVZ, a neurogenic niche, across 26 time points ranging from 3-month-old to 29-month-old mice.[39] To overcome the limited number of nuclei, they implemented bootstrapping and ensemble methods in a regularized linear regression model.[39] Encouraged by this, an exciting future direction will be time-course single-cell profiling of the hypothalamus and implementation of data binarization to build a regression model to quantify the aging trajectory.

While using top HVGs as features in the pipeline is convenient, as domain knowledge is not required for feature selection, these genes are often cell type specific, potentially missing "traditional" age-associated genes. Consistent with this, our model interpretation did not reveal an obvious age-associated pathway or gene set. Given the complexity of defining aging and age-related genes,[76] further experimental perturbations based on age-related readouts are necessary to validate the genes that we identified and their relationship with aging. Additionally, the region-specific nature of models trained on top HVGs suggests the need for alternative feature sets to capture cross-region aging signatures. Incorporating phenotypic features such as behavioral and frailty scores could enhance model generalizability across brain, paving the way for biomarkers that predict mortality rather than relying solely on chronological age.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - The mouse hypothalamus training and test batches
  - Preprocessing and feature selection in Seurat
  - Group-based cross-validation strategy and model testing
  - In silico perturbation
  - Permutation importance
  - Cell-type-specific interpretation
  - Pipeline application to the mouse SVZ control and exercise datasets
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2023.113500.

ACKNOWLEDGMENTS

We thank members of the Webb laboratory and Singh laboratory, especially Kelsey R. Babcock and Dr. Abigail K. Brown, for providing critical feedback. This work was supported by an NSF/DBI 2213824 Biology Integration Institute award to A.E.W. and R.S. and NIH/NIA R21 AG070527 to A.E.W. D.Y. was funded by NIH/NIA F99AG083292, Dr. Achilles Frangistas Fund for Neurodegeneration Research and the Mahoney Fund from the Carney Institute for Brain Science, and the Open Graduate Education Fellowship from the Graduate School at Brown. K.H.H. was funded by a Neustein Graduate Fellowship from the Carney Institute for Brain Science at Brown University. This research was conducted using computational resources and services at the Center for Computation and Visualization at Brown University. Schematic figures were made with BioRender.

AUTHOR CONTRIBUTIONS

D.Y., R.S., and A.E.W. conceptualized and planned the study. D.Y. performed all experiments unless indicated below. M.L., G.L., and Y.H. wrote the scripts for the initial MLP pipeline and implemented binarization. K.H.H. generated and analyzed the mouse hypothalamus data and generated the graphical abstract. A.W. and R.S. provided insights in pipeline optimization and binarization. D.Y., R.S., and A.E.W. wrote the manuscript, and all authors reviewed and provided comments.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: May 4, 2023
Revised: October 20, 2023
Accepted: November 13, 2023
Published: November 30, 2023

REFERENCES

1. Niccoli, T., and Partridge, L. (2012). Ageing as a Risk Factor for Disease. Curr. Biol. *22*, R741–R752.
2. Luu, J., and Palczewski, K. (2018). Human aging and disease: Lessons from age-related macular degeneration. Proc National Acad Sci *115*, 2866–2872.
3. Rutledge, J., Oh, H., and Wyss-Coray, T. (2022). Measuring biological age using omics data. Nat. Rev. Genet. *1*.
4. Hofmann, J.W., Zhao, X., De Cecco, M., Peterson, A.L., Pagliaroli, L., Manivannan, J., Hubbard, G.B., Ikeno, Y., Zhang, Y., Feng, B., et al. (2015). Reduced Expression of MYC Increases Longevity and Enhances Healthspan. Cell *160*, 477–488.
5. Webb, A.E., Kundaje, A., and Brunet, A. (2016). Characterization of the direct targets of FOXO transcription factors throughout evolution. Aging Cell *15*, 673–685.
6. 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. (2020). Meta-Analysis of the Alzheimer's Disease Human Brain Transcriptome and Functional Dissection in Mouse Models. Cell Rep. *32*, 107908.
7. Brown, A.K., Maybury-Lewis, S.Y., and Webb, A.E. (2021). Integrative multi-omics analysis reveals conserved hierarchical mechanisms of FOXO3 pioneer-factor activity. Preprint at bioRxiv.
8. Hajdarovic, K.H., Yu, D., Hassell, L.-A., Evans, S., Packer, S., Neretti, N., and Webb, A.E. (2022). Single-cell analysis of the aging female mouse hypothalamus. Nat. Aging *2*, 662–678.
9. Maybury-Lewis, S.Y., Brown, A.K., Yeary, M., Sloutskin, A., Dhakal, S., Juven-Gershon, T., and Webb, A.E. (2021). Changing and stable chromatin accessibility supports transcriptional overhaul during neural stem cell activation and is altered with age. Aging Cell *20*, e13499.
10. Szulwach, K.E., Li, X., Li, Y., Song, C.-X., Wu, H., Dai, Q., Irier, H., Upadhyay, A.K., Gearing, M., Levey, A.I., et al. (2011). 5-hmC–mediated epigenetic dynamics during postnatal neurodevelopment and aging. Nat. Neurosci. *14*, 1607–1616.
11. Yang, J.-H., Hayano, M., Griffin, P.T., Amorim, J.A., Bonkowski, M.S., Apostolides, J.K., Salfati, E.L., Blanchette, M., Munding, E.M., Bhakta, M., et al. (2023). Loss of epigenetic information as a cause of mammalian aging. Cell *186*, 305–326.e27.
12. Benayoun, B.A., Pollina, E.A., Singh, P.P., Mahmoudi, S., Harel, I., Casey, K.M., Dulken, B.W., Kundaje, A., and Brunet, A. (2019). Remodeling of epigenome and transcriptome landscapes with aging in mice reveals widespread induction of inflammatory responses. Genome Res. *29*, 697–709.
13. Martinez-Miguel, V.E., Lujan, C., Espie–Caullet, T., Martinez-Martinez, D., Moore, S., Backes, C., Gonzalez, S., Galimov, E.R., Brown, A.E.X., Halic, M., et al. (2021). Increased fidelity of protein synthesis extends lifespan. Cell Metab. *33*, 2288–2300.e12.
14. Audesse, A.J., Dhakal, S., Hassell, L.-A., Gardell, Z., Nemtsova, Y., and Webb, A.E. (2019). FOXO3 directly regulates an autophagy network to functionally regulate proteostasis in adult neural stem cells. PLoS Genet. *15*, e1008097.
15. Leeman, D.S., Hebestreit, K., Ruetz, T., Webb, A.E., McKay, A., Pollina, E.A., Dulken, B.W., Zhao, X., Yeo, R.W., Ho, T.T., et al. (2018). Lysosome activation clears aggregates and enhances quiescent neural stem cell activation during aging. Science *359*, 1277–1283.
16. Babcock, K.R., Page, J.S., Fallon, J.R., and Webb, A.E. (2021). Adult hippocampal neurogenesis in aging and Alzheimer's disease. Stem Cell Rep. *16*, 681–693.
17. Kimmel, J.C., Yi, N., Roy, M., Hendrickson, D.G., and Kelley, D.R. (2021). Differentiation reveals latent features of aging and an energy barrier in murine myogenesis. Cell Rep. *35*, 109046.
18. Yeo, R.W., Zhou, O.Y., Zhong, B.L., Sun, E.D., Negredo, P.N., Nair, S., Sharmin, M., Ruetz, T.J., Wilson, M., Kundaje, A., et al. (2023). Chromatin accessibility dynamics of neurogenic niche cells reveal defects in neural stem cell adhesion and migration during aging. Nat. Aging *3*, 866–893. https://doi.org/10.1038/s43587-023-00449-3.
19. Schultz, M.B., Kane, A.E., Mitchell, S.J., MacArthur, M.R., Warner, E., Vogel, D.S., Mitchell, J.R., Howlett, S.E., Bonkowski, M.S., and Sinclair, D.A. (2020). Age and life expectancy clocks based on machine learning analysis of mouse frailty. Nat. Commun. *11*, 4618.
20. Bocklandt, S., Lin, W., Sehl, M.E., Sánchez, F.J., Sinsheimer, J.S., Horvath, S., and Vilain, E. (2011). Epigenetic Predictor of Age. PLoS One *6*, e14821.
21. Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., Klotzle, B., Bibikova, M., Fan, J.-B., Gao, Y., et al. (2013). Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. Mol. Cell *49*, 359–367.
22. Horvath, S. (2013). DNA methylation age of human tissues and cell types. Genome Biol. *14*, R115.
23. Trapp, A., Kerepesi, C., and Gladyshev, V.N. (2021). Profiling epigenetic age in single cells. Nat. Aging *1*, 1189–1201.
24. de Lima Camillo, L.P., Lapierre, L.R., and Singh, R. (2022). A pan-tissue DNA-methylation epigenetic clock based on deep learning. Npj Aging *8*, 4.
25. Yang, Z., Wong, A., Kuh, D., Paul, D.S., Rakyan, V.K., Leslie, R.D., Zheng, S.C., Widschwendter, M., Beck, S., and Teschendorff, A.E. (2016). Correlation of an epigenetic mitotic clock with cancer risk. Genome Biol. *17*, 205.
26. Levine, M.E., Lu, A.T., Quach, A., Chen, B.H., Assimes, T.L., Bandinelli, S., Hou, L., Baccarelli, A.A., Stewart, J.D., Li, Y., et al. (2018). An epigenetic

biomarker of aging for lifespan and healthspan. Aging Albany Ny *10*, 573–591.

27. Lu, A.T., Quach, A., Wilson, J.G., Reiner, A.P., Aviv, A., Raj, K., Hou, L., Baccarelli, A.A., Li, Y., Stewart, J.D., et al. (2019). DNA methylation GrimAge strongly predicts lifespan and healthspan. Aging Albany Ny *11*, 303–327.

28. Peters, M.J., Joehanes, R., Pilling, L.C., Schurmann, C., Conneely, K.N., Powell, J., Reinmaa, E., Sutphin, G.L., Zhernakova, A., Schramm, K., et al. (2015). The transcriptional landscape of age in human peripheral blood. Nat. Commun. *6*, 8570.

29. Meyer, D.H., and Schumacher, B. (2021). BiT age: A transcriptome-based aging clock near the theoretical limit of accuracy. Aging Cell *20*, e13320.

30. Tanaka, T., Biancotto, A., Moaddel, R., Moore, A.Z., Gonzalez-Freire, M., Aon, M.A., Candia, J., Zhang, P., Cheung, F., Fantoni, G., et al. (2018). Plasma proteomic signature of age in healthy humans. Aging Cell *17*, e12799.

31. Lehallier, B., Gate, D., Schaum, N., Nanasi, T., Lee, S.E., Yousef, H., Moran Losada, P., Berdnik, D., Keller, A., Verghese, J., et al. (2019). Undulating changes in human plasma proteome profiles across the lifespan. Nat. Med. *25*, 1843–1850.

32. Deelen, J., Kettunen, J., Fischer, K., van der Spek, A., Trompet, S., Kastenmüller, G., Boyd, A., Zierer, J., van den Akker, E.B., Ala-Korpela, M., et al. (2019). A metabolic profile of all-cause mortality risk identified in an observational study of 44,168 individuals. Nat. Commun. *10*, 3346.

33. Robinson, O., Chadeau Hyam, M., Karaman, I., Climaco Pinto, R., Ala-Korpela, M., Handakas, E., Fiorito, G., Gao, H., Heard, A., Jarvelin, M.R., et al. (2020). Determinants of accelerated metabolomic and epigenetic aging in a UK cohort. Aging Cell *19*, e13149.

34. Schaum, N., Lehallier, B., Hahn, O., Pálovics, R., Hosseinzadeh, S., Lee, S.E., Sit, R., Lee, D.P., Losada, P.M., Zardeneta, M.E., et al. (2020). Ageing hallmarks exhibit organ-specific temporal signatures. Nature *583*, 596–602.

35. Tabula Muris Consortium; Antony, J., Baghel, A.S., Bakerman, I., Bansal, I., Barres, B.A., Beachy, P.A., Berdnik, D., Bilen, B., Brownfield, D., et al. (2020). A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. Nature *583*, 590–595.

36. Işıldak, U., Somel, M., Thornton, J.M., and Dönertaş, H.M. (2020). Temporal changes in the gene expression heterogeneity during brain development and aging. Sci. Rep. *10*, 4080.

37. Singh, S.P., Janjuha, S., Chaudhuri, S., Reinhardt, S., Kränkel, A., Dietz, S., Eugster, A., Bilgin, H., Korkmaz, S., Zararsız, G., et al. (2018). Machine learning based classification of cells into chronological stages using single-cell transcriptomics. Sci. Rep. *8*, 17156.

38. Lu, J., Ahmad, R., Nguyen, T., Cifello, J., Hemani, H., Li, J., Chen, J., Li, S., Wang, J., Achour, A., et al. (2022). Heterogeneity and transcriptome changes of human CD8+ T cells across nine decades of life. Nat. Commun. *13*, 5128.

39. Buckley, M.T., Sun, E.D., George, B.M., Liu, L., Schaum, N., Xu, L., Reyes, J.M., Goodell, M.A., Weissman, I.L., Wyss-Coray, T., et al. (2023). Cell-type-specific aging clocks to quantify aging and rejuvenation in neurogenic regions of the brain. Nat. Aging *3*, 121–137.

40. Hajdarovic, K.H., Yu, D., and Webb, A.E. (2022). Understanding the aging hypothalamus, one cell at a time. Trends Neurosci. *45*, 942–954.

41. Satoh, A., Brace, C.S., Rensing, N., Cliften, P., Wozniak, D.F., Herzog, E.D., Yamada, K.A., and Imai, S.I. (2013). Sirt1 Extends Life Span and Delays Aging in Mice through the Regulation of Nk2 Homeobox 1 in the DMH and LH. Cell Metab. *18*, 416–430.

42. Acosta-Rodríguez, V., Rijo-Ferreira, F., Izumo, M., Xu, P., Wight-Carter, M., Green, C.B., and Takahashi, J.S. (2022). Circadian alignment of early onset caloric restriction promotes longevity in male C57BL/6J mice. Sci New York N Y *376*, 1192–1202.

43. Saito, T., and Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLoS One *10*, e0118432.

44. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. Cell *184*, 3573–3587.e29.

45. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. Cell *177*, 1888–1902.e21.

46. Brennecke, P., Anders, S., Kim, J.K., Kołodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S.A., Marioni, J.C., and Heisler, M.G. (2013). Accounting for technical noise in single-cell RNA-seq experiments. Nat. Methods *10*, 1093–1095.

47. Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. *19*, 15.

48. Ke, Q., Dinalankara, W., Younes, L., Geman, D., and Marchionni, L. (2021). Efficient representations of tumor diversity with paired DNA-RNA aberrations. PLoS Comput. Biol. *17*, e1008944.

49. Dinalankara, W., Ke, Q., Xu, Y., Ji, L., Pagane, N., Lien, A., Matam, T., Fertig, E.J., Price, N.D., Younes, L., et al. (2018). Digitizing omics profiles by divergence from a baseline. Proc. Natl. Acad. Sci. *115*, 4545–4552.

50. Bouland, G.A., Mahfouz, A., and Reinders, M.J.T. (2022). The rise of sparser single-cell RNAseq datasets; consequences and opportunities. preprint at bioRxiv. https://doi.org/10.1101/2022.05.20.492823.

51. Ji, L., Wang, A., Sonthalia, S., Naiman, D.Q., Younes, L., Colantuoni, C., and Geman, D. (2023). CellCover Defines Conserved Cell Types and Temporal Progression in scRNA-seq Data across Mammalian Neocortical Development. preprint at bioRxiv. https://doi.org/10.1101/2023.04.06.535943.

52. Luecken, M.D., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Mueller, M.F., Strobl, D.C., Zappia, L., Dugas, M., Colomé-Tatché, M., and Theis, F.J. (2021). Benchmarking atlas-level data integration in single-cell genomics. Nat. Methods *19*, 41–50.

53. Ruetz, T.J., Kashiwagi, C.M., Morton, B., Yeo, R.W., Leeman, D.S., Morgens, D.W., Tsui, C.K., Li, A., Bassik, M.C., and Brunet, A. (2021). In vitro and in vivo CRISPR-Cas9 screens reveal drivers of aging in neural stem cells of the brain. preprint at bioRxiv. https://doi.org/10.1101/2021.11.23.469762.

54. Sohn, E.H., Flamme-Wiese, M.J., Whitmore, S.S., Wang, K., Tucker, B.A., and Mullins, R.F. (2014). Loss of CD34 Expression in Aging Human Choriocapillaris Endothelial Cells. PLoS One *9*, e86538.

55. Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. J Royal Statistical Soc Ser B Statistical Methodol *67*, 301–320.

56. Codega, P., Silva-Vargas, V., Paul, A., Maldonado-Soto, A.R., DeLeo, A.M., Pastrana, E., and Doetsch, F. (2014). Prospective Identification and Purification of Quiescent Adult Neural Stem Cells from Their In Vivo Niche. Neuron *82*, 545–559.

57. Golomb, S.M., Guldner, I.H., Zhao, A., Wang, Q., Palakurthi, B., Aleksandrovic, E.A., Lopez, J.A., Lee, S.W., Yang, K., and Zhang, S. (2020). Multimodal Single-Cell Analysis Reveals Brain Immune Landscape Plasticity during Aging and Gut Microbiota Dysbiosis. Cell Rep. *33*, 108438.

58. Murase, S.I., Cho, C., White, J.M., and Horwitz, A.F. (2008). ADAM2 promotes migration of neuroblasts in the rostral migratory stream to the olfactory bulb. Eur. J. Neurosci. *27*, 1585–1595.

59. Gentile, G., La Cognata, V., and Cavallaro, S. (2021). The contribution of CNVs to the most common aging-related neurodegenerative diseases. Aging Clin. Exp. Res. *33*, 1187–1195.

60. Liu, L., Kim, S., Buckley, M.T., Reyes, J.M., Kang, J., Tian, L., Wang, M., Lieu, A., Mao, M., Rodriguez-Mateo, C., et al. (2023). Exercise reprograms the inflammatory landscape of multiple stem cell compartments during mammalian aging. Cell Stem Cell *30*, 689–705.e4.

61. van Praag, H., Shubert, T., Zhao, C., and Gage, F.H. (2005). Exercise Enhances Learning and Hippocampal Neurogenesis in Aged Mice. J. Neurosci. *25*, 8680–8685.

62. van Praag, H., Christie, B.R., Sejnowski, T.J., and Gage, F.H. (1999). Running enhances neurogenesis, learning, and long-term potentiation in mice. Proc National Acad Sci *96*, 13427–13431.

63. Jokai, M., Torma, F., McGreevy, K.M., Koltai, E., Bori, Z., Babszki, G., Bakonyi, P., Gombos, Z., Gyorgy, B., Aczel, D., et al. (2023). DNA methylation clock DNAmFitAge shows regular exercise is associated with slower aging and systemic adaptation. GeroScience *45*, 2805–2817.

64. McGreevy, K.M., Radak, Z., Torma, F., Jokai, M., Lu, A.T., Belsky, D.W., Binder, A., Marioni, R.E., Ferrucci, L., Pośpiech, E., et al. (2023). DNAmFitAge: biological age indicator incorporating physical fitness. Aging *15*, 3904–3938.

65. Moignard, V., Woodhouse, S., Haghverdi, L., Lilly, A.J., Tanaka, Y., Wilkinson, A.C., Buettner, F., Macaulay, I.C., Jawaid, W., Diamanti, E., et al. (2015). Decoding the regulatory network of early blood development from single-cell gene expression measurements. Nat. Biotechnol. *33*, 269–276.

66. Qiu, P. (2020). Embracing the dropouts in single-cell RNA-seq analysis. Nat. Commun. *11*, 1169.

67. Li, R., and Quon, G. (2019). scBFA: modeling detection patterns to mitigate technical noise in large-scale single-cell genomics data. Genome Biol. *20*, 193.

68. Bouland, G.A., Mahfouz, A., and Reinders, M.J.T. (2021). Differential analysis of binarized single-cell RNA sequencing data captures biological variation. NAR Genom. Bioinform. *3*, lqab118.

69. Angermueller, C., Clark, S.J., Lee, H.J., Macaulay, I.C., Teng, M.J., Hu, T.X., Krueger, F., Smallwood, S., Ponting, C.P., Voet, T., et al. (2016). Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. Nat. Methods *13*, 229–232.

70. Hu, Y., Huang, K., An, Q., Du, G., Hu, G., Xue, J., Zhu, X., Wang, C.-Y., Xue, Z., and Fan, G. (2016). Simultaneous profiling of transcriptome and DNA methylome from a single cell. Genome Biol. *17*, 88.

71. Clark, S.J., Argelaguet, R., Kapourani, C.-A., Stubbs, T.M., Lee, H.J., Alda-Catalinas, C., Krueger, F., Sanguinetti, G., Kelsey, G., Marioni, J.C., et al. (2018). scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. Nat. Commun. *9*, 781.

72. Hanna, J.H., Saha, K., and Jaenisch, R. (2010). Pluripotency and Cellular Reprogramming: Facts, Hypotheses, Unresolved Issues. Cell *143*, 508–525.

73. Enwere, E., Shingo, T., Gregg, C., Fujikawa, H., Ohta, S., and Weiss, S. (2004). Aging Results in Reduced Epidermal Growth Factor Receptor Signaling, Diminished Olfactory Neurogenesis, and Deficits in Fine Olfactory Discrimination. J. Neurosci. *24*, 8354–8365.

74. Dulken, B.W., Buckley, M.T., Navarro Negredo, P., Saligrama, N., Cayrol, R., Leeman, D.S., George, B.M., Boutet, S.C., Hebestreit, K., Pluvinage, J.V., et al. (2019). Single-cell analysis reveals T cell infiltration in old neurogenic niches. Nature *571*, 205–210.

75. Dulken, B.W., Leeman, D.S., Boutet, S.C., Hebestreit, K., and Brunet, A. (2017). Single-Cell Transcriptomic Analysis Defines Heterogeneity and Transcriptional Dynamics in the Adult Neural Stem Cell Lineage. Cell Rep. *18*, 777–790.

76. Moqri, M., Herzog, C., Poganik, J.R., Biomarkers of Aging Consortium; Justice, J., Belsky, D.W., Higgins-Chen, A., Moskalev, A., Fuellen, G., Cohen, A.A., et al. (2023). Biomarkers of aging for the identification and evaluation of longevity interventions. Cell *186*, 3758–3775.

77. Faraway, J.J. (2019). Extending the Linear Model with R, Generalized Linear, Mixed Effects and Nonparametric Regression Models, Second edition (Chapman and Hall/CRC).

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Mouse hypothalamus snRNAseq data	Hajdarovic et al.[8]	GEO: GSE188646
Mouse exercise SVZ scRNAseq data	Buckley et al.[39]	GEO: GSE196364
Software and algorithms		
R	https://cran.r-project.org/	4.2.1
Tidyverse	https://www.tidyverse.org/	1.3.2
ggplot2	https://ggplot2.tidyverse.org/	3.3.6
Seurat	https://satijalab.org/seurat/articles/install.html	4.1.1
Python	https://www.python.org/	3.7.10, 3.9.13
Scikit-learn	https://scikit-learn.org/stable/	0.24.2
TensorFlow	https://www.tensorflow.org/	1.14.0,
Keras	https://pypi.org/project/keras/	2.9.0
keras-tuner	https://pypi.org/project/keras-tuner/	1.1.3
Seaborn	https://seaborn.pydata.org/	0.11.1
Matplotlib	https://matplotlib.org/	3.4.2
CellBiAge code and virtual environments	This paper	https://github.com/Webb-Laboratory/CellBiAge

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Ashley E. Webb (awebb@buckinstitute.org).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Single-cell RNA-seq data have been deposited at GEO and are publicly available. Accession numbers are listed in the key resources table. The processed csv files for model training and test have been deposited at GitHub and are publicly available.
- All original code has been deposited at https://github.com/Webb-Laboratory/CellBiAge and archived on Zenodo https://doi.org/10.5281/zenodo.10072378and is publicly available.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Datasets used in this study were previously published and are publicly available. See the key resources table for accession numbers.

METHOD DETAILS

The mouse hypothalamus training and test batches

The snRNA-seq dataset from the young and aged mouse hypothalamus was generated and described in Hadjarovic and Yu et al.[8] Briefly, Young (3 month) and aged (19–24 month) C57BL/6 female mice were obtained from the NIA. The test batch was generated using the older 10x Chromium Single Cell 3′ gene expression kit (version 2) and sequenced on the Illumina HiSeq, whereas the training batch was made using the 10x 3′ kits version 3 and Illumina NovaSeq). We assigned the latter dataset to the training batch because of its richer sequencing information compared to the test batch.

Preprocessing and feature selection in Seurat

The training and test sets were preprocessed separately to prevent information leakage between the two sets. Data filtration and quality control were described and published.[8] Samples in training (animals #1–4) and test (animals #5–8 in Figures 1B; S4A) batches were preprocessed and integrated separately (preprocessing.R). Log normalization was applied to the read counts in all nuclei to remove biases in sequencing coverage between nuclei. Data integration was performed to integrate samples on the top 2k HVGs in the same batch to reduce the impact of potential confounding batch variables on the task. Data scaling was applied to enable the mean to equal 0 and standard deviation to equal 1, which standardized the range of features (genes).

For each batch, the top 2k HEGs were selected after ranking the log-normalized gene expression values. The top 2k HVGs were selected in the integration step and served as anchor features. The log-normalized matrix and the scaled log-normalized matrix were stored in the RNA data and scale.data slots of the Seurat object. The scaled log-normalized and integrated gene expression matrix was stored in the integrated scale.data slot. For the gene expression matrix derived from the "integrated" data, the preprocessed matrix was log-normalized, integrated, and then scaled, which is the matrix that was binarized. Data binarization updates the entry values larger than 0 (the mean value after standardization) as 1 and else as 0. This was implemented in the user-defined binarize_data() function in Python.

The intersection of the top 2k genes between the training and test batches were selected in the final pipeline. In total, 1,413 HVGs were shared between the training and test batches.

Group-based cross-validation strategy and model testing

Scikit-learn GridSearchCV was implemented to select the best hyperparameter combinations in ML models, and TensorFlow KerasTuner was used for the MLP model in the customized group-based cross-validation. The target variable is whether the nucleus is from the young or aged category, and the output is the probability of the nucleus originating from the aged category. AUPRC scores were used as the evaluation metrics because of the imbalance of the two classes in the test set.

The range of hyperparameter combinations was fine-tuned and evaluated by the mean AUPRC scores of the four validation sets in the cross-validation step. The combinations that yielded the highest mean AUPRC scores were selected for the final model training and testing. Models were then trained using all four samples in the training set, and tested with the previously unseen held out test batch over 10 trials using different random seeds. Hyperparameters were tuned individually for each preprocessing method combination.

In silico perturbation

The 1,413 HVGs were first ranked by the mean absolute values of their ELN coefficients over 10 trials. Starting from the gene with the lowest absolute coefficient, the entry values of the perturbed gene were shuffled over all nuclei, and AUPRC score of the shuffled matrix was calculated, repeated for 10 runs. The perturbation was repeatedly applied to all genes until the last one with the highest absolute coefficient was perturbed. The perturbation was performed in a cumulative way, such that after 1,413 perturbations, all genes were shuffled. The perturbation pipeline was repeated in a reverse manner where the gene with highest coefficient was shuffled first. To characterize the change of the AUPRC score in the perturbation, a fitted curve was interpolated from the AUPRC curve and its derivatives were calculated.

Permutation importance

The 1,413 HVGs were shuffled individually in a non-cumulative manner. For each shuffled gene, the entry values were shuffled over all nuclei, and AUPRC scores of the shuffled matrix were calculated, repeated for 10 runs.

Cell-type-specific interpretation

Annotations of major cell types were described and published in Hajdarovic and Yu, et al.[8] Cell-type-specific ELN models were trained and tuned using the same group-based cross-validation scheme, by cell types. The type-specific interpretation was performed with the same scripts for the all-cell models. For the Venn diagram, inputs are lists of top 200 genes with the highest absolute values for each cell type, and the result was plotted using https://bioinformatics.psb.ugent.be/webtools/Venn/.

In addition to the cell-type-specific models, for the all-cell trained model, the test performance was broken down by cell types in the ELN and MLP models separately. Briefly, in the held-out test set, specific cell types were subset to test the all-cell trained model and calculate the AUPRC scores.

Pipeline application to the mouse SVZ control and exercise datasets

Fastq files were aligned and preprocessed using Cell Ranger Count and Seurat. Batch integration was performed for training-controls (animals O5, O7, Y5, Y7), training-controls and test-controls (animals O5, O7, Y5, Y7; O2, Y1, Y2), training-controls and test-exercise-Day1 (animals O5, O7, Y5, Y7; O6, O8, Y6, Y8), and training-controls and test-exercise-Day2 (animals O5, O7, Y5, Y7; O3, O4, Y3, Y4) (Figure S8A). The top 2k HVGs from each integrated dataset were selected. The shared 1,617 genes were used for model training and testing. Annotations of major cell types were described and published in Buckley and Sun et al.[39] Only cells that were annotated previously were kept. The training-controls dataset was used for training, and group-based cross validation was performed to select hyperparameters. Testing sets were subset from the integrated data.

QUANTIFICATION AND STATISTICAL ANALYSIS

Pearson correlation coefficients were calculated to determine the correlation between the 1,413 genes in Python. Plots were generated by Matplotlib and Seaborn in Python, and ggplot2 in R. Mixed-effect linear model accounting for group structure (group = animal) was applied to determine if exercise had a significant effect on predicted age probability. This method models both within-group and between-group variations and accounts for the non-independency of cells from the same animals.[77] The assumption of normality of residues was checked before performing the test.