Biolearn, an open-source library for biomarkers of aging

Kejun Ying[1,2,†], Seth Paulson[3,†], Alec Eames[1], Siyuan Li[1], Martin Perez-Guevara[4], Mehrnoosh Emamifar[5], Maximiliano Casas Martínez[6], Dayoon Kwon[7], Michael Snyder[8], Dane Gobel[3], Jesse R. Poganik[1], Mahdi Moqri[1,8,*], Vadim N. Gladyshev[1,*]

[1] Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA

[2] T. H. Chan School of Public Health, Harvard University, Boston, MA, USA

[3] Methuselah Foundation, Washington, DC, USA

[4] Mind Operating Systems, Paris, FR

[5] College of Engineering, Northeastern University, Boston, MA, USA

[6] Division of Exact Sciences, Department of Mathematics, Instituto Tecnológico Autónomo de México, Mexico City, Mexico

[7] Department of Epidemiology, UCLA Fielding School of Public Health, Los Angeles, CA, USA

[8] Department of Genetics, Stanford University, Stanford, CA, USA

[†] Contributed equally

[*] email: vgladyshev@bwh.harvard.edu, mmoqri@bwh.harvard.edu

Abstract

Identifying and validating biomarkers of aging is pivotal for understanding the aging process and testing longevity interventions. Despite the development of numerous aging biomarkers, their clinical validation remains elusive, largely due to the lack of cross-population validation, which is hampered by disparate biomarker designs and inconsistencies in dataset structures. To bridge this gap, we introduce Biolearn, an innovative open-source library dedicated to the implementation and application of aging biomarkers. Biolearn facilitates (1) harmonization of existing aging

biomarkers, while presenting a structured framework for novel biomarkers in standardized formats; (2) unification of public datasets, ensuring coherent structuring and formatting, thus simplifying cross-population validation studies; and (3) provision of computational methodologies to assess any harmonized biomarker against unified datasets. By furnishing a community-driven platform, Biolearn significantly streamlines the development, assessment, and validation of aging biomarkers, paving the way toward more rigorous clinical validation and, ultimately, application in clinical trials targeting healthy longevity. The Biolearn package is open-source and freely available at https://Bio-Learn.github.io/

Introduction

The quest for reliable biomarkers of aging (BoAs) has become a central focus in aging research, driven by the growing recognition of aging as a fundamental driver of chronic diseases and mortality. Numerous biomarkers have been proposed to quantify biological age and elucidate the underlying biological processes of aging. However, their clinical validation remains a significant challenge due to the heterogeneity in their formulations and the disparate structures of validation datasets across populations [1,2].

Since the introduction of composite omic biomarkers of aging, exemplified by Horvath's pioneering work on DNA methylation aging clocks [3], the field has witnessed a rapid expansion in the repertoire of aging biomarkers. These biomarkers now encompass a wide array of omic modalities, including epigenomics, transcriptomics, and proteomics [4–9]. Omic biomarkers provide a comprehensive view of the molecular changes associated with aging, offering valuable insights into the aging process and its impact on human health.

Among the various classes of omic biomarkers, DNA methylation-based clocks have emerged as the most advanced and robust tools for estimating biological age. These clocks, such as the Horvath multi-tissue clock, DunedinPACE [5], GrimAge [10], PhenoAge [11], causality-enriched DamAge/AdaptAge [12], and the PRC2 clock [13], have demonstrated significant associations with age-related conditions and mortality, highlighting the intricate relationship between epigenetic modifications and the aging trajectory [6,10,14,15]. However, the diverse formulations of these biomarkers and the inconsistencies in dataset structures across different populations pose substan-

tial challenges for their systematic cross-population validation and benchmarking, which are crucial steps towards their clinical translation.

Publicly available datasets, such as those from the Gene Expression Omnibus (GEO) [16], the National Health and Nutrition Examination Survey (NHANES), and the Framingham Heart Study (FHS), hold immense potential for accelerating the validation of BoAs. However, the lack of a standardized framework that can accommodate the heterogeneous nature of these datasets hinders their effective utilization for this purpose. There is a pressing need for a unified platform that can seamlessly integrate and analyze various BoAs across datasets with harmonized structures. Such a platform would revolutionize the validation process, facilitate the discovery of novel biomarkers, and provide a structured avenue for community-driven efforts in advancing the field of aging biology.

To address this need, we introduce Biolearn, an open-source library that bridges these gaps by providing a unified framework for the harmonization and computational analysis of BoAs (Figure 1a). Biolearn serves as an innovative tool that harmonizes existing BoAs, structures, and formats human datasets, and offers computational methodologies for assessing biomarkers against these datasets. By enabling the integration and analysis of diverse BoAs and datasets, Biolearn aims to accelerate the development and validation of BoAs, fostering a community-driven approach to aging research.

Results

Harmonization of Biomarkers of Aging and Datasets

In this study, we successfully harmonized a comprehensive set of 33 well-established epigenetic aging clocks and biomarkers (Table 1), as well as two phenotypic biomarkers (Phenotypic Age and Mahalanobis distance metrics), and implemented these BoAs in Biolearn. The epigenetic biomarkers encompass a wide range of categories, including: (1) Chronological clocks: Horvath's multi-tissue clock and Hannum's blood clock [3,17]; (2) Healthspan and mortality-related clocks: GrimAge, GrimAge2, PhenoAge, and Zhang clock [10,11,18,19]; (3) Biomarkers of the rate of aging: DunedinPoAm38 and DunedinPACE [5,20]; (4) Causality-enriched clocks: Ying's CausAge, DamAge, and AdaptAge [12]; and (5) Various other clocks, DNAm-based biomarkers, and disease predictors (Table 1).

The harmonization of these diverse biomarkers is crucial for enabling consistent and reproducible analyses across different datasets, facilitating cross-population validation studies, and advancing our understanding of the aging process. To achieve this, all biomarkers were formatted into standardized input structures, ensuring their consistent application across disparate datasets.

The harmonization process involved collecting and unifying the annotation of clock specifications, such as tissue type, predicted age range, and source references. This meticulous approach guarantees transparent and reproducible analyses, enabling researchers to readily compare and interpret results across different studies and populations.

To further support ongoing research in this field, we developed and implemented an open-source framework. This framework provides a standardized format for epigenetic biomarkers, facilitating the seamless integration and comparison of any future aging clocks and biomarkers that are developed.

By providing a unified platform for the harmonization and analysis of aging biomarkers, this framework aims to foster collaboration and innovation in the field of aging. Researchers can easily contribute new biomarkers, compare their performance against existing ones, and explore their potential applications in various datasets. This collaborative approach is essential for accelerating progress in the field and developing more accurate and robust biomarkers of aging.

To facilitate cross-population validation studies using publicly available data, we harnessed Biolearn's capabilities to integrate and structure multiple public datasets (Table 2). The structured datasets were refined to enable a shared analysis platform, addressing the challenges of data heterogeneity and formatting inconsistencies [21,22]. With this capacity, Biolearn is used as the backend of ClockBase for epigenetic age computation [22], enabling the systemic harmonization of over 200,000 human samples from Gene Expression Omnibus (GEO) array data.

Quality Control, Imputation, and Deconvolution

Biolearn provides a comprehensive toolkit for data preprocessing, normalization, and cell type deconvolution (Figure 1a). Quality control metrics, such as sample deviations from the population mean, missingness, and the number of sites with high percentage missingness, can be readily visualized (Figure 1b). This functionality enables researchers to identify potential outliers or problematic samples and make informed decisions about data inclusion and exclusion criteria.

Moreover, Biolearn enables the prediction of sample sex from DNA methylation data with high accuracy, which can be compared against actual sex distributions (Figure 1c,d)[23]. This feature is particularly useful for identifying potential sample mislabeling or investigating sex-specific effects in aging research. We also implemented common traits predictors, including smoking, BMI, and epigenetic scores for diseases like Down Syndrome (Figure 1e)[24]. These predictors allow researchers to explore the associations between aging biomarkers and various lifestyle factors or disease conditions, providing valuable insights into the complex interplay between aging and health.

Missing DNA methylation data can be easily imputed with different methods in just a few lines of code (Figure 1f,g), ensuring that researchers can make the most of the available data and minimize the impact of missing values on their analyses. Biolearn also facilitates the integration of multiple datasets for large-scale analyses, such as the comparison of DNA methylation levels of CpG sites across different datasets (Figure 1h). This feature allows researchers to investigate the consistency and reproducibility of aging biomarkers across diverse populations and experimental settings, strengthening the robustness and generalizability of their findings.

In addition to these features, Biolearn offers a powerful deconvolution tool that estimates the proportion of cell types in a given sample based on a single bulk-level methylation measurement. Biolearn provides two modes for deconvolution, optimized for the 450K (DeconvoluteBlood450K) and EPIC (DeconvoluteBloodEPIC) methylation platforms, respectively [25]. These modes are designed for estimating cell proportions in blood methylation samples and account for technology-specific biases that can affect the accuracy of deconvolution [26]. The reference methylation matrices for each mode consist of methylation profiles for six cell types representing the most abundant cell types found in blood: neutrophils, monocytes, natural killer cells, B cells, CD4+ T cells, and CD8+ T cells [27,28].

We benchmarked the accuracy of our deconvolution tool using datasets with known cell proportions assessed via fluorescence-activated cell sorting and in vitro cell mixing. Both deconvolution methods generated accurate predictions that matched known cell proportions (Figure 1i,j). These results demonstrate the reliability and utility of Biolearn's deconvolution feature, which can help researchers account for cellular heterogeneity in their analyses and gain insights into the cell type-specific contributions to aging biomarkers.

Assessment of Biomarkers of Aging

Biolearn facilitates extensive evaluations of harmonized biomarkers of aging. As an example, we applied 16 representative epigenetic clock models to the Normative Aging Study (NAS) dataset (N = 1488, 38.8% deceased) and compared their performance in predicting chronological age (Figure 2a-c). We demonstrate that Biolearn can efficiently compute the results for epigenetic biomarkers with a few lines of code (Figure 2b). The top-performing clocks in terms of R^2 with chronological age were the Horvathv2, Hannum, CausAge, Horvathv1, and PhenoAge clocks (Figure 2c).

Biolearn also supports mortality risk analyses. We applied the 16 clock models to the NAS dataset and compared their performance in predicting mortality risk using Cox Proportional Hazards analysis, adjusted for age (Figure 2d). The top-performing clock in terms of hazard ratio (per standard deviation increase of the biomarker) is GrimAgeV2 (HR = 1.42, P = 8.57e-19), followed closely by DunedinPACE (HR = 1.41, P = 8.16e-18) and GrimAgeV1 (HR = 1.36, P = 1.01e-15). Other clocks with strong predictive value include YingDamAge (HR = 1.29, P = 6.38e-11), Zhang_10 (HR = 1.28, P = 4.01e-09), and DunedinPoAm38 (HR = 1.25, P = 2.49e-08). These clocks demonstrate the highest associations between the epigenetic age acceleration they measure and the likelihood of mortality in the NAS cohort. The results highlight the potential of these epigenetic clocks as biomarkers for assessing mortality risk.

Besides cohort data, we also applied models to two harmonized GEO datasets GSE41169 (N = 95) and GSE64495 (N = 113), with DNA methylation profiles across approximately 480,000 CpGs (Figure 3)[29]. We demonstrate that data from GEO can be easily loaded by Biolearn for the downstream analysis (Figure 3b).

To provide further utility in handling clinical data, we implemented the blood-test-based Phenotypic Age [11], and Mahalanobis distance-based biomarker in Biolearn [30]. These clinical biomarkers may be combined with the analysis tools built into Biolearn. For example, we calculated Phenotypic Ages for the NHANES 2010 dataset and illustrated the relationship between biological and chronological age using Biolearn. We further performed a survival analysis that distinguished between individuals with accelerated versus decelerated aging based on biological age discrepancies. The entire analysis was completed with only a few lines of code (Figure 4a-d) [11].

Discussion

Among the most significant challenges in aging biomarker research is the cross-population validation of proposed biomarkers [30]. To take steps to address this need and provide an open-source tool for validation efforts across the field, we built Biolearn to integrate a broad range of datasets, including those with varied formats and structures, and provide a suite of analysis tools. By standardizing data inputs and modeling approaches across multiple populations, Biolearn has demonstrated its potential to facilitate more extensive and robust validation processes that are essential for the clinical translation of biomarkers of aging [6].

With Biolearn, we also harmonized and evaluated several well-established aging clocks, providing the opportunity for these biomarkers to be refined and potentially for new ones to be developed. The modular design of Biolearn encourages the addition of new models and datasets, making it a living library that will grow in tandem with the field itself. By centralizing resources and knowledge, Biolearn considerably reduces redundancy and accelerates biomarker development and validation efforts [5,9].

Our approach emphasizes transparency and reproducibility, core tenets of open science. By making Biolearn publicly available and maintaining detailed documentation and development guidelines, we established an ecosystem that supports open collaboration and knowledge sharing. This open-science framework ensures that findings and tools can be widely accessed, providing equitable opportunities for researchers globally to contribute to and benefit from the collective advances in aging research. Moreover, our hope is that the open-access nature of Biolearn will promote cross-fertilization between aging researchers and scientists currently outside of the field, incentivizing the development of novel and innovative biomarker models and validation approaches.

Previous efforts to harmonize biomarkers of aging, notably methylCIPHER and BioAge [31,32], have been limited in scope, focusing on methylation or blood-based biomarkers only. Furthermore, using R packages limits their scope of use in production. Biolearn supports biomarkers based on multiple different biological data modalities and is written in Python, which has a much broader reach. In comparison to PyAging [33], a preliminary contemporaneous Python biomarker library, Biolearn is focused on ease of use and reproducibility through automated testing against reference data.

While Biolearn represents a significant advance for the field, some limitations remain. Currently, the library is tailored to biomarkers derived from biological samples, predominantly DNA methylation data. Moving forward, the scope of Biolearn will continue to expand to encompass diverse biological modalities—such as proteomics, metabolomics, and microbiomics—broadening its applicability [7,13]. Moreover, integration with larger and more diverse population datasets will be vital in advancing cross-population validation efforts. As new datasets emerge, Biolearn will adapt to incorporate these resources, ensuring ongoing robustness and scalability [34]. Finally, bioinformatics tools, including Biolearn, depend on a user base proficient in programming and data analysis. Efforts to make these tools more accessible to a wider audience, including those with limited computational expertise, will be crucial. This could involve the development of graphical user interfaces (GUIs) or web-based platforms to streamline the user experience.

We anticipate that Biolearn will become a key resource for the field and will transform many facets of aging biomarker studies. Our preliminary survival studies conducted using Biolearn demonstrate not only the power of this new platform, but illuminate the real-world implications of validated biomarkers. Biolearn's standardization and analysis capabilities stand to serve as pivotal tools for researchers seeking to bridge the gap between biomarker discovery and clinical implementation [30].

Methods

Overview of Biolearn Library

Biolearn is an open-source computational suite that facilitates the harmonization and analysis of biomarkers of aging (BoAs). It is written in Python and is readily accessible through the Python Package Index (PyPI). Biolearn is developed using modern software engineering practices, including automated testing to ensure correctness and adherence to software design principles that ensure the safe interchangeability of like components. The library is designed to be user-friendly while offering robust functionalities for researchers across various disciplines involved in aging studies. Biolearn is an open-source computational suite that facilitates the harmonization and analysis of biomarkers of aging (BoAs). It is written in Python and is readily accessible through the Python Package Index (PyPI). Biolearn is developed using modern software engineering practices, including automated testing to ensure correctness and adherence to software design

principles that ensure the safe interchangeability of like components. The library is designed to be user-friendly while offering robust functionalities for researchers across various disciplines involved in aging studies. Additionally, the results of each model in Biolearn are portable and can be easily exported in formats such as CSV, allowing for seamless integration with other tools like R for further analysis and visualization.

System Requirements and Installation

Biolearn requires Python version 3.10 or newer. It can be installed using the Python package manager, pip, with the command pip **install** biolearn. The successful installation of the library can be verified through the import test of Biolearn's core classes. The library is cross-platform and is compatible with major operating systems, including Windows, MacOS, and Linux.

Data Library and Model Gallery

Biolearn incorporates a data library capable of loading and structuring datasets from a multitude of public sources like Gene Expression Omnibus (GEO), National Health and Nutrition Examination Survey (NHANES), and Framingham Heart Study. The model gallery within Biolearn holds reference implementations for various aging clocks and biomarkers, presenting a unified interface for users to apply these models to their datasets. All models were verified to be correct by comparing the outputs on a reference data set against their original implementations where available.

Harmonization Process

We used Biolearn to harmonize several aging clocks. Clock definitions were standardized, specifying the name, publication year, applicable species, target tissue, and the biological aspect they predict (e.g., age, mortality risk). We provided sources for both the original publications and the coefficients necessary for clock applications. Coherence across biological modalities and datasets was assured through Biolearn's systematic approach to data preprocessing, normalization, and imputation procedures.

Integration with Public Datasets

Biolearn's ability to interface seamlessly with public datasets was tested by integrating and formatting data from GEO and NHANES. Preprocessing pipelines were developed to convert raw data into a harmonized format suitable for subsequent analysis. Particular attention was given to

metadata structures, variable normalization, and missing data treatment, ensuring consistent input formats required by the aging models.

Cell Type Deconvolution

Biolearn's deconvolution function estimates cell type proportions within a sample from bulk methylation data. It operates on the principle that bulk methylation is a composite of methylation profiles from various cell types, proportionate to their presence (B = X * P). Here, 'B' represents bulk methylation, 'X' the matrix of cell-type-specific methylation profiles, and 'P' the vector of cell type proportions. With known 'B' and estimated 'X,' 'P' is computable through constrained quadratic programming, ensuring proportions remain within biological plausibility—between zero and one and summing to one [25].

We developed two deconvolution variants tailored for blood methylation analysis: DeconvoluteBlood450K for the 450K platform and DeconvoluteBloodEPIC for the EPIC platform, addressing platform-specific biases [26]. The methylation matrix for each mode derives from the corresponding platform, encapsulating six major blood cell types—neutrophils, monocytes, NK cells, B cells, CD4+ T cells, and CD8+ T cells. Selection of CpG sites for deconvolution relied on identifying the 50 most distinctively hyper- and hypo-methylated sites per cell type, prioritized by the significance of their differential methylation. This yielded 600 reference CpG sites per deconvolution mode [27,28].

Statistical Analysis

All statistical analyses were performed using tools embedded within the Biolearn library or through integration with renowned Python statistics libraries such as statsmodels and seaborn for visualization. The robustness and reproducibility of the analysis were ensured through the use of randomized cross-validation techniques for model assessment and bootstrapping methods for estimating confidence intervals where applicable. Survival analyses were conducted using the Cox Proportional Hazards model, adjusting for age and other relevant covariates. The performance of aging clocks in predicting chronological age and mortality risk was evaluated using metrics such as R^2, hazard ratios, and p-values.

Acknowledgments

We thank all the members of the Biomarkers of Aging Consortium for their valuable feedback and suggestions. We thank S. Horvath and A. Lu for sharing the GrimAgeV1 and V2. This work was inspired by methylCIPHER, an R package for DNA methylation clocks [32]. Supported by grants from the National Institute on Aging, Hevolution Foundation, Methuselah Foundation, and VoLo Foundation.

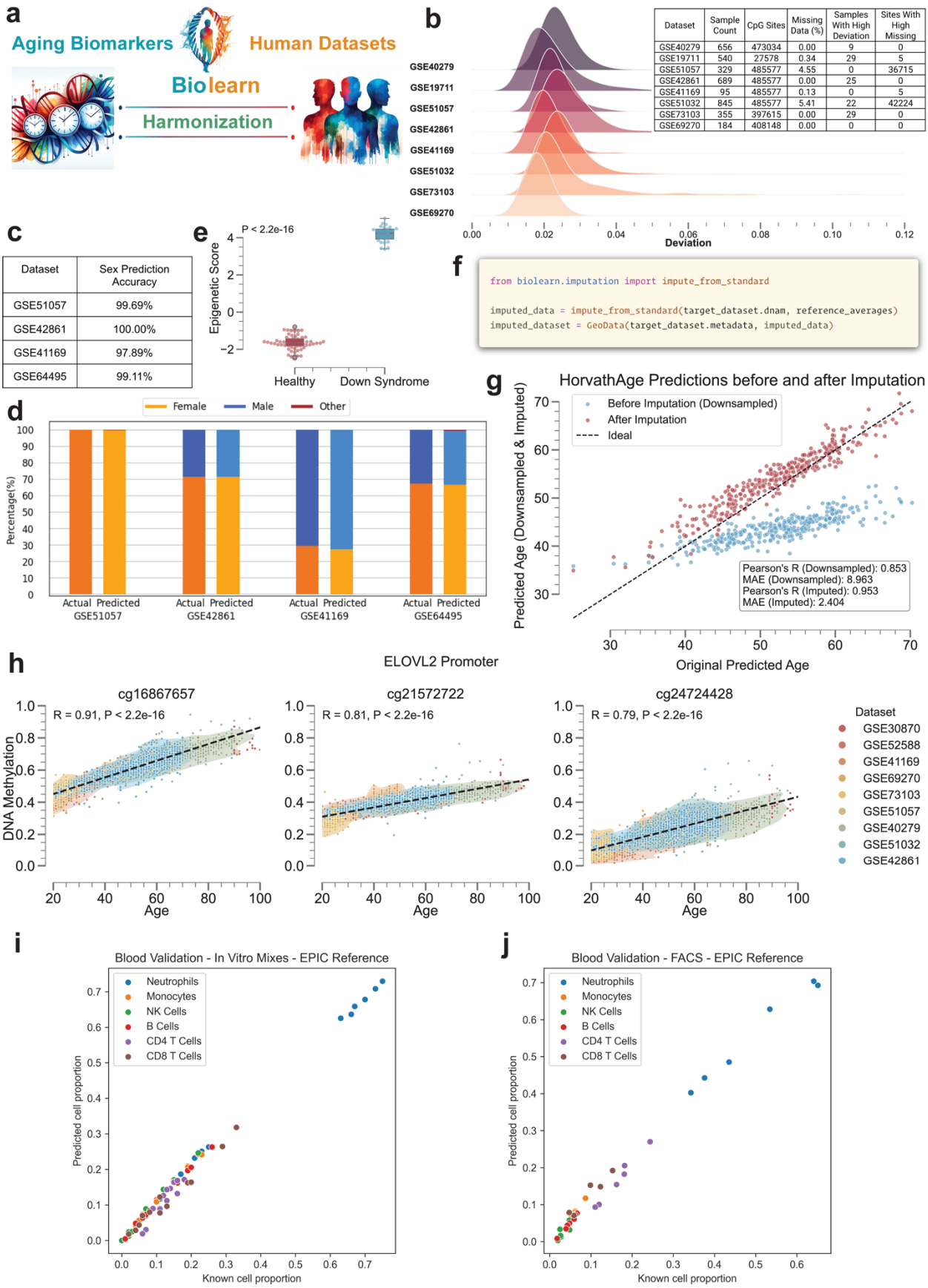
Tables and Figures

Table 1. Harmonized biomarkers in Biolearn.

Name	*Year*	*Species*	*Tissue*	*Predicts*
Horvathv1 [3]	2013	Human	Multi-tissue	Age (Years)
Hannum [17]	2013	Human	Blood	Age (Years)
Lin [35]	2016	Human	Blood	Age (Years)
PhenoAge [11]	2018	Human	Blood	Age (Years)
Horvathv2 [36]	2018	Human	Skin, blood	Age (Years)
PEDBE [37]	2019	Human	Buccal	Age (Years)
Zhang_10 [19]	2019	Human	Blood	Mortality Risk
GrimAge [10]	2019	Human	Blood	Age Adjusted by Mortality Risk (Years)
GrimAge2 [18]	2022	Human	Blood	Age Adjusted by Mortality Risk (Years)
DunedinPoAm38 [20]	2020	Human	Blood	Aging Rate (Years/Year)
DunedinPACE [5]	2022	Human	Blood	Aging Rate (Years/Year)
DNAmTL [38]	2019	Human	Blood, Adipose	Telomere Length
Knight [39]	2016	Human	Cord Blood	Gestational Age
LeeControl [40]	2019	Human	Placenta	Gestational Age
LeeRefinedRobust [40]	2019	Human	Placenta	Gestational Age
LeeRobust [40]	2019	Human	Placenta	Gestational Age
YingCausAge [12]	2022	Human	Blood	Age (Years)
YingDamAge [12]	2022	Human	Blood	Age (Years)
YingAdaptAge [12]	2022	Human	Blood	Age (Years)
SmokingMcCartney [42]	2018	Human	Blood	Smoking Status
AlcoholMcCartney [42]	2018	Human	Blood	Alcohol Consumption
BMI_McCartney [42]	2018	Human	Blood	BMI
EducationMcCartney [42]	2018	Human	Blood	Educational Attainment
TotalCholesterolMcCartney [42]	2018	Human	Blood	Total Cholesterol
HDLCholesterolMcCartney [42]	2018	Human	Blood	HDL Cholesterol
LDLCholesterolMcCartney [42]	2018	Human	Blood	LDL with Remnant Cholesterol
BodyFatMcCartney [42]	2018	Human	Blood	Percentage Body Fat
BMI_Reed [43]	2020	Human	Blood	BMI
ProstateCancerKirby [44]	2017	Human	Prostate	Prostate Cancer Status
HepatoXu [45]	2017	Human	Circulating DNA	Hepatocellular Carcinoma Status
CVD_Westerman [46]	2020	Human	Blood	Coronary Heart Disease Status
AD_Bahado-Singh [47]	2021	Human	Blood	Alzheimer's Disease Status
DepressionBarbu [48]	2021	Human	Blood	Depression Risk

Table 2. Harmonized datasets in Biolearn.

ID	*Title*	*Format*	*Samples*	*Age Present*	*Sex Present*
GSE40279	Genome-wide Methylation Profiles Reveal Quantitative Views o…	Illumina450k	656	Yes	Yes
GSE19711	Genome wide DNA methylation profiling of United Kingdom Ovar…	Illumina27k	540	Yes	No
GSE51057	Methylome Analysis and Epigenetic Changes Associated with Me…	Illumina450k	329	Yes	Yes
GSE42861	Differential DNA methylation in Rheumatoid arthritis	Illumina450k	689	Yes	Yes
GSE41169	Blood DNA methylation profiles in a Dutch population	Illumina450k	95	Yes	Yes
GSE51032	EPIC-Italy at HuGeF	Illumina450k	845	Yes	No
GSE73103	Many obesity-associated SNPs strongly associate with DNA met…	Illumina450k	355	Yes	Yes
GSE69270	Aging-associated DNA methylation changes in middle-aged indi…	Illumina450k	184	Yes	No
GSE36054	Methylation Profiling of Blood DNA from Healthy Children	Illumina450k	192	No	No
GSE64495	DNA methylation profiles of human blood samples from a sever…	Illumina450k	113	Yes	Yes
GSE30870	DNA methylomes of Newborns and Nonagenarians	Illumina450k	40	Yes	No
NHANES	National Health and Nutrition Examination Survey	Phenotypic	2877	Yes	Yes
FHS	Framingham Heart Study	Phenotypic	4434	Yes	Yes



Dataset	Sample Count	CpG Sites	Missing Data (%)	Samples With High Deviation	Sites With High Missing
GSE40279	656	473034	0.00	9	0
GSE19711	540	27578	0.34	29	5
GSE51057	329	485577	4.55	0	36715
GSE42861	689	485577	0.00	25	0
GSE41169	95	485577	0.13	0	5
GSE51032	845	485577	5.41	22	42224
GSE73103	355	397615	0.00	29	0
GSE69270	184	408148	0.00	0	0

Dataset	Sex Prediction Accuracy
GSE51057	99.69%
GSE42861	100.00%
GSE41169	97.89%
GSE64495	99.11%

Figure 1. Biolearn: A harmonization framework for biological datasets. a. Overview of Biolearn's functionalities, including biomarker discovery, dataset harmonization, and imputation. **b.** Quality control results are visualized as a ridge density plot showing the distribution of sample deviations from the mean of the population. The accompanying table lists key metadata for each dataset, including sample count, age statistics, and disease status. **c.** Comparison of actual vs. predicted sex distributions in four datasets (GSE51057, GSE42861, GSE41169, and GSE64495), with prediction accuracies listed in the accompanying table. **d.** Stacked bar graphs displaying the actual and predicted sex distributions for each dataset, with female, male and other categories shown. **e.** Epigenetic score distribution of healthy individuals compared to those with Down Syndrome ($P < 2.2e-16$, Wilcoxon rank-sum test). **f.** Python code snippet illustrating the usage of Biolearn to import, preprocess and impute missing data in a target dataset using a reference dataset. **g.** Scatter plot showing a significant inverse correlation between the Horvath epigenetic age predictions and imputed Hannum age predictions across multiple datasets (Pearson's R = 0.999, MAE = 0.325). **h.** DNA methylation levels of three age-associated CpG sites (cg16867657, cg21572722, cg24724428) plotted against chronological age for 9 datasets, with points colored by dataset. Robust linear regressions are shown, with the correlation coefficient (R) and p-value displayed for each CpG site. **i,j.** Deconvolution results for the DeconvoluteBloodEPIC method applied to a dataset with known cell proportions from in vitro mixes (**i**) and with FACS (**j**). The estimated cell type proportions are compared to the true proportions.

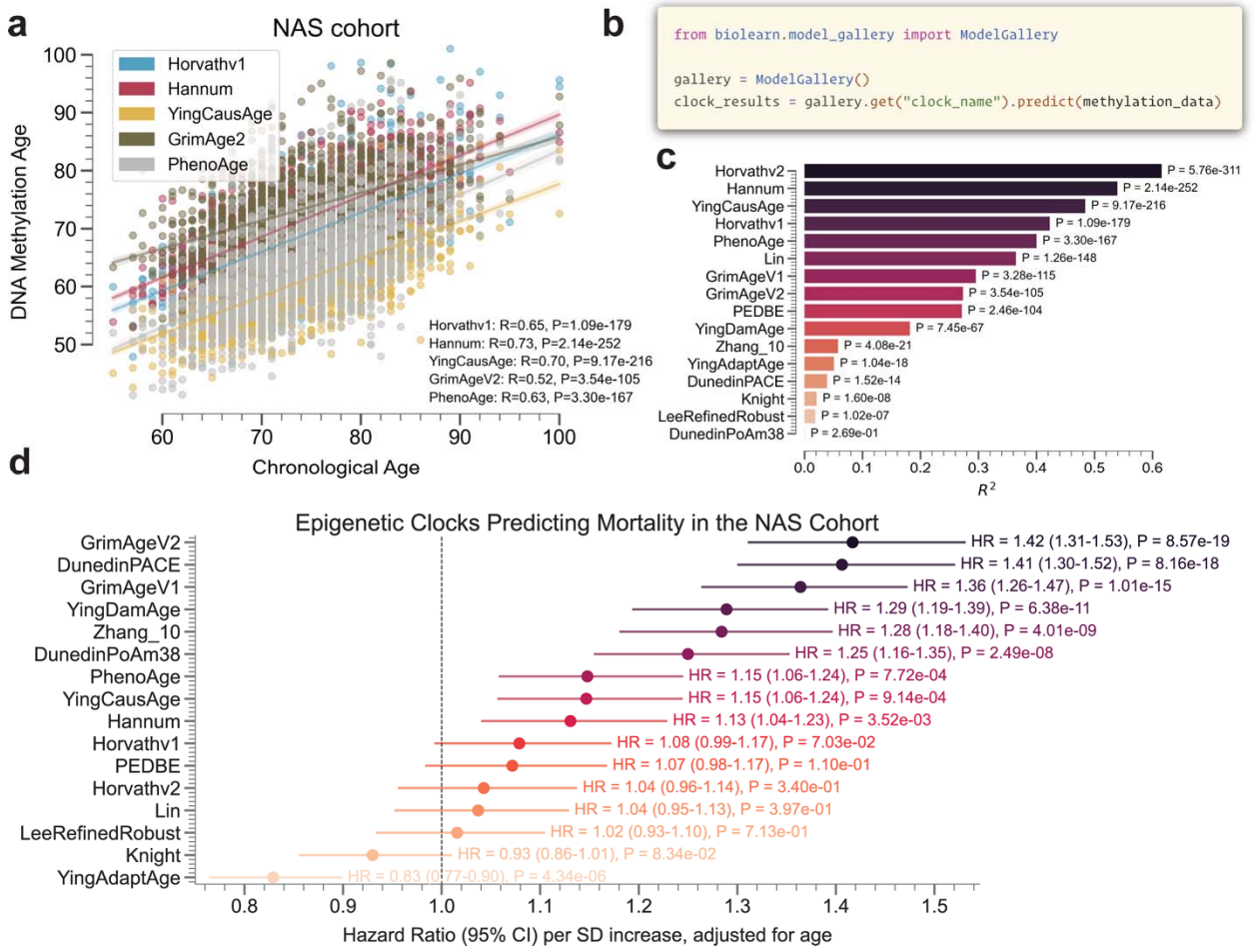
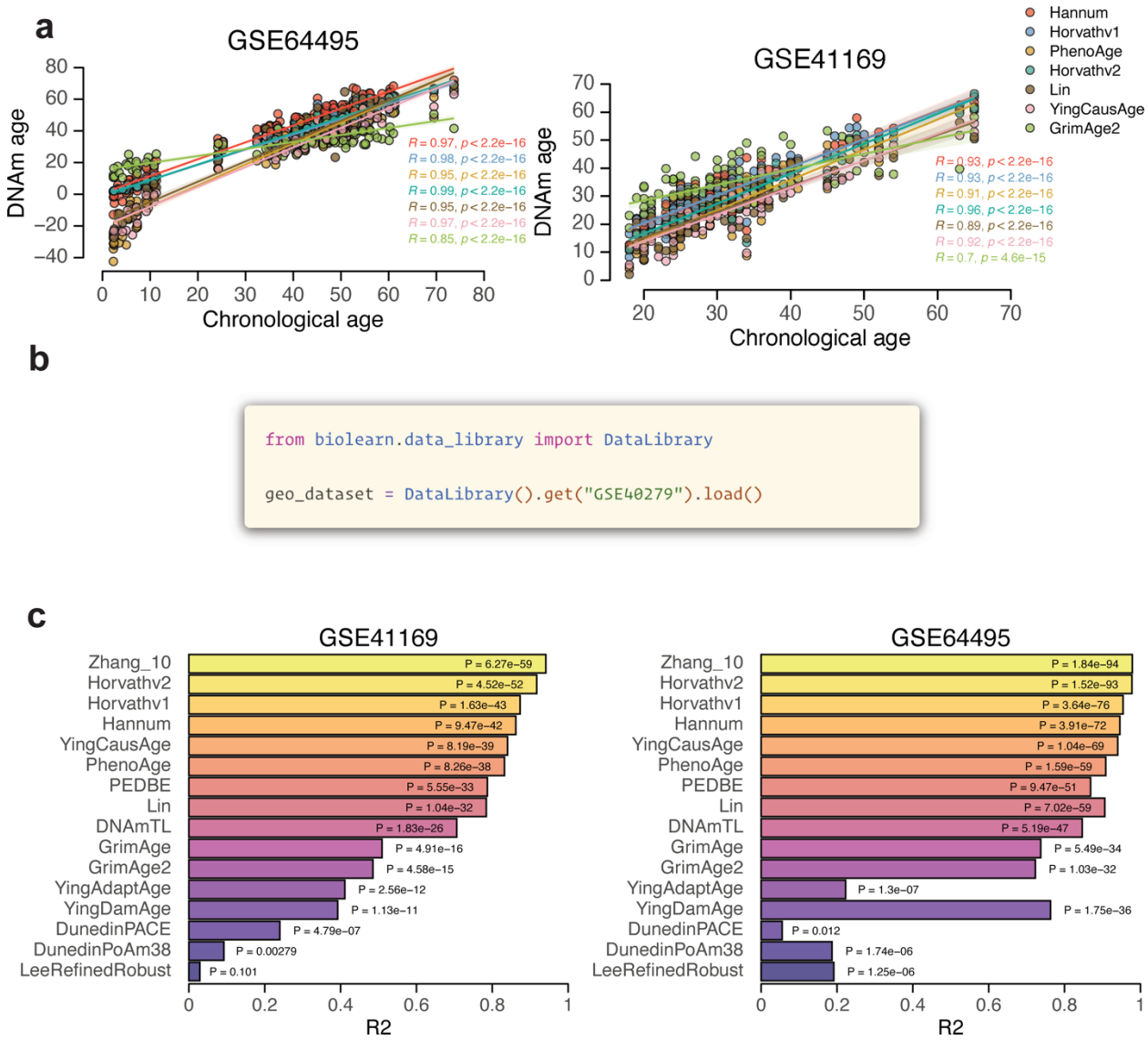


Figure 2. Comprehensive Analysis of Epigenetic Clocks in the Normative Aging Study Using Biolearn. a. Scatter plots illustrating the correlation between predicted epigenetic age and chronological age for the NAS cohort (N = 1488, 38.8% deseased). The 5 most representative clock models are shown. The precision of each clock is indicated by the Pearson's R value. **b.** Code example displaying the simplicity of calculating epigenetic age with the Biolearn library, emphasizing the minimal lines of code required for computation. **c.** Bar graph comparing the performance of the epigenetic clocks in predicting chronological age, ranked by R^2 values. Clocks with higher R^2 are with high correlation with chronological age. **d.** Forest plot presenting the hazard ratios (HR) of mortality risk for the NAS cohort, as predicted by the 16 epigenetic clock models, adjusted for age. The harzard ratios are shown per standard deviation increase of the biomarker. 95% confidence intervals are displayed as error bars. P-values are based on two-sided tests.




```
from biolearn.data_library import DataLibrary

geo_dataset = DataLibrary().get("GSE40279").load()
```

Figure 3. Epigenetic age predict from GEO data. a. Scatter plot of predicted age (y-axis) vs. chronological age (x-axis) for the GSE41169 (N = 95) and GSE64495 (N = 113) datasets. The top 7 clocks with the smallest mean absolute errors are shown in the plot. Pearson's R and unadjusted P-values based on two-sided tests are provided. **b.** The code snippet shows that the DNAm age can be calculated with a few lines of code using Biolearn library. **c.** Bar graph showing the R-square (x-axis) of methylation biomarker prediction vs. chronological age. The color indicates the overall rank of the clock based on the R-square value. The unadjusted P-values based on two-sided tests are shown on the bar.

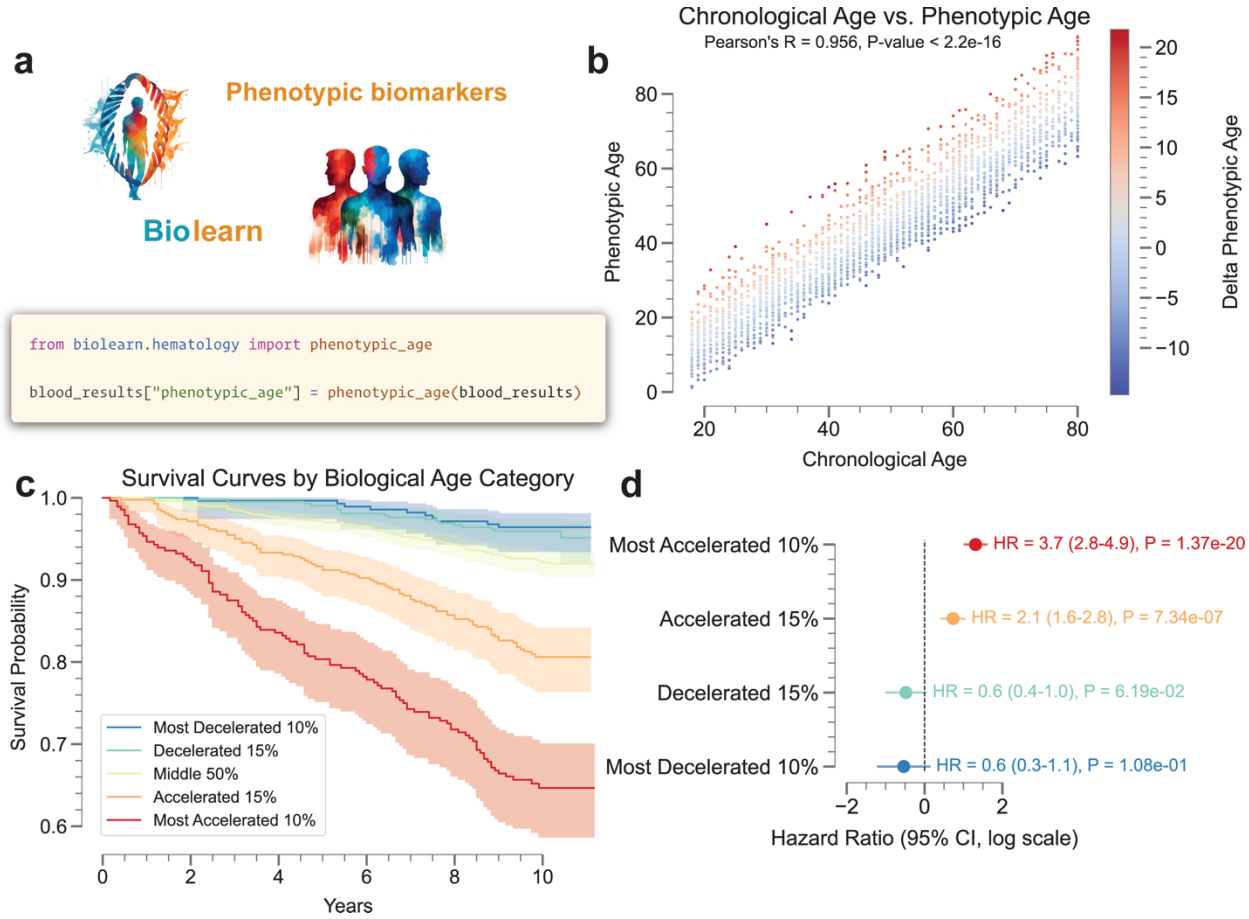


Figure 4. Phenotypic biomarker analysis using Biolearn. a. Overview of Biolearn's phenotypic biomarker functionalities. The code snippet shows that the DNAm age can be calculated with a few lines of code using the Biolearn library. **b.** Scatter plot of chronological age vs. phenotypic age for the NHANES 2010 dataset. **c.** Survival analysis of the NHANES 2010 dataset (N = 2877), stratified by biological age discrepancies (marked by different colors) based on Phenotypic Age. The survival curves demonstrate that individuals with accelerated biological aging (red) have a lower survival probability compared to those with decelerated aging (blue). The shaded areas in c represent the standard error of the survival estimates. **d.** Forest plot displaying the hazard ratios (HR) and 95% confidence intervals (CI) for the association between biological age discrepancies based on the Phenotypic Age metrics and all-cause mortality.

References

1. Moqri, M. *et al.* A framework for validation of omic biomarkers of aging. *Nature Medicine, In Press* (2024).
2. Moqri, M. *et al*. Biomarkers of aging for the identification and evaluation of longevity interventions. *Cell* **186**, 3758–3775 (2023).
3. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biology* **14**, R115 (2013).
4. Bell, C. G. *et al.* DNA methylation aging clocks: challenges and recommendations. *Genome Biol* **20**, 249 (2019).
5. Belsky, D. W. *et al.* DunedinPACE, a DNA methylation biomarker of the pace of aging. *Elife* **11**, e73420 (2022).
6. Chen, B. H. *et al*. DNA methylation-based measures of biological age: meta-analysis predicting time to death. *Aging* **8**, 1844–1865 (2016).
7. Galkin, F. *et al.* Biohorology and biomarkers of aging: Current state-of-the-art, challenges and opportunities. *Ageing Research Reviews* **60**, 101050 (2020).
8. Horvath, S. & Raj, K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat Rev Genet* **19**, 371–384 (2018).
9. Jylhävä, J., Pedersen, N. L. & Hägg, S. Biological Age Predictors. *EBioMedicine* **21**, 29–36 (2017).
10. Lu, A. T. *et al.* DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging* **11**, 303–327 (2019).
11. Levine, M. E. *et al.* An epigenetic biomarker of aging for lifespan and healthspan. *Aging* **10**, 573–591 (2018).

12. Ying, K. *et al.* Causality-enriched epigenetic age uncouples damage and adaptation. *Nat Aging* 1–16 (2024) doi:10.1038/s43587-023-00557-0.
13. Moqri, M. *et al.* PRC2 clock: a universal epigenetic biomarker of aging and rejuvenation. 2022.06.03.494609 Preprint at https://doi.org/10.1101/2022.06.03.494609 (2022).
14. Field, A. E. *et al.* DNA Methylation Clocks in Aging: Categories, Causes, and Consequences. *Molecular Cell* **71**, 882–895 (2018).
15. Marioni, R. E. *et al.* The epigenetic clock is correlated with physical and cognitive fitness in the Lothian Birth Cohort 1936. *Int. J. Epidemiol.* **44**, 1388–1396 (2015).
16. Edgar, R. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210 (2002).
17. Hannum, G. *et al.* Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. *Molecular Cell* **49**, 359–367 (2013).
18. Lu, A. T. *et al.* DNA methylation GrimAge version 2. *Aging* **14**, 9484–9549 (2022).
19. Zhang, Q. *et al.* Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing. *Genome Med* **11**, 54 (2019).
20. Belsky, D. W. *et al.* Quantification of the pace of biological aging in humans through a blood test, the DunedinPoAm DNA methylation algorithm. *eLife* **9**, e54870 (2020).
21. Thompson, M. J. *et al.* A multi-tissue full lifespan epigenetic clock for mice. *Aging* **10**, 2832–2854 (2018).
22. Ying, K. *et al. ClockBase*□: a comprehensive platform for biological age profiling in human and mouse. Preprint at https://doi.org/10.1101/2023.02.28.530532 (2023).
23. Wang, Y. *et al.* DNA methylation-based sex classifier to predict sex and identify sex chromosome aneuploidy. *BMC Genomics* **22**, 484 (2021).

24. Horvath, S. *et al.* Accelerated epigenetic aging in Down syndrome. *Aging Cell* **14**, 491–495 (2015).

25. Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**, 86 (2012).

26. Hicks, S. C. & Irizarry, R. A. methylCC: technology-independent estimation of cell type composition using differentially methylated regions. *Genome Biology* **20**, 261 (2019).

27. Reinius, L. E. *et al.* Differential DNA Methylation in Purified Human Blood Cells: Implications for Cell Lineage and Studies on Disease Susceptibility. *PLoS One* **7**, e41361 (2012).

28. Salas, L. A. *et al.* An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray. *Genome Biology* **19**, 64 (2018).

29. Horvath, S. *et al.* Aging effects on DNA methylation modules in human brain and blood tissue. *Genome Biol* **13**, R97 (2012).

30. Li, Q. *et al.* Biomarkers of aging for the identification and evaluation of longevity interventions. *Popul Health Metrics* **186**, 3758–3775 (2023).

31. Kwon, D. & Belsky, D. W. A toolkit for quantification of biological age from blood chemistry and organ function test data: BioAge. *GeroScience* **43**, 2795–2808 (2021).

32. Thrush, K. L., Higgins-Chen, A. T., Liu, Z. & Levine, M. E. R methylCIPHER: A Methylation Clock Investigational Package for Hypothesis-Driven Evaluation & Research. 2022.07.13.499978 Preprint at https://doi.org/10.1101/2022.07.13.499978 (2022).

33. Camillo, L. P. de L. pyaging: a Python-based compendium of GPU-optimized aging clocks. 2023.11.28.569069 Preprint at https://doi.org/10.1101/2023.11.28.569069 (2023).

34. Yang, J. *et al.* Prevalence of comorbidities and its effects in patients infected with SARS-CoV-2: a systematic review and meta-analysis. *International Journal of Infectious Diseases* **94**, 91–95 (2020).

35. Lin, Q. *et al.* DNA methylation levels at individual age-associated CpG sites can be indicative for life expectancy. *Aging* **8**, 394–401 (2016).

36. Horvath, S. *et al.* Epigenetic clock for skin and blood cells applied to Hutchinson Gilford Progeria Syndrome and ex vivo studies. *Aging (Albany NY)* **10**, 1758–1775 (2018).

37. McEwen, L. M. *et al.* The PedBE clock accurately estimates DNA methylation age in pediatric buccal cells. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 23329–23335 (2020).

38. Lu, A. T. *et al.* DNA methylation-based estimator of telomere length. *Aging* **11**, 5895–5923 (2019).

39. Knight, A. K. *et al.* An epigenetic clock for gestational age at birth based on blood methylation data. *Genome Biol* **17**, 206 (2016).

40. Lee, Y. *et al.* Placental epigenetic clocks: estimating gestational age using placental DNA methylation levels. *Aging* **11**, 4238–4253 (2019).

41. Ying, K. *et al.* Causal Epigenetic Age Uncouples Damage and Adaptation. 2022.10.07.511382 Preprint at https://doi.org/10.1101/2022.10.07.511382 (2022).

42. McCartney, D. L. *et al.* Epigenetic prediction of complex traits and death. *Genome Biol* **19**, 136 (2018).

43. Reed, Z. E., Suderman, M. J., Relton, C. L., Davis, O. S. P. & Hemani, G. The association of DNA methylation with body mass index: distinguishing between predictors and biomarkers. *Clinical Epigenetics* **12**, 50 (2020).

44. Kirby, M. K. *et al.* Genome-wide DNA methylation measurements in prostate tissues uncovers novel prostate cancer diagnostic biomarkers and transcription factor binding patterns. *BMC Cancer* **17**, 273 (2017).

45. Xu, R. *et al*. Circulating tumour DNA methylation markers for diagnosis and prognosis of hepatocellular carcinoma. *Nature Mater* **16**, 1155–1161 (2017).

46. Westerman, K. *et al*. Epigenomic Assessment of Cardiovascular Disease Risk and Interactions With Traditional Risk Metrics. *J Am Heart Assoc* **9**, e015299 (2020).

47. Bahado-Singh, R. O. *et al.* Artificial intelligence and leukocyte epigenomics: Evaluation and prediction of late-onset Alzheimer's disease. *PLOS ONE* **16**, e0248375 (2021).

48. Barbu, M. C. *et al.* Epigenetic prediction of major depressive disorder. *Mol Psychiatry* **26**, 5112–5123 (2021).