Biolearn, an open-source library for biomarkers of aging

Kejun Ying[1,2,*], Seth Paulson[3,*], Martin Perez-Guevara[4], Mehrnoosh Emamifar[5], Maximiliano Casas Martínez[6], Dayoon Kwon[7], 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] Independent Researcher, Boston, MA, 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@stanford.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 augments the development, assessment, and validation trajectories 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) is an essential focus in aging research, driven by a growing understanding of aging as a fundamental factor in chronic diseases and mortality. A plethora of biomarkers have been proposed to quantify biological age or the pace of aging and elucidate the underlying biological intricacies; however, their clinical validation remains a significant challenge due to varied formulations and disparate structures in validation datasets across populations. Since the introduction of composite omic biomarkers of aging, epitomized by Horvath's seminal work on DNA methylation aging clocks [1], subsequent efforts have expanded the repertoire of aging biomarkers, which now span a wide array of omic modalities, including epigenomics, transcriptomics, and proteomics [2–7].

DNA methylation-based clocks are currently the most developed class of omic biomarkers of aging and represent robust tools for biological age estimation, offering insights into age-related changes at the molecular level and their implications for human health and longevity [1,2,4]. For instance, the Horvath multi-tissue clock and the subsequent DNA methylation-based clocks, such as DunedinPACE [3], GrimAge [8], causality-enriched DamAge/AdaptAge [9], PRC2 clock [10], and others, have shown promising correlations with various age-related conditions and mortality, reflecting the intricate interplay between epigenetic modifications and the aging trajectory [4,8,11,12]. However, the diverse formulation of these biomarkers and the inconsistency in structure across different validation datasets pose substantial hurdles for systematic cross-population validation and benchmarking of these biomarkers, critical steps towards their clinical translation.

Publicly available datasets, such as those from the Gene Expression Omnibus [13], the National Health and Nutrition Examination Survey (NHANES), and the Framingham Heart Study (FHS) have the potential to greatly accelerate the validation of biomarkers of aging. However, their use for this purpose is complicated by the lack of a standardized framework that can accommodate the heterogeneous nature of their datasets. Thus, there is a clear need for a unified platform that can seamlessly integrate and analyze various biomarkers of aging across datasets with harmonized structures. Such a platform would transform the validation process, foster 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 designed to bridge 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 fostering a community-driven approach, Biolearn aims to propel the development and validation process of BoAs, ultimately contributing to a deeper understanding of the aging process, facilitating the discovery of interventions for age-related diseases, and bringing BoAs to the clinic.

Results

Harmonization of Biomarkers of Aging

We successfully harmonized 22 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 include chronological clocks like Horvath's multi-tissue clock, Hannum's blood clock [1,14]; healthspan and mortality-related clocks like GrimAge, GrimAge2, PhenoAge, and Zhang clock [8,15–17]; biomarkers of the rate of aging including DunedinPoAm38 and DunedinPACE [3,18]; causality-enriched clocks including Ying's CausAge, DamAge, and AdaptAge [9]; as well as many other clocks and DNAm-based biomarkers (Table 1). All of the biomarkers were formatted into standardized input structures, allowing for consistent application across disparate datasets. This harmonization process involved collecting and unifying the annotation of clock specifications,

such as tissue type, predicted age range, and source references, ensuring transparent and reproducible analyses. To further support ongoing research in this field, we developed and implemented an open-source framework (https://github.com/bio-learn/biolearn/blob/master/biolearn/model.py). This framework is designed to provide a standardized format for epigenetic biomarkers, facilitating the seamless integration and comparison of any future aging clocks that are developed. This addition aims to foster collaboration and innovation in the study of aging biomarkers.

Integration of Diverse Human Datasets

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 [13,19]. With this capacity, the Biolearn is used as the backend of ClockBase for epigenetic age computation [13], enabling the systemic harmonization of over 200,000 human samples from Gene Expression Omnibus (GEO) array data.

Computational Analysis and Biomarker Assessment

Using Biolearn's computational methodologies, we conducted extensive evaluations of the harmonized biomarkers of aging. As an example, we applied several models to two GEO dataset GSE41169 (N = 95) and GSE64495 (N = 113), with DNA methylation profiles across approximately 480,000 CpGs [20]. We show that with a few lines of code, Biolearn is able to efficiently compute results for over 20 epigenetic biomarkers across hundreds of samples on the order of seconds (Figure 1b-c). Biolearn also allows rapid comparison of the performance of different biomarkers in independent datasets (Figure 1d). For example, in both testing datasets, the top 5 performing clocks, in terms of R^2 with chronological age, are the Zhang clock, Horvath v2 and v1, Hannum clock, and YingCausAge (Figure 1d).

Beyond aging biomarkers, we also integrated several common epigenetic predictors in Biolearn. For instance, sex can be inferred (Wang et al. [21]) from DNAm profiles with high accuracy (Figure 1e).

Clinical Data Harmonization

To provide further utility in handling clinical data, we implemented the blood-test-based Phenotypic Age [16], and Mahalanobis distance-based biomarker in Biolearn [22]. 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 2a-b) [16].

Discussion

Among the most significant challenges in aging biomarker research is the cross-population validation of proposed biomarkers [23]. 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 [4].

With Biolearn, we have 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 [3,7].

Our approach emphasizes transparency and reproducibility, core tenets of open science. By making Biolearn publicly available and maintaining detailed documentation and development guidelines, we have 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 [24,25], have been limited in scope, focusing on methylation or blood-based biomarkers only. Furthermore, being 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 [26], 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, certain 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 [5,10]. 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 [27]. 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 [23].

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.

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. The more detailed instructions can be found here: https://Biolearn.github.io/quickstart.html

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.

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.

Acknowledgments

We thank all the members of the Biomarkers of Aging Consortium for their valuable feedback and suggestions. This work was heavily inspired by methylCIPHER, an R package for DNA methylation clocks [24]. Supported by grants from the National Institute on Aging, Hevolution Foundation, Methuselah Foundation, and VoLo Foundation.

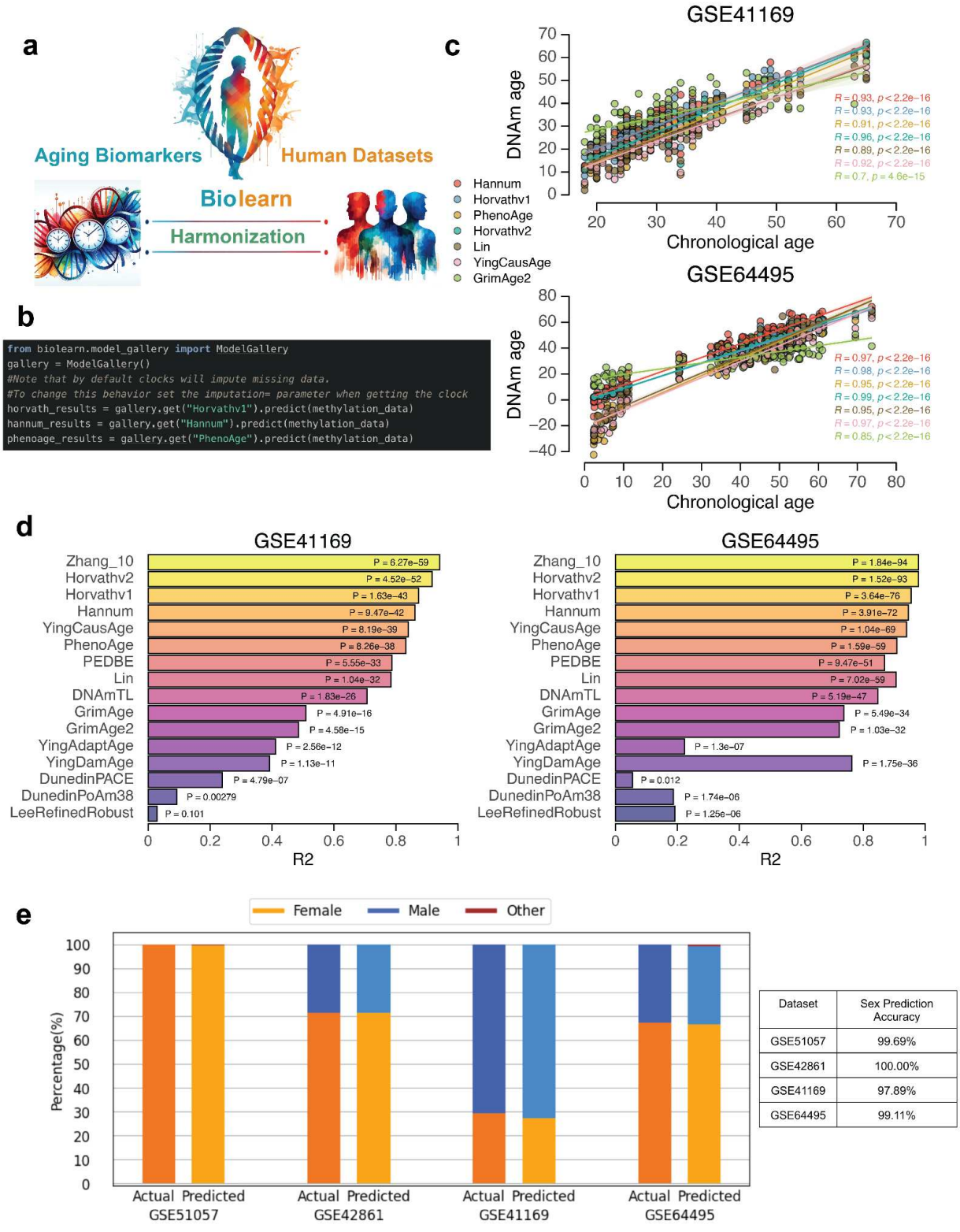
Tables and Figures

Table 1. Harmonized biomarkers of aging in Biolearn.

Name	Year	Species	Tissue	Predicts
Horvathv1 [1]	2013	Human	Multi-tissue	Age (Years)
Hannum [14]	2013	Human	Blood	Age (Years)
Lin [28]	2016	Human	Blood	Age (Years)
PhenoAge [16]	2018	Human	Blood	Age (Years)
Horvathv2[29]	2018	Human	Skin + blood	Age (Years)
PEDBE [30]	2019	Human	Buccal	Age (Years)
Zhang_10 [17]	2019	Human	Blood	Mortality Risk
GrimAge [8]	2019	Human	Blood	Age Adjusted by Mortality Risk (Years)
GrimAge2 [15]	2022	Human	Blood	Age Adjusted by Mortality Risk (Years)
DunedinPoAm38 [18]	2020	Human	Blood	Aging Rate (Years/Year)
DunedinPACE [3]	2022	Human	Unknown	Aging Rate (Years/Year)
AlcoholMcCartney [31]	2018	Human	Blood	Alcohol Consumption
BMI_McCartney [31]	2018	Human	Blood	BMI
DNAmTL [32]	2019	Human	Blood, Adipose	Telomere Length
Knight [33]	2016	Human	Cord Blood	Gestational Age
LeeControl [34]	2019	Human	Placenta	Gestational Age
LeeRefinedRobust [34]	2019	Human	Placenta	Gestational Age
LeeRobust [34]	2019	Human	Placenta	Gestational Age
YingCausAge/DamAge/AdaptAge [9]	2022	Human	Blood	Age (Years)
SmokingMcCartney [31]	2018	Human	Blood	Smoking Status

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

Figure 1. a. Overview of Biolearn's functionalities. **b.** The code snippet showing that the DNAm age can be calculated with a few lines of code using Biolearn library. **c.** 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. **d.** 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. **e.** Stacked bar graph shows the actual vs. predicted sex distribution in four different datasets: GSE51057, GSE42861, GSE41169, and GSE64495. The accompanying table provides the sex prediction accuracy for each dataset.

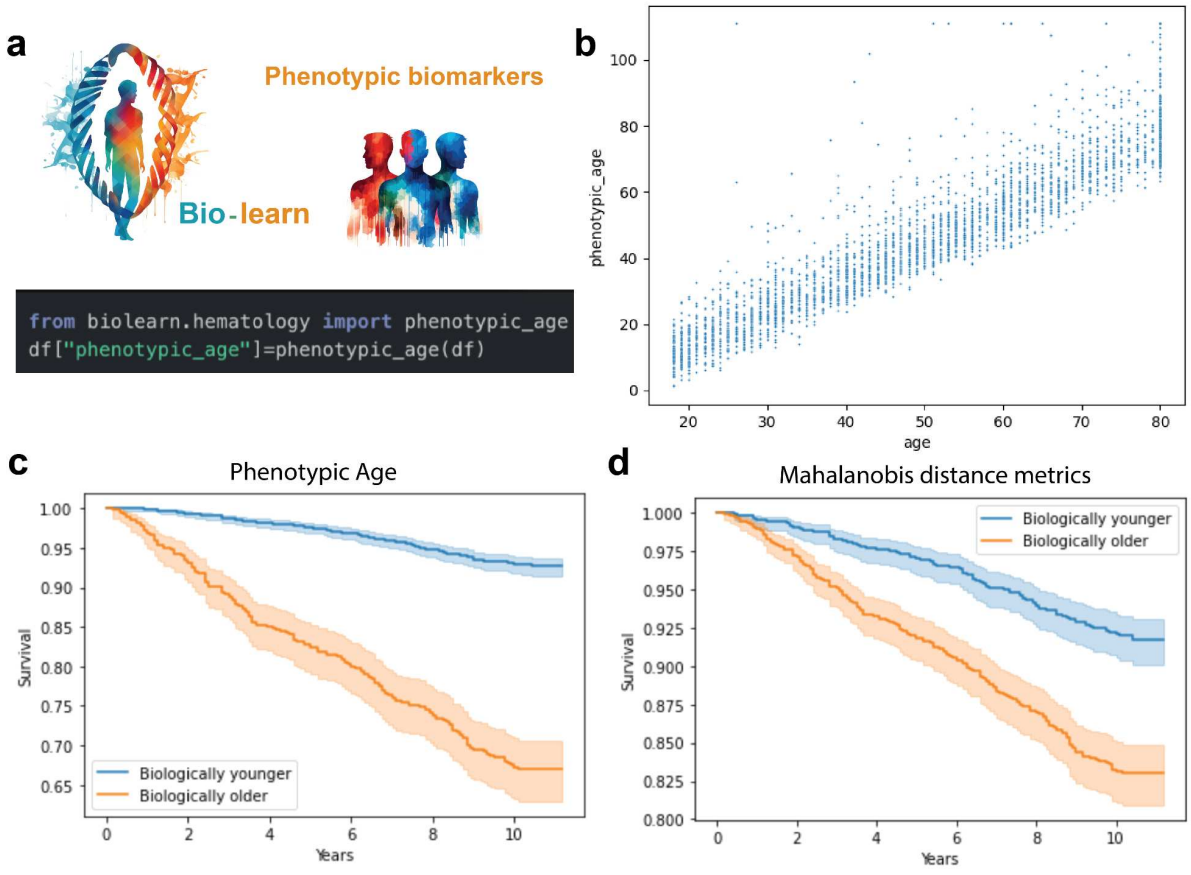


Figure 2. a. Overview of Biolearn's phenotypic biomarker functionalities. The code snippet shows that the DNAm age can be calculated with a few lines of code using the Biolearn library. **b.** Scatter plot of chronological age vs. phenotypic age for the NHANES 2010 dataset. **c, d.** Survival analysis of the NHANES 2010 dataset (N = 2877), stratified by biological age discrepancies (marked by different colors) based on Phenotypic Age (c) and Mahalanobis distance metrics (d). Individuals with biological age higher than chronological age are marked as biologically older and vice versa. For the purpose of demonstration, the result is not adjusted by chronological age. The shadow shows the standard error.

References

1. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biology* **14**, R115 (2013).
2. Bell, C. G. *et al.* DNA methylation aging clocks: challenges and recommendations. *Genome Biol* **20**, 249 (2019).
3. Belsky, D. W. *et al.* DunedinPACE, a DNA methylation biomarker of the pace of aging. *Elife* **11**, e73420 (2022).
4. Chen, B. H. *et al.* DNA methylation-based measures of biological age: meta-analysis predicting time to death. *Aging* **8**, 1844–1865 (2016).
5. Galkin, F. *et al.* Biohorology and biomarkers of aging: Current state-of-the-art, challenges and opportunities. *Ageing Research Reviews* **60**, 101050 (2020).
6. Horvath, S. & Raj, K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat Rev Genet* **19**, 371–384 (2018).
7. Jylhävä, J., Pedersen, N. L. & Hägg, S. Biological Age Predictors. *EBioMedicine* **21**, 29–36 (2017).
8. Lu, A. T. *et al.* DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging* **11**, 303–327 (2019).
9. 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).
10. 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).
11. 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).
12. Field, A. E. *et al.* DNA Methylation Clocks in Aging: Categories, Causes, and Consequences. *Molecular Cell* **71**, 882–895 (2018).
13. Ying, K. *et al. ClockBase* : a comprehensive platform for biological age profiling in human and mouse. 2023.02.28.530532 Preprint at https://doi.org/10.1101/2023.02.28.530532 (2023).
14. Hannum, G. *et al.* Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. *Molecular Cell* **49**, 359–367 (2013).
15. Lu, A. T. *et al.* DNA methylation GrimAge version 2. *Aging* **14**, 9484–9549 (2022).
16. Levine, M. E. *et al.* An epigenetic biomarker of aging for lifespan and healthspan. *Aging* **10**, 573–591 (2018).
17. Zhang, Q. *et al.* Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing. *Genome Med* **11**, 54 (2019).
18. 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).
19. Thompson, M. J. *et al.* A multi-tissue full lifespan epigenetic clock for mice. *Aging* **10**, 2832–2854 (2018).
20. Horvath, S. *et al.* Aging effects on DNA methylation modules in human brain and blood tissue. *Genome Biol* **13**, R97 (2012).
21. Wang, Y. *et al.* DNA methylation-based sex classifier to predict sex and identify sex chromosome aneuploidy. *BMC Genomics* **22**, 484 (2021).
22. Li, Q. *et al.* An objective metric of individual health and aging for population surveys. *Popul Health Metr* **20**, 11 (2022).
23. Moqri, M. *et al.* Biomarkers of aging for the identification and evaluation of longevity interventions. *Cell* **186**, 3758–3775 (2023).
24. 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).

25. 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).

26. 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).

27. 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).

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

29. 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).

30. 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).

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

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

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

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