Understanding Aging at Multi-scale Using Explainable AI

Wei Qiu

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee:

Su-In Lee, Chair

Linda Shapiro

Sheng Wang

Program Authorized to Offer Degree:

Computer Science & Engineering

University of Washington

Abstract

Understanding Aging at Multi-scale Using Explainable AI

Wei Qiu

Chair of the Supervisory Committee:

Su-In Lee

Paul G. Allen School of Computer Science Engineering

As human lifespans increase, understanding the mechanisms shaping aging has become essential. This dissertation introduces explainable AI (XAI) frameworks that characterize aging from population-level health data to bulk and single-cell transcriptomics. IMPACT improves all-cause mortality prediction in NHANES and uses XAI to uncover overlooked risk factors and interactions. ENABL Age extends this foundation to estimate biological age and quantify how lifestyle, clinical, and biochemical features influence aging. At the molecular level, DeepProfile learns latent spaces from 50,211 cancer transcriptomes across 18 tumor types, revealing immune activation patterns, subtype structures, and links among mutation burden, cell cycle, antigen presentation, and survival. ACE disentangles aging-related expression changes in single-cell RNA-seq data from mouse, fly, and human, recovering tissue- and cell-type–specific signatures, conserved pathways, and regulators such as Uba52 validated in C. elegans. Together, these contributions form a multi-scale XAI framework advancing mechanistic aging biology and interpretable approaches for improving healthspan.

The work presented in this thesis was funded by the National Science Foundation (DBI-1759487, DBI-1552309, DBI-1355899, DGE-1762114), and National Institutes of Health (R35 GM 128638, R01 NIA AG 061132, P30 AG 013280).

Do the Right Thing; Treat People Well; Relentlessly Pursue Excellence.

PUBLICATIONS

Much of the material presented in this thesis is drawn from the following publications. Here $*$ denotes equal contribution.

- Wei Qiu, Hugh Chen, Ayse Berceste Dincer, Scott Lundberg, Matt Kaeberlein, and Su-In Lee. "Interpretable machine learning prediction of all-cause mortality." In: *Communications medicine* 2.1 (2022), p. 125
- Wei Qiu, Hugh Chen, Matt Kaeberlein, and Su-In Lee. "ExplaiNAble BioLogical Age (ENABL Age): an artificial intelligence framework for interpretable biological age." In: *The lancet Healthy longevity* 4.12 (2023), e711–e723
- Wei Qiu, Ayse B Dincer, Joseph D Janizek, Safiye Celik, Mikael J Pittet, Kamila Naxerova*, and Su-In Lee*. "Deep profiling of gene expression across 18 human cancers." In: *Nature biomedical engineering* 9.3 (2025), pp. 333–355
- Wei Qiu, Chris Arian, Ethan Weinberger, Soo R Yun, Alexander R Mendenhall, Jessica E Young, Maria Brbic*, and Su-In Lee*. "An explainable AI framework for identifying universal aging signatures in cell embeddings." In: (Under review)

ACKNOWLEDGMENTS

This thesis would not have been possible without the support, guidance, and friendship of many people.

I would first like to express my deepest gratitude to my advisor, Prof. Su-In Lee, for her extraordinary mentorship throughout my doctoral studies. I am sincerely grateful for the opportunity to join her lab and for the freedom she gave me to explore my research interests. Our first meeting six years ago left a lasting impression on me, and her passion for scientific discovery has continued to inspire me every day since. Her encouragement was especially meaningful when I began working on aging, a challenging and relatively unexplored direction within computer science at the time. Her unwavering support, thoughtful advice, and willingness to connect me with exceptional collaborators were essential to the development of this dissertation. I am also deeply appreciative of her patience, openness during moments of disagreement, and the tremendous amount of guidance she offered during my job search. I could not have completed this work without her mentorship.

I am grateful to my committee members, Prof. Linda Shapiro, Prof. Sheng Wang, and Prof. Meliha Yetisgen, for their time, feedback, and insightful suggestions during both my general examination and dissertation defense. I owe special thanks to Prof. Sheng Wang for his continued guidance, for encouraging me to pursue the faculty job market, and for his thoughtful advice on my job talk and application materials. His support has been instrumental to my professional growth.

I would also like to thank my labmates, including Gabriel Erion, Nicasia Beebe-Wang, Ayse Dincer, Ian Covert, Joseph Janizek, Pascal Sturmfels, Hugh Chen, Ethan Weinberger, Alex DeGrave, Chanwoo Kim, Chris Lin, Mingyu Lu, Soham Gadgil, Patrick Yu, Allison Li, Xiaojian Chen, and Chenhao Zhang. I am especially grateful to the senior students for their mentorship during the early stage of my PhD. The COVID-19 pandemic began only a few months after I joined the lab, and the strong sense of community, collaboration, and enthusiasm in the group kept me motivated during an exceptionally challenging period. I have learned so much from each of my labmates through research discussions, day-to-day interactions, and their exemplary dedication. I cherish the friendships we formed and look forward to future opportunities to work together.

I am deeply thankful to my collaborators for their expertise, generosity, and enthusiasm. Dr. Matt Kaeberlein introduced me to aging research and greatly shaped the direction of my doctoral work. I am grateful to Dr. Matt Kaeberlein, Prof. Kamila Naxerova, and Prof. Maria Brbić for their support during my job search. I am especially thankful to Prof. Maria Brbić for her extensive and thoughtful feedback on my job talk, which greatly strengthened its clarity and impact. I also thank Prof. Jessica Young, Prof. Alexander

Mendenhall, and Dr. Christopher Arian for their guidance on the biological aspects of my research and for the opportunity to learn from their expertise.

I would like to acknowledge my friends in Seattle, including Liwei Jiang, Yilin Song, Ruotong Wang, Yue Guo, Judy Kong, Yuxuan Mei, Raymond Fok, Hao Peng, and Zhaoqi Li, for their companionship and encouragement throughout my PhD. Their support, shared experiences, and friendship greatly enriched my life in Seattle and helped me navigate many difficult moments. I am also grateful to my friends in the UW CSE computational biology community, including Xinming Tu, Zixuan Liu, Hanwen Xu, Xiao Wang, Yilun Sheng, Tong Chen, and Shengqi Hang. Our frequent research discussions and shared struggles during stressful periods were invaluable. These conversations helped shape my research directions and broaden my understanding of the field.

I am also grateful to my friends in Beijing, Xinyao Li, Yue Chen, Yao Chen, Haoran Chang, Yuan Yao, Wenzhu Wu, and Jingxuan Li, for always welcoming me home with warmth and familiarity. We explored new restaurants, traveled together, and supported one another through both the pandemic and the challenges of graduate school. I am also thankful to Yi Wang and Shilei Ding, now in Zurich, for their continued friendship despite the distance. Their support throughout the pandemic and during the most stressful parts of the PhD meant a great deal to me.

I would also like to thank the friends I met during my internship at Genentech, including Yilin Song, Yuheng Fu, Xinming Tu, Jiahui Peng, Wancen Mu, Xin Huang, Yiwei Gong, Ziyi Song, and Wenbin Guo. Their optimism, kindness, and shared experiences made that summer both memorable and meaningful, and I am grateful that our friendships have continued long after the internship ended.

Finally, I would like to express my deepest appreciation to my family. I am profoundly grateful to my parents for their unconditional love, trust, and encouragement. They have always supported my curiosity and never pressured me about my studies, giving me the freedom to pursue what I love. My mother, who is also my closest friend, has always been caring, patient, and understanding, and I am endlessly grateful for her presence in my life. I also thank my grandparents for their warmth and unwavering support, even though they may not fully understand the details of my work. The pandemic kept me from returning home for nearly four years, and I deeply regret that I could not see my grandfather before he passed away. I wish I could tell him that I completed my PhD, found a job, and will continue working toward meaningful scientific contributions. His love and encouragement continue to inspire me, and I hope to make him proud.

To my family

CONTENTS

Part I

PRELIMINARIES

1

UNDERSTANDING AGING AT MULTIPLE SCALES

Human lifespan has increased dramatically over the last century, largely due to improvements in public health, medical care, and socioeconomic conditions [55]. However, this increase in *lifespan* has not been matched by a corresponding increase in *healthspan*, defined as the period of life spent in good health without chronic disease or disability [137]. Epidemiological data show that while people are living longer, many experience prolonged periods of illness in later life, leading to increased healthcare costs and diminished quality of life [65]. For example, mortality curves from 1900 and 2016 reveal a clear extension of lifespan, but the period of high morbidity in late life remains largely unchanged [208]. This demographic transition poses profound implications for public health systems, as the prevalence of chronic, age-related diseases continues to rise. Thus, there is a critical need to not only extend lifespan but also improve healthspan.

To address this challenge, it is essential to focus on the underlying biological process of aging itself. Aging is the greatest risk factor for many of the leading causes of death, including cardiovascular disease [331], cancer [316], Alzheimer's disease [201], and diabetes [209]. Unlike disease-specific approaches that target individual pathologies, interventions that delay the aging process could simultaneously reduce the burden of multiple diseases [137]. Indeed, interventions such as caloric restriction, rapamycin treatment, and senolytic therapies have demonstrated the ability to modulate aging pathways and improve healthspan across multiple organ systems in model organisms [40, 212].

At its core, aging is a progressive loss of physiological integrity, leading to impaired function and increased vulnerability to death [166]. The biological mechanisms that drive aging are multifactorial and deeply interconnected. The "hallmarks of aging" framework describes several conserved processes, including genomic instability, telomere attrition, epigenetic alterations, loss of proteostasis, mitochondrial dysfunction, deregulated nutrient sensing, cellular senescence, stem cell exhaustion, and altered intercellular communication [39, 166]. These hallmarks represent molecular signatures that together shape the complex trajectory of biological aging.

Despite decades of research, a comprehensive understanding of aging remains elusive, partly because it operates across multiple biological scales, from molecules and cells to tissues, organs, and the whole organism. Traditional studies have often focused on a single scale, such as cellular senescence or organismal lifespan, without capturing the hierarchical nature of these processes. However, emerging evidence from systems biology and multi-omics approaches indicates that aging is not a uniform process; rather, it is a multi-scale and multi-dimensional phenomenon [257, 280, 285].

Aging manifests differently depending on the biological scale under investigation. At the molecular level, DNA damage, protein misfolding, and metabolic imbalance accumulate over time. At the cellular level, these molecular perturbations alter transcriptional programs and functional states, giving rise to senescent, apoptotic, or metabolically altered cells. At the tissue and organ level, changes in cellular composition and communication lead to structural remodeling, functional decline, and altered inter-organ signaling. At the organismal level, these cumulative effects shape systemic aging phenotypes such as frailty, cognitive decline, and metabolic dysfunction.

Therefore, studying aging at a single scale provides only a partial view. To truly understand the biology of aging, it is necessary to integrate evidence across these scales, capturing both localized changes and systemic interactions.

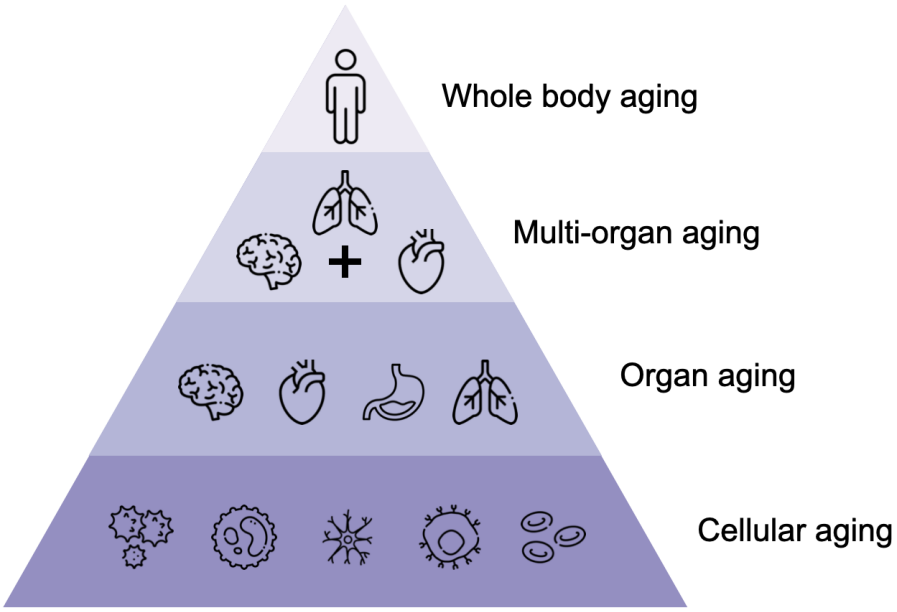


Figure 1.1: **Multi-scale framework of aging studied in this thesis.** In this thesis, I focus on understanding aging at four biological scales: whole-body aging, multi-organ aging, organ aging, and cellular aging. These levels represent a hierarchy from organismal health decline to molecular and cellular changes, each providing complementary insights into how aging manifests and progresses across systems.

In this thesis, I focus on understanding aging at four interconnected scales: **whole-body aging**, **multi-organ aging**, **organ aging**, and **cellular aging** (Figure **??**).

At the whole-body level, individuals of the same chronological age can differ substantially in their *biological age*, which reflects their physiological state and overall functional capacity [131]. This biological age can be estimated through diverse biomarkers, such as DNA methylation clocks [117], transcriptomic and proteomic age predictors [13, 194], or clinical data [162]. The concept of biological age emphasizes that aging is not purely

chronological but rather a quantifiable, dynamic process that reflects cumulative biological damage and resilience across systems. Understanding whole-body aging helps identify global determinants of longevity and inter-individual variability.

At the multi-organ scale, aging manifests as coordinated yet heterogeneous changes across physiological systems. For example, cardiovascular aging leads to vascular stiffening and cardiac decline, impairing circulation and accelerating dysfunction in other tissues [147]. Likewise, immune aging drives chronic inflammation and reduced regenerative capacity across organs [85]. Capturing these cross-system interactions is essential for identifying universal versus organ-specific aging mechanisms.

At the organ level, aging manifests as a gradual deterioration of structure and function driven by cumulative cellular and molecular changes [285]. Each organ follows a distinct aging trajectory shaped by its metabolic activity, regenerative capacity, and cellular composition. For example, cardiac aging involves fibrosis, impaired contractility, and mitochondrial decline; brain aging is characterized by synaptic loss, neuroinflammation, and protein aggregation; and hepatic aging alters detoxification and metabolic regulation. Understanding organ-specific aging is crucial for identifying targeted interventions that preserve function and delay disease onset.

Finally, at the cellular level, individual cells within the same tissue can exhibit distinct aging states or trajectories. Cellular heterogeneity arises from stochastic DNA damage, epigenetic drift, metabolic reprogramming, and microenvironmental factors [280, 290]. Recent advances in single-cell technologies have made it possible to profile these changes with unprecedented resolution, revealing how subpopulations of cells resist or succumb to aging processes. Cellular-level insights are essential for linking molecular mechanisms to organismal outcomes and identifying potential rejuvenation targets such as senescent cell clearance or transcriptional reprogramming.

The multi-scale nature of aging presents both a challenge and an opportunity for computational biology. Each scale, ranging from cellular to organ and whole-body levels, is characterized by high-dimensional, heterogeneous data such as genomic, transcriptomic, proteomic, and clinical measurements. The relationships within and across these biological features are inherently complex and often nonlinear, making traditional statistical approaches insufficient to fully capture aging patterns. These challenges motivate the use of advanced machine learning and explainable AI methods, which can model nonlinear dependencies, extract informative representations, and provide interpretable insights into the biological processes underlying aging.

2

AI AND EXPLAINABLE AI IN AGING RESEARCH

Recent advances in artificial intelligence (AI) and machine learning (ML) have transformed how complex biological systems can be studied. These approaches provide powerful tools for uncovering hidden structure in data that are high-dimensional, nonlinear, and noisy—features that are characteristic of nearly all biological measurements. Unlike traditional statistical models that rely on linear assumptions or hand-crafted features, AI models can automatically learn multivariate dependencies directly from data. This ability makes them particularly valuable for studying aging, a process governed by intricate molecular networks and physiological feedback loops that evolve over time and vary widely across individuals.

The biological study of aging increasingly depends on large-scale datasets collected at multiple levels of organization: single-cell transcriptomes, proteomic and metabolomic profiles, clinical biomarkers, and longitudinal health records. Each dataset captures a different view of the aging process, yet together they create a rich, high-dimensional space that is difficult to interpret with classical tools. AI methods enable efficient representation learning in such spaces. They can extract latent factors that summarize cellular states, infer molecular signatures of biological age, and integrate heterogeneous modalities to predict physiological decline. In recent years, AI has achieved notable success in genomics [168], proteomics [313], and clinical phenotyping [237], demonstrating that data-driven models can complement biological intuition by identifying patterns invisible to human observation.

Applying AI to aging, however, introduces unique conceptual and technical challenges. Aging is not a single process but a multifaceted and dynamic phenomenon that manifests differently across biological scales. Individuals of the same chronological age often show striking differences in health status, reflecting variation in their biological age and resilience. At the organ and tissue levels, aging trajectories diverge across systems, influenced by both intrinsic genetic programs and external environmental exposures [285]. Even within a single tissue, cellular populations display heterogeneous transcriptional and epigenetic profiles that correspond to distinct aging trajectories [280, 290]. Modeling the diversity of aging patterns requires algorithms capable of handling high-dimensional, nonlinear, and heterogeneous data. Relationships among features within each biological scale are often complex, context-dependent, and difficult to capture using traditional regression or correlation analyses. Addressing this challenge requires AI models that can effectively model complex and nonlinear relationships in biological data.

Another major challenge lies in *interpretability*. Although AI models can capture complex patterns and achieve superior predictive performance, they are often regarded as

"black boxes." Their internal mechanisms can be biologically opaque. This opacity limits scientific discovery and makes it difficult to translate model predictions into mechanistic understanding or actionable hypotheses. In aging research, interpretability is not optional: understanding why a model predicts accelerated aging, and identifying which biomarkers or molecular pathways contribute most strongly to aging, is critical for experimental validation, hypothesis generation, and clinical translation. Without interpretability, predictive models risk becoming descriptive rather than explanatory [204, 287].

Explainable AI (XAI) has emerged as a promising framework to overcome this limitation. XAI techniques are designed to make AI models more transparent by quantifying the contribution of input features to model predictions, disentangling complex signals into interpretable variables, and mapping data-driven features to biologically meaningful entities. Examples include feature attribution methods such as Integrated Gradients [272], SHAP values [178], and disentangled representation learning in generative models [311]. When applied to aging research, XAI serves as a bridge between computational modeling and biological interpretation. It enables the identification of risk factors that influence aging trajectories, the discovery of genes associated with age-related changes, and the elucidation of molecular pathways underlying the aging process. Through these methods, AI becomes not merely a predictive model but a hypothesis generation tool that can guide experimental and translational research.

The integration of AI and XAI provides a powerful opportunity to deepen our understanding of aging. By combining predictive modeling with interpretability, researchers can build computational frameworks that are both accurate and biologically meaningful. This paradigm shifts aging research from static descriptions of biomarkers toward dynamic, data-driven models that explain how biological systems evolve over time.

In this thesis, I apply AI to study aging across multiple biological scales while ensuring interpretability within each level. Rather than modeling all scales simultaneously, I focus on each scale individually to disentangle the mechanisms most relevant to that level. Through explainable modeling, I aim to uncover fundamental aging processes and generate insights that connect computational learning with biological understanding. The frameworks introduced in this work demonstrate how interpretable AI can reveal the principles underlying aging while remaining transparent, robust, and grounded in biomedicine.

3 OUR CONTRIBUTIONS

In the previous chapter, we discussed the importance of understanding aging from a multi-scale perspective and the growing role of AI and XAI in uncovering its underlying mechanisms. Aging manifests at different biological scales, from organismal health decline to molecular and cellular changes, and no single level of analysis can capture its full complexity. To address this, it is essential to develop models that are both predictive and interpretable, capable of identifying not just what changes with age but why those changes occur.

This thesis integrates AI and XAI to study aging across four complementary scales: (1) whole-body aging, (2) multi-organ aging, (3) organ-specific aging, and (4) cellular aging (Figure **??**). The models developed in this work collectively aim to explore aging from these different perspectives, each emphasizing distinct biological contexts and data modalities. Together, they illustrate how explainable modeling can provide interpretable insights into the mechanisms that link individual variation to broader biological processes of aging.

This section provides a brief overview of models developed as part of this thesis.

IMPACT (**??**).

Relation to multi-scale aging.

At the whole-body level, aging can be characterized by cumulative physiological decline that ultimately influences survival. Mortality risk thus provides an integrated outcome reflecting how biological, behavioral, and environmental factors interact over the lifespan. Studying mortality offers a data-driven way to understand whole-body aging and identify predictors of health resilience or vulnerability. To analyze these complex relationships, the IMPACT framework applies explainable AI to large population datasets, enabling interpretable modeling of all-cause mortality and revealing factors that link population-level health outcomes with the biology of aging.

Abstract.

Unlike linear models which are traditionally used to study all-cause mortality, complex machine learning models can capture non-linear interrelations and provide opportunities to identify unexplored risk factors. Explainable artificial intelligence can improve prediction accuracy over linear models and reveal great insights into outcomes like mortality. This paper comprehensively analyzes all-cause mortality by explaining complex machine learning models.

We propose the IMPACT framework that uses XAI technique to explain a state-of-the-art tree ensemble mortality prediction model. We apply IMPACT to understand all-cause mortality for 1-, 3-, 5-, and 10-year follow-up times within the NHANES dataset, which contains 47,261 samples and 151 features.

We show that IMPACT models achieve higher accuracy than linear models and neural networks. Using IMPACT, we identify several overlooked risk factors and interaction effects. Furthermore, we identify relationships between laboratory features and mortality that may suggest adjusting established reference intervals. Finally, we develop highly accurate, efficient and interpretable mortality risk scores that can be used by medical professionals and individuals without medical expertise. We ensure generalizability by performing temporal validation of the mortality risk scores and external validation of important findings with the UK Biobank dataset.

IMPACT's unique strength is the explainable prediction, which provides insights into the complex, non-linear relationships between mortality and features, while maintaining high accuracy. Our explainable risk scores could help individuals improve self-awareness of their health status and help clinicians identify patients with high risk. IMPACT takes a consequential step towards bringing contemporary developments in XAI to epidemiology.

This work was published in *Nature Communications medicine* [231].

ENABL AGE (**??**).

Relation to multi-scale aging.

To better understand aging, the most important first step is measuring it. Building on mortality risk modeling, the next step is to quantify aging itself through the concept of biological age, a metric that reflects an individual's physiological state and health status rather than chronological age. Biological age provides a direct measure of overall health and functional decline, offering a more meaningful assessment of aging than chronological age. The ENABL Age framework extends the modeling of all-cause mortality to estimate whole-body biological age and applies the same explainable AI approach to cause-specific mortality to capture multi-organ aging. By combining machine learning with interpretability, ENABL Age produces accurate and transparent age estimators that reveal how different physiological systems collectively shape the aging process.

Abstract.

Biological age is a measure of health that offers insights into ageing. The existing age clocks, although valuable, often trade off accuracy and interpretability. We introduce ExplaiNAble BioLogical Age (ENABL Age), a computational framework that combines machine-learning models with explainable artificial intelligence (XAI) methods to accurately estimate biological age with individualised explanations.

To construct the ENABL Age clock, we first predicted an age-related outcome (eg, all-cause or cause-specific mortality), and then rescaled these predictions to estimate biological age, using UK Biobank and National Health and Nutrition Examination Survey (NHANES) datasets. We adapted existing XAI methods to decompose individual ENABL Ages into contributing risk factors. For broad accessibility, we developed two versions: ENABL Age-L, based on blood tests, and ENABL Age-Q, based on questionnaire characteristics. Finally, we validated diverse ageing mechanisms captured by each ENABL Age clock through genome-wide association studies (GWAS) association analyses.

Our ENABL Age clock was significantly correlated with chronological age (r=0.7867, p<0.0001 for UK Biobank; r=0.7126, p<0.0001 for NHANES). These clocks distinguish individuals who are healthy (ie, their ENABL Age is lower than their chronological age) from those who are unhealthy (ie, their ENABL Age is higher than their chronological age), predicting mortality more effectively than existing clocks. Groups of individuals who were unhealthy showed approximately three to 12 times higher log hazard ratio than healthy groups, as per ENABL Age. The clocks achieved high mortality prediction power with an area under the receiver operating characteristic curve of 0.8179 for 5-year mortality and 0.8115 for 10-year mortality on the UK Biobank dataset, and 0.8935 for 5-year mortality and 0.9107 for 10-year mortality on the NHANES dataset. The individualised explanations that revealed the contribution of specific characteristics to ENABL Age provided insights into the important characteristics for ageing. An association analysis with risk factors and ageing-related morbidities and GWAS results on ENABL Age clocks trained on different mortality causes showed that each clock captures distinct ageing mechanisms.

ENABL Age brings an important leap forward in the application of XAI for interpreting biological age clocks. ENABL Age also carries substantial potential in practical settings, assisting medical professionals in untangling the complexity of ageing mechanisms, and potentially becoming a valuable tool in informed clinical decision-making processes.

This work was published in *The lancet Healthy longevity* [232].

DEEPPROFILE (**??**).

Relation to multi-scale aging.

At the organ level, aging manifests through progressive functional decline that increases vulnerability to disease. Cancer exemplifies this process, as the risk of tumor development rises with age due to the accumulation of molecular damage, impaired repair mechanisms, and immune dysregulation. Studying cancer therefore offers an opportunity to understand organ-specific aspects of aging and the biological pathways that drive tissue deterioration. The DeepProfile framework leverages large-scale transcriptomic data across multiple cancer types to model these processes, us-

ing unsupervised deep learning to uncover shared and distinct molecular signatures of organ aging. By connecting cancer biology to the broader context of aging, this work reveals how age-related molecular changes contribute to organ dysfunction and disease progression.

Abstract.

Clinical and biological information in large datasets of gene expression across cancers could be tapped with unsupervised deep learning. However, difficulties associated with biological interpretability and methodological robustness have made this impractical. Here we describe an unsupervised deep-learning framework for the generation of low-dimensional latent spaces for gene-expression data from 50,211 transcriptomes across 18 human cancers. The framework, which we named DeepProfile, outperformed dimensionality-reduction methods with respect to biological interpretability and allowed us to unveil that genes that are universally important in defining latent spaces across cancer types control immune cell activation, whereas cancer-type-specific genes and pathways define molecular disease subtypes. By linking latent variables in DeepProfile to secondary characteristics of tumours, we discovered that mutation burden is closely associated with the expression of cell-cycle-related genes, and that the activity of biological pathways for DNA-mismatch repair and MHC class II antigen presentation are consistently associated with patient survival. We also found that tumour-associated macrophages are a source of survival-correlated MHC class II transcripts. Unsupervised learning can facilitate the discovery of biological insight from gene-expression data.

This work was published in *Nature Biomedical Engineering* [233].

ACE (**??**).

Relation to multi-scale aging.

At the cellular level, aging arises from molecular changes that accumulate over time, leading to disrupted gene expression, loss of homeostasis, and functional decline. While different organs deteriorate at different rates, aging is even more heterogeneous at the cellular level—cells within the same tissue, and even of the same type, can age at different speeds. Understanding this variability is essential for uncovering the mechanisms that drive tissue- and organism-level aging. Advances in single-cell RNA sequencing (scRNA-seq) now allow aging to be studied with such fine resolution, yet age-related transcriptional signals are often obscured by dominant factors like cell type and tissue identity. To address this challenge, the ACE (Aging Cell Embeddings) framework uses a deep generative and explainable modeling approach to disentangle aging-related gene expression patterns from confounding biological variation. This approach identifies both global and cell-type-specific aging signatures, offering a high-resolution and interpretable view of cellular aging across tissues and species.

Abstract.

Aging is a complex biological process marked by progressive physiological decline and increased disease vulnerability. Single-cell RNA sequencing offers unprecedented resolution for studying aging, yet isolating aging-related signatures remains challenging because gene expression is primarily shaped by other factors such as cell type, tissue, and sex. We present **ACE (Aging Cell Embeddings)**, an explainable deep generative framework that disentangles aging-related gene expression changes from background biological variation. ACE employs two latent representations: one capturing aging-related signatures and another representing non-aging-related variation in the data. Through explainable AI, ACE identifies key genes and pathways associated with aging amid dominant non-aging-related variations. Applied to large-scale mouse, fly, and human datasets, ACE uncovers aging signatures both within specific tissue–cell-type contexts and across all tissues and cell types, enabling accurate prediction of biological age. Moreover, ACE identifies aging genes conserved across species, highlighting its ability to reveal shared biological mechanisms of aging. Experimental RNAi knockdowns in *C. elegans* validate ACE's findings, confirming its ability to prioritize novel aging genes affecting lifespan. ACE reveals key pathways involved in proteostasis, immune regulation, and extracellular matrix remodeling, and identifies *Uba52* through the cross-species model as an important aging gene, whose knockdown in *C. elegans* significantly shortens lifespan. By providing interpretable and generalizable aging embeddings, ACE establishes a foundation for cross-species single-cell aging studies and translational geroscience.

Part II

OUR CONTRIBUTIONS

4

INTERPRETABLE MACHINE LEARNING PREDICTION OF ALL-CAUSE MORTALITY

4.1 INTRODUCTION

Identification of risk factors and prediction of all-cause mortality have long been important issues in epidemiology. Most prior studies identify risk factors using associations between each predictor and mortality [46, 143, 184]; only a few papers use multi-variate linear models to predict mortality and identify risk factors [90, 302]. In terms of prediction, a variety of linear mortality risk scores have been proposed to help characterize unhealthy individuals [88, 116, 255]. Although linear models have historically been popular because they are interpretable, modern complex machine learning (ML) models often achieve higher predictive accuracy because they can capture interactions among variables in addition to non-linear relationships (e.g., "U-shaped" relationships).

The field of artificial intelligence (AI) has seen considerable advances in supervised learning problems, which involve predicting an outcome variable (e.g., all-cause mortality) based on a set of features (e.g., individual-level characteristics). Notable applications of AI in healthcare include diabetic retinopathy detection in ophthalmology images [103], red blood cells classification [234], Alzheimer's disease prediction [104], lung cancer classification from histopathology images [63], and skin cancer classification [77]. Despite this progress, a major obstacle to the adoption of AI applications in healthcare is that many of them are considered "black box," which refers to their lack of interpretability. The inability to understand why a model makes a prediction is especially harmful in healthcare applications, where the patterns a model discovers can be even more important than its predictive accuracy. This is especially true in epidemiology, which aims to identify important variables to guide public health policy or detect risk predictors that warrant further study. To address this need, we turn to a variety of techniques to help us better understand complex ML models from the emerging area of explainable AI (XAI) [176, 178, 241].

In this paper, we present the IMPACT (Interpretable Machine learning Prediction of All-Cause morTality) framework (Figure **??**), which improves the interpretability of complex machine learning models for mortality prediction. We combine an accurate, complex ML model and a state-of-the-art XAI technique to predict all-cause mortality and conduct a systematic and integrated study of the relationships among many variables and all-cause mortality. We apply IMPACT to the NHANES (1999-2014) dataset to reveal important all-cause mortality findings. First, using explainable complex ML models rather than linear models, we identify risk predictors that are highly informative of future mortality. Second,

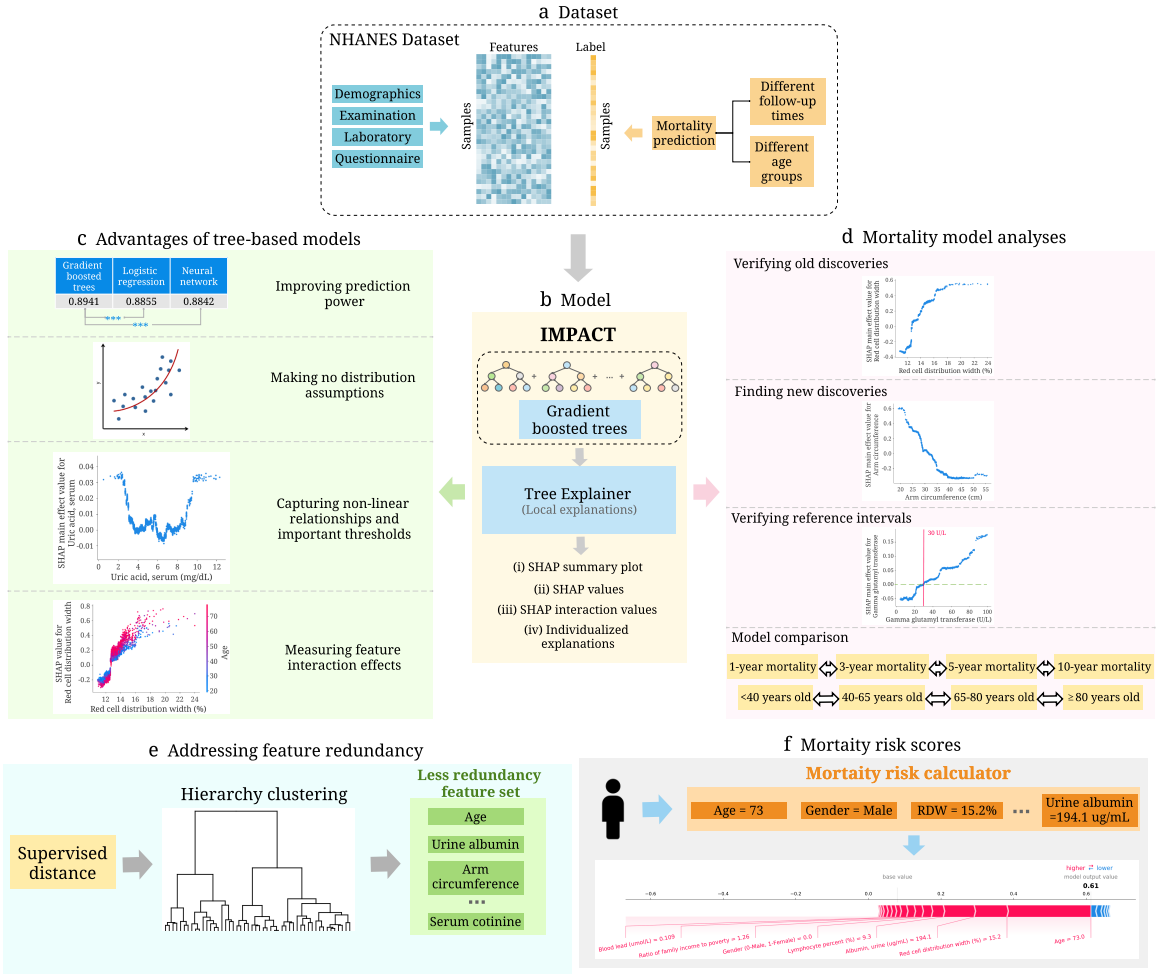


Figure 4.1: **Overview of the IMPACT model and analyses.** (a) We use the NHANES (1999-2014) dataset, which includes 151 variables and 47,261 samples. The variables can be categorized into four groups: demographics, examination, laboratory and questionnaire. We train the model using different follow-up times and different age groups. (b) IMPACT combines tree-based models with an explainable AI method. Specifically, IMPACT (1) trains tree-based models for mortality prediction using the NHANES dataset, and (2) uses TreeExplainer to provide local explanations for our models. (c) We illustrate the advantages of interpretable tree-based models compared to traditional linear models in epidemiological studies. (d) We further analyze all mortality models and demonstrate the effectiveness of IMPACT at verifying existing findings, identifying new discoveries, verifying reference intervals, obtaining individualized explanations, and comparing models using different follow-up times and age groups. (e) We propose a supervised distance to help us explore feature redundancy. We further develop a supervised distances-based feature selection method which helps us select predictive and less-redundant features. (f) We build mortality risk scores that are applicable to professional and non-professional individuals with different cost-vs-accuracy tradeoffs. The individualized explanations of IMPACT show the impact of each risk factor for the overall risk score.

	Task	Age	AUROC	AUROC of IMPACT-20	AUROC of IMPACT-20 (temporal validation)
Mortality risk scores					
Intermountain [116]	1-year mortality	18+	0.84	0.92	0.88
Gagne Index [88]	1-year mortality	65+	0.79	0.85	0.85
Intermountain [116]	5-year mortality	18+	0.87	0.89	0.88
Prognostic score [90]	5-year mortality	40-70	Male: 0.80	Male: 0.85	Male: 0.80
			Female: 0.79	Female: 0.83	Female: 0.80
Schonberg Index [255]	5-year mortality	65+	0.75	0.80	0.83
Biological ages					
Horvath DNAm Age [117, 153]	10-year mortality	21-84	0.56	0.90	0.89
Hannum DNAm Age [108, 153]	10-year mortality	21-84	0.57	0.90	0.89
DNAm PhenoAge [153]	10-year mortality	21-84	0.62	0.90	0.89
Phenotypic Age [153, 162]	10-year mortality	20-85	0.88	0.90	0.89

Table 4.1: **Comparing the AUROCs between an existing mortality score or a biological age as reported in the original paper and the IMPACT-20 model tested for the corresponding follow-up time and age ranges in the NHANES dataset.** The "AUROC" column shows the AUROCs reported in the original paper. The "AUROC of IMPACT-20" column shows the performance of IMPACT models trained with the selected top 20 features (Supplementary Tables 2 and 3). The "AUROC of IMPACT-20 (temporal validation)" column shows the performance of the IMPACT-20 models evaluated on the temporal validation set (Supplementary Methods).

our flexible models capture non-linear relationships, which provide more comprehensive information about the relationship between feature values and mortality risk: for example, the "inflection" points of risk predictors could provide a unique perspective of reference intervals that has consequential implications in public health. Third, understanding which features are the most important enables us to develop highly accurate, efficient (using less features) and interpretable mortality risk scores. Furthermore, the individualized explanation of risk scores can help users understand their most important risk factors and adjust their lifestyle. In Table **??**, we compare the AUROCs between an existing mortality score or a biological age as reported in the original paper and the IMPACT-20 model tested for the corresponding follow-up time and age ranges in the NHANES dataset. We find that IMPACT risk scores (Supplementary Methods) have higher predictive power than popular mortality risk scores [88, 90, 116, 255] and biological ages [108, 117, 153, 172]. We ensure generalizability by performing temporal validation of the mortality risk scores and external validation of feature importances and important relationships with the UK Biobank dataset. All our results and risk scores are available on an interactive website (https://suinleelab.github.io/IMPACT) to encourage exploration of important risk predictors and support the use of interpretable individual risk scores for individuals with and without medical expertise. The IMPACT framework can also be applied to other health outcomes and diseases to improve the predictive accuracy and interpretability of complex ML models in epidemiological studies.

4.2 METHODS

4.2.1 *Data cohorts*

This study primarily focuses on NHANES [67, 68, 125] (http://www.cdc.gov/nchs/nhanes.htm) data based on samples collected between 1999-2014. We include demographic, laboratory, examination, and questionnaire features that could be automatically matched across different NHANES cycles. The National Center for Health Statistics Research Ethics Review Board approved all NHANES protocols, and all participants gave informed consent. After data preprocessing (Supplementary Methods), 47,261 samples with 151 features remain. Follow-up mortality data is provided from the date of survey participation through December 31, 2015. We predict all-cause mortality for two broad categories: (1) follow-up times of 1-year, 3-year, 5-year, and 10-year, and (2) age groups of $<$40, 40-65, 65-80, and $\geqslant$80 years old. For mortality prediction with different follow-up times, we use samples of all ages. For different age groups, we fix the follow-up time to predict 5-year mortality and divide all samples for 5-year mortality prediction into four sets based on age. The dataset is randomly divided into training (80%) and testing (20%) sets. Demographic characteristics and sample size of the data for different tasks are shown in Supplementary Figure

?? and Supplementary Table **??**. The histogram of the the samples' age in different data collection cycles are shown in Supplementary Figure **??**.

In additioin, we use UK Biobank (https://www.ukbiobank.ac.uk/) samples as an external validation dataset. Ethics approval for the UK Biobank study was obtained from the North West - Haydock Research Ethics Committee (21/NW/0157). Informed consent was obtained from all UK Biobank participants (the consent form is available at https://www.ukbiobank.ac.uk/consent). For UK Biobank data, we include the 51 features that overlap between the NHANES and UK Biobank datasets and have 384,762 samples with confirmed 5-year mortality status. All-cause mortality included deaths occurring before May, 2021. The dataset is randomly divided into training (80%) and testing (20%) sets. More detail about UK Biobank dataset is in Supplementary Methods and Supplementary Figure **??**.

4.2.2 *IMPACT framework*

To achieve high accuracy and explainable mortality prediction models, we developed the IMPACT (Figure **??**) framework, which combines tree-based models and TreeExplainer [177]. To model all-cause mortality, we use gradient boosted trees (GBTs). GBTs are non-parametric models composed of iteratively trained decision trees. The final ensemble of trees can capture non-linear and interaction effects between predictors. The hyperparameters are chosen by GridSearch and 5-fold cross-validation (Supplementary Methods). Model performance is measured using the area under the receiver operator characteristic curve (AUROC).

In our previous work, we introduced TreeExplainer [177], which provides a local (i.e., for each subject) explanation of the impact of input features on individual predictions for GBT models (Supplementary Methods). Specifically, TreeExplainer calculates exact SHAP [178] (SHapley Additive exPlanations) values for GBT models, which guarantee a set of desirable theoretical properties. SHAP values are additive; they sum to the model's output, i.e., the log-odds for GBTs. They are also consistent, which means features that are unambiguously more important are guaranteed to have a higher SHAP value. Therefore, SHAP values are consistent and accurate calculations of each feature's contribution to the model's prediction. TreeExplainer also extends local explanations to capture pairwise feature interactions directly. In this work, we utilize TreeExplainer to conduct a systematic and integrated study of associations between a large number of variables and all-cause mortality. Here, higher SHAP values imply large contributions to mortality risk. By showing the impact of each variable and interactions among variables for local, sample-specific explanations, we can obtain a comprehensive understanding of why the model made a specific mortality prediction.

In addition to studying the relationships between risk factors and all-cause mortality, we further propose a technique, "relative risk percentage", to identify sub-optimal refer-

ence intervals and a metric, "supervised distance", to measure feature redundancy and identify redundant groups of features given a specific prediction task. Building on supervised distance, we also propose a recursive feature selection strategy to select feature sets that are both predictive and less redundant. We additionally propose a recursive feature selection method to train accurate and efficient (low-cost) interpretable mortality risk scores.

4.2.3 *Supervised distance*

4.2.3.1 *Supervised distance and hierarchical clustering*

Supervised distance can accurately measure feature redundancy based on a specific prediction task. To calculate the supervised distance between feature i and feature j, we first train a uni-variate GBT model to predict the label (e.g. 5-year mortality in our study) using feature i. Then, we can obtain the $\mathtt{Prediction}_i$ which is the output of the fitted uni-variate GBT. Next, we fit another uni-variate GBT to predict $\mathtt{Prediction}_i$ using feature j. We define the output of the new GBT as $\mathtt{Prediction}_i^j$. All hyperparameter values of the uni-variate GBTs are set to their default values. Following the same above steps, we can obtain $\mathtt{Prediction}_j^i$. The supervised distance between feature i and feature j (supervised distance(i,j)) is defined as:

$$\mathrm{supervised}\ R^2(i,j) = max(0, 1 - mean(\frac{(\mathtt{Prediction}_i - \mathtt{Prediction}_i^j)^2}{var(\mathtt{Prediction}_i)})) \tag{4.1}$$

$$\mathrm{supervised\ distance}(i,j) = max(1 - \mathrm{supervised}\ R^2(i,j), 1 - \mathrm{supervised}\ R^2(j,i)) \tag{4.2}$$

where $var(x)$ is the variance of the vector x, $mean(x)$ is the average of the vector x. Supervised distance is scaled roughly between 0 and 1, where 0 distance means the features perfectly redundant and 1 means they are completely independent.

To explore the redundant feature groups, we hierarchically cluster all features according to the supervised distance. Specifically, we use complete linkage hierarchical clustering which merges in each step the two clusters whose merger has the smallest diameter.

4.2.3.2 *Supervised distance-based feature selection*

We propose a supervised distance-based feature selection method to select predictive and less-redundant feature sets. Firstly, we fit a GBT for 5-year mortality prediction on all features using the training set and rank the features by mean absolute SHAP values from TreeExplainer. We cluster features except age and gender into a specific number of groups using supervised distances-based hierarchical clustering and select the most important feature in each cluster. Then, we add age and gender to the selected feature set and refit the model. Next, we rerun the clustering using the new feature set except age and

gender. This process is repeated until all remaining features cluster to a single group. In every iteration, we remove 5 features. The models are evaluated on the testing set with bootstrapping for 1,000 times. We report the average of the AUROCs and the minimum supervised distance within the selected feature sets.

4.2.4 *5-year mortality risk scores*

IMPACT mortality risk scores are defined to be the prediction of the 5-year mortality prediction models. To compare with Intermountain gender-specific risk scores, we evaluate the models on different gender groups. The models are trained on the whole training set and evaluate on different gender groups in the testing set. Furthermore, considering the different feature collection cost for the general public and medical professionals, we build the risk scores starting from different feature sets. For the general public, the models are trained on all demographics, questionnaire features and examination features that are accessible at home for general public, For medical professionals, the models are trained on all demographics and laboratory features. We implement recursive feature selection to reduce the number of features included in the risk scores. Recursive feature elimination works by searching for a subset of features by starting with all features in the training dataset and successively removing features until the desired number of features remains. Firstly, we train a model on the full dataset with all features. Then we rank features by importance (mean absolute SHAP values) and remove the least important features. Another model is trained on the resulting feature set, and the process iterates until only the desired number of features are left. We remove 5 features in each iteration. We bootstrap the test set for 1,000 times and assess the predictive performance. We report the average of the AUROCs within the selected feature sets.

4.3 RESULTS

4.3.1 *Advantages of tree-based models*

Linear models are commonly used in epidemiology because their coefficients indicate each feature's contribution to the model's prediction [195]. However, more expressive models, such as tree-based models, can achieve higher predictive accuracy across many datasets by learning non-linear relationships between features and the outcome variable. Gradient boosted trees (GBTs) have achieved state-of-the-art performance in many domains [79, 240, 288, 335]. We observe the same trend in our study: tree-based models outperform both linear models and neural networks across almost all tasks we consider (Figure **??**a, Supplementary Figure **??**). The superior prediction performance of tree models indicates that we can capture signals relevant to mortality, which alternative approaches could not. Besides predictive power, tree-based models have more advantages compared

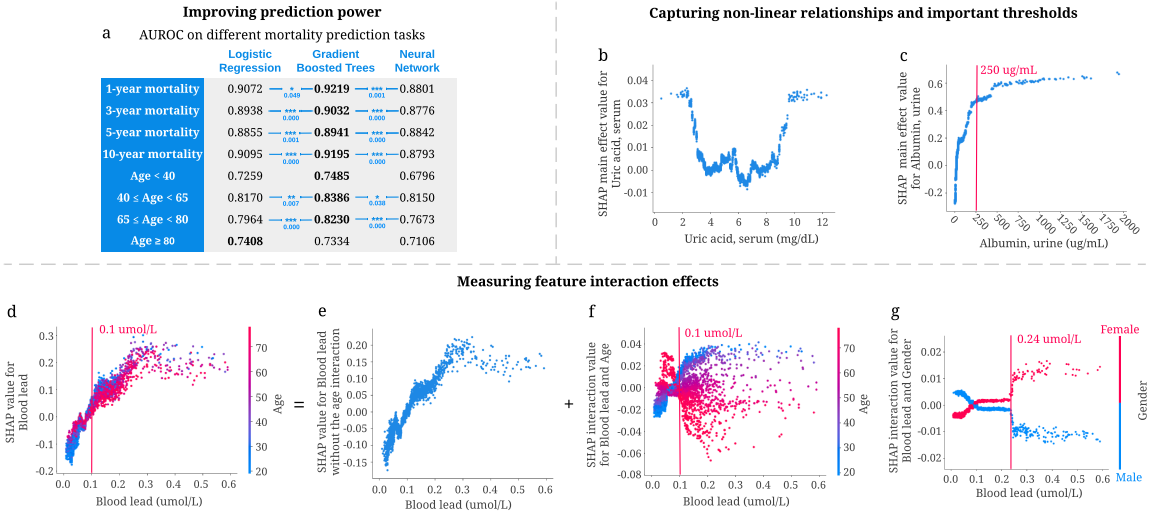


Figure 4.2: **Advantages of tree-based models for mortality prediction.** (a) The area under the ROC curve (AUROC) of gradient boosted tree models outperforms both linear models and neural networks for seven of our prediction models.$(***)$ represents a p-value < 0.001, $(**)$ represents a p-value < 0.01, and $(*)$ represents a p-value < 0.05. P values highlighted in blue are computed using bootstrap resampling over the tested time points while measuring the difference in area between the curves with $n = 1000$ independently resampling. (b,c) Tree-based models can capture non-linear relationships and important thresholds. (b) The main effect of uric acid on 5-year mortality. Higher SHAP value leads to higher mortality risk. (c) The main effect of urine albumin on 5-year mortality. (d–g) Tree-based models can measure feature interaction effects. (d) SHAP value for blood lead level in the 5-year mortality model. Each dot corresponds to an individual. The color corresponds to the value of a second feature (i.e., age) that has an interaction effect with blood lead. (e) We can use SHAP interaction values to remove the interaction effect of age from the model and obtain the SHAP value of blood lead without the age interaction on 5-year mortality. (f) Plotting just the interaction effect of blood lead with age shows how the effect of blood lead on mortality risk varies with age. (g) The SHAP interaction value of blood lead vs. gender in the 5-year mortality model.

with traditional linear models. Our study illustrates the advantages of tree-based models in epidemiology, including making minimal assumptions, capturing non-linear relationships, important thresholds, and interaction effects.

4.3.1.1 *Tree-based models make minimal assumptions about the data distribution.*

Several assumptions associated with linear models (e.g., linearity, independence, normality, etc.) constrain the features they can use. To satisfy these assumptions, scientists often manually transform non-linear variables before fitting a model (e.g., log-transformation, discretization of continuous variables, etc.). For instance, to explore the effect of blood lead on mortality, researchers first discretized blood lead using different thresholds. They observed that individuals with blood lead levels higher than the threshold had increased mortality risk compared to those with lower blood lead levels [179, 192, 254]. In comparison, tree-based models make minimal assumptions about the data distribution and need no data transformations. Figure **??**d shows a positive relationship between blood lead and 5-year mortality risk. Tree-based models can capture complex relationships directly without needing to manually transform the variables.

4.3.1.2 *Tree-based models capture non-linear relationships and important thresholds.*

Discovering non-linear relationships is important but challenging for epidemiological research using traditional linear models. J-shaped and U-shaped associations are two common and meaningful non-linear relationships [188]. However, linear models must use manually transformed features to capture non-linear relationships. As an example, Suliman, Johnson, García-López, Qureshi, Molinaei, Carrero, Heimbürger, Bárány, Axelsson, and Lindholm [271] used a linear model to show a J-shaped relationship between uric acid levels and mortality in patients with stage 5 chronic kidney disease (CKD) by dividing uric acid level into three categories and calculating the hazard ratio for each. Unlike linear models, tree-based approaches can directly capture non-linear relationships. We observe a U-shaped relationship between uric acid level and all-cause 5-year mortality predictions in Figure **??**b. This relationship differs from the J-shaped one in previous work, possibly because of categorization, which loses essential information about values within the categories.

Additionally, discovering thresholds (i.e., inflection points beyond which changing a feature's value has diminishing returns) is important in epidemiological analysis. Figure **??**c shows that 250 μg/mL is an important threshold: according to our model, increasing urine albumin generally increases 5-year mortality risk; however, urine albumin higher than this threshold has almost the same impact on mortality risk.

4.3.1.3 *Tree-based models capture feature interaction effects.*

Feature interaction examines how the effect of one feature on the outcome differs across strata of another feature, highlighting the complex relationship of two features on the outcome [70]. Tree-based models can naturally capture interaction effects by splitting on different features in the same tree. As shown in Figures **??**d-f, SHAP dependence plots can be decomposed into main effects and interaction effects for each sample. Figure **??**f highlights a specific interaction: the relationship of blood lead level to mortality presents differently for young and old individuals. Specifically, for those with blood lead higher than 0.1 µmol/L, younger individuals have a higher 5-year mortality risk than older individuals. Figure **??**g shows the SHAP interaction effects of gender with blood lead level: females have a higher 5-year mortality risk than males with blood lead levels higher than 0.24 µmol/L. The interaction effects of age and gender with blood lead level cannot be clearly identified without SHAP interaction values because being male or older generally increases mortality risk. These findings underscore how being able to detect interaction effects can expose opportunities for further research.

4.3.2 *Discoveries from 5-year mortality prediction*

Figure **??**a shows a summary plot that displays the magnitude, prevalence, and direction of the effect of the top 20 most impactful features on 5-year mortality prediction (Supplementary Methods). This summary plot provides an integrated explanation of the 5-year IMPACT model. Several features are known to be associated with mortality in epidemiological studies. Our results examine and support these studies' conclusions and surface additional discoveries, including features, thresholds, and non-linear relationships.

4.3.2.1 *IMPACT verifies well-studied features associated with mortality.*

Some of the top 20 most important features for our 5-year mortality prediction models have been previously identified. For example, red cell distribution width (RDW), the second most important feature of the 5-year IMPACT model, has been shown to have a strong positive relationship with mortality in many studies under several conditions [81, 213, 214, 217]. We also find a positive relationship between RDW and risk of mortality (Figure **??**b); moreover, 12.7% is an important threshold over which RDW manifests a positive effect on mortality. Serum albumin level's relation to mortality is also well-studied; previous studies show that serum albumin is negatively associated with mortality risk [62, 96, 221]. The relationship shown in Figure **??**c matches this trend. Furthermore, Corti, Guralnik, Salive, and Sorkin [62] showed that serum albumin level<35 g/L was associated with an increased risk of mortality compared to serum albumin levels greater than 43 g/L [62]. We observe that 35 g/L and 43 g/L are indeed key inflection points (Figure **??**c): serum

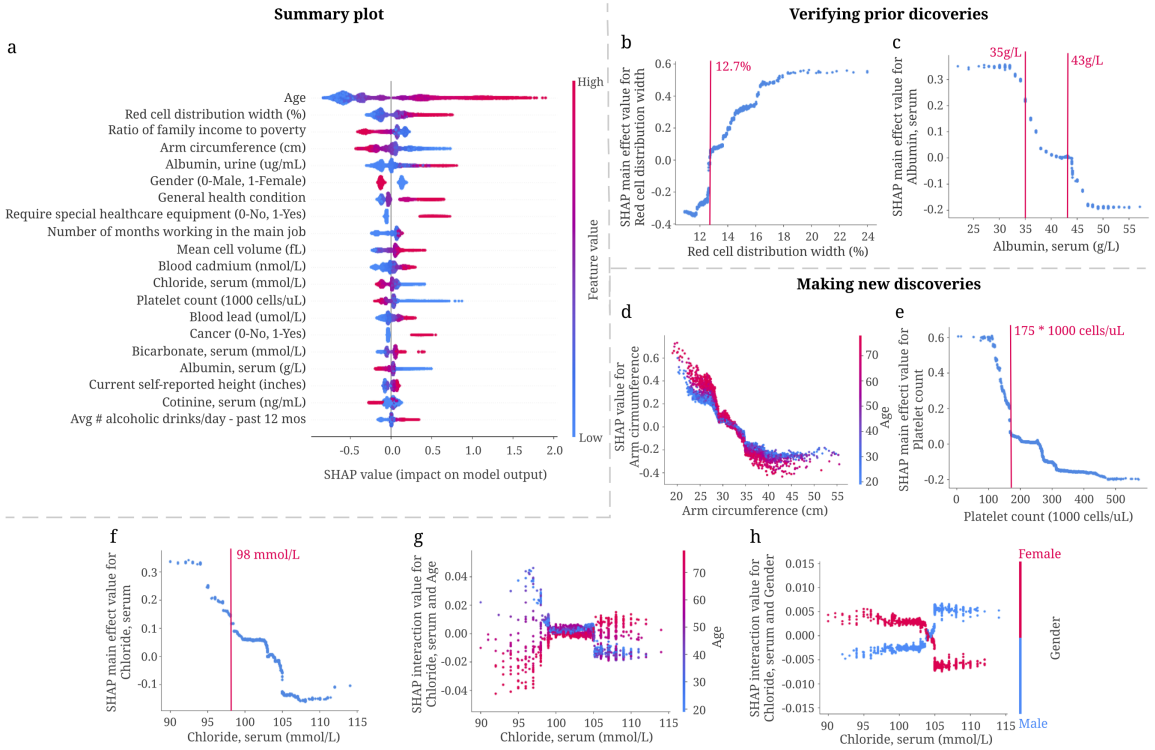


Figure 4.3: **Combining 5-year mortality prediction gradient boosted trees models and local explanations to achieve significant discoveries about the entire model and individual features.** (a) SHAP summary plot for the gradient boosted trees trained on the 5-year mortality prediction task. The plot shows the most impactful features on prediction (ranked from most to least important) and the distribution of the impacts of each feature on model output, which includes a set of plots where each dot corresponds to an individual. The colors represent feature values for numeric features: red for larger values, and blue for smaller. The thickness of the line comprised of individual dots is determined by the number of examples at a given value. A negative SHAP value (extending to the left) indicates reduced mortality risk, while a positive one (extending to the right) indicates increased mortality risk. (b,c) IMPACT can verify well-studied features associated with mortality. (b) The main effect of red cell distribution width on 5-year mortality. (c) The main effect of serum albumin on 5-year mortality. (d-h) IMPACT can identify less well-studied features associated with mortality. (d) The SHAP value for arm circumference in 5-year mortality model. (e) The main effect of platelet count on 5-year mortality. (f) The main effect of serum chloride on 5-year mortality. (g) The SHAP interaction value of serum chloride vs. age in the 5-year mortality model. (h) The SHAP interaction value of serum chloride vs. gender in the 5-year mortality model.

albumin levels lower than 43 g/L have a positive relationship with mortality prediction, while those around 35 g/L are associated with a dramatically increased mortality risk.

4.3.2.2 *IMPACT identifies less well-studied features associated with mortality.*

Some of the top 20 most important features identified by IMPACT are less appreciated as mortality risk factors in the existing epidemiological literature. Three of these are arm circumference, platelet count, and serum chloride level. Figure **??**d shows a negative relationship between arm circumference and 5-year mortality, especially for older people. This negative relationship is consistent with previous work [9, 340]. IMPACT ranks arm circumference as the fourth most important feature for 5-year mortality prediction, with an importance ranking that greatly exceeds that of BMI (the 56th). This suggests that smaller arm circumference is more predictive than BMI for modeling mortality, as in [291].

Figure **??**e shows a negative relationship between platelet count, the 13th most important feature, and 5-year mortality. $175 \times 1,000$ cells/μL is an important threshold; platelet count lower than that level is associated with dramatically increased mortality risk. Serum chloride is also inversely related to 5-year mortality (Figure **??**f). The normal adult value for chloride is 98-106 mmol/L. We observe that serum chloride lower than 98 mmol/L is associated with sharply increased mortality risk. In Figures **??**g-h, we plot the interaction effect of age and sex with serum chloride level. This analysis reveals that younger people and females with low serum chloride have a higher mortality risk than older people and males. The interaction effect of age and serum chloride shows that early rather than late-onset low chloride level has a greater effect on the model.

4.3.2.3 *IMPACT can provide an additional perspective to laboratory reference intervals.*

A reference interval (RI) is the range of values that is deemed normal for a physiologic measurement in healthy persons [135]. It is the most commonly used decision support tool to interpret patient laboratory test results. RIs enable differentiation of healthy and unhealthy individuals [127, 210]. Hence, the quality of the RIs is as crucial as the quality of the result itself. RIs in use today are most commonly defined as the central 95% of laboratory test results in a reference population. Unfortunately, this definition does not consider mortality or disease risk, which may lead to misdiagnosis since RIs are often used to identify unhealthy individuals. The partial dependence plots (Supplementary Methods) of IMPACT models directly reflect the effects of the features on mortality risk, which provides an alternative perspective for identifying inappropriate reference intervals with mortality/disease relevance.

We define the relative risk percentage (RRP; Supplementary Methods) that measures the relative risk of the feature values within the reference interval compared to the relative risk of all values (Table **??**). A higher RRP indicates that the feature values within the

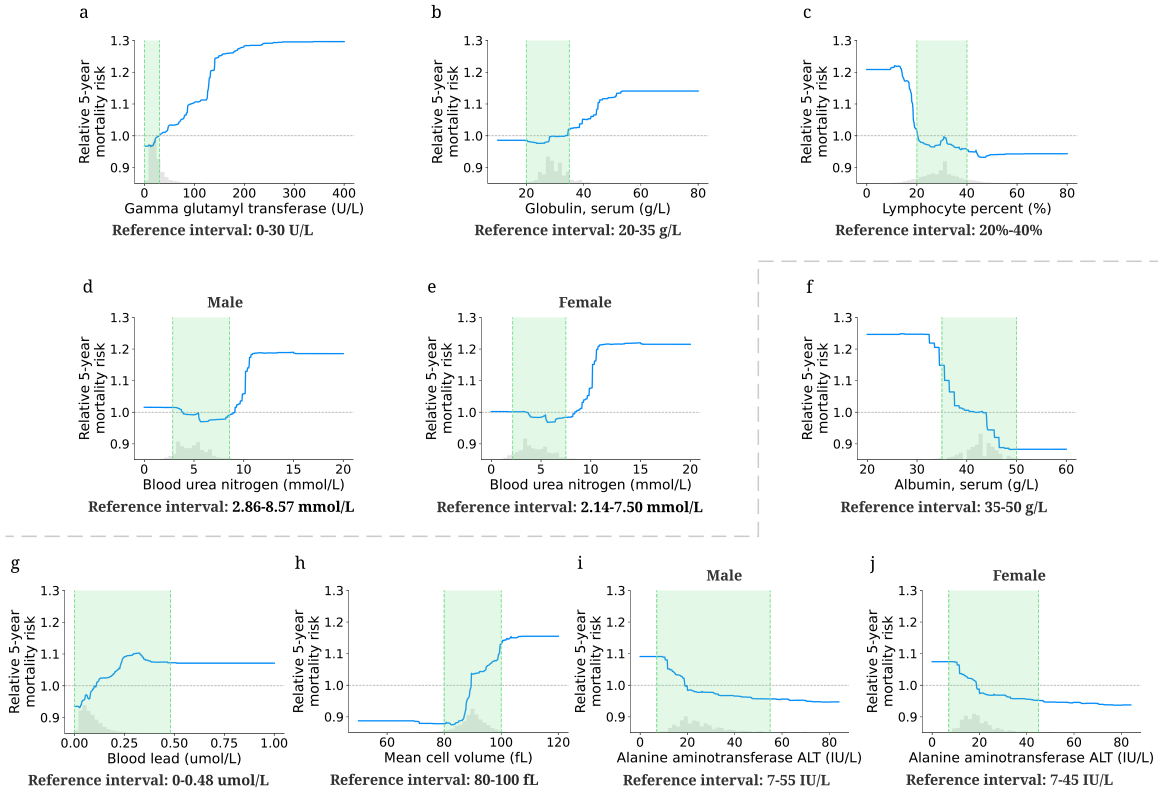


Figure 4.4: **Effect of varying laboratory feature values on 5-year mortality risk.** The partial dependence plots show the change in relative 5-year mortality risk for all values of a given laboratory feature. The grey histograms on each plot show the distribution of values for that feature in the test set. The green shaded region shows the reference interval of each feature. The grey dotted line shows the average value of the model predicted probability (y=1). (a-e) The partial dependence plots of the features whose reference intervals are optimal for mortality risk. (f-j) The partial dependence plots of the features whose reference intervals are sub-optimal for mortality risk.

Feature	Reference Interval	Relative Risk Percentage (RRP)			
		1-year	3-year	5-year	10-year
Gamma glutamyl transferase	0-30 U/L	16.93%	-4.57%	-0.97%	-6.04%
Globulin, serum	20-35 g/L	5.39%	7.95%	14.73%	4.59%
Lymphocyte percent	20%-40%	15.63%	7.02%	6.55%	10.81%
Blood urea nitrogen (Male)	2.86-8.57 mmol/L	8.12%	2.92%	8.02%	21.08%
Blood urea nitrogen (Female)	2.14-7.50 mmol/L	-0.15%	3.07%	0.40%	12.16%
Albumin, serum	35-50 g/L	28.56%	49.70%	59.77%	93.48%
Blood lead	0-0.48 umol/L	100.00%	94.71%	100.00%	100.00%
Mean cell volume	80-100 fL	82.80%	75.82%	83.92%	57.26%
Alanine aminotransferase ALT (Male)	7-55 IU/L	100.00%	100.00%	100.00%	100.00%
Alanine aminotransferase ALT (Female)	7-45 IU/L	100.00%	100.00%	100.00%	100.00%

Table 4.2: **Providing additional perspective to laboratory reference intervals.** The table lists the reference interval and relative risk percentage (RRP) of the selected laboratory features. RRP measures the relative risk of the feature values within the reference interval compared to the relative risk of all values. A higher RRP indicates that the current reference interval is relatively more inappropriate. The negative value indicates that the reference interval of that laboratory feature is optimal for mortality risk. The 100% value suggests that the reference interval may be sub-optimal for mortality risk.

reference interval may lead to high mortality risk, which call for special attention. The first four features in Table **??** have relatively low 5-year mortality RRP. From Figures **??**a-e, we observe that the values of these features within the reference interval have a low 5-year relative mortality risk; the values outside the reference interval may lead to increased 5-year mortality risk. Therefore, IMPACT confirms the reference intervals of these four features as optimal for mortality risk. In contrast, the RRPs of the last four features in Table **??** are high. Figures **??**f-j also shows that the relative 5-year mortality risk of the values within the reference interval is high compared to the maximum relative risk of all values. Hence, IMPACT identified the divergence where reference intervals appear to be poorly tuned to mortality risk, suggesting that these reference intervals may in fact be sub-optimal for health. Note that our goal is not to suggest the optimal reference range: to find the optimal reference interval, more careful sample and study design need to occur. The partial dependence plots for the 1-, 3- and 10-year mortality prediction models are shown in Supplementary Figure **??**.

4.3.2.4 *External Validation of IMPACT on UK Biobank (UKB) dataset.*

We validate the key findings of the 5-year mortality prediction IMPACT model using the UKB dataset. Our external validation includes two aspects. The first aims to validate the entire IMPACT framework using a new dataset by checking whether the explanations from a model trained on the NHANES dataset can also be found in a model trained on the UKB dataset. The second aims to test the generalizability of the mortality prediction model trained on the NHANES dataset.

To validate the IMPACT framework, we train a tree-based 5-year mortality prediction model on the UKB dataset using the 51 overlapping features between NHANES and UKB. Then, we calculate the SHAP values of the UKB mortality prediction model using the UKB samples. Figure **??**a shows the relative global feature importances of the 51 overlapping features of the NHANES model (trained on all 151 features) and the UKB model (trained on 51 features). We can see that the top 20 most important features are largely consistent, where 14 features are the same for both models. The p-value of the Fisher's exact test (p=0.0004) shows that the overlap between the top 20 most important features of NHANES (151 features) and UKB (51 features) model is significant. The Spearman's correlation coefficient of the NHANES and UKB model's feature importance is 0.6654 (p-value < 0.0001), showing the significant positive correlation between the ranking of the overlapping features in NHANES and UKB. It is worth mentioning that waist circumference is more important than BMI in the UKB model, which further validates that some anthropometric measures (i.e., arm circumference in the NHANES model, waist circumference in the UKB model) are more predictive than BMI for modeling mortality. Figures **??**b-d show the relationship between 5-year mortality and three important features: red cell distribution width, serum albumin, and serum uric acid. The trends discovered by the SHAP main effects in the UKB model corroborate previous findings from the NHANES model.

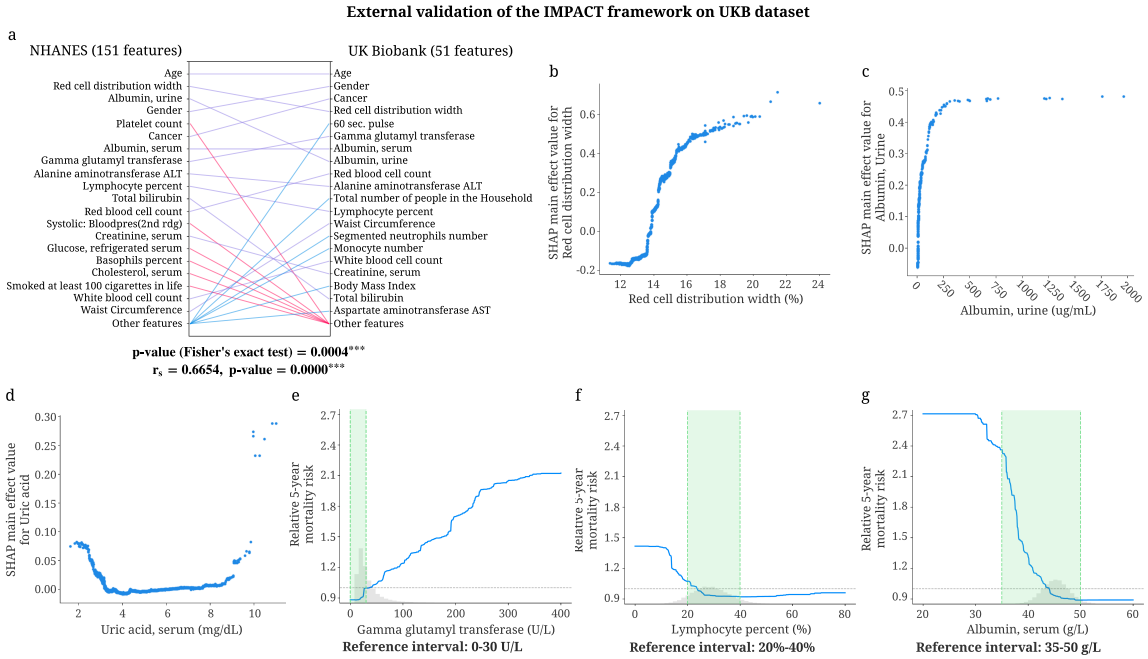


Figure 4.5: **External validation of IMPACT framework on the UKB dataset.** (a) Relative importance of 51 overlapping features in the 5-year mortality prediction models trained on the NHANES (151 features) and UKB (51 features) datasets. For each model, the figure shows the 20 most important features of prediction (ordered by importance). The purple line indicates that the feature is in the top 20 features of both models. Blue and red lines indicate the feature is in the top 20 features of one model but not the other. The p-value of the Fisher's exact test examines the overlap between the top 20 most important overlapping features in the NHANES and UKB models (the contingency table in Supplementary Figure **??**f). The Spearman's correlation coefficient is calculated using the feature importance of the overlapping features in NHANES and UKB ($n = 51$ features). $(***)$ represents a p-value < 0.001. (b–d) The main effect of red cell distribution width, urine albumin and serum uric acid on 5-year mortality in the model trained on UKB (51 features) dataset. (e–g) The relative 5-year mortality risk of gamma glutamyl transferase, lymphocyte percent, and serum albumin in the model trained on the UKB (51 features) dataset.

In Figures **??**e-f, the values of gamma glutamyl transferase and lymphocyte percent in the reference interval have a low 5-year relative mortality risk, which demonstrates that the reference intervals of these two features are optimal for mortality risk. In contrast, Figure **??**g shows that the relative 5-year mortality risk of the values of serum albumin in the reference interval is high, which suggests that the reference interval may be suboptimal for health. These results are consistent with our findings from the NHANES model trained on 151 features. More validation results of IMPACT framework on the UKB dataset are in Supplementary Figure **??**.

Furthermore, we would like to validate whether the performance and explanations of the NHANES prediction model generalize to an unseen population (UKB). Training details and results are described in Supplementary Note **??**, Supplementary Figures **??**–**??**. Our external validation results show that the NHANES mortality prediction model generalizes well to the UKB dataset in terms of both mortality prediction performance and key relationships between features and mortality.

4.3.3 *Discoveries for mortality prediction using different follow-up times*

The relationship between each feature and mortality may change for different models. For instance, comparing important features between IMPACT models using different follow-up times can reveal features that are predictive only for short-term mortality, not longer-term mortality (and vice versa).

Figure **??**a shows the top 20 most important features and relative importance of input features in IMPACT's 1-year, 3-year, 5-year, and 10-year mortality prediction models. Feature importance rankings change greatly between these four models. Some features are important for all four (e.g., age, RDW, and urine albumin level). Some features become more important over time (e.g., platelet count, whose importance ranking is 75 for the 1-year model and 12 for the 10-year model). Other features become less important over time (e.g., serum potassium, whose importance ranking is 17 for the 1-year model and 42 for the 10-year model). These results provide a more comprehensive understanding of shorter- and longer-term mortality risk, which can facilitate the investigation of mechanisms underlying risk predictors and potentially help validate interventions.

The relationship between each feature and mortality may change for models that predict different mortality outcomes or utilize different subsamples of the general population. For instance, Figures **??**b,c show the SHAP value for serum potassium in IMPACT's 1-year and 5-year mortality prediction models. The finding that serum potassium lower than 3.5 mmol/L and higher than 4.0 mmol/L are associated with increased mortality risk has been previously observed [6, 97, 198]. Interestingly, for the 1-year model, hyperkalemia (high potassium) has a higher mortality risk than hypokalemia (low potassium). For the 5-year model, hypokalemia has the same or higher mortality risk than hyperkalemia. Figure **??**d shows that serum sodium higher than 139 mmol/L increases 1-year mortality risk,

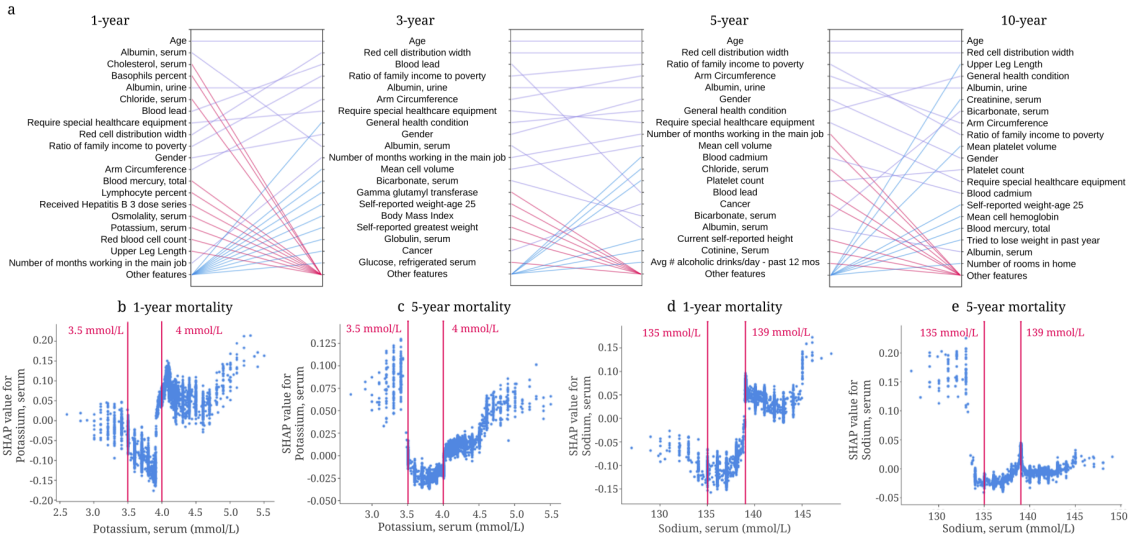


Figure 4.6: **Understanding important risk factors for mortality prediction from tree-based models based on different follow-up times.** (a) Relative importance of input features in 1-, 3-, 5- and 10-year mortality models. For each model, the figure shows the 20 most important features of prediction (ordered by importance). The purple line indicates that the feature is in the top 20 features of two models. Blue and red lines indicate that the feature is in the top 20 features of one model, but not in the top 20 features of the other. (b) The SHAP value of serum potassium in the 1-year mortality model. (c) The SHAP value of serum potassium in the 5-year mortality model. (d) The SHAP value of serum sodium in the 1-year mortality model. (e) The SHAP value of serum sodium in the 5-year mortality model.

and low serum sodium with negative SHAP values decreases mortality risk. However, the relationship differs completely in the 5-year mortality prediction model (Figure **??**e): hyponatremia (serum sodium <135 mmol/L) is associated with a higher 5-year mortality risk. This type of insight, especially regarding the differences of non-linear trends, is not apparent using linear models.

Likewise, we can compare models trained on distinct subpopulations (e.g., samples in different age groups). The differences between these models can help researchers identify risk predictors relevant to each subpopulation and provide epidemiological insights that may guide policy for specific at-risk populations. The discoveries for mortality prediction using different age groups are discussed in Supplementary Note **??** and Supplementary Figure **??**. We further explore explaining the mortality predictions using different age distributions in Supplementary Note **??**, Supplementary Figure **??**.

4.3.4 *Exploring feature redundancy using supervised distance*

Features in datasets are often partially or fully redundant with each other, such that a model could use either feature and achieve the same accuracy. It is important to be aware of redundant features when we interpret a model because these features may include the same information about the output and thereby split the importance of this information. To this end, we propose a supervised distance, which helps us explore and better understand redundant features (Supplementary Methods). Building upon supervised distance, we develop a feature selection method to maximize accuracy and minimize redundancy.

4.3.4.1 *Supervised distances measures feature redundancy and identifies redundant groups of features.*

Researchers often use unsupervised methods, such as some form of correlation-based clustering, to identify dependent features [293, 332]. However, when we have a specific prediction task in mind, we would like to measure feature redundancy with respect to outcome. This can be done using supervised distance, which measures the similarity of two features' information about the prediction task by training one uni-variate model to predict the outcome of another (Supplementary Figure **??**; Supplementary Methods). Supervised distance is scaled roughly between 0 and 1, where 0 distance means the features are perfectly redundant regarding the prediction task and 1 means they are not redundant at all.

To identify groups of redundant features, we hierarchically cluster all features according to supervised distance (Supplementary Figure **??**; Supplementary Methods). Redundant features with the same information about the output group together. For example, arm circumference, the fourth most important feature of the 5-year IMPACT model, is grouped with weight-related features: BMI, waist circumference, weight, etc. These weight-related features all contain similar information about 5-year mortality. To further explore the pre-

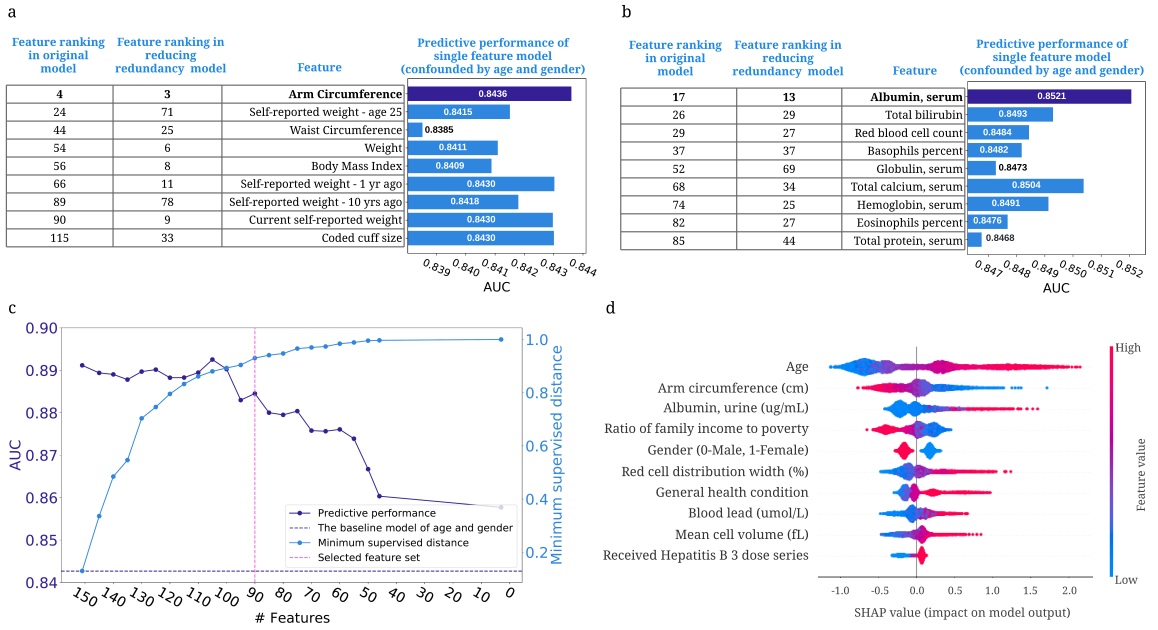


Figure 4.7: **Exploring feature redundancy using supervised distance.** (a) The feature importance ranking of the BMI-related features in original models and reducing redundancy models (models using one weight-related feature and all non-weight-related features), and the AUC of the single-feature models controlling for age and gender. (b) The feature importance ranking of the selected laboratory features in original models and reducing redundancy models, and the AUROC of the single-feature models confounded by age and gender. (c) The AUROC of the models using the selected feature sets and minimum feature redundancy within the selected feature sets when running supervised distance-based feature selection. The purple dashed line shows the AUROC of the model trained on age and gender. The pink dashed line indicates the feature set we select for further analysis. (d) The SHAP summary plot for the gradient boosted trees trained on the selected 90 features for the 5-year mortality prediction.

dictive ability of the features, we train models using one weight-related feature and all non-weight-related features (reducing redundancy models) and models using one weight-related feature in addition to age and gender (single-feature models) (Supplementary Methods). Arm circumference is the most predictive weight-related feature across all settings (Figure **??**a), and may be more informative than other weight-related features with respect to all-cause mortality. Another example is the cluster that includes many blood test features (Figure **??**b). Similar to arm circumference, serum albumin is the most predictive feature among these blood test features. In summary, using supervised distance, we can easily identify redundant feature groups and select the most representative feature based on predictive power. These selected features can be the strongest risk predictors because they have strong predictive power and can represent a number of features.

4.3.4.2 *Supervised distance-based less-redundant feature selection.*

To address feature redundancy more rigorously, we propose a recursive feature selection method to select predictive and less redundant feature sets based on supervised distance (Supplementary Methods). Figure **??**c shows the predictive power and minimum supervised distance of subsets of features refined by our feature selection approach. We observe that as the number of features declines, the predictive performance drops, and the feature redundancy reduces (as indicated by an increasing minimum supervised distance). The figure shows that when using 90 features, the model can achieve good predictive performance (AUROC = 0.8845), and the minimum supervised distance within the features is high (0.9301). Figure **??**d shows the summary plot of the top 10 features in the 5-year mortality prediction model using the selected 90 features. Since there is less redundancy in the selected features, we mitigate the issue of redundant features splitting credit. This lets us explore more richly the effect of important risk predictors on mortality. In our low redundancy model, arm circumference is selected to represent the weight-related features and still receives high importance. Furthermore, we find that "requiring special healthcare equipment," a top 10 feature in the model trained on all features, is removed from the feature list because it is redundant with "general health condition." In summary, our feature selection method helps remove redundant features while retaining highly predictive features, thereby balancing accuracy and interpretability.

4.3.5 *Highly accurate and efficient interpretable mortality risk scores*

A mortality risk score can help individuals monitor their health status, clinicians stratify high-risk patients, and public health organizations guide policy. Most prior mortality risk scores are built with linear models, such as logistic regression and linear hazard models [90, 116]. However, compared with traditional models, tree-based models achieve higher predictive performance, which can stratify patients better than linear models (Table **??**). Besides predictive performance, we must also consider the feature collection cost. There is

a tradeoff between collecting fewer features (which is less costly) and model performance (cost-vs-accuracy). Moreover, the cost of features differs for different users. For example, blood test features are easily collected by clinicians, but, for the public, questionnaire features and examination features are easy to obtain at home. Furthermore, in addition to calculating their risk scores, users may want to know which features contributed more or less to their risk. To address these problems, we build interpretable tree-based mortality risk scores with different cost-vs-accuracy tradeoffs and different types of features for the general public (demographic, examination, and questionnaire features) and medical professionals (demographic, laboratory features and features from common test panels) to use (Supplementary Methods). Compared with previous mortality risk scores, ours are more interpretable, more accurate, applicable to more users, and flexible with respect to different cost-vs-accuracy tradeoffs.

4.3.5.1 *IMPACT develops highly accurate and efficient 5-year mortality risk scores.*

The predicted probability of IMPACT models can be directly used as mortality risk scores (IMPACT risk scores). We did a temporal validation of the risk scores by training and validating them in samples from NHANES 1999-2008 and assessing their performances in NHANES 2009-2014. The sample size, the number of deceased samples and the histogram of age in the training set with the testing set and the temporal validation set are shown in Supplementary Figure **??**. For comparison, we train linear and tree-based Cox proportional hazard models widely used in previous work (Supplementary Methods). To find less costly but nearly as accurate models, we select the features using recursive feature elimination (RFE; Supplementary Methods). Moreover, we compare IMPACT risk scores with Intermountain sex-specific risk scores [116] (Supplementary Methods). The models are evaluated on different gender groups.

In Figures **??**a-b, we show the AUROC of the 5-year mortality risk scores of female samples (Supplementary Figure **??** for male results) in the test set and the temporal validation set. We see that the IMPACT model with only 20 features obtains an AUROC of 0.8971, which is almost as same as the performance of the model using all features (AUROC = 0.9030); using fewer than 20 features leads to a dramatic accuracy drop. Figures **??**a,b also show that IMPACT models achieve better performance than linear and tree-based Cox proportional hazard models. Furthermore, we see that the IMPACT risk score using the laboratory features (AUROC = 0.8881) and the risk score using the questionnaire and examination features (AUROC = 0.8835) both achieve acceptable predictive performance. The IMPACT risk score using the features from common test panels achieves higher AUROCs than the intermountain risk score, which uses CBC and BMP panels features. With the models trained with different cost-vs-accuracy tradeoffs, users who cannot measure certain features (i.e., high-cost features) can still calculate accurate mortality risk scores. Figure **??**b shows that the performance of our models drops only a little on the temporal validation set, which can indicate that our risk scores generalize fairly well. The selected

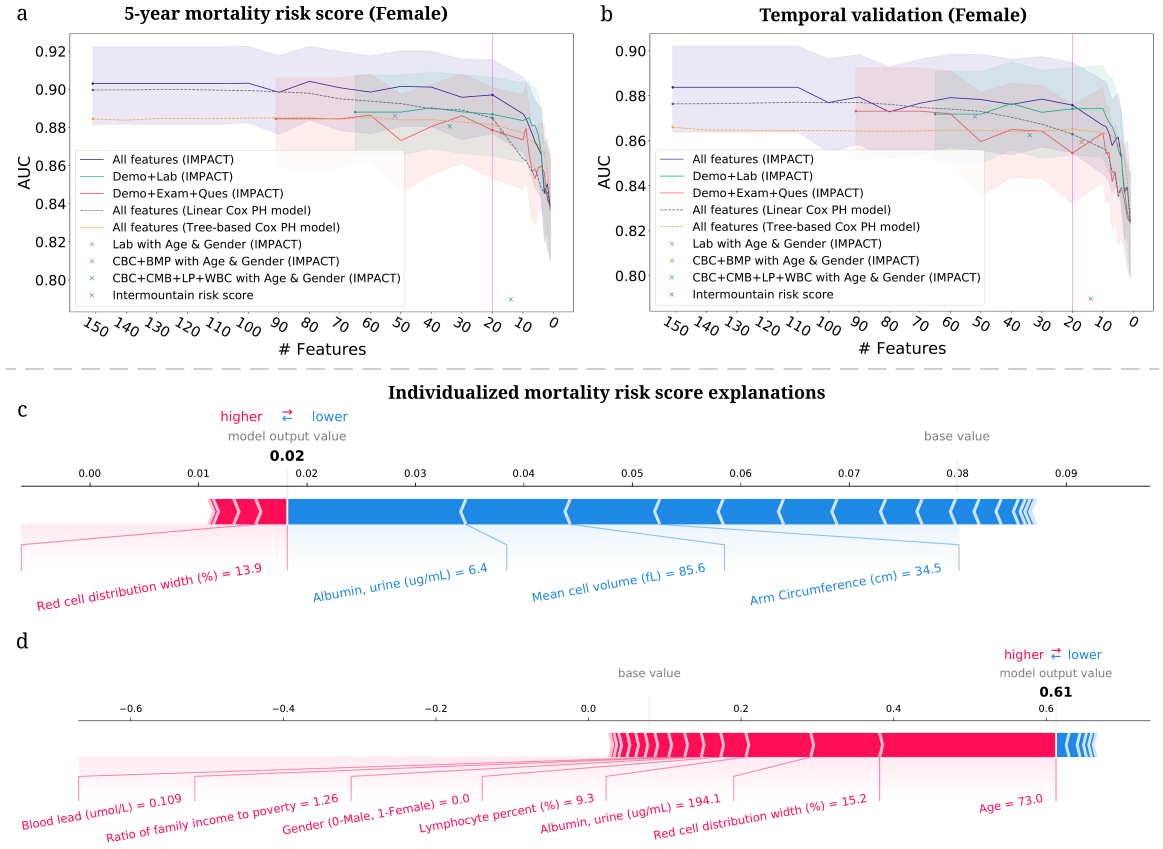


Figure 4.8: **Developing highly accurate and efficient interpretable 5-year mortality risk scores.** (a,b) The AUROC of the models using different feature sets after recursive feature elimination. Lines are mean performance over 1000 random train/test splits, and shaded bands are 95 percent normal confidence intervals. (a) The AUROC of the models tested on the female group in the test set of NHANES 1999-2008. (b) The AUROC of the models tested on the female group in the temporal validation set (NHANES 2009-2014). (c,d) IMPACT can analyze individualized mortality risk scores. (c) The individualized explanation for an individual who is alive after 5 years. The output value is the risk score for that individual. The base value is the mean risk score, i.e., the score that would be predicted if we did not know any features for the current output. The features in red increase mortality risk, and those in blue decrease it. (d) The individualized explanation for a sample who is deceased after 5 years.

top 20 features and features from CBC and BMP panels are listed in Supplementary Table **??**. In summary, we build IMPACT risk scores that are applicable to professional and non-professional individuals with different cost-vs-accuracy tradeoffs.

4.3.5.2 *IMPACT exposes individualized mortality risk score explanations.*

TreeExplainer can help researchers analyze the prediction for each individual and illustrate each features' contribution to the mortality risk score. We explain the mortality prediction model in terms of its probability predictions (risk scores). Figures **??**c,d show individualized explanations for two people in the model using the top 20 features (Supplementary Methods). The first individual (Figure **??**c) was alive after 5 years. From the figure, we observe that IMPACT predicted that the individual's 5-year mortality risk score was 0.02, lower than the average predicted risk (i.e., base value). Certain features can increase mortality risk, such as red cell distribution width, and others can decrease it, such as urine albumin level. For this individual, the features that drive down mortality risk outweigh those that increase it. The second individual (Figure **??**d) was deceased after 5 years, and the model's predicted mortality probability is 0.61, much higher than the average predicted risk. The top three features that increase this individual's risk are high age, high red cell distribution width, and high urine albumin concentration. The interpretable risk score can both help individuals improve health awareness and understand their health status, and it can help health professionals identify high-risk individuals.

4.4 DISCUSSION

IMPACT combines high-accuracy complex ML models and state-of-the-art local explanation methods to allow the systematic study of all-cause mortality. In epidemiology, high accuracy is neccessary but insufficient; explaining models to humans is also essential for drawing epidemiological hypotheses [314, 315]. IMPACT's combination of accuracy and explanation aims to optimize accuracy while also gaining insight into complex interrelations between mortality and an individual's features.

Using 151 features in NHANES 1999-2014, we build tree-based mortality prediction models and explore the effect of those features on mortality for different follow-up times and age groups. Importantly, we demonstrate the value and significance of explaining complex ML prognostic models. IMPACT lets us to capture both non-linear and interaction effects that are difficult to uncover with linear models. These results help us verify well-studied findings (e.g. the relationship of red cell distribution width and serum albumin with mortality) as well as identify less well-studied ones (e.g. the important risk predictors arm circumference, platelet count and serum chloride, and the complex interactions among the features). One pitfall to inferring relationships between determinants and an outcome are relationships between the determinants themselves (redundancy). To address this, we propose a supervised distance and feature selection approach, which we

utilize to select the minimally redundant feature sets. Finally, we build easy-to-use and explainable mortality risk scores for use by both the general public and medical professionals with different tradeoffs between feature collection cost and model performance. These scores can help individuals improve self-awareness of their health status and help clinicians identify patients with high mortality risk to target with specific interventions. In this paper, we present only a small part of our findings. All our results and risk scores are available for public use in an interactive website (https://suinleelab.github.io/IMPACT), where associations and interactions can be explored in detail to generate new research hypotheses.

In terms of epidemiological findings, this study shows a negative relationship between arm circumference and mortality. Our clustering method groups arm circumference with BMI and other weight-related features, indicating that these features share information about mortality. Several prior studies have found a U-shaped association between BMI and mortality, where very low or very high BMI is associated with greater mortality risk [9, 109]. This U-shaped relationship may be the result of compound effects from body fat and fat-free mass. Since upper arm circumference is an indicator of fat-free mass [9, 340], it may be the case that fat-free mass is driving the inverse correlation between arm circumference and mortality risk. Larger arm circumference is expected to be associated with greater muscle mass, while smaller arm circumference may reflect muscle deterioration along with diminished nutritional status or malnutrition [251, 321]. The importance of arm circumference in IMPACT is consistent with previous studies, which show that low arm circumference was more effective than low BMI in predicting follow-up mortality risk in older people [251, 292, 317].

One limitation of IMPACT is that the relationships and interactions detected by our model cannot be claimed to be causal. This is not unique to our method and poses a fundamental problem in epidemiological studies using observational data. The purpose of this study is not to address causality, but rather to conduct a systematic study of mortality associations with the NHANES population. In particular, a primary obstacle in capturing causal effects with observational data and predictive models are confounding variables. In order to condition on confounders (and potential surrogate confounders), it is often desirable to include as many features as possible in the model [253]. Conversely, we may want to remove colliders and mediators that skew the real effect of treatment features of interest. Our solution to redundancy, i.e., supervised distance, can potentially help narrow down related features for which domain experts can identify colliders, mediators, and confounders. This is a potential future research question that takes a step in the direction of making explanations from complex models causal.

Our study is performed on NHANES 1999-2014 data, which is designed to assess the health status of participants in the United States. We perform temporal validation within the NHANES samples to evaluate the performance of our mortality risk scores. To evaluate the generalizability of important features and relationships, we implement

the IMPACT model on a geographically distinct dataset with samples exclusively from the United Kingdom (UK Biobank). Although our qualitative findings were consistent between NHANES and UK Biobank, there are differences between both populations, primarily in terms of age (37-72 in UKB vs. 18-80+ in NHANES), which also affects the base rates of mortality in each data set. As such, further external validation of our mortality models on datasets with similar distribution of variables and mortality rates should be undertaken to further increase the generalizability of the findings.

Over the past several years, a variety of ML approaches have been applied in the field of aging research to develop "clocks" that can predict the chronological age of an individual based on different phenotypic features [338]. The most common of these are the epigenetic clocks that have identified patterns of methylation on DNA that change with age and can be used to predict chronological age with high accuracy across a variety of different species and tissue types [118, 203]. Other clocks based on gene expression, metabolites, facial features, telomere length, etc., have also been described [323]. Efforts have also been made to use these clocks to predict an individual's biological age, which may differ from their chronological age if they are aging more rapidly or slowly than the general population. Such "biological aging clocks" are expected to reflect the underlying health status of the individual and be useful for predicting future health outcomes and mortality. Although we have not yet attempted to validate IMPACT as a tool for assessing biological age, those individuals with lower IMPACT mortality risk than expected for their chronological age would be predicted to have a lower biological age, and vice-versa. Because IMPACT is trained to predict all-cause mortality rather than fit to chronological age, it will be of interest to determine how IMPACT compares to these various clocks in predictive capacity, particularly if done for the same cohort of individuals.

Prognosis research using complex ML models will likely increase over the coming years as ML techniques continue to rapidly develop. However, "black box" ML models that predict without explaining are difficult for clinicians to trust and difficult to extract meaningful information from. Therefore, the combination of complex ML models and 'explainable artificial intelligence' (XAI) is necessary and urgent. IMPACT takes a consequential step towards XAI for mortality prediction. This study's improvement in predictive accuracy and explanation of complex ML models warrants further exploration for other epidemiological outcomes.

4.A SUPPLEMENTARY METHODS

4.A.1 *Data collection and processing*

The National Health and Nutrition Examination Survey (NHANES) from the National Center for Health Statistics (NCHS)[1] conducts interviews and physical examinations to

1 http://www.cdc.gov/nchs/nhanes.htm

assess the health and nutrition data for all ages in the United States. The interviews include demographic, socioeconomic, dietary, and health-related questions. The examinations include medical, dental, physiological measurements, and laboratory tests administered by highly trained medical personnel. Since 1999, data were collected and released at 2-year intervals. Each year NHANES examines a nationally representative sample of roughly 5,000 individuals across the Unites States. The design of the sample changed periodically. Oversampled subgroups for 1999–2006 included non-Hispanic black persons, Mexican-American persons, low-income white persons (beginning in 2000), adolescents aged 12–19, and persons aged 70 and over. Oversampled subgroups for 2007–2010 included all Hispanic persons, non-Hispanic black persons, low-income white persons, and persons aged 80 and over[1]. In this study, we include NHANES data sampled between 1999 and 2014. All-cause mortality is ascertained by a linked NHANES mortality file that provides follow-up mortality data from the date of survey participation through December 31, 2015. We exclude participants under age 18 because they are not eligible for public release mortality data[2].

Our study includes samples with known mortality status who participated in NHANES 1999-2014 ($n = 47,261$). In the raw data, individuals 85 and over are topcoded at 85 years of age in NHANES 1999-2006 and individuals 80 and over are topcoded at 80 years of age in NHANES 2007-2014. To keep consistency, we topcode individuals 80 and over at 80 years of age. The histogram of the the samples' age in different data collection cycles are shown in Supplementary Figure **??**. We include all demographic, laboratory, examination, and questionnaire features that could be automatically matched across different NHANES cycles. We exclude variables that are missing for more than 50% of the participants and highly correlated features with correlations greater than 0.98; after filtering and one-hot encoding, 151 features remain. We impute missing data using MissForest [267], a nonparametric random forest-based multiple imputation method for mixed-type data, with seven iterations. We predict all-cause mortality for two broad categories: (1) follow-up times of 1-year, 3-year, and 5-year and (2) age groups of $<$40, 40-65, 65-80, and $\geqslant$80 years old. For different follow-up times, we remove samples with unconfirmed mortality status. For different age groups, we predict 5-year mortality. The demographic characteristics and sample size of the data for different tasks are shown in Supplementary Table **??**.

We use UK Biobank samples as an external validation dataset. Participants were enrolled in the UK Biobank from April, 2007, to July, 2010, from 21 assessment centres across England, Wales, and Scotland using standardised procedures. When participants agreed to take part in UK Biobank, they visited their closest assessment centre to provide baseline information, physical measures, and biological samples. We include the 51 features that are overlapping between NHANES and UK Biobank dataset. We exclude samples with

1 https://www.cdc.gov/nchs/data/series/sr_01/sr01_056.pdf
2 https://www.cdc.gov/nchs/data/datalinkage/public-use-2015-linked-mortality-file-description.pdf

missing values. All-cause mortality included all deaths occurring before May, 2021. We include 384,762 samples aged 37-72 years with confirmed 5-year mortality status. Of these samples, 6,336 died after 5 years. The histograms of age, gender and body mass index of UK Biobank samples are shown in Supplementary Figure **??**.

4.A.2 *Predictive modeling*

To model mortality, we use gradient boosted trees (GBTs). GBTs are nonparametric methods composed of iteratively trained decision trees. The final ensemble of trees captures non-linearity and interactions between predictors. The dataset is randomly divided into training (80%) and testing (20%) sets. We use the implementation XGBoost [52][2] with a learning rate set to 0.002 , subsample ratio set to 0.5 and 10,000 trees of max depth 3. For comparison, we also train logistic regression models and deep neural networks. For logistic regression, we use L2 regularization. The L2 regularization weight was set to 100. For neural networks, we use a single layer with 1,000 nodes, and max iteration set to 1,000. The hyperparameters specified above are chosen by GridSearch and 5-fold cross validation. Other hyperparameter values are left at their default values. Models' performance is measured with the area under the receiver operator characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC). We bootstrap the test set for 1,000 times to assess the statistical significance of the difference in AUROC and AUPRC for pairs of models. Specifically, we resample with replacement from the test set 1,000 times and compare the models' performance on resampled test sets. We report a p-value which is the percentage of time that logistic regression or the neural network's performance is better than or equal to gradient boosted trees, divided by the number of resampled test sets. All models are built using the Scikit-learn package in Python 3.7.

4.A.3 *Model interpretation*

To explain the GBT models, we utilize TreeExplainer [177], which provides a local explanation of the impact of input features on individual predictions. Specifically, TreeExplainer calculates exact SHAP [178] (SHapley Additive exPlanations) values for tree-based models. When explaining the mortality prediction models, we randomly select 10,000 background samples from the training set and 5,000 foreground samples from the test set.

4.A.3.1 *SHAP (SHapley Additive exPlanation) values*

SHAP (SHapley Additive exPlanation) values attribute to each feature the change in the expected model prediction when conditioning on that feature. The change of the model's prediction when the feature is masked is recorded across all possible subsets of features,

2 https://xgboost.readthedocs.io/en/latest/python/index.html

yielding an average change in prediction resulting from the inclusion of a feature in the model:

$$\phi_i(f, x) = \sum_{R \in \mathcal{R}} \frac{1}{M!} \left[f_x(P_i^R \cup i) - f_x(P_i^R) \right], \tag{4.3}$$

where ϕ_i is the feature attribution (SHAP value) of feature i in model f for data point x, $\mathcal{R}$ is the set of all feature permutations, P_i^R is the set of all features before i in the ordering R, M is the number of input features, and f_x is an estimate of the conditional expectation of the model's prediction: $f_x(S) \approx E[f(x) \mid x_S]$ where x_S is the set of observed features.

SHAP values which guarantee a set of desirable theoretical properties, including additivity and consistency. Additivity states that when approximating the original model f for a specific input x, the SHAP values sum up to the output $f(x)$:

$$f(x) = \phi_0(f) + \sum_{i=1}^{M} \phi_i(f, x), \tag{4.4}$$

The sum of feature attributions (SHAP values) matches the original model output f(x), where $\phi_0(f) = E[f(z)] = f_x(\emptyset)$. Consistency states that if a model changes so that some feature's contribution increases or stays the same regardless of the other inputs, that input's attribution should not decrease. Therefore, SHAP values are consistent and accurate calculations of each feature's contribution to the model's prediction.

4.A.3.2 *SHAP interaction values and main effects*

The SHAP interaction effects is based on the Shapley interaction index from game theory. While standard feature attribution results in a vector of values, one for each feature, attributions based on the Shapley interaction index result in a matrix of feature attributions. The main effects are on the diagonal and the interaction effects on the off-diagonal. The **SHAP interaction values** are defined as:

$$\Phi_{i,j}(f, x) = \sum_{S \subseteq \mathcal{M} \setminus \{i,j\}} \frac{|S|!(M - |S| - 2)!}{2(M-1)!} \nabla_{i,j}(f, x, S), \tag{4.5}$$

when $i \neq j$, and

$$\nabla_{ij}(f, x, S) = f_x(S \cup \{i, j\}) - f_x(S \cup \{i\}) - f_x(S \cup \{j\}) + f_x(S). \tag{4.6}$$

where $\mathcal{M}$ is the set of all M input features. In Equation **??** the SHAP interaction value between feature i and feature j is split equally between each feature so $\Phi_{i,f}(f, x) = \Phi_{j,i}(f, x)$ and the total interaction effect is $\Phi_{i,f}(f, x) + \Phi_{j,i}(f, x)$.

The **main effects** for a prediction can then be defined as the difference between the SHAP values and the off-diagonal SHAP interaction values for a feature:

$$\Phi_{i,i}(f, x) = \phi_i(f, x) - \sum_{j \neq i} \Phi_{i,j}(f, x). \tag{4.7}$$

4.A.3.3 *Partial dependence plots and additional perspective to reference interval*

We use partial dependence plots to show the change in mortality risk for all values of a laboratory feature. Partial dependence plots show the marginal effect a set of features has on the prediction of an ML model. The partial function f_S is estimated by:

$$f_S(x_S) = \frac{1}{n}\sum_{i=1}^{n} f(x_S, x_C^{(i)}). \tag{4.8}$$

In this formula, f is a ML model and the S are features for which the partial dependence function should be plotted. In our study, S is the laboratory feature of interest and x_S is the given value of the feature. $x_C^{(i)}$ is actual feature values for the features of no interest in the test set, and n is the number of instances in the test set. The partial function tells us the average marginal effect on the prediction for given value(s) of features S. We extend the partial function to the relative mortality risk RR_S:

$$RR_S(x_S) = f_S(x_S)/(\frac{1}{n}\sum_{i=1}^{n} f(x^{(i)})). \tag{4.9}$$

In other words, the relative mortality risk is defined as the average value of the model's predicted probability when we fix a specific feature to a given value divided by the average value of the model's predicted probability. We further define the relative risk percentage (RRP) as follows:

$$RRP_S = \frac{max(RR_S(x_S), x_S \text{ in RI}) - 1}{max(RR_S(x_S)) - 1}, \tag{4.10}$$

where RI stands for reference interval. High relative risk percentage indicates that the values within the reference interval have a relatively high mortality risk. The partial dependence plots of selected laboratory feature values on 1-, 3-, and 10-year mortality risk are shown in Supplementary Figure **??**.

4.A.4 *Model interpretation plots*

In this section we describe a number of plotting types for model explanation visualization.

SHAP value, SHAP main effect value and SHAP interaction value plots In SHAP value/SHAP main effect value/SHAP interaction value plots, every point corresponds to a single sample where the x-axis is the value of the feature and the y-axis is the SHAP value/SHAP main effect value/SHAP interaction value. The coloring of the points often denotes the value of a separate feature.

Summary plot Summary plots show the feature attributions (SHAP values) for many samples and multiple features in order of global feature importance (the mean absolute

SHAP values). Summary plots stack multiple subplots for each feature. For the feature plots, every point corresponds to a single sample where the x-axis is the feature attribution value and the y-axis is vertical dispersion representing the frequency of samples with a particular feature attribution value. Finally, the color of each point represents the normalized feature value, with red representing a high value and blue representing a low one. Intermediary feature values are interpolations between red and blue.

Individualized explanation plot Individualized explanation plots show the feature attributions (SHAP values) for an individual in terms of how they drive the model's prediction for the individual away from the average model prediction across the baseline distribution. The width of the bars indicate the SHAP value with red indicating a positive affect and blue indicating a negative one. The features corresponding to the largest bars are below with their actual values for the individual.

4.A.5 *Supervised distance*

4.A.5.1 *Supervised distance and hierarchical clustering*

Supervised distance can accurately measure feature redundancy based on a specific prediction task. As Supplementary Figure **??** shows, to calculate the supervised distance between feature i and feature j, we first train a uni-variate GBT model to predict the label (e.g. 5-year mortality in our study) using feature i. Then, we can obtain the $\mathtt{Prediction}_i$ which is the output of the fitted uni-variate GBT. Next, we fit another uni-variate GBT to predict $\mathtt{Prediction}_i$ using feature j. We define the output of the new GBT as $\mathtt{Prediction}_i^j$. All hyperparameter values of the uni-variate GBTs are set to their default values. Following the same above steps, we can obtain $\mathtt{Prediction}_j^i$. The supervised distance between feature i and feature j (supervised distance(i,j)) is defined as:

$$\mathit{supervised}\, R^2(i,j) = max(0, 1 - mean(\frac{(\mathtt{Prediction}_i - \mathtt{Prediction}_i^j)^2}{var(\mathtt{Prediction}_i)})) \tag{4.11}$$

$$\mathit{supervised\ distance}(i,j) = max(1 - \mathit{supervised}\, R^2(i,j), 1 - \mathit{supervised}\, R^2(j,i)) \tag{4.12}$$

where $var(x)$ is the variance of the vector x, $mean(x)$ is the average of the vector x. Supervised distance is scaled roughly between 0 and 1, where 0 distance means the features perfectly redundant and 1 means they are completely independent.

To explore the redundant feature groups, we hierarchically cluster all features according to the supervised distance. Specifically, we use complete linkage hierarchical clustering which merges in each step the two clusters whose merger has the smallest diameter. The hierarchical clustering tree is shown in Supplementary Figure **??**.

4.A.5.2 *Redundant feature groups experiments training details*

Reducing redundancy model To identify the most representative feature in a redundant feature group, we train GBTs using one feature in the redundancy group and all features outside the group for 5-year mortality prediction. Then we compare the feature importance ranking of the redundant features by calculating the mean absolute SHAP values using TreeExplainer. The hyperparameters of the GBTs are chosen by GridSearch and 5-fold cross validation. The max depth is selected from $\{1, 3, 5, 7, 9\}$ and the subsample ratio is selected from $\{0.2, 0.5, 0.8, 1.0\}$. Other hyperparameter values are left at their default values.

Single feature model We further analyze the predictive power of the redundant features by fitting 5-year mortality prediction GBTs using one feature in the redundant feature group. Specifically, we use one feature in the redundant feature group and two important confounders, age and gender, to train a GBTs for 5-year mortality prediction. All hyperparameter values are set to their default values. We compare the AUCs of the models. We bootstrap the test set for 1,000 times and compare the models' performance on resampled test sets. The averages of the AUCs are reported.

4.A.5.3 *Supervised distance-based feature selection*

We propose a supervised distance-based feature selection method to select predictive and less-redundant feature sets. The workflow of our feature selection method is shown in Supplementary Figure **??**. The dataset is randomly divided into training (80%) and testing (20%) sets. Firstly, we fit a GBT for 5-year mortality prediction on all features using the training set and rank the features by mean absolute SHAP values from TreeExplainer. The hyperparameters of the GBTs are chosen by GridSearch and 5-fold cross validation. The max depth is selected from $\{1, 3, 5, 7, 9\}$ and the subsample ratio is selected from $\{0.2, 0.5, 0.8, 1.0\}$. The max number of trees is set to 1000. We use 20% of the training samples as validation set for early stopping. The number of early stopping rounds is set to 100. Since age and gender are important confounders, we would like to keep them in the selected feature set. Therefore, we cluster features except age and gender into a specific number of groups using supervised distances-based hierarchical clustering and select the most important feature in each cluster. Then, we add age and gender to the selected feature set and re-fit the model. Next, we rerun the clustering using the new feature set except age and gender. This process is repeated until all remaining features cluster to a single group. In every iteration, we remove 5 features. The models are evaluated on the testing set with bootstrapping for 1,000 times. We report the average of the AUROCs and the minimum supervised distance within the selected feature sets.

4.A.6 *5-year mortality risk scores*

4.A.6.1 *Mortality risk scores training details*

IMPACT mortality risk scores are defined to be the prediction of the 5-year mortality prediction models. For comparison, we train linear[3] and gradient boosted tree-based Cox proportional hazard models[4]. We do a temporal validation of the risk scores by assessing their performances in the samples collected in 2009-2014 ($N = 7,034$). Specifically, the samples collected in 1999-2008 ($N = 28,820$) are randomly divided into training (80%) and testing (20%) sets. The sample size, the number of deceased samples and the histogram of age in the training set with the testing set and the temporal validation set are shown in Supplementary Figure **??**. To compare with Intermountain gender-specific risk scores, we evaluate the models on different gender groups. The models are trained on the whole training set and evaluate on different gender groups in the testing set. Furthermore, considering the different feature collection cost for the general public and medical professionals, we build the risk scores starting from different feature sets. For the general public, the models are trained on all demographics, questionnaire features and examination features that are accessible at home for general public, For medical professionals, the models are trained on all demographics and laboratory features. All trained models are evaluated on different gender groups of the samples collected in 2009-2014 for temporal validation.

The hyperparameters are chosen by GridSearch and 5-fold cross validation. For XGBoost 5-year mortality prediction models, the max depth is selected from $\{1, 3, 5, 7, 9\}$ and the subsample ratio is selected from $\{0.2, 0.5, 0.8, 1.0\}$. The max number of trees is set to 1000. We use 20% of the training samples as validation set for early stopping. The number of early stopping rounds is set to 100. For linear Cox proportional hazard models, the regularization parameter α is selected from $\{0.01, 0.1, 1, 10, 100\}$. For tree-based Cox proportional hazard models, the max depth is selected from $\{1, 3, 5, 7, 9\}$ and the subsample ratio is selected from $\{0.2, 0.5, 0.8, 1.0\}$. Other hyperparameter values are left at their default values.

We explain the mortality prediction model in terms of its probability predictions. Specifically, we rescale the SHAP values (in the log-odds space) to be in the probability space directly. The rescaled SHAP values now sum to the probability output of the model.

4.A.6.2 *Recursive feature elimination*

Recursive feature elimination works by searching for a subset of features by starting with all features in the training dataset and successively removing features until the desired number of features remains. Firstly, we train a model on the full dataset with all features.

3 https://scikit-survival.readthedocs.io/en/latest/api/generated/sksurv.linear_model.CoxPHSurvivalAnalysis.html
4 https://scikit-survival.readthedocs.io/en/latest/api/generated/sksurv.ensemble.GradientBoostingSurvivalAnalysis.html

Then we rank features by importance (mean absolute SHAP values) and remove the least important features. Another model is trained on the resulting feature set, and the process iterates until only the desired number of features are left. Starting from 151 features, we remove 6 features at the first iteration. Then, we remove 5 features in each iteration until only one feature is left. We bootstrap the test set for 1,000 times and assess the predictive performance. Specifically, we resample with replacement from the test set 1,000 times and report the average and the 95% confidence interval of the AUROCs.

4.A.6.3 *Intermountain mortality risk score*

Intermountain mortality risk scores [116] are built using complete blood count and basic metabolic profile. Specifically, 13 laboratory features are used to predict 30 days, 1-year and 5-year mortality. Logistic regression was used to model the risk prediction equations with adjustment for age and sex. Dummy variables modeled each category, with the referent defined as the lowest risk group (except for age categories: 18-29, 30-39, 40-49 [referent], 50-59, 60-69, 70-79, and $\geqslant$80 years). A scalar score value was derived for each variable category by multiplying its β-coefficient by 3 and rounding to the nearest integer (referent value = zero). Each individual's risk score became the sum of the score values based on his or her individual data. Since all of the features used in the Intermountain risk scores are included in our NHANES dataset, we evaluate the Intermountain risk score on our NHANES testing set with bootstrapping for 1,000 times.

4.A.6.4 *Comparing the predictive power of popular mortality risk scores and biological ages with IMPACT*

Since not all features used in the popular mortality risk scores and biological ages are included in the NHANES dataset (except for Intermountain risk scores; see **??**), it would not be fair to compare the existing mortality scores and biological ages computed based on a partial set of features with the IMPACT model based on the NHANES dataset. Therefore, we chose to show the AUROCs reported in the original papers. As the AUROCs are not sensitive to the base rate, we assume that these scores would be consistent among different datasets if the risk scores and biological ages generalize well.

Table **??** compares the AUROCs between an existing mortality score or a biological age as reported in the original paper and the IMPACT-20 model tested for the corresponding follow-up time and age ranges in the NHANES dataset. Here, IMPACT-20 means the IMPACT model when the top 20 features were used; we chose 20 features because in Figures **??**a,b, the IMPACT model with 20 features obtains an AUROC that is almost the same as the performance of the model using all features, and using fewer than 20 features leads to a dramatic decline in accuracy.

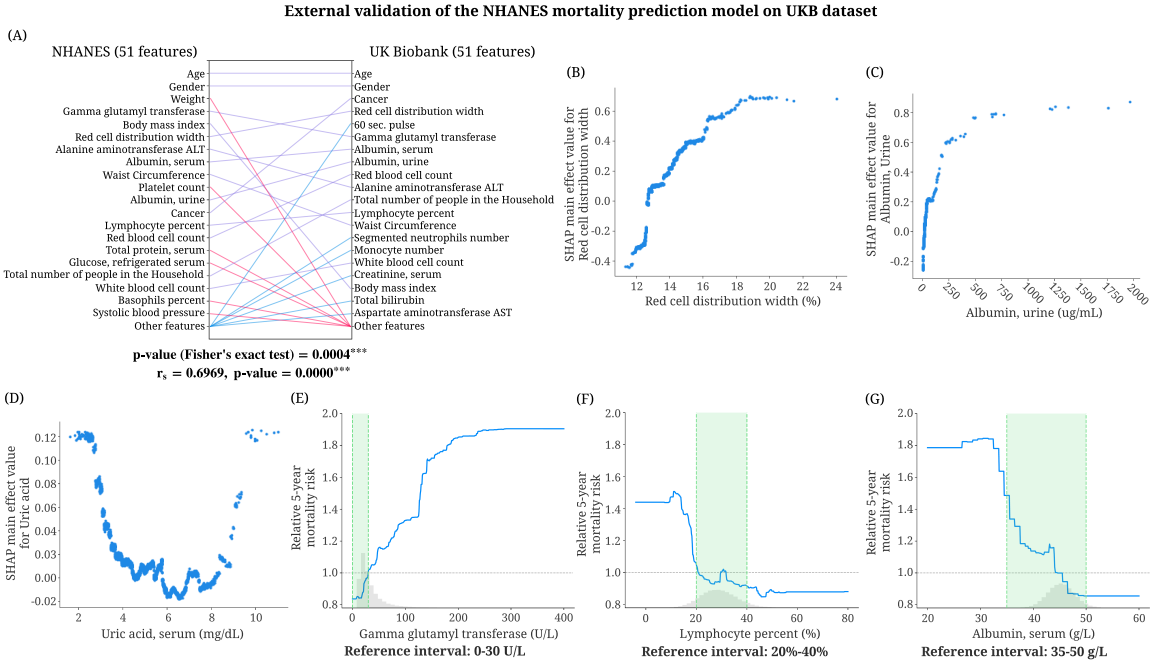
To get the top 20 most important features for 1-year and 10-year mortality predictions, we repeat the same mortality risk scores training and recursive feature elimination process

for 1-year and 10-year predictions. We perform temporal validation to show the generalizability of the IMPACT-20 risk scores on data collected at different time periods. To have similar base rates and age distributions in the test set and temporal validation set, we use the samples from different collection cycles as the temporal validation set for different follow-up times. For 1-year mortality prediction, we use the samples collected in 1999-2012 as the training/testing set and the samples collected in 2013-2014 as the temporal validation set. For 5-year mortality prediction, we use the samples collected in 1999-2008 as the training/testing set and those collected in 2009-2014 as the temporal validation set. For 10-year mortality prediction, we use the samples collected in 1999-2000 as the training/testing set and those collected in 2001-2014 as the temporal validation set. With respect to the 5-year mortality risk scores, samples that are not included in the temporal validation set are randomly split into 80% for training and 20% for testing. The sample size, number of deceased samples, and histogram of age in the training set, with the testing and temporal validation sets, are shown in Supplementary Figure **??**. In Table **??**, 1,the "AUROC" column shows the AUROCs reported in the original paper. The "AUROC of IMPACT-20" column shows the performance of 1-year, 5-year and 10-year IMPACT models trained with the selected top 20 features (listed in Supplementary Tables **??** and **??**). The IMPACT-20 models are trained on samples of all ages and evaluated on the samples within the same age range in the original paper. We bootstrap the test set and the temporal validation set for 1,000 times when measuring the AUROCs.

4.B SUPPLEMENTARY APPENDIX

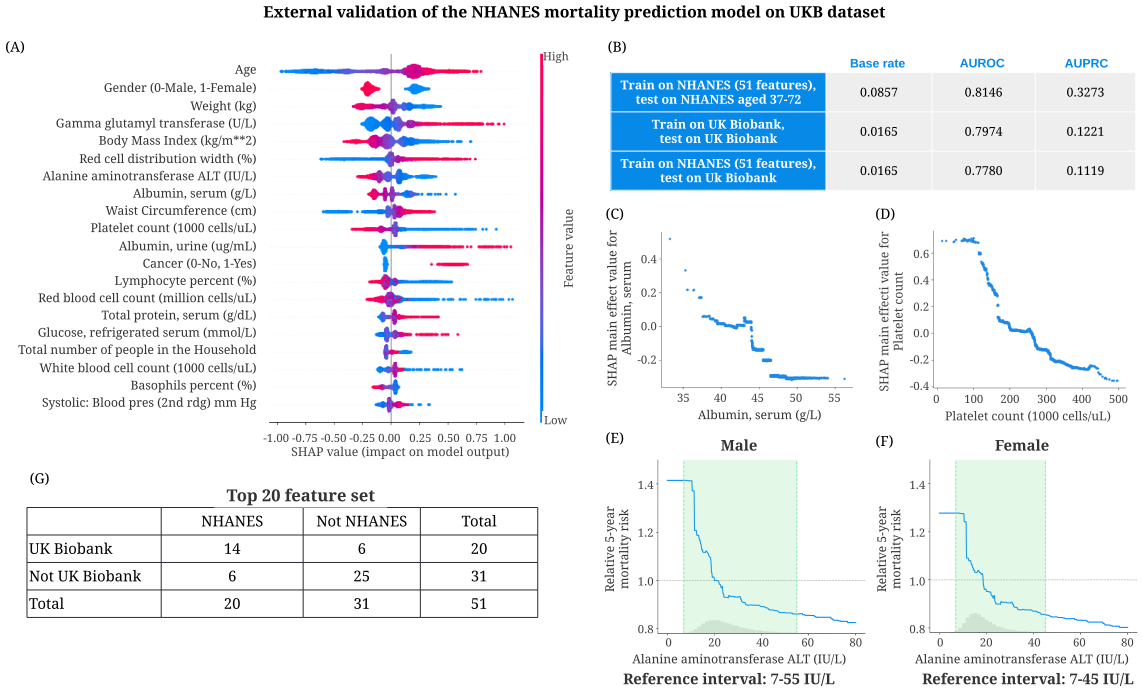
4.B.1 *External validation of the NHANES mortality prediction model on the UK Biobank (UKB) dataset*

We aim to validate whether the performance and explanations of the NHANES mortality prediction model generalize to an unseen population (UKB). To do so, we train a new tree-based 5-year mortality prediction model on the NHANES dataset using the 51 overlapping features between NHANES and UKB. As shown in Supplementary Figure **??**h, the classification accuracy on the UKB test set of the model trained on NHANES samples (AUROC = 0.7780) and UKB samples (AUROC = 0.7974) are close, which shows the generalizability of the NHANES model. Supplementary Figure **??**a shows the feature importances of the 51 features of the NHANES (51 features) and UKB models. *The SHAP values of both models are calculated using the same UKB samples.* We observe that the top 20 most important features are largely consistent, with 14 features the same for both models. The p-value of the Fisher's exact test (p-value = 0.0004) shows that the overlap between the top 20 most important features of both models is significant. The Spearman's correlation coefficient of both models' feature importance is 0.6969 (p-value < 0.0001). Supplementary Figures **??**b–g show noteworthy results of the NHANES (51 features) model explained by



Supplementary Figure A.4.9: **External validation of the NHANES mortality prediction model on the UKB dataset.** (a) Feature importance ranking of models trained on the NHANES (51 features) dataset and the UKB (51 features) dataset. *The SHAP values are calculated using UKB samples.* For each model, the figure shows the 20 most important features of prediction (ordered by importance). The purple line indicates that the feature is in the top 20 features of two models. Blue and red lines indicate the feature is in the top 20 features of one model but not in the top 20 features of the other. The p-value of the Fisher's exact test examines the overlap between the top 20 most important overlapping features in NHANES and UKB models (the contingency table in Supplementary Figure **??**g). The Spearman's correlation coefficient is calculated using the feature importance of the overlapping features in NHANES and UKB. $(***)$ represents a p-value < 0.001. (b–d) The main effect of red cell distribution width, urine albumin and serum uric acid on 5-year mortality of the model trained on the NHANES (51 features) dataset and explained using UKB samples. (e–g) The relative 5-year mortality risk of gamma glutamyl transferase, lymphocyte percent and serum albumin of the model trained on the NHANES (51 features) dataset and explained using UKB samples.

UKB samples: the SHAP main effect of red cell distribution width, serum albumin and serum uric acid, and the relative 5-year mortality risk of gamma glutamyl transferase, lymphocyte percent and serum albumin. The trends shown in these figures are consis-



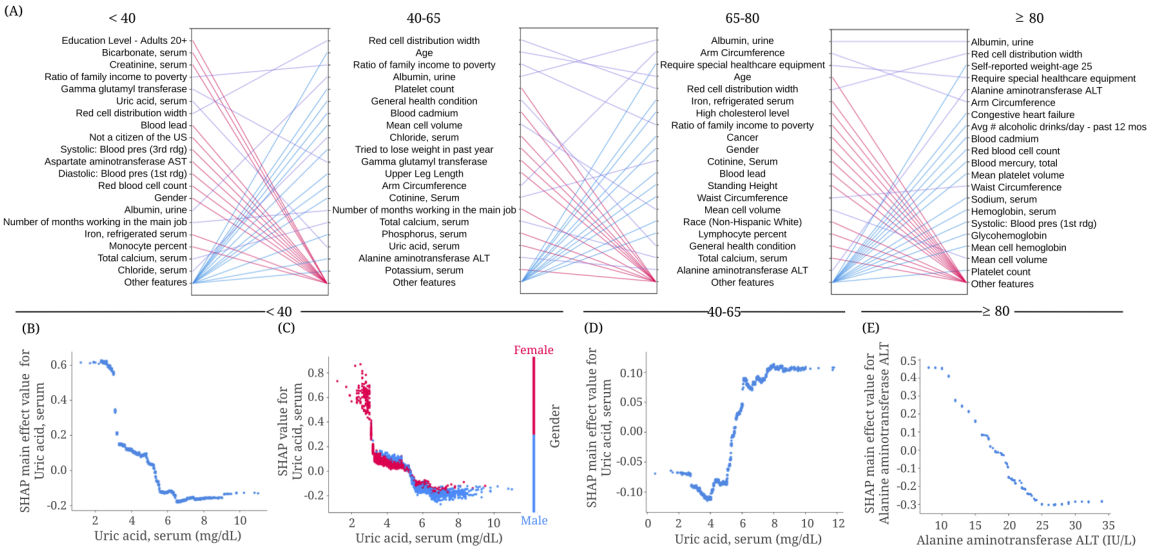
(G)

Top 20 feature set

	NHANES	Not NHANES	Total
UK Biobank	14	6	20
Not UK Biobank	6	25	31
Total	20	31	51

Supplementary Figure A.4.10: **External validation of the NHANES mortality prediction model on the UKB dataset.** (a) SHAP summary plot for the 5-year mortality prediction model trained on NHANES (51 features) dataset and explained using UKB samples. (b) The predictive performance of the models trained on the NHANES (51 features) and UKB (51 features) datasets. The AUROCs are calculated on the testing set by bootstrapping 1,000 times. (c,d) The main effect of serum albumin and platelet count on 5-year mortality of the model trained on the NHANES (51 features) dataset and explained using UKB samples. (e,f) The relative 5-year mortality risk of alanine aminotransferase ALT on male and female samples of the model trained on the NHANES (51 features) dataset and explained using UKB samples. (g) The contingency table of the Fisher's exact test that evaluates the significance of the overlap between the top 20 most important overlapping features in the model trained on the NHANES (51 features) dataset and the model trained on the UKB (51 features) dataset. Both models are explained using UKB samples.

tent with previous findings from both the NHANES (151 features) and UKB (51 features) models. Additional validation results on the UKB dataset are presented in Supplementary Figure **??**.



Supplementary Figure A.4.11: **Understanding important risk factors for mortality prediction in different age groups.** (a) Relative importance of input features in <40, 40-65, 65-80 and ⩾80 age groups. For each model, the figure shows the 20 most impactful features on prediction (ranked from most to least important). The purple line indicates that the feature is in the top 20 features of two models. Blue and red lines indicate the feature is in the top 20 features of one model, but not in the top 20 features of the other. (b) The main effect of serum uric acid on 5-year mortality in the <40 age group. (c) The SHAP value of serum uric acid in the <40 age group 5-year mortality model. (d) The main effect of serum uric acid on 5-year mortality in the 40-65 age group. (e) The main effect of alanine aminotransferase on 5-year mortality in the ⩾80 age group.

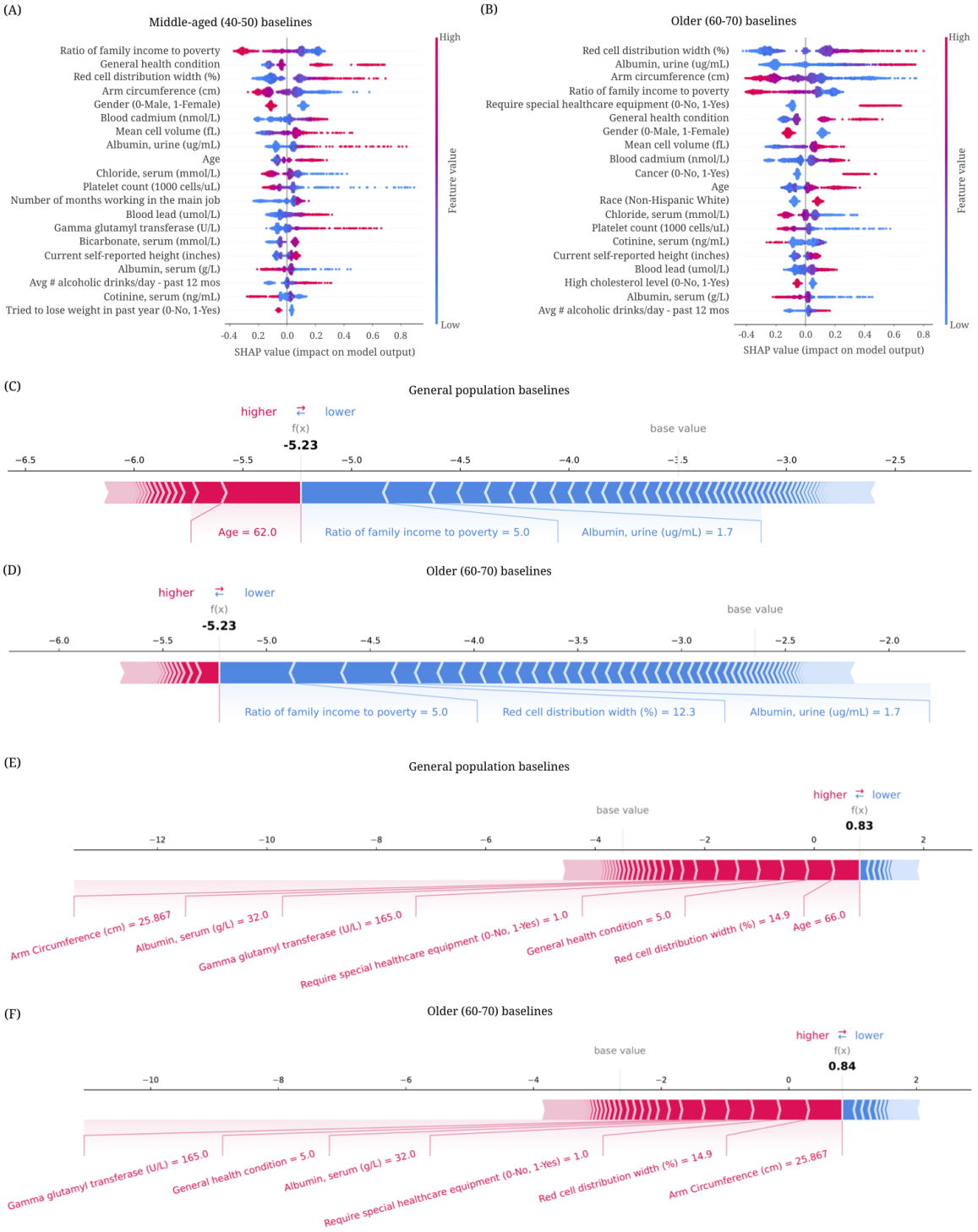
4.B.2 *Discoveries for mortality prediction using different age groups*

IMPACT identifies important features for mortality prediction in different age groups. Supplementary Figure **??**a shows the top 20 most important features and relative importance in 5-year mortality prediction models using different age groups (<40, 40-65, 65-80 and ⩾80). Some features become more important for older subpopulations, such as alanine aminotransferase (ALT), the fifth most important feature in the model using samples over 80 years. Supplementary Figure **??**E shows the main effect of ALT for age⩾80, which shows the negative relationship between ALT and 5-year mortality. Moreover, some features are less important for older subpopulations than younger ones. One example is uric acid level, the sixth most important feature in the age<40 model and the 59th most important in the age⩾80 model. Supplementary Figure **??**b plot the main effect and SHAP value of uric acid in the age<40 model, showing that low uric acid levels increase mortality risk

prediction. However, in the age 40-65 model, higher uric acid is associated with higher mortality risk (Supplementary Figure **??**d). Previous work shows that low uric acid in blood serum can injure the endothelium and induce oxidative stress-related disease [268, 301], and that hyperuricemia (high uric acid) is associated with various adverse health outcomes, including hypertension, stroke, cardiovascular disease and cancer [80, 84, 150, 269]. The numerous downstream effects of high uric acid and low uric acid might explain the different relationship between uric acid and mortality in different age groups. Moreover, the reference range of uric acid differs for males and females (2.4-6.0 mg/dL for females and 3.4-7.0 mg/dL for males). This difference is shown in Figure **??**c, where women have lower uric acid, which can increase mortality risk.

4.B.3 *Explaining the mortality predictions using different baseline distributions*

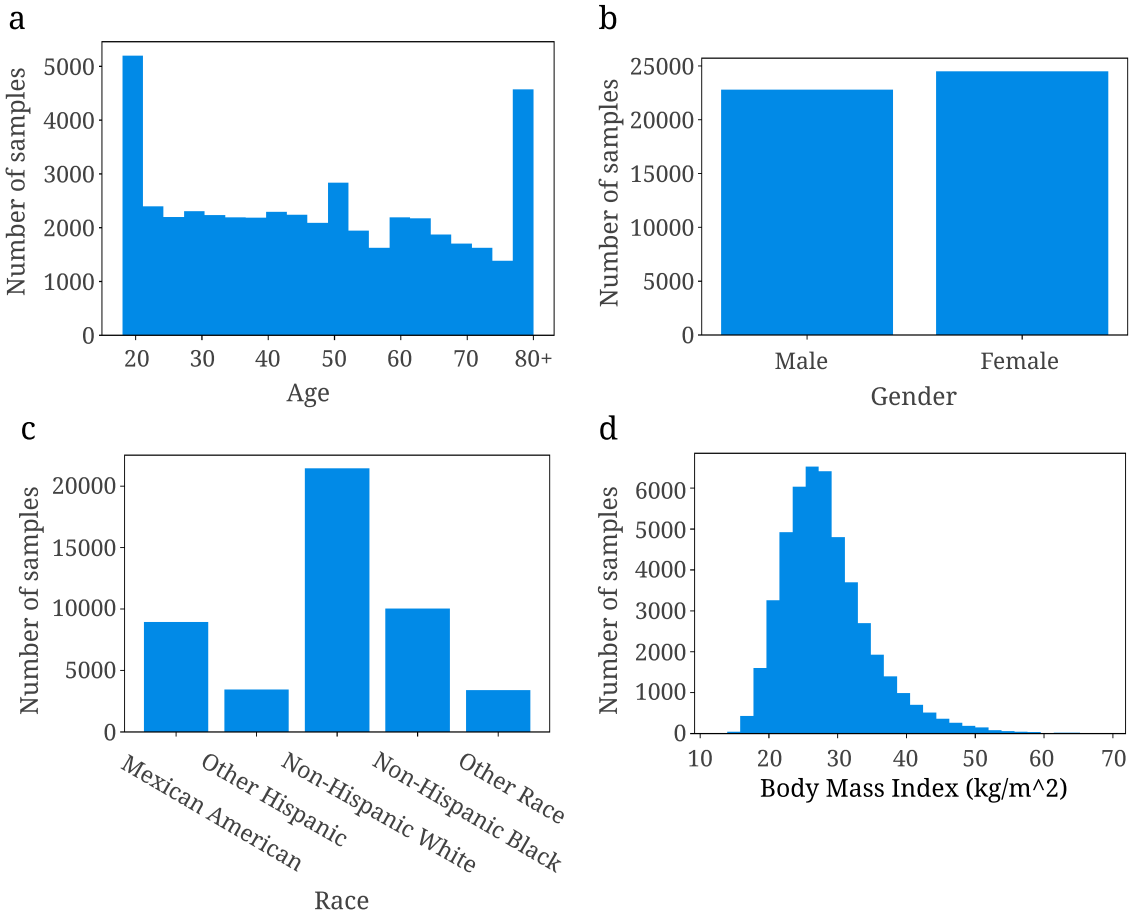
In the Results section, we use TreeExplainer to explain an explicand relative to a baseline distribution drawn uniformly from all training samples (Figure **??**a). This explanation substantially emphasizes age because it compares the explicand to the general population baselines that include individuals of all ages. However, in practice, epidemiologists are more interested in an individual's strong risk factors compared with people of the same age. To show this, we can manually select baselines from the samples that have similar age with the explicand. We take the middle-aged (40-50) baseline distribution and the older (60-70) baseline distribution as two examples. Specifically, we use the testing samples in the specific age range as the explicands (i.e., samples being explained) and training samples in the same age range as the baselines (i.e., background samples) when calculating the SHAP values. The SHAP summary plots are shown in Supplementary Figures **??**a,b. From the figures, we observe that age is no longer the most important feature. Also, compared with Figure **??**, the SHAP value ranges are relatively similar. Therefore, we can identify the strong mortality predictors other than age for different age groups using different baseline distributions. Supplementary Figures **??**c,e show the individualized explanations of a healthier vs unhealthier sample using baselines from the general population. We observe that age contributes a lot to the prediction. However, as shown in Supplementary Figures **??**d,f, the contribution of other important risk factors increases when we use older baselines. These examples illustrate that using the baselines with similar age can help identify strong risk factors besides age.



Supplementary Figure A.4.12: **Explaining the 5-year mortality predictions using different baseline distributions.** (a) Explaining the middle-aged subpopulation (40-50 years old) with the baselines of the same age range. (b) Explaining the older subpopulation (60-70 years old) with the baselines of the same age range. (c,d) The individualized explanation for an individual aged 62 using the general population baselines and the older (60-70) baselines. (e,f) the individualized explanation for an individual aged 66 using the general population baselines and the older (60-70) baselines.

Different follow-up times					
		1-year (n=47,261)	3-year (n=41,434)	5-year (n=35,854)	10-year (n=21,986)
Number of deaths		577 (1.22%)	1,897 (4.58%)	3,074 (8.57%)	5,295 (24.08%)
Age,years		46 (30-63)	46 (30-64)	46 (30-64)	48 (30-67)
Sex					
	Male	22,778 (48.20%)	19,998 (48.26%)	17,276 (48.18%)	10,630 (48.35%)
	Female	24,483 (51.80%)	21,436 (51.74%)	18,578 (51.82%)	11,356 (51.65%)
Ethnicity					
	Mexican American	8,947 (19.93%)	8,164 (19.70%)	7,543 (21.04%)	4,844 (22.03%)
	Other Hispanic	3,452 (7.03%)	2,929 (7.07%)	2,335 (6.51%)	979 (4.45%)
	Non-Hispanic White	21,428 (45.34%)	18,990 (45.83%)	17,081 (47.64%)	10,921 (49.67%)
	Non-Hispanic Black	10,039 (21.24%)	8,821 (21.29%)	7,337 (20.46%)	4,353 (19.78%)
	Other Race	3,395 (7.18%)	2,530 (6.11%)	1,558 (4.35%)	889 (4.04%)
Different age groups (follow-up time = 5-year)					
		Age $<$ 40 (n=14,221)	40 $\leqslant$ Age $<$ 65 (n=12,705)	65 $\leqslant$ Age $<$ 80 (n=6,004)	Age $\geqslant$ 80 (n=2,924)
Number of deaths		183 (1.29%)	648 (5.10%)	961 (16.01%)	1,282 (43.84%)
Age,years		27 (21-33)	51 (45-58)	71 (68-75)	80 (80-80)
Sex					
	Male	6,629 (46.61%)	6,263 (49.30%)	3,056 (50.90%)	1,328 (45.42%)
	Female	7,592 (53.39%)	6,442 (50.70%)	2,948 (49.10%)	1,596 (54.58%)
Ethnicity					
	Mexican American	3,663 (25.76%)	2,622 (20.64%)	1,046 (17.52%)	212 (7.25%)
	Other Hispanic	993 (6.98%)	906 (7.13%)	329 (5.48%)	107 (3.66%)
	Non-Hispanic White	5,686 (39.98%)	5,825 (45.85%)	3,319 (55.28%)	2,251 (76.98)
	Non-Hispanic Black	3,143 (22.10%)	2,793 (21.98%)	1,125 (18.74%)	276 (9.44)
	Other Race	736 (5.18%)	559 (4.40%)	185 (3.08%)	78 (2.67%)

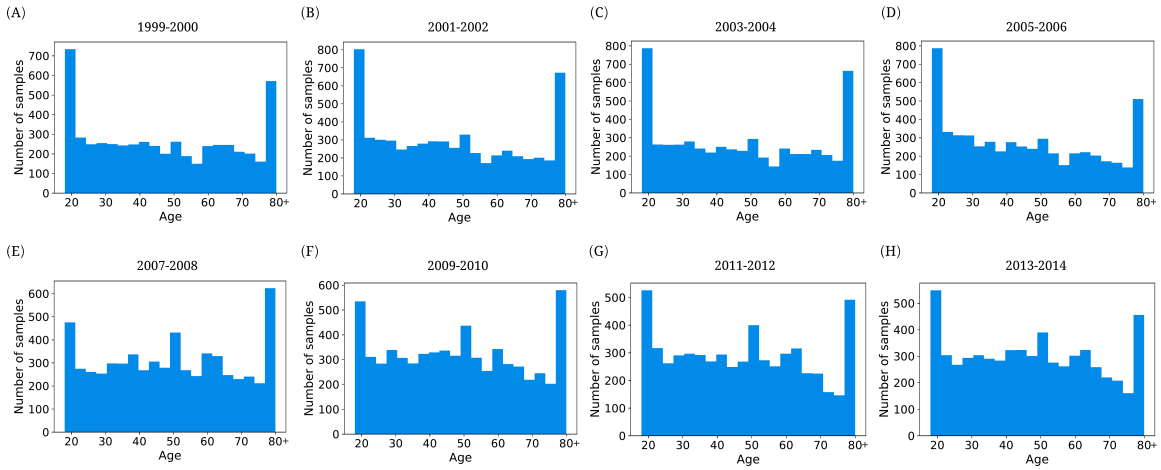
Supplementary Table A.4.1: Population characteristics for the study cohorts. Data are median (IQR), or n/N (%).



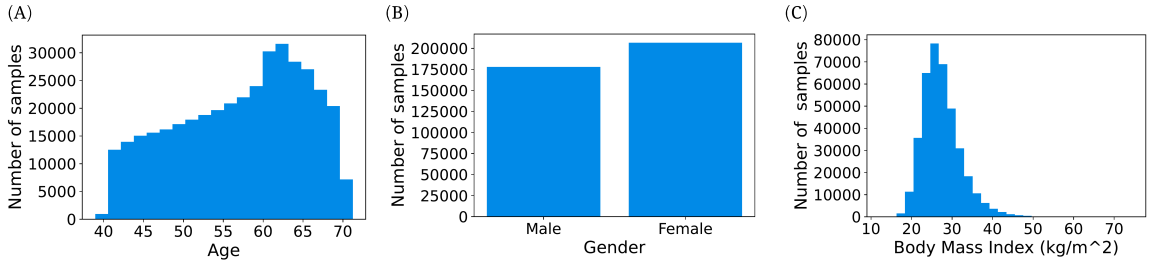
e

	Number of samples	Number of alive samples	Number of deceased samples
1-year mortality	47,261	46,684	577
3-year mortality	41,434	39,537	1,897
5-year mortality	35,854	32,780	3,074
10-year mortality	21,986	5,295	16,691
Age < 40	14,221	14,038	183
40 ≤ Age < 65	12,705	12,057	648
65 ≤ Age < 80	6,004	5,043	961
Age ≥ 80	2,924	1,642	1,282

Supplementary Figure A.4.1: (a–d) Histograms of age, gender, race, and body mass index in the NHANES dataset. (e) The sample size and number of living and deceased samples for different follow-up times and different age groups. For different age groups, the follow-up time is set to 5 years.



Supplementary Figure A.4.2: (a–h) Histograms of age in different two-year data collection cycles.



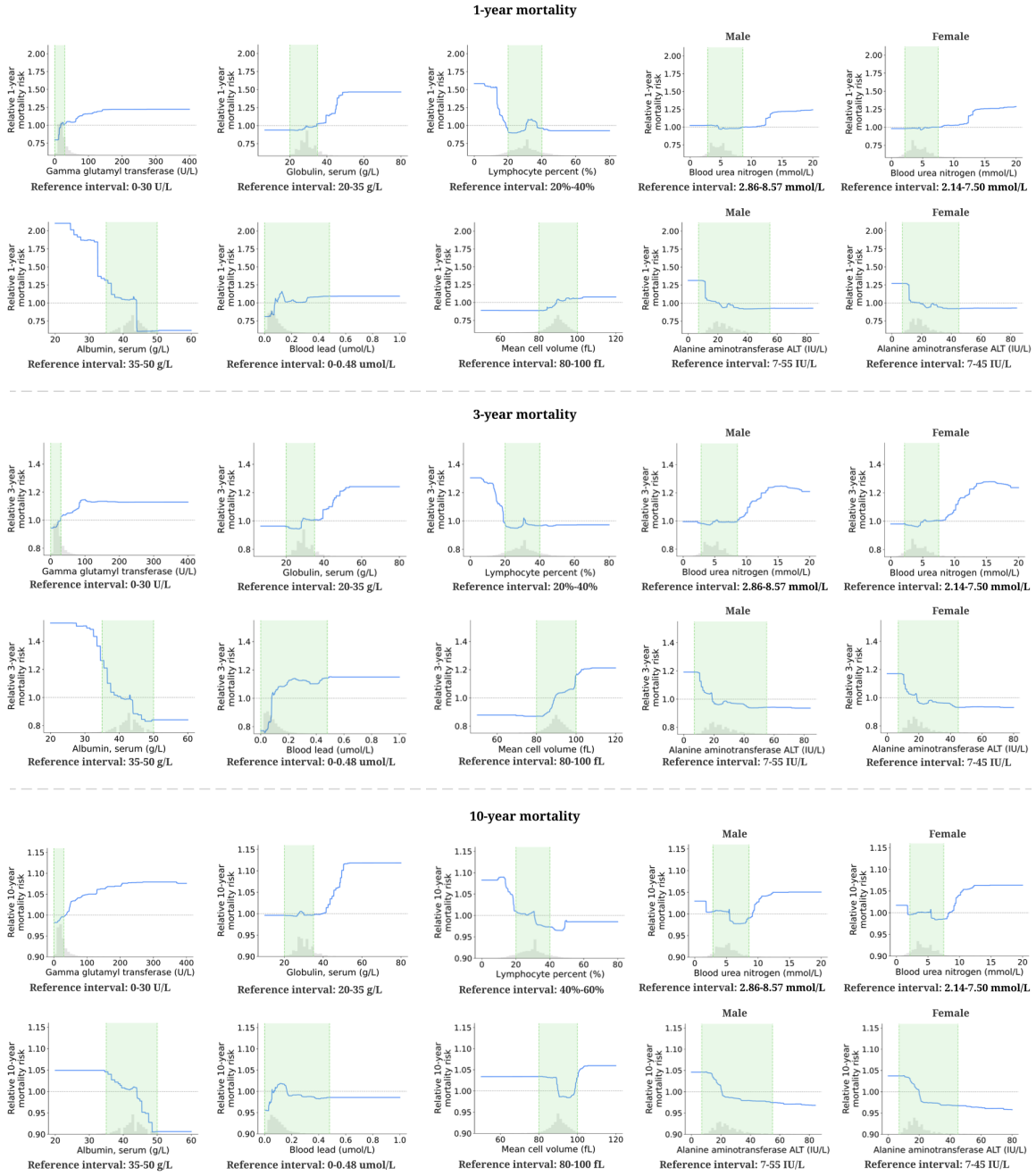
Supplementary Figure A.4.3: (a–c) Histograms of age, gender, and body mass index in the UK Biobank dataset.

Improving prediction power

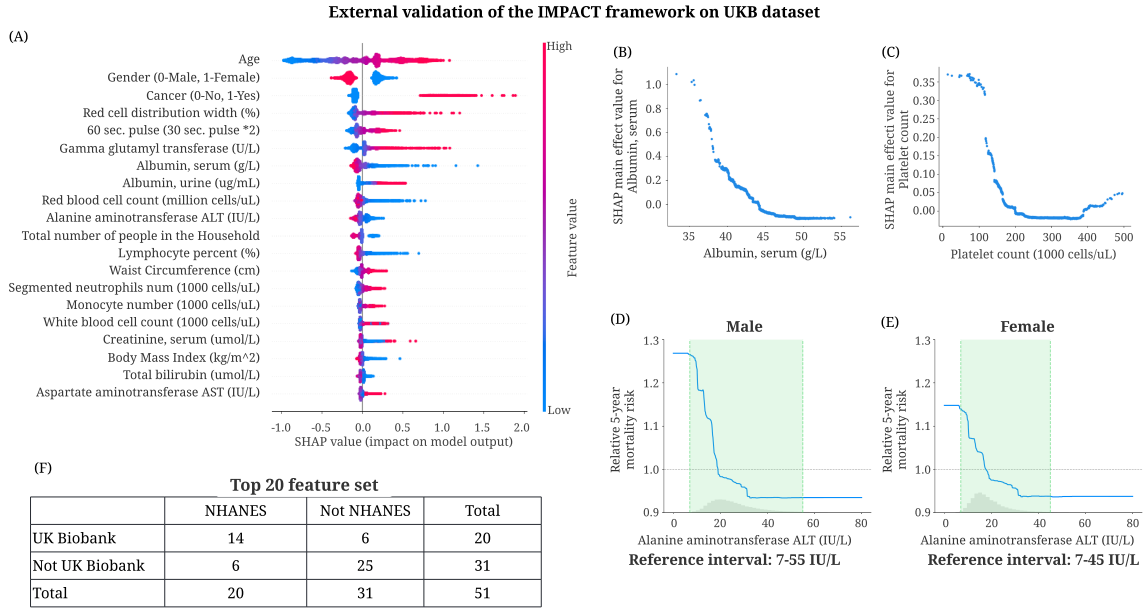
AUPRC on different mortality prediction tasks

	Logistic Regression		Gradient Boosted Trees		Neural Network
1-year mortality	0.1405	***	**0.2307**	***	0.1016
3-year mortality	0.3787	***	**0.4397**	***	0.3507
5-year mortality	0.5131	***	**0.5464**	***	0.4838
10-year mortality	0.7980	***	**0.8212**	***	0.7066
Age < 40	0.0441		**0.1047**	*	0.0423
40 ≤ Age < 65	0.3436		**0.3823**	***	0.2931
65 ≤ Age < 80	0.5263	**	**0.5790**	***	0.4717
Age ≥ 80	**0.7447**		0.7071		0.6766

Supplementary Figure A.4.4: The area under the precision-recall curve (AUPRC) of gradient boosted tree models outperforms both linear models and neural networks for seven of our prediction models. $(***)$ represents a p-value < 0.001, $(**)$ represents a p-value < 0.01, and $(*)$ represents a p-value < 0.05. P-values are computed using bootstrap resampling over the tested time points while measuring the difference in area between the curves.



Supplementary Figure A.4.5: **Effect of varying laboratory feature values on 1-, 3- and 10-year mortality risk.** These partial dependence plots show the change in relative 1-, 3- and 10-year mortality risk for all values of a given laboratory feature. The grey histograms on each plot show the distribution of values for that feature in the test set. The green shaded region shows the reference interval of each feature.

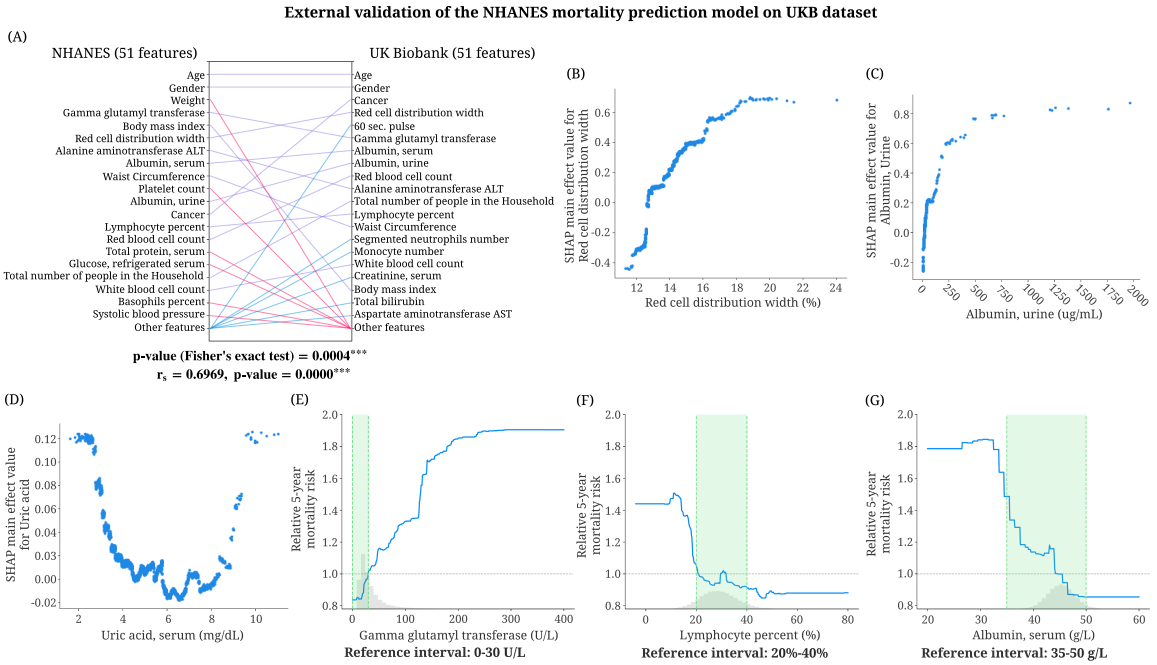


(F)

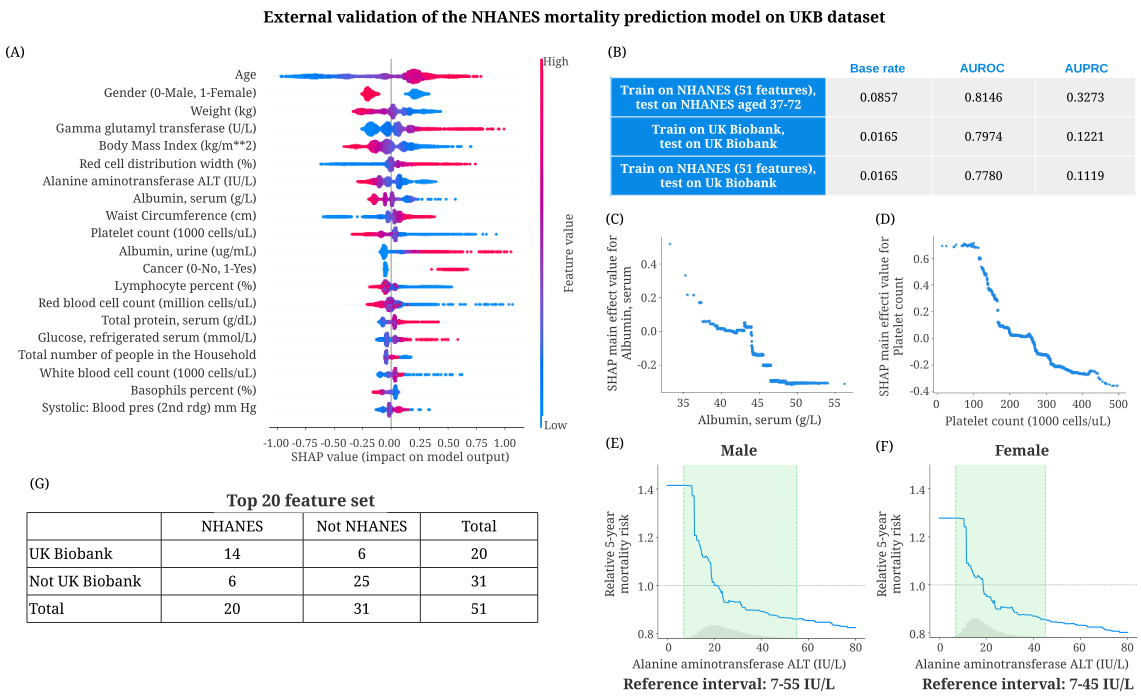
Top 20 feature set

	NHANES	Not NHANES	Total
UK Biobank	14	6	20
Not UK Biobank	6	25	31
Total	20	31	51

Supplementary Figure A.4.6: **External validation of the IMPACT framework on UKB dataset.** (a) SHAP summary plot for the 5-year mortality prediction model trained on UKB (51 features) dataset. (b,c) The main effect of serum albumin and platelet count on 5-year mortality of the model trained on UKB (51 features) dataset. (d,e) The relative 5-year mortality risk of alanine aminotransferase ALT on male and female samples of the model trained on UKB (51 features) dataset. (f) The contingency table of the Fisher's exact test that evaluates the significance of the overlap between the top 20 most important overlapping features in the model trained on NHANES (151 features) dataset and the model trained on UKB (51 features) dataset.



Supplementary Figure A.4.7: **External validation of the NHANES mortality prediction model on the UKB dataset.** (a) Feature importance ranking of models trained on the NHANES (51 features) dataset and the UKB (51 features) dataset. *The SHAP values are calculated using UKB samples.* For each model, the figure shows the 20 most important features of prediction (ordered by importance). The purple line indicates that the feature is in the top 20 features of two models. Blue and red lines indicate the feature is in the top 20 features of one model but not in the top 20 features of the other. The p-value of the Fisher's exact test examines the overlap between the top 20 most important overlapping features in NHANES and UKB models (the contingency table in Supplementary Figure **??**g). The Spearman's correlation coefficient is calculated using the feature importance of the overlapping features in NHANES and UKB. $(***)$ represents a p-value < 0.001. (b–d) The main effect of red cell distribution width, urine albumin and serum uric acid on 5-year mortality of the model trained on the NHANES (51 features) dataset and explained using UKB samples. (e–g) The relative 5-year mortality risk of gamma glutamyl transferase, lymphocyte percent and serum albumin of the model trained on the NHANES (51 features) dataset and explained using UKB samples.

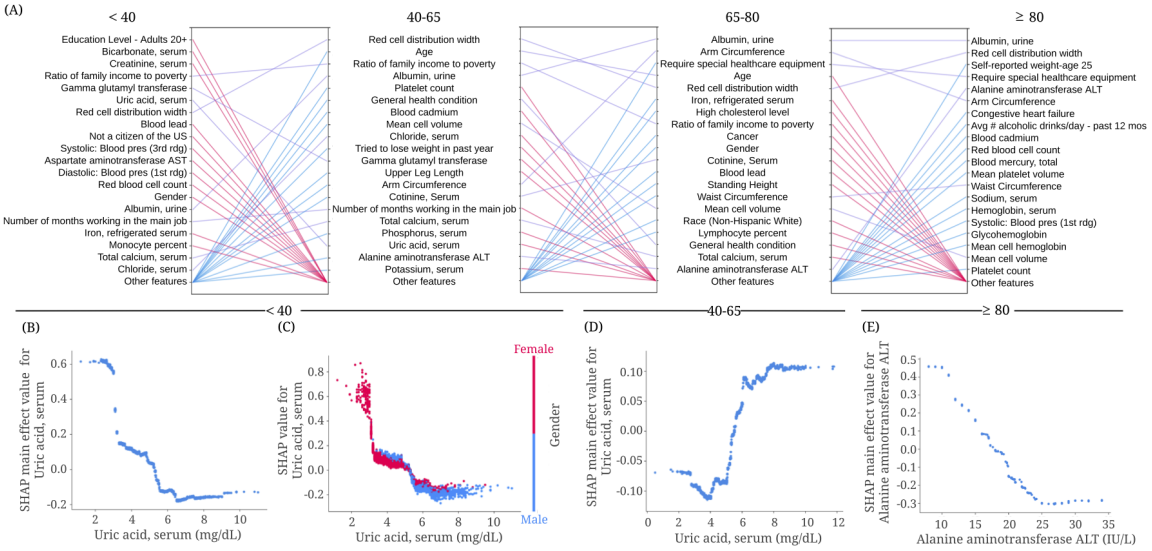


(G)

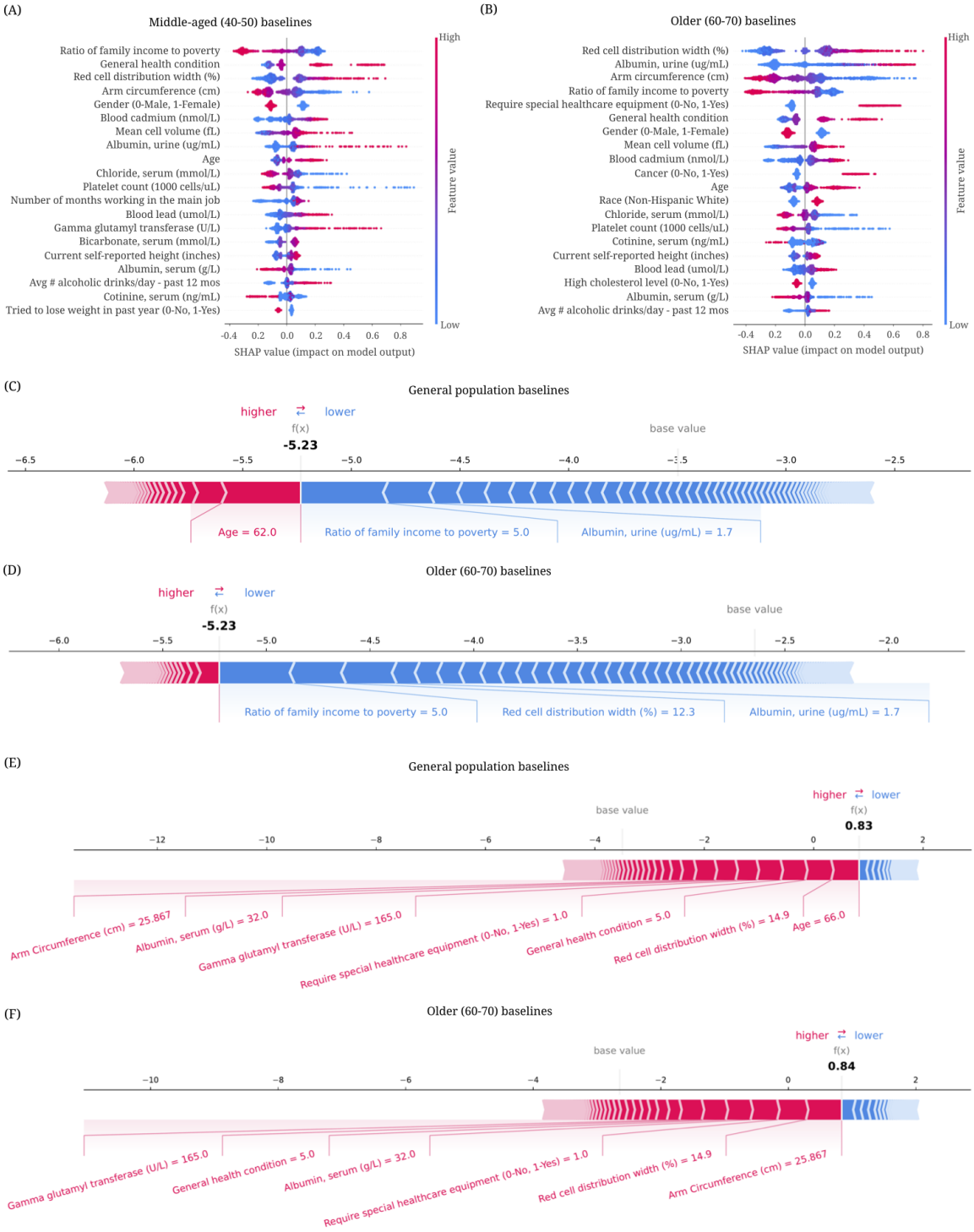
Top 20 feature set

	NHANES	Not NHANES	Total
UK Biobank	14	6	20
Not UK Biobank	6	25	31
Total	20	31	51

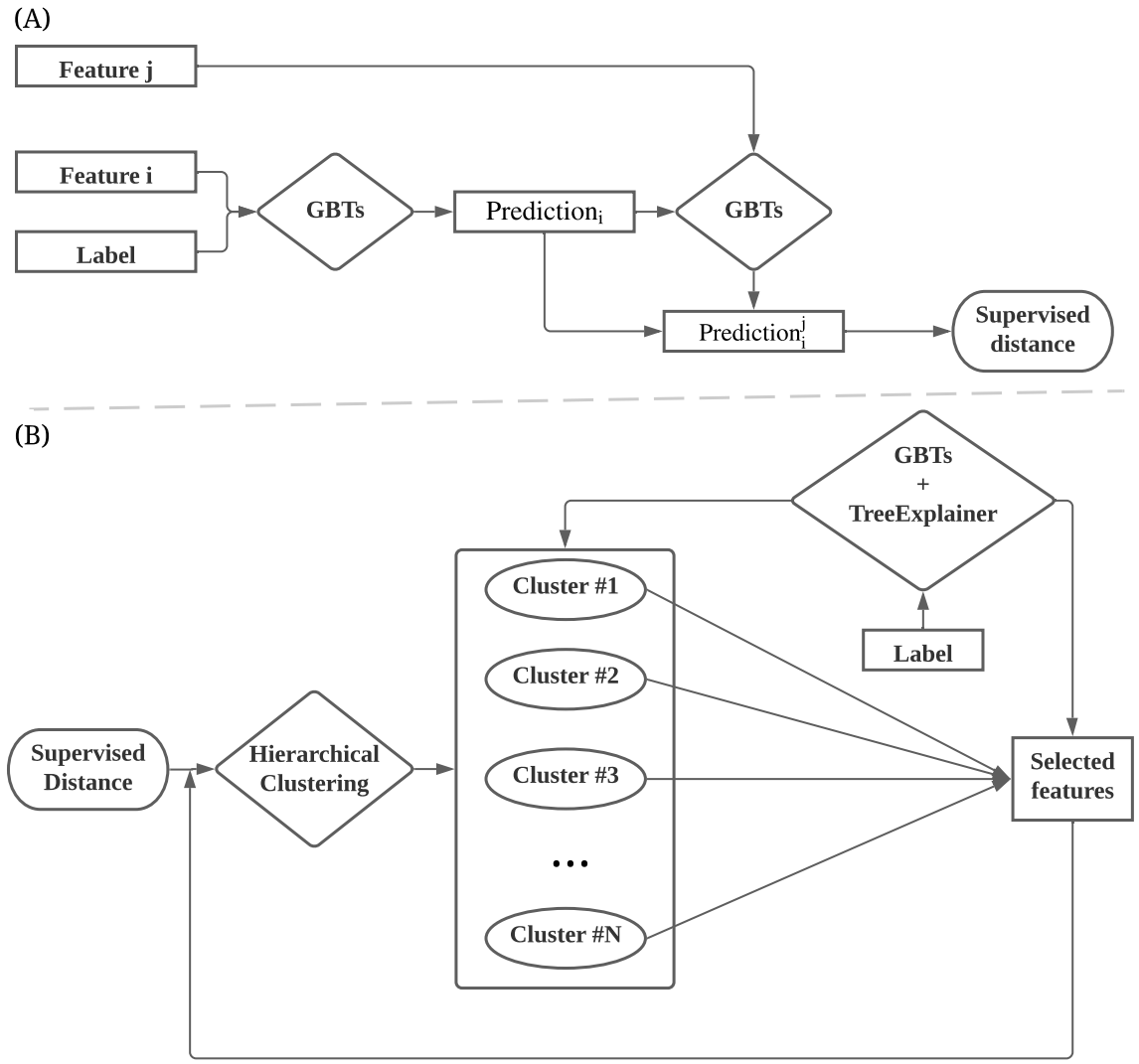
Supplementary Figure A.4.8: **External validation of the NHANES mortality prediction model on the UKB dataset.** (a) SHAP summary plot for the 5-year mortality prediction model trained on NHANES (51 features) dataset and explained using UKB samples. (b) The predictive performance of the models trained on the NHANES (51 features) and UKB (51 features) datasets. The AUROCs are calculated on the testing set by bootstrapping 1,000 times. (c,d) The main effect of serum albumin and platelet count on 5-year mortality of the model trained on the NHANES (51 features) dataset and explained using UKB samples. (e,f) The relative 5-year mortality risk of alanine aminotransferase ALT on male and female samples of the model trained on the NHANES (51 features) dataset and explained using UKB samples. (g) The contingency table of the Fisher's exact test that evaluates the significance of the overlap between the top 20 most important overlapping features in the model trained on the NHANES (51 features) dataset and the model trained on the UKB (51 features) dataset. Both models are explained using UKB samples.



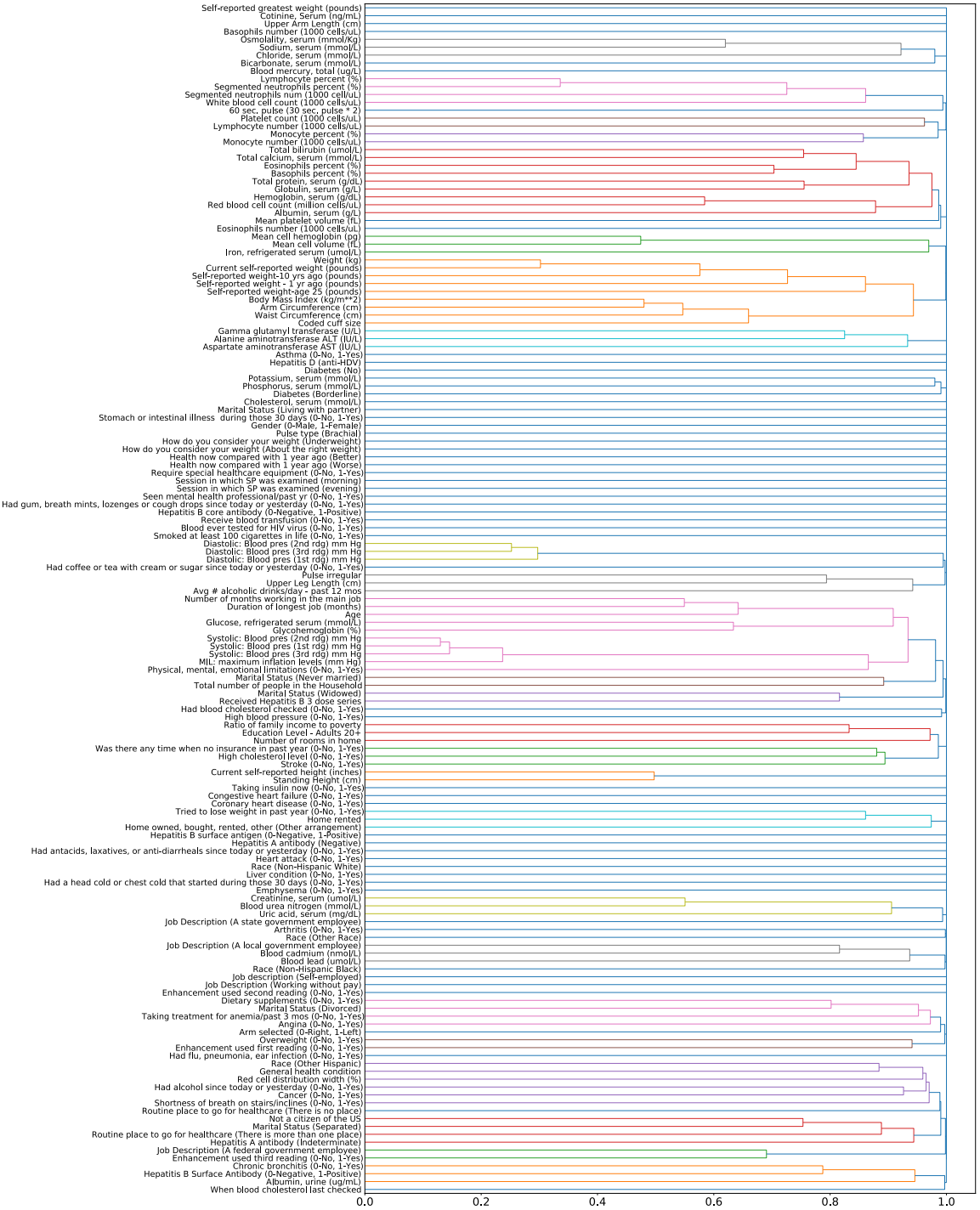
Supplementary Figure A.4.9: **Understanding important risk factors for mortality prediction in different age groups.** (a) Relative importance of input features in <40, 40-65, 65-80 and ⩾80 age groups. For each model, the figure shows the 20 most impactful features on prediction (ranked from most to least important). The purple line indicates that the feature is in the top 20 features of two models. Blue and red lines indicate the feature is in the top 20 features of one model, but not in the top 20 features of the other. (b) The main effect of serum uric acid on 5-year mortality in the <40 age group. (c) The SHAP value of serum uric acid in the <40 age group 5-year mortality model. (d) The main effect of serum uric acid on 5-year mortality in the 40-65 age group. (e) The main effect of alanine aminotransferase on 5-year mortality in the ⩾80 age group.



Supplementary Figure A.4.10: **Explaining the 5-year mortality predictions using different baseline distributions.** (a) Explaining the middle-aged subpopulation (40-50 years old) with the baselines of the same age range. (b) Explaining the older subpopulation (60-70 years old) with the baselines of the same age range. (c,d) The individualized explanation for an individual aged 62 using the general population baselines and the older (60-70) baselines. (e,f) the individualized explanation for an individual aged 66 using the general population baselines and the older (60-70) baselines.



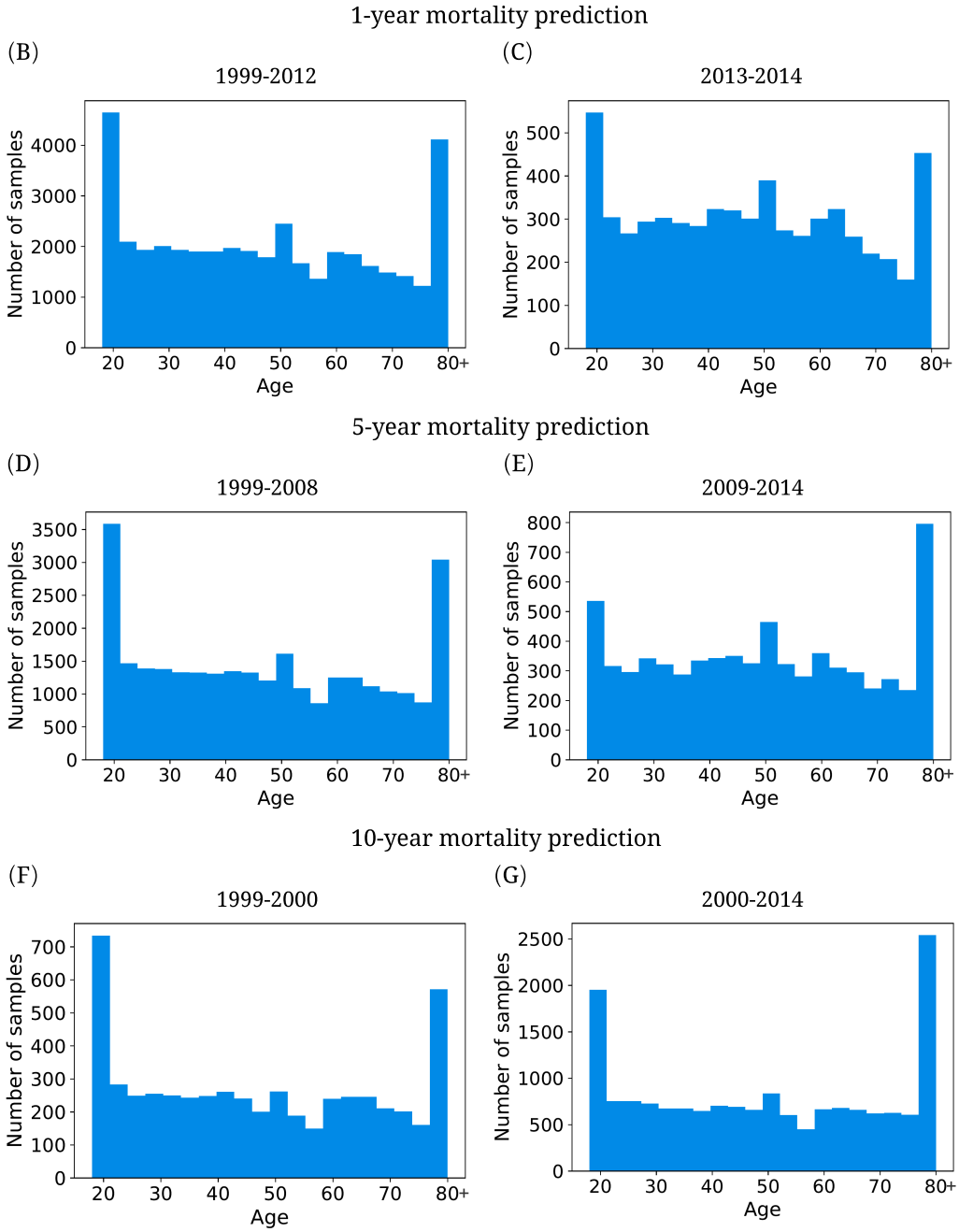
Supplementary Figure A.4.11: (a) The workflow of supervised distance calculation. (b) The workflow of supervised-distance feature selection.



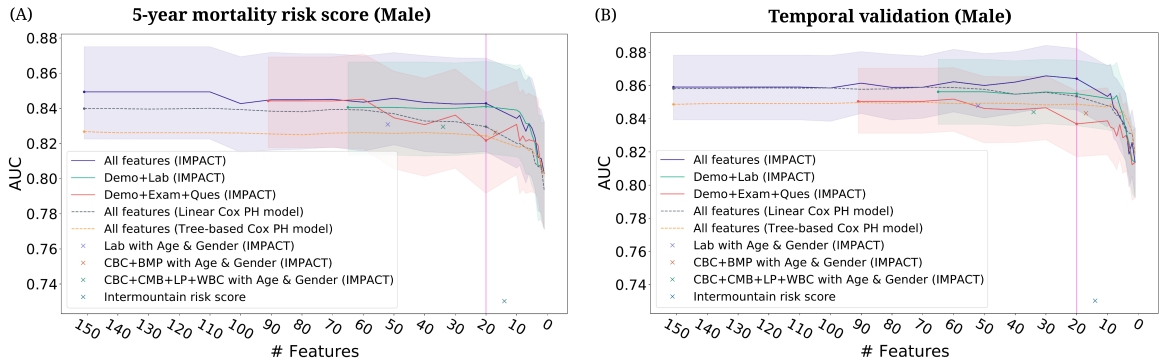
Supplementary Figure A.4.12: The cluster tree of supervised distance based hierarchical clustering. The color threshold is set to 0.98.

(A)

Follow-up time	Training set (80%) + Testing set (20%)				Temporal validation set			
	Collection cycles	Number of samples	Number of deaths	Base rate	Collection cycles	Number of samples	Number of deaths	Base rate
1-year	1999-2012	41,179	524	1.27%	2013-2014	6,082	53	0.81%
5-year	1999-2008	28,820	2,247	7.80%	2009-2014	7,034	827	11.76%
10-year	1999-2000	5,444	931	17.10%	2001-2014	16,542	4,364	26.38%



Supplementary Figure A.4.13: (a) Population characteristics of the training/testing and temporal validation sets with different follow-up times. (b–g) Histograms of age in the training/testing set and temporal validation set with different follow-up times.



Supplementary Figure A.4.14: (a,b) The AUROC of the models using different feature sets after recursive feature elimination. Lines are mean performance over 1000 random train/test splits, and shaded bands are 95 percent normal confidence intervals. (a) The AUROC of the models tested on the male group in the test set of NHANES 1999-2008. (b) The AUROC of the models testing on the male group in the temporal validation set (NHANES 2009-2014).

Importance Ranking	IMPACT-20	IMPACT-20 (Demo+Lab)
1	Age	Age
2	Albumin, urine (ug/mL)	Blood lead (umol/L)
3	Arm Circumference (cm)	Albumin, urine (ug/mL)
4	Gender (0-Male, 1-Female)	Ratio of family income to poverty
5	Blood lead (umol/L)	Education Level - Adults 20+
6	Ratio of family income to poverty	Red cell distribution width (%)
7	Albumin, serum (g/L)	Chloride, serum (mmol/L)
8	Red cell distribution width (%)	Blood cadmium (nmol/L)
9	Received Hepatitis B 3 dose series	Lymphocyte percent (%)
10	General health condition	Mean cell volume (fL)
11	Mean cell volume (fL)	Red blood cell count (million cells/uL)
12	Number of months working in the main job	Albumin, serum (g/L)
13	Self-reported greatest weight (pounds)	Creatinine, serum (umol/L)
14	Education Level - Adults 20+	Cotinine, Serum (ng/mL)
15	Lymphocyte percent (%)	Platelet count (1000 cells/uL)
16	Require special healthcare equipment (0-No, 1-Yes)	Potassium, serum (mmol/L)
17	Chloride, serum (mmol/L)	Sodium, serum (mmol/L)
18	Blood cadmium (nmol/L)	Alanine aminotransferase ALT (IU/L)
19	Weight (kg)	Blood urea nitrogen (mmol/L)
20	Shortness of breath on stairs/inclines (0-No, 1-Yes)	Race (Non-Hispanic White)

Importance Ranking	IMPACT-20 (Demo+Exam+Ques)	IMPACT (CBC+BMP with age and gender)
1	Age	Age
2	Require special healthcare equipment (0-No, 1-Yes)	Red cell distribution width (%)
3	Arm Circumference (cm)	Mean cell volume (fL)
4	General health condition	Chloride, serum (mmol/L)
5	Education Level - Adults 20+	Gender (0-Male, 1-Female)
6	Gender (0-Male, 1-Female)	Glucose, refrigerated serum (mmol/L)
7	Congestive heart failure (0-No, 1-Yes)	Red blood cell count (million cells/uL)
8	Ratio of family income to poverty	White blood cell count (1000 cells/uL)
9	Diastolic: Blood pres (2nd rdg) mm Hg	Potassium, serum (mmol/L)
10	Systolic: Blood pres (2nd rdg) mm Hg	Creatinine, serum (umol/L)
11	Avg # alcoholic drinks/day - past 12 mos	Platelet count (1000 cells/uL)
12	Cancer (0-No, 1-Yes)	Blood urea nitrogen (mmol/L)
13	Self-reported weight-age 25 (pounds)	Sodium, serum (mmol/L)
14	Number of months working in the main job	Hemoglobin, serum (g/dL)
15	Self-reported greatest weight (pounds)	Mean cell hemoglobin (pg)
16	Duration of longest job (months)	Total calcium, serum (mmol/L)
17	Smoked at least 100 cigarettes in life (0-No, 1-Yes)	Mean platelet volume (fL)
18	Shortness of breath on stairs/inclines (0-No, 1-Yes)	
19	60 sec. pulse (30 sec. pulse * 2)	
20	Current self-reported height (inches)	

Supplementary Table A.4.2: The selected top 20 features of the **5-year** mortality risk scores using different feature types and the features included in CBC and BMP panels.

Importance Ranking	IMPACT-20 (1-year mortality prediction)	IMPACT-20 (10-year mortality prediction)
1	Age	Age
2	Albumin, serum (g/L)	Albumin, urine (ug/mL)
3	Albumin, urine (ug/mL)	Blood lead (umol/L)
4	Lymphocyte percent (%)	General health condition
5	Blood lead (umol/L)	Albumin, serum (g/L)
6	Education Level - Adults 20+	Arm Circumference (cm)
7	Red cell distribution width (%)	Red cell distribution width (%)
8	Cholesterol, serum (mmol/L)	Chloride, serum (mmol/L)
9	Blood mercury, total (ug/L)	Education Level - Adults 20+
10	General health condition	Blood cadmium (nmol/L)
11	Red blood cell count (million cells/uL)	Creatinine, serum (umol/L)
12	Basophils percent (%)	Received Hepatitis B 3 dose series
13	Require special healthcare equipment (0-No, 1-Yes)	Self-reported greatest weight (pounds)
14	Arm Circumference (cm)	Body Mass Index (kg/m**2)
15	Upper Arm Length (cm)	Systolic: Blood pres (2nd rdg) mm Hg
16	Blood cadmium (nmol/L)	Mean cell hemoglobin (pg)
17	Chloride, serum (mmol/L)	Gamma glutamyl transferase (U/L)
18	Avg # alcoholic drinks/day - past 12 mos	Potassium, serum (mmol/L)
19	Systolic: Blood pres (1st rdg) mm Hg	Blood mercury, total (ug/L)
20	Blood urea nitrogen (mmol/L)	How do you consider your weight?

Supplementary Table A.4.3: Selected top 20 features of the 1-year and the 10-year mortality risk scores.

5

EXPLAINABLE BIOLOGICAL AGE (ENABL AGE): AN ARTIFICIAL INTELLIGENCE FRAMEWORK FOR INTERPRETABLE BIOLOGICAL AGE

5.1 INTRODUCTION

Ageing is a major risk factor for many age-related diseases and disorders, such as heart disease, neurodegeneration, and cancer [200]. Chronological age is the time elapsed since birth, whereas ageing refers to the gradual decline in biological function that leads to an increased risk of death or disease. Measuring the state of ageing of an individual (ie, their biological age) is a crucial step toward understanding and addressing age-related diseases and extending lifespans. Healthy ageing includes more than disease susceptibility and mortality, it involves independence, quality of life, living disability free, and more [215]. This broader perspective highlights the importance of accurately assessing and interpreting biological age to better understand the ageing process and develop interventions for promoting healthy ageing.

Biological age, a measure of the biological functioning of an organism compared with an expected level for a certain chronological age, reflects the general health status of an individual. First-generation biological age clocks use chronological age as a surrogate, predicting it with ageing biomarkers [82, 108, 117, 149, 279, 320]. However, as people of the same chronological age can exhibit varying rates of ageing, this method imperfectly captures biological ageing, potentially underemphasising underlying ageing mechanisms and weakening the correlation between biological age and mortality risk as prediction accuracy increases [337]. Such clocks generally show variable associations with ageing outcomes [86] and weak associations with mortality risk [153, 172]. These findings led to the development of second-generation clocks, which integrate information from age-related outcomes, such as mortality, during their training process [152, 153, 172]. These clocks are powerful morbidity and mortality predictors [172, 190], and have strong associations with ageing traits [129, 172, 190]. Many of them are based on epigenetic profiles [108, 117, 152, 153, 172]. There are also clocks derived from other data types, including blood markers [153, 320], transcriptomic [82], and proteomic [279] profiles.

Existing biological age clocks have three main limitations. First, they necessitate a trade-off between accuracy (ie, predictive performance for chronological age or mortality) and interpretability (ie, understanding each feature's contribution to the prediction). Most of them use linear models that offer interpretability but weaker predictive power for chronological age prediction (related to first-generation biological age clocks) [227, 250] and

mortality prediction (related to second-generation biological age clocks) [177, 231] than complex machine-learning models. This choice is natural given that interpretability is a key goal of biological age clocks: identifying biomarkers of biological age can improve our understanding of the ageing process and help develop drugs that target ageing-related dysfunction. Although advanced machine-learning models have created first-generation biological age models using diverse data types such as epigenetic features [89], blood markers [17, 95, 182, 227], electrocardiogram features [160], brain MRI features [58], and transcriptomic features [115], these models are hard to interpret and do not have individualised explanations. To build models that are both accurate and interpretable, we turn to the emerging area of explainable artificial intelligence (XAI) [178].

The second limitation is that interpretations of previous biological age clocks might not address important scientific questions. Previous biological age studies primarily explain the model as a whole (global explanation). However, given the substantial variations in ageing processes among individuals, individualised explanations are crucial for comprehending complex ageing mechanisms. We leveraged recent XAI methods to provide principled individualised (local) explanations on the basis of feature attributions. Typically, feature attributions can be difficult to understand for non-machine-learning practitioners because they are usually in units of predicted probability or logits units. To make our biological age explanations more accessible, we rescaled our attributions to the age scale in units of years so that the rescaled attributions sum to the biological age acceleration (AgeAccel) of an individual.

The third limitation of current biological age clocks is their inability to incorporate several age-related outcomes, such as cause-specific mortalities. Their inability to account for these factors restricts our understanding of important features for different ageing processes. This shortcoming is problematic because biological ageing is enormously complex and thought to be driven by many biological processes [137, 166]. Previous studies noted low agreement between biological age clocks in terms of their correlations with each other and associations with ageing traits [26, 156], implying that they measure different aspects of biological age. To solve this, we developed our biological age clocks by predicting diverse age-related outcomes, such as specific mortalities and morbidities, allowing us to target and specify particular underlying ageing mechanisms that our clocks capture.

We introduce ExplaiNAble BioLogical Age (ENABL Age), a new approach to estimate and interpret biological age that combines complex machine learning and XAI methods. We did a comprehensive validation of ENABL Age using the UK Biobank and National Health and Nutrition Examination Survey (NHANES) datasets, assessing its ability to capture ageing mechanisms and offering concrete examples of its interpretability. Our interactive website (`https://suinleelab.github.io/ENABLAge`) allows users to calculate and interpret their own ENABL Age, and the code for our study can be found at `https://github.com/suinleelab/ENABLAge`.

5.2 METHODS

5.2.1 *Data sources and study population*

Our study used the UK Biobank and the NHANES datasets. UK Biobank participants were enrolled between 2007 and 2014 across England, Wales, and Scotland. We excluded features missing in more than 80% of samples, highly correlated features (correlations >0.98), and individuals deceased from external causes of mortality. After preprocessing, our UK Biobank data had 501,366 samples from people aged 40–70 years with 825 features from numerous categories, including demographics, blood assays, medical history, and lifestyle. The samples from two Scottish centres were left out as the geographical validation set (n=35,735). We predicted all-cause mortality and five cause-specific mortality categories, comprising neoplasm, circulatory disease, respiratory disease, digestive disease, and other diseases. Ethical approval for the UK Biobank study was obtained from the North West Haydock Research Ethics Committee (21/NW/0157). Informed consent was obtained from all UK Biobank participants.

The NHANES data were collected from individuals in the USA between 1999 and 2014. We excluded features missing in more than 50% of samples, highly correlated features, and individuals deceased from external causes of mortality. After excluding features, 47,084 samples from people aged 18 years to 80 years or older with 158 features remained. We predicted all-cause mortality and nine cause-specific mortality categories, comprising heart disease, malignant neoplasms, chronic lower-respiratory disease, cerebrovascular diseases, Alzheimer's disease, diabetes, pneumonia and influenza, kidney disease, and other diseases. When predicting cause-specific mortality, we excluded the individuals deceased from all other causes. The National Centre for Health Statistics Research Ethics Review Board approved all NHANES protocols, and all participants gave informed consent. Demographic characteristics and sample size of the data for different tasks are shown in Supplementary Tables **??**–**??**. More details are in Supplementary Methods **??**.

5.2.2 *Overview of the ENABL Age framework*

We constructed ENABL Age clocks in two stages (Figure **??**A). First, we used Cox proportional hazard gradient boosted trees (GBTs) to create predictors for all-cause and cause-specific mortality (Supplementary Methods **??**). The GBTs capture non-linear and interaction effects between features. In the second stage, we fitted an exponential curve to the GBT predictions and chronological ages of the training samples. Using the inverse function, we calculated ENABL Age from the predictions (Supplementary Methods **??**). We defined ENABL Age acceleration (AgeAccel) as the difference between ENABL Age and chronological age. Lastly, to obtain the contribution of different mortality causes to

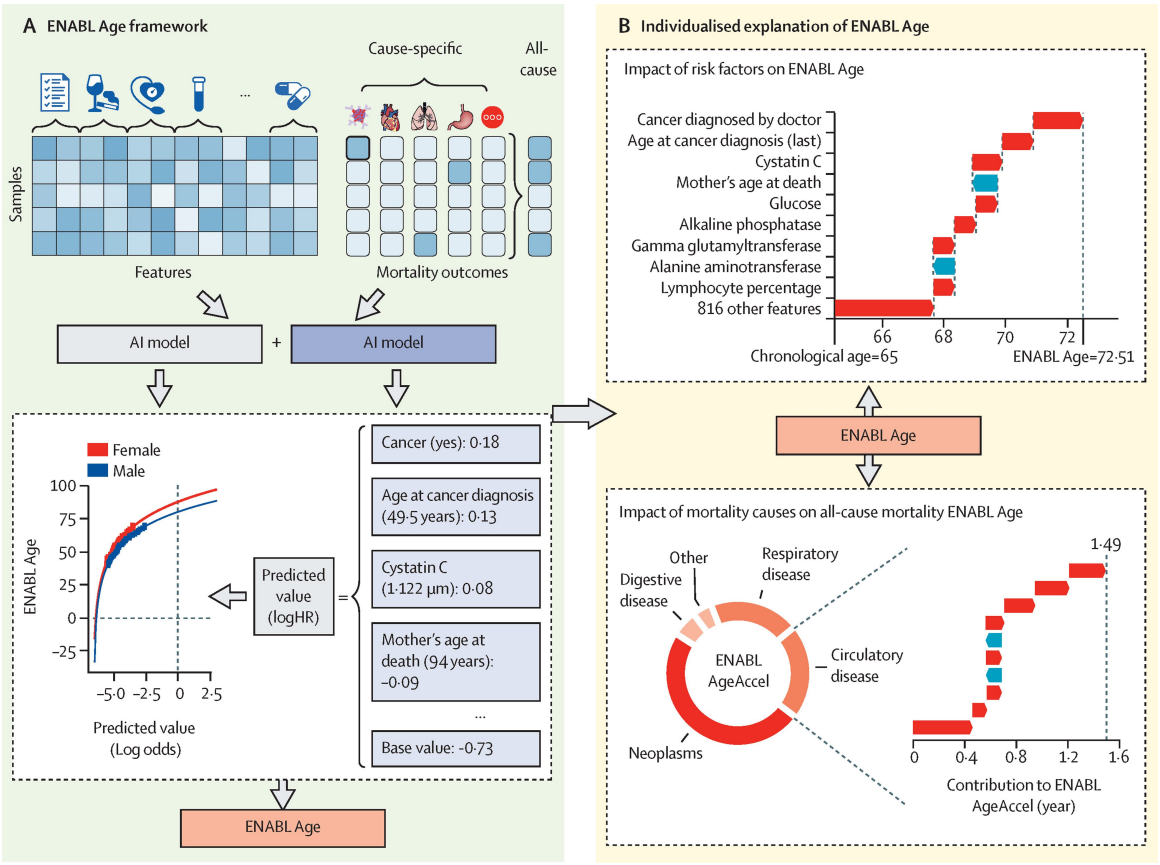


Figure 5.1: Overview of the ENABL Age framework. (A) ENABL Age directly predicts an age-related outcome of interest (eg, all-cause mortality or cause-specific mortality) using gradient-boosted trees and then non- linearly rescales the predictions to be in units of age. (B) We extended existing explainable machine-learning methods to calculate the contributions of ENABL Age input features in units of years, which makes it more interpretable by humans. We also provided the two-layer explanation of ENABL Age: ENABL Age computed based on all-cause mortality is explained based on different mortality causes first, which is, in turn, explained based on the features of an individual.

all-cause ENABL Age, we trained GBT models to predict all-cause ENABL Age using chronological age and cause-specific ENABL Ages (Supplementary Methods **??**).

5.2.3 *ENABL Age interpretability*

We used TreeExplainer [177] to determine the effect of each feature or cause on ENABL Age (Figure **??**B). TreeExplainer provides a local (ie, for each subject) explanation of the effect of input features on individual predictions for GBT models. TreeExplainer calculates the exact Shapley additive explanations (SHAP) values [178], which sum to the output of the model and rank the importance of features. To identify the strong risk factors of an individual, we calculated SHAP values using samples of the same age and gender for comparison. These values were then rescaled to ENABL Age space, resulting in units of years that sum to the ENABL AgeAccel, following the generalised rescale rule from our previous work [50]. This method allowed us to see the contribution of each feature to ENABL Age. Similarly, we calculated SHAP values of different mortality causes to all-cause mortality and rescaled them. Consequently, the ENABL Age framework provides insights into how features contribute to different mortality causes and to all-cause mortality (Supplementary Methods **??**).

5.2.4 *Statistical analysis*

Datasets were split into 80% training and 20% testing sets. The performance of Cox proportional hazard models was assessed via the concordance index. To directly compare the mortality prediction of ENABL Age with chronological age, we re-estimated the weights of PhenoAge [153] and BioAge [152] (two second-generation biological age clocks built with phenotypic features) on the UK Biobank and NHANES datasets (Supplementary Methods **??**). We computed PhenoAge and BioAge with original and re-estimated weights and trained ENABL Age clocks using the same features. We also used other subsets of features to build all-cause mortality ENABL Age clocks. Then, we trained 5-year and 10-year mortality-prediction models using different biological age accelerations adjusted for chronological age and sex, and we compared the area under the receiver operating characteristic curves (AUROCs; Supplementary Methods **??**). Associations between ENABL AgeAccels, PhenoAgeAccel, and BioAgeAccel and risk factors and age-related morbidity outcomes were examined. Risk factors and age-related morbidity outcomes were regressed on each of the biological age accelerations, adjusted for chronological age and sex (Supplementary Methods **??**).

A genome-wide association study (GWAS) on all-cause and cause-specific mortality ENABL AgeAccels was done using UK Biobank data to assess whether they capture different ageing mechanisms. BOLT-LMM [163] was used to examine the association between ENABL AgeAccel and each single nucleotide polymorphism (SNP). We also did a stepwise

model selection procedure on the genome-wide SNP summary statistics using the COJO algorithm. SNP p values lower than 8.3×10^{-9} were deemed to be statistically significant (Supplementary Methods **??**). Lastly, genetic correlations between ENABL AgeAccels and several age and health-related traits, including anthropometric traits, adiposity, longevity, lifestyle, and several diseases, were calculated. We also assessed the genetic correlations among various ENABL AgeAccels and their correlations with age accelerations on the basis of diverse biomedical data sources (Supplementary Methods **??**).

5.3 RESULTS

The GBTs showed superior performance compared with linear models for the UK Biobank dataset, with significant improvements across almost all mortality prediction tasks considered ($p = 0.003$ for circulatory, $p = 0.005$ for respiratory, $p = 0.42$ for digestive, $p < 0.0001$ for all cause, $p < 0.0001$ for neoplasms, and $p < 0.0001$ for other; Supplementary Figure **??**). For the NHANES dataset, GBTs outperformed linear models in nearly all (seven of ten) mortality prediction tasks considered, with significant improvements observed for three tasks ($p < 0.0001$; Supplementary Figure **??**). The superior prediction performance of tree models indicates that they can effectively capture signals relevant to mortality, which are also strongly related to ageing.

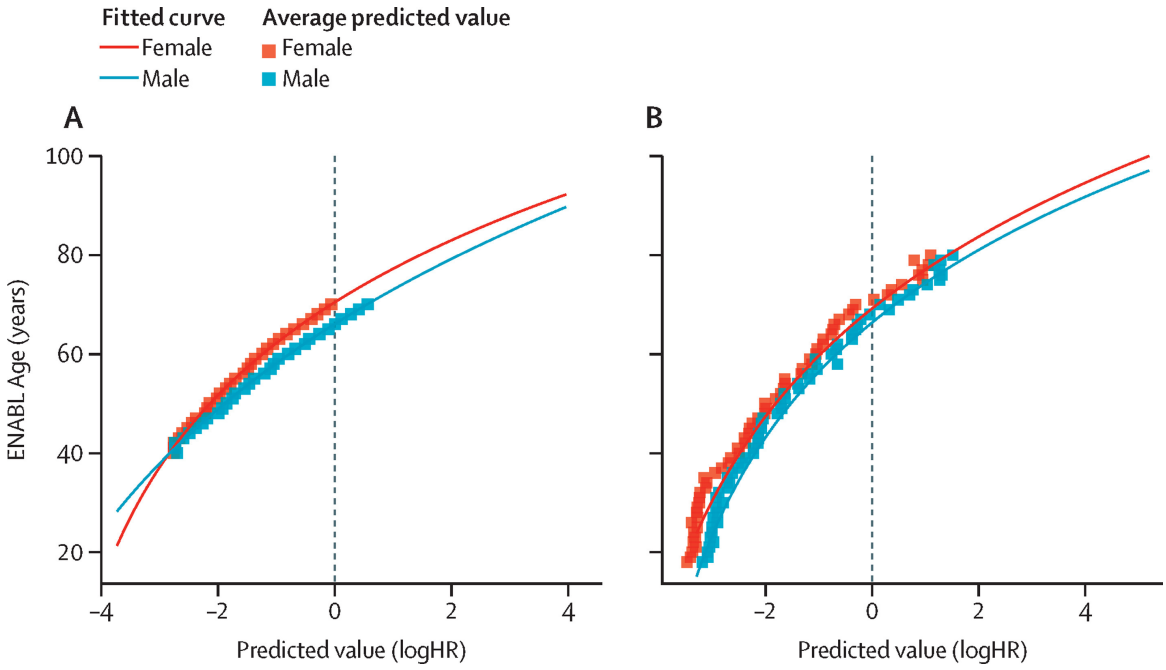


Figure 5.2: ENABL Age calculation. (A–B) The curves that transform the predicted value of GBTs to ENABL Age on the UK Biobank (A) and NHANES (B) datasets.

The inverse functions of the exponential curves accurately model the association between the average predicted values and chronological ages (Figure **??**A, B). ENABL Age was significantly correlated with chronological age ($r = 0.7867$, $p < 0.0001$ for UK Biobank; $r = 0.7126$, $p < 0.0001$ for NHANES; Supplementary Figure **??**). The median error between ENABL Age and chronological age was 4.1 (IQR 2.1–6.6) for the UK Biobank dataset and 5.9 (2.8–9.9) for the NHANES dataset. Individuals who were alive (at mortality data collection time) had a lower average ENABL Age than deceased individuals for all ages (Supplementary Figure **??**), illustrating the effectiveness of our model.

In the UK Biobank and NHANES samples, when individuals were divided according to ENABL AgeAccels into healthy age groups (ie, their ENABL Age was lower than their chronological age) and unhealthy age groups (ie, their ENABL Age was higher than their chronological age), the unhealthy groups showed three to 12 times higher log hazard ratio than the healthy groups (Supplementary Figure **??**). Individuals with the highest ENABL AgeAccels had much steeper declines in survival over the 14 years (UK Biobank) and 16 years (NHANES) years of follow-up. For neoplasm-caused mortality, which has a low base rate, the difference in survival probability between healthy ageing and unhealthy ageing individuals is especially important when considering the rarity of this mortality cause. These findings collectively indicate the effectiveness of ENABL Age in distinguishing between individuals who are healthy and unhealthy, validating its accuracy in reflecting health status.

Then, we directly evaluated the mortality prediction power of ENABL AgeAccels, PhenoAgeAccel, and BioAgeAccel by training 5-year and 10-year all-cause mortality prediction models (Supplementary Methods **??**). ENABL Age-L (using laboratory features in the four most popular blood panels), ENABL Age-Q (using the top 20 most important questionnaire features), ENABL Age-20 (using the top 20 most important features), and ENABL Age using all features significantly outperformed BioAge (NHANES dataset: $p < 0.0001$ for all comparisons with BioAge; UK Biobank dataset: $p < 0.0001$ for all comparisons with BioAge) and PhenoAge (NHANES dataset: $p = 0.082$ for ENABL Age-L vs PhenoAge, $p = 0.001$ for ENABL Age-Q vs PhenoAge, and $p < 0.0001$ for all other comparisons with PhenoAge; UK Biobank dataset: $p < 0.0001$ for all comparisons with PhenoAge; Figures **??**A, B; Supplementary Figure **??**). Furthermore, using identical features, the ENABL Age clocks (ENABL AgeAccel [PhenoAge features] and ENABL AgeAccel [BioAge features]) significantly outperformed PhenoAgeAccel ($p < 0.0001$) and BioAgeAccel ($p < 0.0001$ for the UK Biobank dataset and $p = 0.0080$ for the NHANES dataset) in terms of AUROCs, highlighting their enhanced prediction ability for mortality. Comparing AUROC differences between models incorporating biological ages and those using only chronological age and sex, ENABL Ages showed about double the improvement over BioAge and PhenoAge (Supplementary Tables **??**,**??**,**??**).

Importantly, both PhenoAge and BioAge use C-reactive protein as an input, which is not collected in the most popular blood panels. ENABL Age-L has a similar mortality

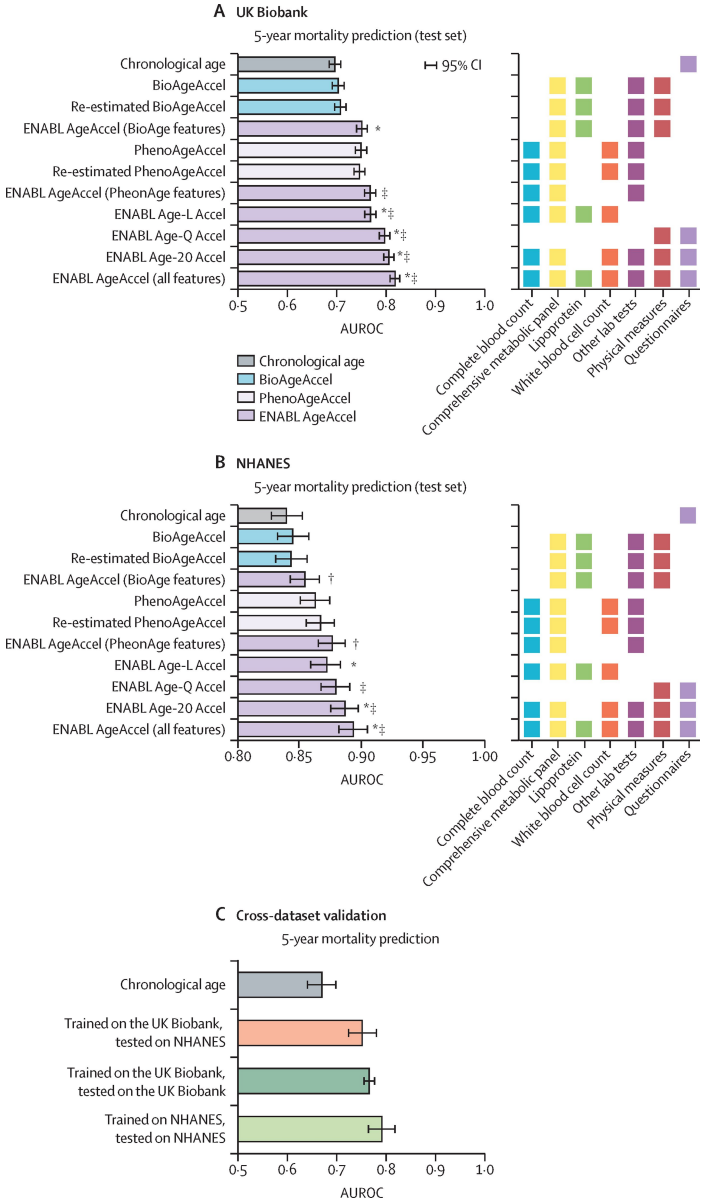


Figure 5.3: ENABL AgeAccels have high mortality prediction power. (A–B) The AUROCs of the 5-year mortality prediction GBT models of ENABL AgeAccels, PhenoAgeAccel, and BioAgeAccel adjusted by chronological age and sex trained and tested on the UK Biobank and NHANES datasets. Chronological age refers to the GBT models trained by chronological age and sex. BioAgeAccel and PhenoAgeAccel were calculated using the formulae in their original papers. Re-estimated BioAgeAccel and re-estimated PhenoAgeAccel were calculated using the re-estimated weights on the UK Biobank and NHANES datasets. ENABL AgeAccel (BioAge features) and ENABL AgeAccel (PhenoAge features) were calculated by the ENABL Age framework using only the features included in BioAge and PhenoAge, respectively. ENABL Age-L Accel refers to the ENABL Age clock that uses laboratory features in the four most popular blood panels, comprising complete blood count, comprehensive metabolic panel, lipoprotein, and white-blood-cell count. ENABL Age-Q Accel refers to the ENABL Age clock that uses the top 20 most important questionnaire features selected by recursive feature elimination. ENABL Age-20 Accel refers to the ENABL Age clock that uses the top 20 most important features of all possible features selected by recursive feature elimination. Coloured squares indicate the feature types and laboratory panels (complete blood count, comprehensive metabolic panel, lipoprotein, and white-blood-cell count) used to calculate biological ages. CIs were computed using bootstrap resampling of the test set 1000 times. p values, derived from the bootstrap results, were used to assess the performance improvement of ENABL AgeAccels over re-estimated BioAgeAccel and re-estimated PhenoAgeAccel.

prediction power to ENABL Age clock using the features that PhenoAge and BioAge rely on, suggesting that C-reactive protein can be replaced by features collected in the most popular blood panels. ENABL Age-Q and ENABL Age-20 achieved performance similar to the ENABL Age clock using all the features, showcasing the effectiveness of the selected features. In particular, ENABL Age-L and ENABL Age-Q, with their reduced feature sets, are both computationally efficient and highly accurate compared with the comprehensive ENABL Age that uses 825 features. ENABL Age-L relies only on popular blood tests and is usable by medical professionals, whereas ENABL Age-Q depends only on questionnaire features and can be used by non-professional health-care consumers.

The ENABL Age clock successfully generalises to a geographically distinct UK Biobank validation set (ie, samples collected in two Scottish centres; Supplementary Figure **??**). We also did a cross-dataset validation using ENABL Age-L by training it on UK Biobank samples and evaluating it on NHANES samples of individuals aged 40–70 years (Supplementary Methods **??**). The AUROCs of the ENABL Age clock trained on UK Biobank and tested on NHANES were similar to those trained and tested within the same dataset, confirming the external generalisability of the ENABL Age clock (Figure **??**C; Supplementary Figure **??**).

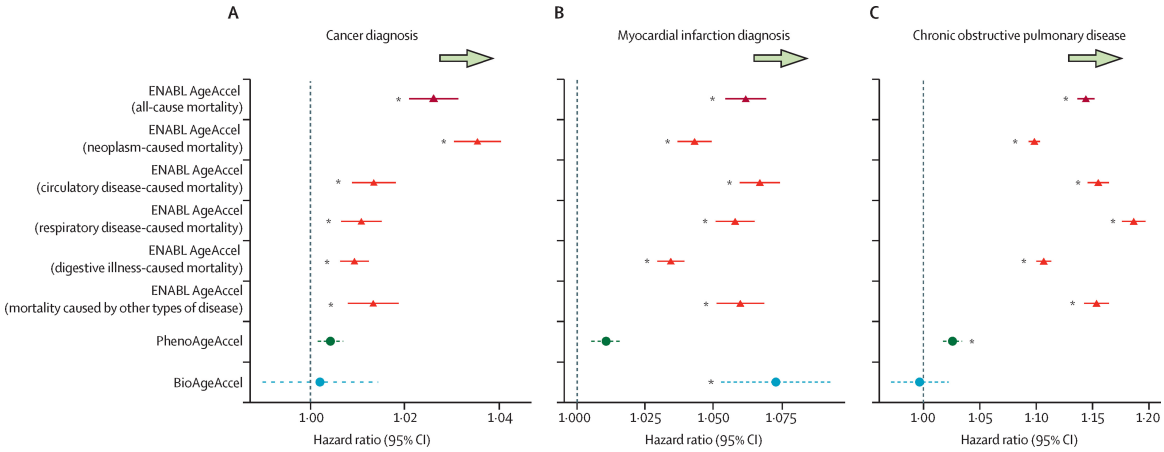


Figure 5.4: ENABL AgeAccels are strongly associated with diverse age-related morbidity outcomes. (A–C) Associations between age accelerations (ENABL AgeAccel, PhenoAgeAccel, and BioAgeAccel) with age-related morbidity outcomes (ie, cancer, myocardial infarction, and chronic obstructive pulmonary disease). The outcomes were regressed separately on each of the biological age accelerations, adjusting for chronological age and sex using Cox regression. The significance threshold was adjusted using the Bonferroni correction method. *$p < 0.00001$.

Given that biological age should reflect the health status of an individual, we examined whether ENABL Age clocks relate to diverse risk factors and age-related morbidity outcomes. We examined the associations of ENABL AgeAccels, PhenoAgeAccel, and

BioAgeAccel with a range of risk factors on the UK Biobank geographical validation data (Supplementary Methods **??**). We observed strong and significant ($p < 0.00001$, Bonferroni corrected) associations between ENABL AgeAccels and all risk factors and age-related outcomes we considered (Figure **??**; Supplementary Figures **??**,**??**); the associations were stronger than PhenoAgeAccel and BioAgeAccel for most of the risk factors and age-related outcomes. Although BioAgeAccel has stronger associations with myocardial infarction than ENABL AgeAccels, it showed no significant associations with four of seven of the age-related morbidity outcomes we consider (ie, cancer, chronic obstructive pulmonary disease, all-cause dementia, and asthma). As an approach that estimates second-generation biological age, the ENABL Age framework incorporates age-related outcomes by directly predicting them, thereby powerfully and robustly capturing ageing morbidity and mortality information. Furthermore, the ENABL Age framework can use any available features; thus, it can capture more comprehensive ageing mechanisms than PhenoAge and BioAge, which use only a small number of features correlated with ageing or mortality. In summary, the ENABL Age clocks exhibit strong and consistent performance in identifying individuals at risk, predicting mortality, and demonstrating significant associations with various risk factors and age-related morbidity outcomes.

Individualised explanations of biological age (ie, the individualised contributions of features to biological age) are important for understanding complex ageing mechanisms on a personal basis and for further guiding clinical decision making. The features that most affect all-cause mortality ENABL Age and the distribution of the effects of each feature on ENABL AgeAccel can be shown in units of years (Figure **??**A). With the global explanations of ENABL Age, we can better identify the effects of different features on ageing for a specific age and sex group. The individualised explanation of the all-cause mortality ENABL Age of a woman aged 65 years is provided (Figure **??**B). We observed that the all-cause mortality ENABL Age of the individual was 73.33 years and her neoplasm-caused mortality was 74.27 years (see neoplasm-cause mortality ENABL Age in Supplementary Figure **??**), which was higher than her chronological age (ie, 65 years). Features that increased the ENABL Ages of this individual included age at cancer diagnosis, sex-hormone-binding globulin, and long-standing illness, disability, or infirmity; features that decreased her ENABL Ages included never having any past tobacco smoking and her type of cancer tumour (epithelial).

Overall, for all-cause mortality ENABL Age, we identified nine features that appear in the top 20 most important risk factors for more than half of the randomly selected samples (n=30,000) from the test set. These features, and their respective percentages of individuals in which they appear among the top 20 most important risk factors, are as follows: long-standing illness, disability, or infirmity (29,376 [97.92%] of 30,000); cystatin C (23,487 [78.29%] of 30,000); overall health rating (21,372 [71.24%] of 30,000); average total household income before tax (21,024 [70.08%] of 30,000); red-blood-cell distribution width (20,370 [67.90%] of 30,000); pack years of smoking (18,072 [60.24%] of 30,000); past to-

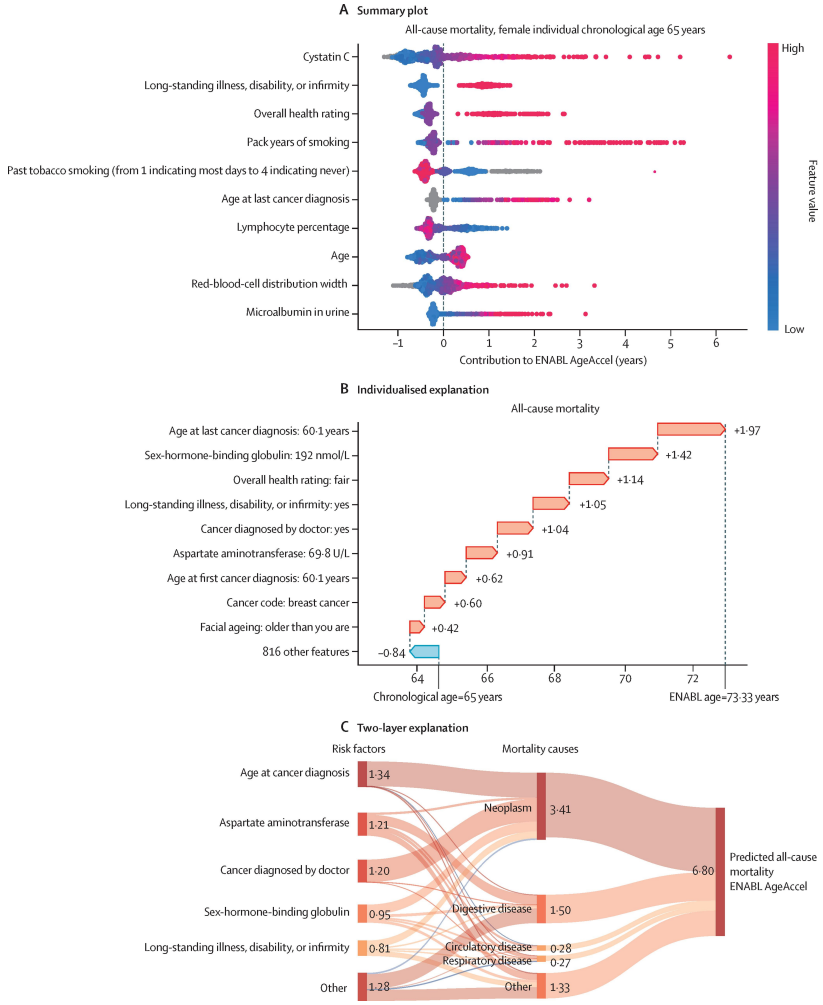


Figure 5.5: ENABL Age exposes individualised ageing explanations (A) SHAP summary plot for ENABL Age clocks trained on all-cause mortality using UK Biobank datasets for samples from female individuals aged 65 years. The features that most affect ENABL Age (ranked from the most to least important) are shown. The distribution of the effects of each feature on ENABL AgeAccel, which includes a set of plots for which each dot corresponds to an individual, is also shown. The colours represent feature values for numeric features, with red for larger values, blue for smaller values, and grey to indicate not applicable (eg, currently smoking on most or all days for past tobacco smoking). The thickness of the line (which actually comprises many individual dots) is determined by the number of examples at a given value, for which dots are spread out vertically if there are many examples. A negative rescaled SHAP value (extending to the left) indicates reduced ENABL AgeAccel, whereas a positive SHAP value (extending to the right) indicates increased ENABL AgeAccel. SHAP values are calculated using explicands and baselines that have the same age and sex (ie, female individuals aged 65 years). (B) Individualised explanation of all-cause mortality ENABL Age for a single female individual aged 65 years. The output value (the grey dashed line with the number at the top of the plot) shows ENABL Age for that individual. The base value (the grey dashed line with the number at the bottom of the plot) approximates chronological age (ie, 65 years). The features in red increase ENABL AgeAccel, and those in blue decrease it. (C) The two-layer all-cause mortality ENABL AgeAccel explanations depict the contributions of features to different mortality causes and of mortality causes to all-cause mortality ENABL Age. Flows in red represent positive rescaled SHAP values, which increase ENABL AgeAccel, whereas flows in blue indicate negative rescaled SHAP values, which decrease ENABL AgeAccel. The width of each flow and the grey numbers correspond to the contribution of risk factors to mortality causes and mortality causes to all-cause mortality ENABL Age in units of years. Predicted all-cause mortality ENABL AgeAccel refers to the predicted all-cause mortality ENABL Age using the cause-specific ENABL Ages minus chronological age.

bacco smoking (17,646 [58.82%] of 30,000); sex-hormone-binding globulin (15,534 [51.78%] of 30,000); and lymphocyte percentage (15,462 [51.54%] of 30,000). For neoplasm-caused mortality ENABL Age, we identified seven features that appeared in the top 20 most important risk factors for more than half of the same random sample: past tobacco smoking (27,279 [90.93%] of 30,000), pack years of smoking (24,384 [81.28%] of 30,000), cystatin C (22,929 [76.43%] of 30,000), red-blood-cell distribution width (21,357 [71.19%] of 30,000), C-reactive protein (21,261 [70.87%] of 30,000), age at last cancer diagnosis (20,793 [69.31%] of 30,000), and platelet distribution width (19,239 [64.13%] of 30,000).

The two-layer explanations for the same individual (Figure **??**B), provides a more comprehensive insight into ageing processes (Figure **??**C). For this individual, all mortality causes, particularly neoplasm, increased their all-cause mortality ENABL AgeAccel. We also analysed the two-layer explanations of a healthier individual (Supplementary Figure **??**), for whom most causes decreased their all-cause mortality ENABL AgeAccel (examples from the NHANES dataset are in Supplementary Figure **??**). Interpreting personal ENABL Age helps people improve their health awareness and adjust their lifestyle accordingly. This interpretation also assists researchers and medical professionals in efforts to elucidate complex and interrelated ageing mechanisms and potentially guide clinical decision making.

Previous studies have suggested that different biological age clocks for the same individual might be capturing distinct aspects of the ageing process, therefore understanding what ageing signals are being captured is vital [131, 145, 247]. In our ENABL Age model, we predicted age-related outcomes directly, and mapped these predictions to age, enabling us to determine the ageing signals that our clock captures. ENABL AgeAccels, trained on distinct causes of mortality, exhibited stronger associations with respective diseases than other ENABL Ages: neoplasm-caused mortality ENABL AgeAccel showed a stronger association with cancer; circulatory disease-caused mortality ENABL AgeAccel showed a stronger association with myocardial infarction; and respiratory disease-caused mortality ENABL AgeAccel showed a stronger association with chronic obstructive pulmonary disease. These results suggest that each ENABL Age clock, specifically trained on a unique cause of mortality, successfully captures the relevant ageing mechanisms causing these diseases.

To validate whether ENABL Age clocks trained on different tasks can capture different ageing mechanisms, we did a GWAS on all ENABL AgeAccels using UK Biobank data. Given that genes can be tied to biological ageing processes, showing the different genetic architectures of ENABL Age clocks can help us understand the ageing signals they capture (Supplementary Table **??**, Manhattan plots Supplementary Figures **??**,**??**). For all-cause mortality ENABL AgeAccel, the mapped genes were associated with anthropometric measures (eg, BMI and waist–hip ratio), blood-count measures (eg, platelet count and neutrophil count), alcohol consumption, smoking behaviour, and different cancers, which were all age-related and health-related traits [247]. For neoplasm-caused mortality

ENABL AgeAccel, the mapped genes were associated with different cancers (eg, breast cancer and lung cancer), smoking behaviour, BMI, and blood-count measures. For circulatory disease-caused mortality ENABL AgeAccel, the mapped genes were associated with circulatory diseases (eg, coronary artery disease and myocardial infarction) and protein levels.

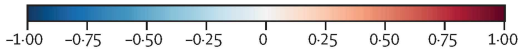
We also identified genes associated with both all-cause mortality ENABL Ages and cause-specific mortality ENABL Ages, underscoring the unique ageing mechanisms that these age-related genes contribute to. AK5 and FTO, for instance, are associated with both all-cause mortality ENABL Age and neoplasm-caused or circulatory disease-caused mortality ENABL Ages respectively, suggesting their involvement in cancer-related and circulatory disease-related ageing mechanisms. We also found overlapping genetic variants such as APOE, associated with all-cause specific mortality ENABL AgeAccels, and also associated with PhenoAgeAccel, BioAgeAccel, and longevity [145, 286]. APOE is reported to be associated with several ageing-related diseases [98, 174, 281]. Therefore, APOE might be a shared genetic factor underlying various ageing mechanisms. The genes identified by ENABL AgeAccels are also associated with other biological ages, multivariate ageing traits, and longevity (Supplementary Table **??**) [145, 189, 286], suggesting that our ENABL Age clocks concur with and potentially complement previous biological age clocks.

The Pearson correlations between cause-specific mortality ENABL AgeAccels are predominantly lower than 0.8 and those from genetic correlations are predominantly lower than 0.9 (Supplementary Figure **??**), suggesting that these cause-specific mortality ENABL Ages might capture distinct age-related signals. The all-cause mortality ENABL AgeAccel has significant correlations ($p < 0.0003$, Bonferroni corrected) with all examined health-related traits, outperforming other biological age accelerations (Figure **??**A). Notably, cancer and smoking showed the highest genetic correlation with neoplasm-caused mortality ENABL AgeAccel, whereas circulatory diseases (eg, heart attack and stroke) correlated most with circulatory disease-caused mortality ENABL AgeAccel. Results from the GWAS and genetic correlation emphasised the distinct genetic architectures of different ENABL AgeAccels, providing further evidence that our clocks capture meaningful and diverse ageing mechanisms.

There are few significant genetic correlations between ENABL AgeAccels and first-generation biological ageAccels (ie, Hannum AgeAccel, Horvath AgeAccel, and BrainAgeAccel; Figure **??**B). By contrast, strong, significant genetic correlations ($p < 0.0017$, Bonferroni corrected) are observed with second-generation biological ageAccels (ie, GrimAgeAccel and DNAmPhenoAgeAccel) indicating that ENABL AgeAccels exhibit stronger genomic overlaps with second-generation biological age measures. Notably, although the GrimAge, DNAmPhenoAge, and ENABL Ages all use mortality labels during training, the genetic correlations of ENABL Ages with other ageing clocks are all lower than 0.65, suggesting ENABL Ages capture unique ageing mechanisms that are not represented by other clocks.

A

	All-cause mortality ENABL AgeAccel	Neoplasm-caused mortality ENABL AgeAccel	Circulatory disease-caused mortality ENABL AgeAccel	Respiratory disease-caused mortality ENABL AgeAccel	Digestive disease-caused mortality ENABL AgeAccel	Other causes of mortality ENABL AgeAccel	GrimAgeAccel	DNAmPhenoAgeAccel	Hannum AgeAccel	Horvath AgeAccel	BrainAgeAccel
BMI	0·6216	0·5426	0·6612	0·4957	0·5834	0·5545	0·3574	0·2382	0·1058	0·0740	0·0838
Whole body fat mass	0·6184	0·5993	0·6194	0·4613	0·5504	0·5094	0·3324	0·2415	0·0899	0·0765	0·0287
Leg fat mass (right)	0·6375	0·6033	0·6485	0·4876	0·5818	0·5349	0·3423	0·2407	0·0889	0·0722	0·0427
Arm fat mass (right)	0·6319	0·5907	0·6468	0·4703	0·5639	0·5359	0·3417	0·2480	0·0985	0·0793	0·0430
Trunk fat mass	0·5884	0·5828	0·5813	0·4351	0·5179	0·4749	0·3187	0·2287	0·0867	0·0703	0·0145
Mother's age at death	−0·8209	−0·7600	−0·8071	−0·7864	−0·8132	−0·7526	−0·5499	−0·2520	−0·0771	−0·0304	−0·2590
Father's age at death	−0·7392	−0·6648	−0·7438	−0·6642	−0·6924	−0·6589	−0·5727	−0·3180	−0·2271	−0·1291	−0·0380
Current tobacco smoking	0·7153	0·7629	0·6069	0·6557	0·6947	0·5913	0·6309	0·2366	0·0251	0·0036	0·1244
Cancer	0·2686	0·3868	0·1333	0·1362	0·1771	0·1814	0·2015	0·2474	0·2264	0·0351	−0·0338
Heart attack	0·5115	0·4129	0·5925	0·4838	0·4975	0·4918	0·3381	0·2527	−0·0305	−0·0873	0·0746
Stroke	0·8127	0·6951	0·8483	0·7715	0·7724	0·7421	0·5011	0·3853	0·0327	0·2227	−0·0788
Angina	0·5694	0·4372	0·6609	0·5533	0·5669	0·5424	0·2809	0·1815	−0·0037	−0·0750	0·1728
Hypertension	0·4666	0·3507	0·6037	0·3875	0·4282	0·4340	0·2062	0·2411	0·0796	0·1099	0·1253



B

	GrimAge Accel	DNAmPhenoAgeAccel	Hannum AgeAccel	Horvath AgeAccel	BrainAgeAccel
All-cause mortality ENABL AgeAccel	0·6215	0·4305	0·1834	0·1151	0·1421
Neoplasm-caused mortality ENABL AgeAccel	0·6134	0·3752	0·1539	0·0712	0·0746
Circulatory disease-caused mortality ENABL AgeAccel	0·5541	0·4044	0·1478	0·1131	0·1388
Respiratory disease-caused mortality ENABL AgeAccel	0·5613	0·3995	0·1385	0·0919	0·1550
Digestive disease-caused mortality ENABL AgeAccel	0·5837	0·3785	0·1216	0·0821	0·1567
Other causes of mortality ENABL AgeAccel	0·5640	0·4024	0·1637	0·0916	0·1570

Figure 5.6: Genetic correlations of ENABL Age. (A) Genetic correlations between ENABL AgeAccels, other biological age accelerations, and health-related traits, including anthropometric traits (eg, BMI), adiposity (eg, whole-body, leg, arm, or trunk fat mass), longevity (mother and father's longevity), lifestyle (eg, smoking), and several diseases (eg, heart attack, heart failure, angina, stroke, and hypertension). The genetic correlation values that passed a significance threshold ($p < 0.0003$) are highlighted with bold text. The significance threshold was corrected using the Bonferroni correction method, accounting for a total of 143 tests (0.05 divided by 143, approximately equal to 0.0003). (B) Genetic correlations between ENABL AgeAccels and several age accelerations on the basis of diverse biomedical data sources, such as epigenetic age acceleration (eg, GrimAgeAccel, DNAmPhenoAgeAccel, HannumAccel, and HorvathAccel) [98] and brain MRI-based age acceleration (ie, BrainAgeAccel) [189]. The genetic correlation values that passed a significance threshold ($p < 0.0017$) are highlighted with bold text. The significance threshold was corrected using the Bonferroni correction method, accounting for a total of 30 tests (0.05 divided by 30, approximately equal to 0.0017). AgeAccel=age acceleration. DNAm=DNA methylation. ENABL Age=Explainable Biological Age.

5.4 DISCUSSION

In this paper, we introduced ENABL Age, a novel approach that combines complex machine-learning models and XAI methods to measure and interpret biological age. Validated through the UK Biobank and NHANES datasets, our ENABL Age clocks accurately reflect health status, predict mortality, and capture various ageing mechanisms. We enhanced the interpretability of ENABL Age by extending TreeExplainer, an XAI method, to calculate feature contributions in years. Our association analysis with risk factors and ageing-related morbidities, as well as GWAS results, reveal the unique ageing mechanisms each ENABL Age clock captures.

For all-cause mortality ENABL Age, features such as pack years of smoking and past tobacco smoking are top influencers for more than half the individuals. Pack years of smoking was also employed in the GrimAge calculation, in which DNA methylation-based surrogate biomarkers of ageing were selected [200]. Our findings corroborate the importance of smoking-related features in biological age estimation. We also studied the global feature importance of commonly used blood markers. Top influential markers for all-cause mortality ENABL Age included cystatin C, red-blood-cell distribution width, sex-hormone-binding globulin, lymphocyte percentage, alanine aminotransferase, C-reactive protein, creatinine, and glycated haemoglobin. Cystatin C, in particular, has been identified as a prominent feature of biological ageing in several studies [95, 172], underscoring its role in the ageing process. Other crucial features, such as red-blood-cell distribution width [227], lymphocyte percentage [227], alanine aminotransferase [95], creatinine [17, 95], and glycated haemoglobin [17], have also been recognised as key features in other biological age models, further substantiating their importance in the ageing process.

Genes identified by ENABL AgeAccels align with those linked to other biological ages, multivariate ageing traits, and longevity from previous research [145, 189, 286]. Our findings reveal the associations between ENABL AgeAccels and APOE, which is also associated with other biological ages and ageing traits, thereby underscoring its crucial role in the ageing process. Similarly, FTO, MED24, and RPN1, associated with various biological ages, overlap with ENABL AgeAccels findings, suggesting the consistency of our ENABL Age clocks with existing biological age clocks and their potential for complementation. Further, our cause-specific mortality ENABL Age clocks pinpoint specific ageing mechanisms to which age-related genes contribute; for instance, AK5 is possibly involved in neoplasm-related ageing and FTO potentially influences circulatory disease-related ageing mechanisms. Our genetic correlation results highlight the strong correlations between ENABL Ages, second-generation biological ages (ie, GrimAge and DNAmPhenoAge), and health traits, emphasising the effectiveness of incorporating morbidity and mortality information in biological age measurements. We observed stronger correlations between ENABL Ages and second-generation ages than first-generation counterparts. This finding aligns with the study by McCartney and colleagues [136], further indicating that clocks

integrating morbidity and mortality data capture more similar biological ageing mechanisms.

ENABL Age offers several key benefits compared with previous methods. First, it is interpretable, which is a crucial attribute of biological age clocks. ENABL Age can provide individualised explanations of accurate models by attributing contributions of input features to the biological age in years. This understanding can enhance knowledge of the ageing process, help individuals adjust their lifestyles, and potentially accelerate ageing-related drug development. Furthermore, unlike traditional mortality predictors using logHR or predicted probabilities, ENABL Age estimates mortality as a biological age, providing an easily comprehendible risk score. For instance, an ENABL Age of 65 implies similar risk score to individuals aged 65, providing a contextualised mortality and disease risk estimate.

Additionally, ENABL Age is flexible in terms of input features, the age-related phenotype or outcome being predicted, and the choice of model. The complex machine-learning models can capture high-dimensional and non-linear relationships, allowing for more comprehensive and accurate models. By directly predicting the age-related outcome, ENABL Age eliminates the need for complex feature selection as in many second-generation biological age clocks [152, 153, 172], enabling the inclusion of all available features to capture a broader range of ageing signals. This adaptability also facilitates the creation of models that can depend on specific types of features, such as questionnaires or lab tests.

Many previous biological age clocks have been built by analysis of omics data, such as epigenetics [108, 117, 172], transcriptomics [82, 194], proteomics [149, 279], and more. Because of flexibility, we can easily extend ENABL Age to omics data by using these data types as the input features. Moreover, ENABL Age allows us to determine what ageing signals to capture by defining various age-related prediction tasks. We demonstrate ENABL Age's applicability to other age-related outcomes by accurately capturing dementia-related ageing signals in the ROSMAP dataset (Supplementary Results **??**, Supplementary Figure **??**). Therefore, we can measure different biological ages for different ageing aspects, providing a comprehensive ageing process understanding and answering targeted questions about specific ageing aspects.

Although we used GBTs in this paper because of their superior performance on tabular datasets, particularly mortality predictions [177, 231], the flexibility of ENABL Age enables the use of other complex machine-learning models, such as deep neural networks. Despite their initial absence of interpretability, advancements in feature-attribution methods are making these models more insightful. Future work will leverage this flexibility to incorporate neural networks and feature-attribution methods for multiomic data and multitask learning, which is not possible with GBTs. This approach will help discern shared or unique ageing biomarkers across omics layers and simultaneously measure biological age for different ageing aspects.

Despite advancements, ENABL Age has limitations. First, ENABL Age does not provide causal insights, which is a common issue even in many of the previous biological age models [247]. However, ENABL Age explanations can reveal meaningful biological associations. Moreover, incorporating additional mortality information, as in PhenoAge, increases the likelihood of identifying potential causal factors of ageing [199]. Because ENABL Age includes additional ageing phenotypes and outcomes, it similarly has a greater potential to identify causal factors. Furthermore, for a more profound understanding of causality, careful feature selection will be necessary in future research. Second, our study primarily applies to older individuals and those who are middle aged in the UK Biobank dataset; generalising the findings requires further exploration. Third, we did not apply our approach to omics data; DNA methylation data, for instance, are commonly used in previous studies to determine biological age. However, our flexible ENABL Age approach can be readily extended to various omics data in the future. Lastly, high genetic correlations between different ENABL AgeAccels observed could be caused by shared genetic factors or unaccounted secondary mortality causes, warranting further investigation with curated mortality-cause datasets. In summary, we developed and validated a new approach, ENABL Age, to measure and interpret biological age related to different ageing mechanisms using the combination of a complex machine-learning model and XAI methods. ENABL Age takes a consequential step towards applying XAI to interpret biological age models. Its flexibility allows for many future extensions to omics data, even multiomic data, and multitask learning.

5.A SUPPLEMENTARY METHODS

5.A.1 *Data collection and processing*

5.A.1.1 *UK Biobank*

The participants were enrolled in the UK Biobank from April, 2007, to July, 2010, from 21 assessment centres across England, Wales, and Scotland using standardised procedures. Ethics approval for the UK Biobank study was obtained from the North West - Haydock Research Ethics Committee (21/NW/0157). Informed consent was obtained from all UK Biobank participants (the consent form is available at `https://www.ukbiobank.ac.uk/consent`). The participants visited their closest assessment centre to provide baseline information, physical measures, and biological samples. In this study, we include all measurements available on November 19, 2020. We exclude (1) features that are missing in more than 80% of the samples and (2) one feature from pairs of highly correlated features with correlations greater than 0.98. Specifically, we calculate the pairwise correlation between all pairs of features and removed one of the features in each pair if their correlation coefficient exceeded 0.98. After preprocessing, our UKB data has 501,366 samples aged 40-70 with 825 features from numerous categories: demographics, blood assays, health and

medical history, lifestyle and environment, physical measures, etc. We use the 35,735 samples from two Scottish centers (Edinburgh and Glasgow) as the geographical validation set. We impute missing data using MissForest [267], a nonparametric random forest-based multiple imputation method for mixed-type data.

The mortality data we use includes all deaths occurring before March 8, 2022. Detailed information about the mortality data is available at https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/DeathLinkage.pdf. As with Ganna and Ingelsson [91], we defined five cause-specific mortality categories using the International Classification of Diseases, edition 10 (ICD-10), classification as follows: neoplasms, C00–D48; diseases of the circulatory system, I05–I89; diseases of the respiratory system, J09–J99; diseases of the digestive system, K20–K93; and other diseases, all remaining ICD-10 codes. We remove deceased individuals due to external causes (V01–Y84) of mortality and morbidity. We consider only the primary cause of death for all participants. When predicting cause-specific mortality, the individuals deceased due to all other causes are excluded to prevent incorporating individuals whose contributing (secondary) causes of death include the cause predicted in the study. The demographic characteristics and sample size of the data for different tasks are shown in Supplementary Table **??**. The sample size flow chart is shown in Supplementary Figure **??**.

5.A.1.2 *NHANES*

The National Health and Nutrition Examination Survey (NHANES) from the National Center for Health Statistics (NCHS)(http://www.cdc.gov/nchs/nhanes.htm) conducts interviews and physical examinations to assess the health and nutrition data for all ages in the United States. The interviews include demographic, socioeconomic, dietary, and health-related questions. The examinations include medical, dental, physiological measurements, and laboratory tests administered by highly trained medical personnel. Since 1999, data were collected and released at 2-year intervals. Each year NHANES examines a nationally representative sample of roughly 5,000 individuals across the Unites States. In this study, we include NHANES data collected between 1999 and 2014. We exclude participants under age 18 because they are not eligible for public release mortality data. Our study includes samples with known mortality status who participated in NHANES 1999-2014 ($n = 47,261$). In the raw data, individuals 85 and over are topcoded at 85 years of age in NHANES 1999-2006 and individuals 80 and over are topcoded at 80 years of age in NHANES 2007-2014. To keep consistency, we topcode individuals 80 and over at 80 years of age. We include all demographic, laboratory, examination, and questionnaire features that could be automatically matched across different NHANES cycles. We exclude variables that are missing for more than 50% of the participants and one feature from pairs of highly correlated features with correlations greater than 0.98; after filtering and one-hot encoding, 158 features remain. We also impute missing data using MissForest [267]. We use a smaller missing rate threshold for NHANES data because the training data size

could be insufficient for effectively training imputation models for variables with a high missing rate.

All-cause mortality is ascertained by a linked NHANES mortality file that provides follow-up mortality data from the date of survey participation through December 31, 2015. For NHANES 1999-2014, the eight cause-specific death categories in the linked mortality files include the following groups selected from the NCHS primary cause-of-death records: heart disease, cancer (malignant neoplasms), chronic lower respiratory disease, cerebrovascular diseases, Alzheimer's disease, diabetes, pneumonia and influenza, and kidney disease. We remove deceased individuals of unintentional injuries. All remaining deaths are categorized into other causes. When predicting cause-specific mortality, the individuals deceased due to all other causes are excluded to prevent incorporating individuals whose contributing (secondary) causes of death include the cause predicted in the study. The demographic characteristics and sample size of the data for different tasks are shown in Supplementary Table **??**. The sample size flow chart is shown in Supplementary Figure **??**.

5.A.1.3 *ROSMAP*

The Religious Orders Study (ROS) [1] and Memory Aging Project (MAP) [2] are complementary epidemiological studies that each enroll persons without dementia who agree to annual evaluations and eventual organ donation. ROS enrolls clergy living communally from 40 Catholic groups across the US. As a complementary study, MAP recruited participants from a wider range of life experiences throughout northeastern Illinois. Clinical data collection procedures were consistent between both studies to allow the data to be merged for analyses [2]. Due to their recruitment strategies, followup rates of survivors reached around 95% for both studies.

We use the data from Beebe-Wang, Okeson, Althoff, and Lee [24]. Our prediction task is dementia onset in individuals with no history of dementia within the next three years. We have a sample size of 9,110 samples, of which 1,244 are labeled as positive (dementia onset within the next three years). The data is derived from 1,597 individuals, of which 521 developed dementia. We use 53 features including demographics features, medical history features, lifestyle factors, cognitive tests and genetic features (APOE genotype). We use the data collected in one year to predict dementia onset within the next three years, without using repeated measurements.

5.A.2 *ENABL Age approach*

We construct ENABL Age clocks in two stages. First, we develop predictors of all-cause mortality and cause-specific mortality on all or a subset of the variables using Cox proportional hazard (CoxPH) models. Second, we fit an exponential curve to the training samples' chronological ages and the predictions of the GBT models. Then, we use the

inverse function of the exponential curve to calculate our ENABL Age given a prediction. We define ENABL Age acceleration (AgeAccel) as the difference between the ENABL Age estimate and chronological age.

5.A.2.1 *Age-related outcome prediction models*

To model mortality, we use gradient boosted trees (GBTs) with the Cox regression objective. GBTs are nonparametric methods composed of iteratively trained decision trees. The final ensemble of trees captures non-linearity and interactions between predictors. The datasets are randomly divided into training (64%), validation (16%), and testing (20%) sets. The hyperparameters are chosen by GridSearch using the training and validation set. We use the XGBoost implementation [52] (https://xgboost.readthedocs.io/en/latest/python/index.html). The hyperparameters are chosen from the following values:

- Learning rate: 0.01
- Maximum number of trees: 1,000
- Early stopping rounds: 100
- Maximum tree depth: {1, 3, 5, 7, 9}
- Subsampling: {0.2, 0.5, 0.8, 1.0}

Parameter values not specified above are left at their default values.

For comparison, we also train linear Cox's proportional hazard models. We use the implementation Lifelines (https://lifelines.readthedocs.io/en/latest/fitters/regression/CoxPHFitter.html). The hyperparameters specified below are chosen by GridSearch using the train and validation set. The hyperparameters are chosen from the following values:

- Penalizer: {0, 0.001, 0.01, 0.1, 1, 10, 100}
- L1 ratio: {0, 0.001, 0.01, 0.1, 1, 10, 100}

Parameter values not specified above are left at their default values. The penalty term is

$$penalizer(\frac{1-l1_{ratio}}{2}\|\beta\|_2^2) + l1_{ratio}\|\beta\|_1.$$

Therefore, when the $l1_{ratio}$ is set to 1, the model is lasso; when the $l1_{ratio}$ is set to 0, the model is ridge; in other cases, the model is an elastic net.

Models' performance is measured with the concordance index (C-index) which is the proportion of concordant pairs divided by the total number of possible evaluation pairs. We bootstrap the test set for 1,000 times and assess the statistical significance of the difference in C-index for pairs of models. Specifically, we resample with replacement from the test set 1,000 times and compare the models' performance on resampled test sets. We

report a p-value which is the percentage of times that the linear Cox proportional hazard model is better than or equal to gradient boosted trees, divided by the number of resampled test sets. All models are built in Python 3.7.

To predict dementia onset using ROSMAP data, we use GBTs with a logistic regression objective, but all other training procedures are same an in the mortality prediction models. Models' performance is measured with AUROC. We also bootstrap the test set 1,000 times when evaluating this model. We also conduct significance tests comparing the C-indices of the GBT and linear models.

5.A.2.2 *Recursive feature elimination*

To find less costly but nearly as accurate models, we select features using recursive feature elimination. Recursive feature elimination works by searching for a subset of features by starting with all features in the training dataset and successively removing features until the desired number of features remains. Firstly, we train a model on the full dataset with all features. Then we rank features by importance (according to mean absolute SHAP values) and remove the least important features. Another model is trained on the resulting feature set, and the process iterates until only the desired number of features are left. Starting from all features, we remove about 5 features in each iteration until only one feature is left. The model's performance is shown in Supplementary Figure **??**. Specifically, we (bootstrap) resample with replacement from the test set 1,000 times and report the average and the 95% confidence interval of the C-index.

5.A.2.3 *Rescaling predictions to ENABL Age*

The prediction (i.e., logHR) of the mortality/dementia onset prediction models can be interpreted as a version of ENABL Age which is not contextualized in terms of chronological age. To do so, we non-linearly transform the models' predictions to produce biological age estimates, which are in units of years. Specifically, we fit an exponential curve, with parameters a, b, and c:

$$\begin{aligned} &\mathrm{pred} = e^{a \times \mathrm{age} + b} + \min(\mathrm{pred}) + c, \\ &c = -0.1, \text{ if } c \geqslant 0, \end{aligned}$$

to the training samples' chronological ages (age) and the predictions (pred) of the GBT models using the Levenberg-Marquardt algorithm [151]. Here, $\min(\mathrm{pred})$ is the minimum value of the prediction of training samples. We use the inverse function of the exponential curve:

$$\text{ENABL Age} = \frac{\ln(\mathrm{pred} - \min(\mathrm{pred}) - c) - b}{a}$$

to calculate our ENABL Age given a prediction.

We choose an exponential curve to model the relationship between chronological age and logHR because it provides a better fit to the data compared to a linear function used by GrimAge [172]. Supplementary Figure **??** shows the root mean squared errors (RMSEs) and fitting plots of both exponential and linear curves for the UKB dataset and NHANES dataset. These results demonstrate that the exponential curves yield lower RMSEs and better fitting performance on both datasets. Moreover, for the exponential curve we use, the parameter "a" controls the slope and concavity of the curve. When "a" is extremely small, the curve is close to a linear curve. Therefore, using this exponential curve could capture more relationships between logHR and chronological age than a simple linear curve.

5.A.2.4 *All-cause mortality ENABL Age prediction using cause-specific ENABL Ages*

To obtain the contribution of different cause-specific ENABL Age estimate to the all-cause ENABL Age estimate, we train GBTs models to predict the all-cause mortality ENABL Age estimate using chronological age and cause-specific mortality ENABL Age estimates for different sex using UKB and NHANES dataset. We use GBTs with the mean squared error objective. The dataset is randomly divided into training (64%), validation (16%) and testing (20%) sets. We use the following parameters:

- Learning rate: 0.01
- Maximum number of trees: 1,000
- Early stopping rounds: 100

Parameter values not specified above are left at their default values. The trained models achieve high precision (R^2 = 0.9722 for female, R^2 = 0.9721 for male).

5.A.2.5 *Interpretation of ENABL Age*

To interpret the ENABL Age estimate, we utilize TreeExplainer [177], which provides a local explanation of the impact of input features on individual predictions. TreeExplainer calculates exact SHAP [178] (SHapley Additive exPlanations) values for tree-based models.

SHAP (SHAPLEY ADDITIVE EXPLANATION) VALUES Firstly, we calculate the SHAP values for the mortality prediction models. SHAP (SHapley Additive exPlanation) values attribute to each feature the change in the expected model prediction when conditioning on that feature. The change of the model's prediction when the feature is masked is

recorded across all possible subsets of features, yielding an average change in prediction resulting from the inclusion of a feature in the model:

$$\phi_i(f, x) = \sum_{R \in \mathcal{R}} \frac{1}{M!} \left[f_x(P_i^R \cup i) - f_x(P_i^R) \right],$$

where ϕ_i is the feature attribution (SHAP value) of feature i in model f for data point x, $\mathcal{R}$ is the set of all feature permutations, P_i^R is the set of all features before i in the ordering R, M is the number of input features. In this paper, we use baseline shapley values and marginal shapley values. For baseline shapley values ($\phi(f, x^e, x^b)$), $f_x(S) = f(x_S^e, x_{\bar{S}}^b)$, where $f(x_S^e, x_{\bar{S}}^b)$ denotes evaluating f on a hybrid sample where present features are taken from the explicand x^e and absent features are taken from the baseline x^b. Then we estimate marginal Shapley values by first estimating baseline Shapley values for many baselines and then averaging them [48].

SHAP values guarantee a set of desirable theoretical properties, including additivity and consistency. In particular, additivity states that when approximating the original model f for a specific input x, the SHAP values sum up to the output $f(x)$:

$$f(x) = \phi_0(f) + \sum_{i=1}^{M} \phi_i(f, x),$$

where $\phi_0(f) = E[f(x)] = f_x(\emptyset)$ that is the average model prediction on the baseline samples. Consistency states that if a model changes so that some feature's contribution increases or stays the same regardless of the other inputs, that input's attribution should not decrease.

For the age-related outcome (e.g., mortality and dementia onset) prediction models, we use the baselines and explicands (i.e., samples being explained) of the same age and sex. The baselines (i.e., background samples) are from the training set and the explicands are from the testing set. Therefore, the $\phi_0(f)$ for a specific age equals to the average predictions of the samples of that age in the training set.

RESCALING SHAP VALUES TO THE ENABL AGE SPACE We rescale the SHAP values to the ENABL Age space so that the rescaled SHAP values are in units of years and sum to the ENABL AgeAccel. To do so, we use the generalized rescale rule proposed in our previous work [49]. Suppose we have a age-related outcome prediction model f, and a non-linear model g that rescales the prediction to ENABL Age. We can calculate ENABL Age using $h(x) = g(f(x)), x \in \mathbb{R}^m$. We first calculate the baseline shapley values $\phi(f, x^e, x^b)$ for the age-related outcome prediction model f given an explicand x^e and a baseline x^b. Then, based on the generalized rescale rule, we can calculate the baseline shapley values for h as follows:

$$\phi(h, x^e, x^b) = \phi(f, x^e, x^b)((h(x^e) - h(x^b)) \oslash (f(x^e) - f(x^b))),$$

where $\oslash$ denotes Hadamard division. After rescaling, the $\phi_0(h)$ for a specific age equals to the average ENABL Age of the samples of that age in the training set, which is approximate to that age. At the end, we average the baseline shapley values to produce an estimate of the SHAP values (marginal shapley values). Therefore, the sum of the rescaled SHAP values is approximate to the ENABL Age acceleration. As a consequence, we can decompose the ENABL Age to the contribution of different features in units of years.

TWO-LAYER EXPLANATION OF ALL-CAUSE MORTLALITY ENABL AGEACCEL We train models to predict the all-cause ENABL Age estimates using chronological age and cause-specific ENABL Age estimates for different sexes, so that we can rescale the SHAP values of different cause-specific mortality predictions to all-cause mortality ENABL AgeAccel. We use t(chronological age, $h_0(x)$, $h_1(x)$, $\cdots$, $h_{k-1}(x)$) to denote the all-cause ENABL Age prediction model, where k is the number of mortality causes we consider. We can rescale the SHAP values as follows:

$$\begin{aligned}\phi(t, x^e, x^b) &= \phi(h, x^e, x^b)(\phi(t, x^e, x^b) \oslash (h(x^e) - h(x^b))) \\ &= \phi(f, x^e, x^b)(\phi(t, x^e, x^b) \oslash (f(x^e) - f(x^b)))\end{aligned}$$

Consequently, we achieve the two-layer explanations of ENABL AgeAccel, i.e., the features to different cause-specific mortality as well as cause-specific mortality to all-cause mortality ENABL Age estimates.

5.A.3 *BioAge and PhenoAge measures*

5.A.3.1 *BioAge*

BioAge is a biological age estimator developed by Levine [152] using data from the National Health and Nutrition Examination Survey III (NHANES III) data. Based on prior knowledge regarding the features' role or dependency in the aging process and significant correlation with chronological age, 21 biomarkers are preselected for BioAge. Then, 10 features are correlated with chronological age with Pearson correlation coefficients greater than 0.1 or less than -0.1. Three biomarkers were further removed for not being significantly loaded on the first principal component in men or in women based on the principal components analysis results using the 10 biomarkers. Seven biomarkers and chronological age were used to calculate BioAge by applying an algorithm previously proposed by Klemera and Doubal [142]:

$$\frac{\sum_{i=1}^{7}(x_i - q_i)(\frac{k_i}{s_i^2}) + \frac{age}{\alpha}}{\sum_{i=1}^{7}(\frac{k_i}{s_i})^2 + \frac{1}{\alpha}},$$

BioAge = where x_i represents the values of the features. We compare ENABL Ages with BioAge using the weights from the original paper as well as re-estimated weights based

on our UK Biobank and NHANES datasets. The weights are provided in Supplementary Table **??**.

5.A.3.2 *PhenoAge*

PhenoAge is a biological age estimator developed by Levine, Lu, Quach, Chen, Assimes, Bandinelli, Hou, Baccarelli, Stewart, Li, et al. [153] also using data from the NHANES III dataset for mortality prediction. The model was trained on 41 available biomarkers, and nine key biomarkers were selected through Cox penalized regression. PhenoAge incorporates these nine biomarkers along with chronological age to construct a Gompertz distribution-based parametric proportional hazards model. The 10-year mortality risk is then converted into years to derive PhenoAge. The formula for PhenoAge is given by:

$$\text{PhenoAge} = \alpha + \frac{\ln(\beta \times \ln(1 - \text{mortality risk}))}{\theta},$$

$$\text{mortality risk} = 1 - e^{\frac{-e^{\sum_{i=1}^{10} x_i b_i + b_0}(e^{(120\gamma)} - 1)}{\gamma}},$$

where x_i and b_i represent the values and linear weights of the features. We compare ENABL Ages with PhenoAge using the weights from the original paper as well as re-estimated weights based on our UK Biobank (UKB) and NHANES datasets. The weights are provided in Supplementary Table **??**.

5.A.4 *5-year and 10-year mortality prediction models*

5.A.4.1 *UK Biobank and NHANES mortality prediction models*

We directly evaluate the mortality prediction power of ENABL AgeAccel, PhenoAgeAccel and BioAgeAccel by training 5- and 10-year mortality prediction models adjusted by chronological age and sex. Firstly, We calculate PhenoAge and BioAge using the formulae in their original papers. Then, we train all-cause mortality ENABL Age clocks using the features included in PhenoAge and BioAge to show the effectiveness of ENABL Age framework. We further build all-cause mortality ENABL Age clocks using other subsets of features: laboratory features in the four most popular blood panels (ENABL Age-L; Supplementary Table **??**) – CBC (Complete Blood Count), CMP (Comprehensive Metabolic Panel), LP(Lipoprotein (a)) and WBC (White Blood Cell Count), the top 20 most important questionnaire features (ENABL Age-Q; Supplementary Table **??**), and the top 20 most important features (ENABL Age-20; Supplementary Table **??**). We train GBTs to predict 5- and 10-year all-cause mortality using ENBAL AgeAccel, PhenoAgeAccel, and BioAgeAccel, chronological age and sex. To compare with chronological age, we also train 5- and 10-year mortality prediction models using only chronological age and sex. The dataset is

randomly divided into training (80%) and testing (20%) sets. We use GBTs using the logistic regression loss with default hyperparameters. Models' performance is measured with AUROC. We bootstrap the test sets/geographical validation set 1,000 times and obtain the confidence intervals of the AUROCs. We also calculate the p-values using the bootstrap results to assess the performance improvement of ENABL AgeAccels over re-estimated BioAgeAccel and re-estimated PhenoAgeAccel.

5.A.4.2 *Cross dataset validation*

We perform a cross-dataset validation using ENABL Age-L by training it on UKB samples using the features in the four popular blood panels that overlap between UKB and NHANES datasets and evaluating the 5- and 10-year mortality prediction power on NHANES testing samples aged 40-70. Specifically, the ENABL Age-L is trained on UKB samples and used to calculate the ENABL Age-L for NHANES samples. Then, we train and test 5- and 10-year mortality prediction models using the ENABL Age-L, chronological age and sex on NHANES samples aged 40-70. For "Train on NHANES, test on NHANES", we train the ENABL Age-L on NHANES training samples aged 40-70 and used it to calculate the ENABL Age-L for NHANES testing samples. Then, we train and test 5- and 10-year mortality prediction models using the ENABL Age-L, chronological age and sex on NHANES samples aged 40-70. For "Train on UKB, test on UKB", we train the ENABL Age-L on UKB training samples aged 40-70 and used it to calculate the ENABL Age-L for UKB testing samples. Then, we train and test 5- and 10-year mortality prediction models using the ENABL Age-L, chronological age and sex on UKB samples. To compare with chronological age, we also train 5- and 10-year mortality prediction models using only chronological age and sex on NHANES samples aged 40-70.

5.A.5 *Association analysis*

We examine the associations of ENABL AgeAccels, PhenoAgeAccel, BioAgeAccel with a range of risk factors and age-related morbidity outcomes using UKB test set. We consider the following risk factors and age-related morbidity outcomes:

- Pack years of smoking: Pack years calculated for individuals who have smoked. The general definition of a pack year is the number of cigarettes smoked per day, divided by twenty, multiplied by the number of years of smoking. The number of years of smoking is calculated by substracting the age of starting smoking from the age smoking was stopped. The pack years of smoking is available in data field 20161.
- Walking pace: The walking pace is collected from the touchscreen question "How would you describe your usual walking pace?" from all participants except those who indicated they were unable to walk. We consider "less than 4 miles per hour"

as slow pace and "more than 4 miles per hour" as fast pace. The walking pace is available in data field 924.

- Grip strength: We use the maximum for the left hand grip strength and right strength grip strength. The units of grip strength is Kg. The grip length measurements are available in data field 46 (left hand) and 47 (right hand).
- Forced expiratory volume in 1-second: We use the highest measure from the array of values for Forced Expiratory Volume in 1-second which is in units of liter. The forced expiratory volume in 1-second is available in data field 20150.
- Waist-hip ratio: We calculate the waist-hip ratio using waist circumference divided by the hip circumference. The waist circumference is available in data field 48 and the hip circumference is available in data field 49.
- Cancer diagnosis: The time-to-event is calculated by subtracting the date of attending the assessment center from the date of cancer diagnosis. The date of cancer diagnosis is available in data field 40005.
- Myocardial infarction diagnosis: The date of myocardial infarction diagnosis is available in data field 42000.
- Stroke diagnosis: The date of stroke diagnosis is available in data field 42006.
- COPD diagnosis: The date of COPD diagnosis is available in data field 42016.
- Asthma diagnosis: The date of asthma diagnosis is available in data field 42014.
- All-cause dementia diagnosis: The date of all-cause dementia diagnosis is available in data field 42018.
- End-stage renal disease diagnosis: The date of end-stage renal disease diagnosis is available in data field 42026.

We retrain the ENABL Age models to avoid circular analysis. Specifically, we first filtered out the outcome feature of interest, i.e., the risk factors and morbidity outcomes, and the related features from the feature set. We then retrained the mortality prediction models and obtained the new ENBAL AgeAccels, which are used to fit the regression models for association analyses. We separately regress risk factors and outcomes on each of the biological age accelerations, adjusting for chronological age and sex using ordinary least squares regression with biological age accelerations as the dependent variable (pack years of smoking and waist-hip ratio), ordinary least squares regression with biological age accelerations as the independent variable (grip strength and forced expiratory volume in 1-second), logistic regression with biological age accelerations as the independent variable (walking pace), and Cox regression with biological age accelerations as the independent

variable (time to cancer, myocardial infarction, stroke, COPD, asthma, all-cause dementia, and end-stage renal disease diagnosis), as appropriate. We report the change (i.e., the beta for ordinary least squares regression; the odds ratio for logistic regression; and the hazard ratio for Cox regression) in each of the outcome measures associated with a 1-year increase in age acceleration for each of the four biological aging measures or a standard unit increase in risk factors (i.e., pack years of smoking and waist-hip ratio). We include height as a covariate when fitting our regression models for FEV1, grip strength, and walking pace. The significance thresholds for p-values are adjusted using the Bonferroni correction method, accounting for a total of 96 tests (8 biological ages and 12 traits). We use the implementation Statsmodels (https://www.statsmodels.org/stable/index.html) for ordinary least squares regression and logistic regression and Lifelines (https://lifelines.readthedocs.io/en/latest/fitters/regression/CoxPHFitter.html) for Cox regression. All analyses are conducted using Pythob 3.7.

5.A.6 *Genome-wide association analysis*

5.A.6.1 *Quality control*

In our GWAS analysis, we include all white British that have very similar genetic ancestry based on a principal components analysis of the genotypes in UK Biobank, identified using the data field 22006. Additionally, one in third-degree or closer pairs are removed, identified via pairwise kinship coefficients. SNPs are excluded if meeting any of the criteria: (1) missing rates exceeding 0.01 (2) MaCH Rsq imputation quality metric < 0.3, (3) minor allele frequency $< 0.1\%$, (4) Hardy–Weinberg equilibrium test p-value $< 10^{-6}$, (5) missing imputation information score, minor allele frequency, or Hardy–Weinberg equilibrium test result. Overall, 12,755,286 SNPs passed the quality control. The quality control is implemented using PLINK 2.0 https://www.cog-genomics.org/plink/2.0/. We use LD-based pruning with an r2 threshold of $r^2 < 0.8$ and identified 199,637 independent genomic regions from the UK Biobank data.

5.A.6.2 *GWAS analysis*

The association between ENABL Age accelerations with each SNP is examined using an efficient Bayesian linear mixed effects model (BOLT-LMM software version 2.4; https://alkesgroup.broadinstitute.org/BOLT-LMM/BOLT-LMM_manual.html) [163] adjusted for chronological age, sex, genotyping array type, and assessment center, and top 20 genetic principal components. By default, the LD scores included in the BOLT-LMM for European-ancestry samples are used to calibrate the BOLT-LMM statistic. 5×10^{-8} is a commonly used significance threshold in genome-wide association studies based on the estimation of approximately 1 million independent SNPs in a population [216]. To account for the multiple testing of the AgeAccel estimators, we use the Bonferroni correction, which adjusts

for the number of tests performed. We divided the significance threshold ($5 \times 10{-}8$) by the number of AgeAccel estimators tested (6), yielding an adjusted p-value of 8.3×10^{-9}. Thus, SNP p-values smaller than 8.3×10^{-9} are deemed to be statistically significant. Manhattan plots are created using the CMplot R package. We use the LD score regression intercept to evaluate the presence of genomic inflation due to population stratification or cryptic relatedness. The LD score regression intercept minus one is an estimator of the mean contribution of confounding bias (i.e., cryptic relatedness and population stratification) to the inflation in the test statistics [36].

We also perform a stepwise model selection procedure on the genome-wide SNP summary statistics to identify independent signals using the COJO (Conditional and Joint association analysis) [329] model in the GCTA (Genome-wide Complex Trait Analysis) software (https://yanglab.westlake.edu.cn/software/gcta/#Overview) [330]. SNPs more than 10,000 kb away from each other are assumed to be in complete linkage equilibrium. As SNPs are selected, the SNPs with multiple regression R2 greater than 0.9 with already pre-selected SNPs are excluded, so that redundant signals from high LD are excluded. We identify 62 conditionally independent associations for all-cause mortality ENABL AgeAccel, 34 for neoplasm-cause mortality ENABL AgeAccel, 133 for circulatory disease-cause mortality ENABL AgeAccel, 107 for respiratory disease-cause mortality ENABL AgeAccel, 101 for digestive disease-cause mortality ENABL AgeAccel, and 29 for other causes mortality ENABL AgeAccel. (p-value $< 8.3 \times 10^{-9}$ (Bonferroni corrected), Manhattan plots in Supplementary Figure **??** and Supplementary Figure **??**). The LD score regression intercepts for the GWAS on ENABL AgeAccels are smaller than 1.1 (Supplementary Table **??**), indicating that there is no genomic inflation due to population stratification or cryptic relatedness. The significant SNPs are mapped to genes based on GRCh37/hg19 coordinates, and were used in searches for published GWAS associations based on GWAS catalog [37].

5.A.6.3 *Genetic correlations*

We perform cross-trait LD score regression (LDSC) using GWAS summary statistics [36] to relate ENABL AgeAccel to various health-related traits: anthropometric traits (e.g., BMI), adiposity (e.g., whole body/leg/arm/trunk fat mass), longevity (mother/father's longevity), lifestyle (e.g., smoking) and several diseases (e.g., heart attack, heart failure, angina, stoke and hypertension). GWAS summary statistics for health-related traits are downloaded from the Ben Neale Lab round 2 http://www.nealelab.is/uk-biobank. The significance threshold is corrected using the Bonferroni correction method, accounting for a total of 143 tests ($0.05/143 \approx 0.0003$). In addition, we assess the genetic correlations among various ENABL AgeAccels and their genetic correlations with several age accelerations based on diverse biomedical data sources, such as epigenetic age accelerations (e.g., GrimAgeAccel, DNAmPhenoAgeAccel, HannumAccel, and HorvathAccel), as well as brain MRI-based age acceleration. (i.e., BrainAgeAccel [136]). We utilized publicly available GWAS summary statistics of epigenetic age accelerations and

BrainAge from previous studies [136, 189]. It is noted that the GWAS summary statistics of GrimAgeAccel, DNAmPhenoAgeAccel, HannumAccel, and HorvathAccel [189] are conducted in smaller cohorts (n=34,710) than UKB (n=382,152) and using a meta-analysis approach. The significance threshold is corrected using the Bonferroni correction method for a total of 30 tests ($0.05/30 \approx 0.0017$). We use the LDSC (v1.0.1) implementation https://github.com/bulik/ldsc. As recommend by LDSC, we filtered to HapMap3 SNPs for each GWAS summary data, which could help align allele codes of our GWAS results with other GWAS results for the genetic correlation analysis.

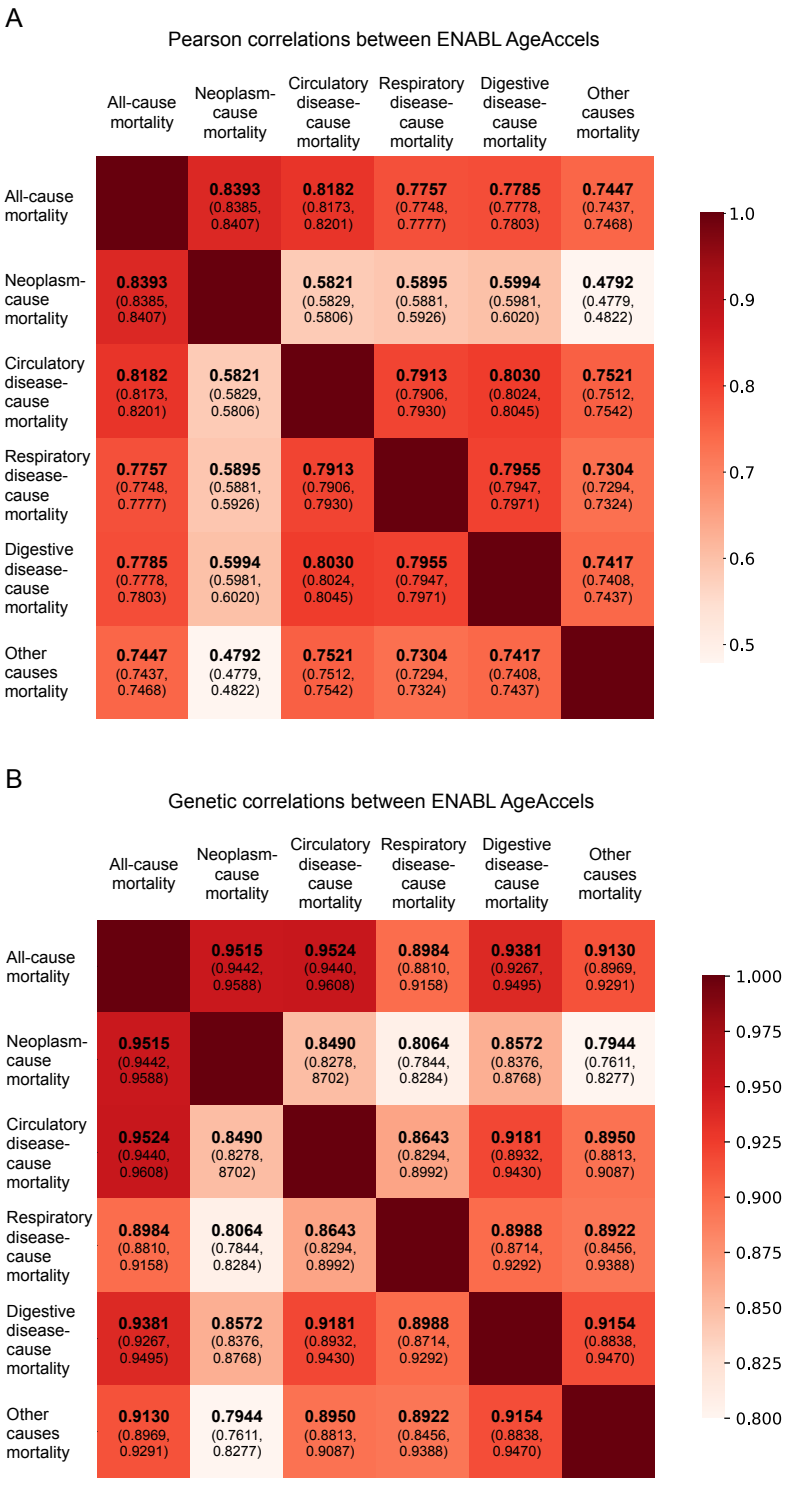
5.B SUPPLEMENTARY RESULTS

5.B.1 *ENABL Age can be applied to other age-related tasks.*

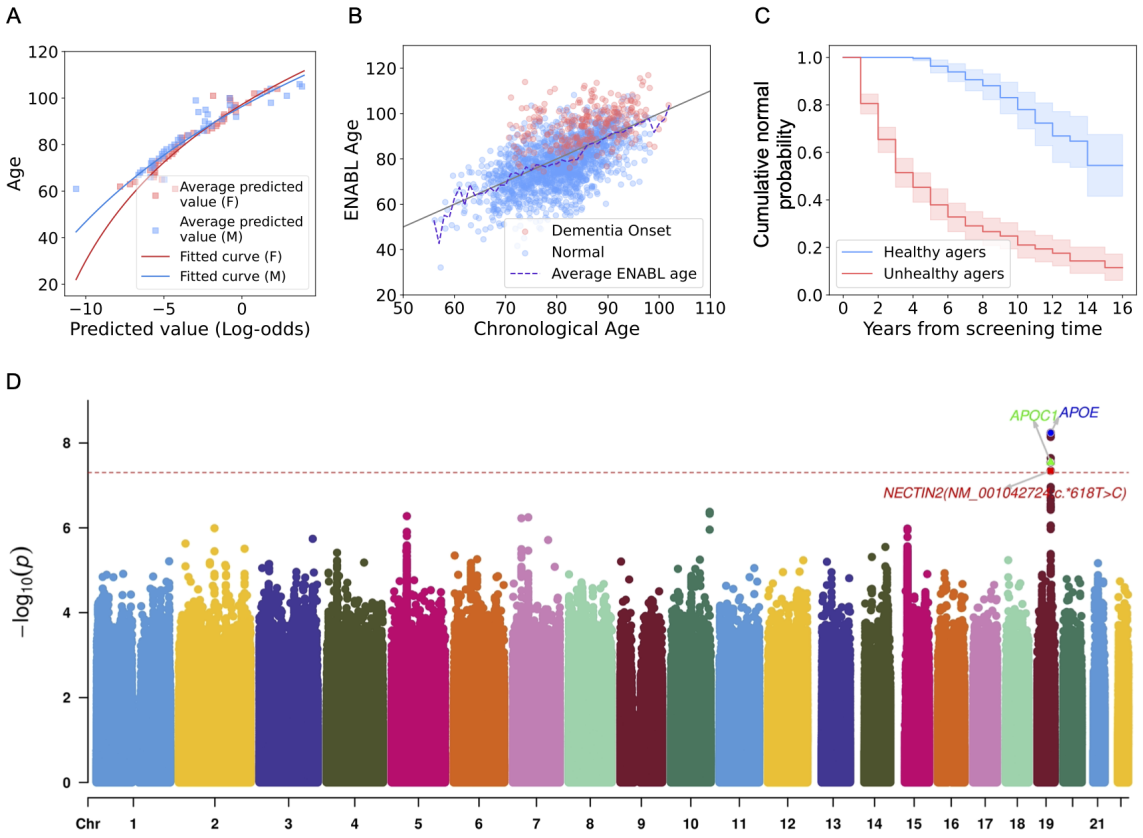
To validate that ENABL Age can be extended to age-related tasks other than mortality, we apply ENABL Age on dementia prediction using the ROSMAP dataset from two longitudinal aging cohort studies. We generate samples with no dementia history and have a sample size of 9,103 samples aged 56 to 106 derived from 1,597 individuals. We train a GBT classification model to predict imminent dementia onset (i.e., a diagnosis within the next three years) while learning the ENABL Age and then fit exponential curves between chronological ages and dementia predictions (Supplementary Figure **??**A).

In Supplementary Figure **??**B, we show that the ENABL Ages of the dementia-onset samples are generally higher than their corresponding chronological ages and the ENABL Ages of the normal samples. Supplementary Figure **??**C shows the Kaplan-Meier curve of dementia onset for the individuals, stratified into the healthy (lowest 25%) and unhealthy (highest 25%) agers according to the ENABL Age accelerations of the first visits. The unhealthy agers have much steeper declines in dementia-onset over the 16 years of follow-up, suggesting the dementia ENABL Age's effectiveness on identifying high dementia risk individuals. These results demonstrate that the dementia ENABL Age reflects individuals' dementia risk comparing with the samples of the same age and suggests that the dementia ENABL Age may capture the aging signal related to dementia.

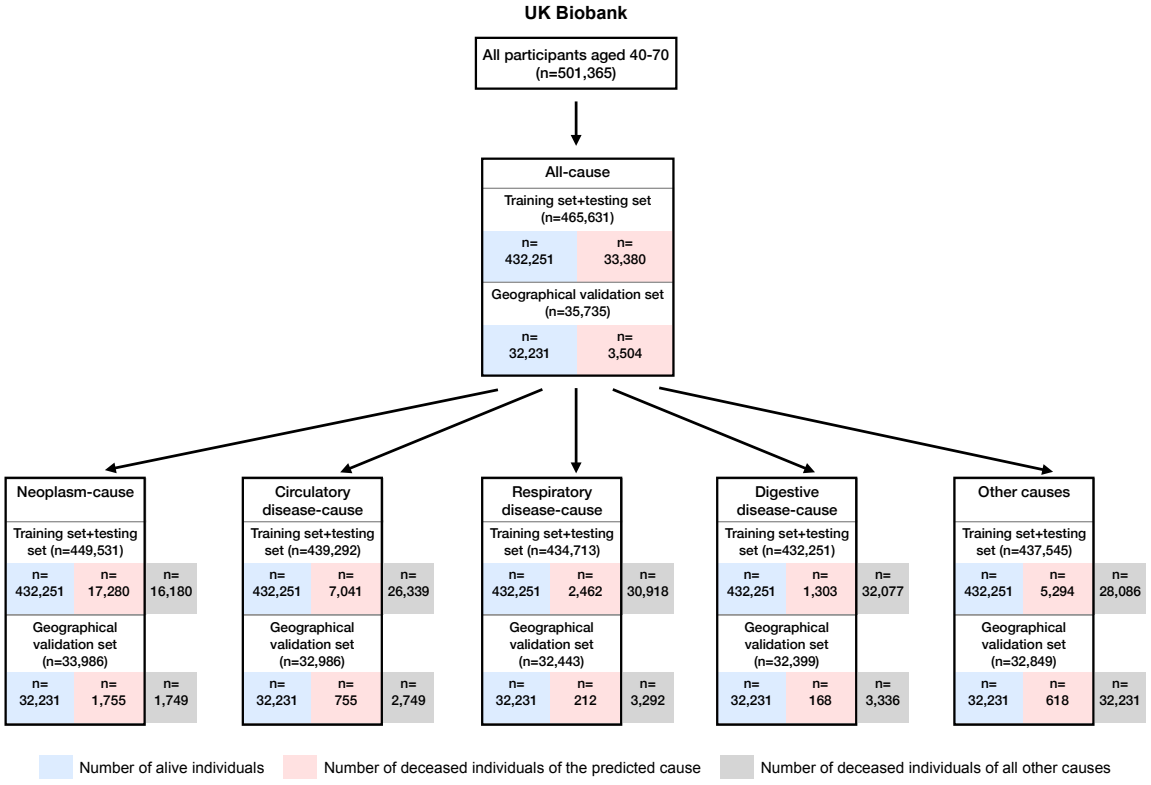
We also perform GWAS on dementia ENABL Age to validate that the dementia ENABL Age captures demential status. Supplementary Figure **??**D shows the Manhattan plot and the selected mapped genes of the lead SNPs. All three mapped genes, APOE, APOC1 and NECTIN2, are associated with Alzheimer's disease, indicating that the dementia ENABL Age captures the aging signal related to dementia/Alzheimer's disease, further indicating that ENABL Age can be successfully applied to age-related task beyond mortality.



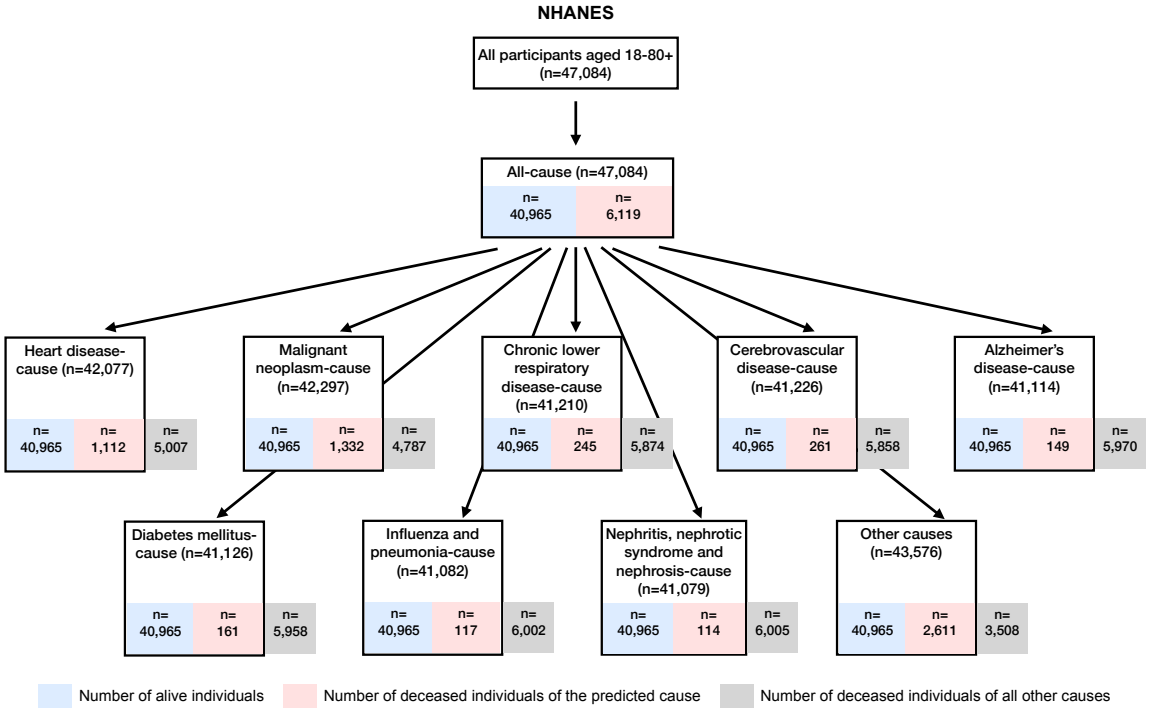
Supplementary Figure A.5.7: The Pearson correlations and genetic correlations between different ENABL AgeAccels, with the numbers in brackets representing the corresponding 95% confidence intervals.



Supplementary Figure A.5.1: **The results of ENABL Age framework on the ROSMAP dataset.** (A) The curves that transform GBTs' predicted value to ENABL Age predictions on the ROSMAP dataset. (B) The scatter plot of ENABL Age predictions versus chronological ages on the ROSMAP dataset. (C) Kaplan–Meier curves for persons in the highest 25% (unhealthy agers) versus the lowest 25% (healthy agers) of dementia onset ENABL age acceleration in the ROSMAP test set. (D) Manhattan plots for dementia onset ENABL Age.



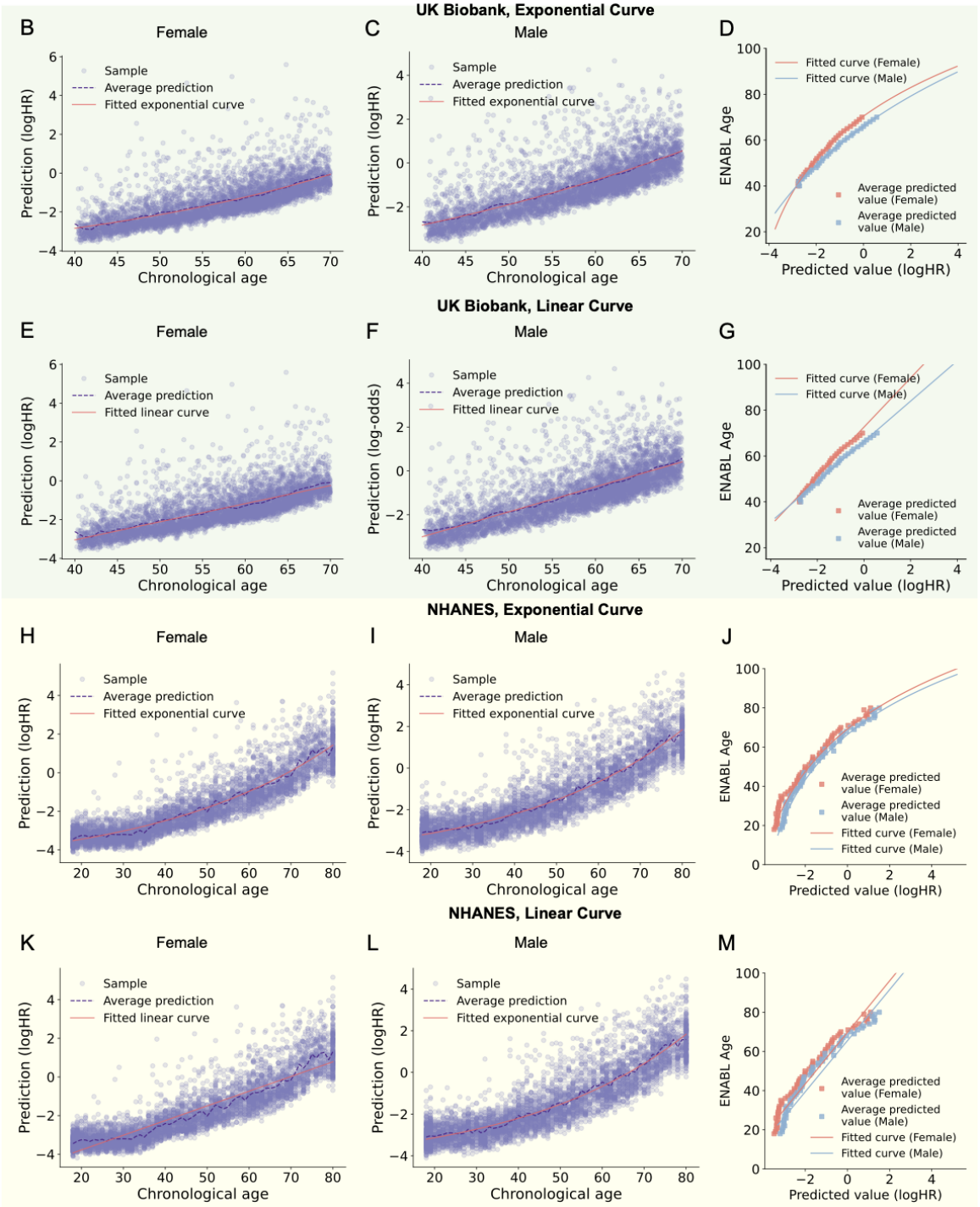
Supplementary Figure A.5.2: Sample size flow chart for the UK Biobank dataset. When predicting cause-specific mortality, individuals deceased due to all other causes are excluded.



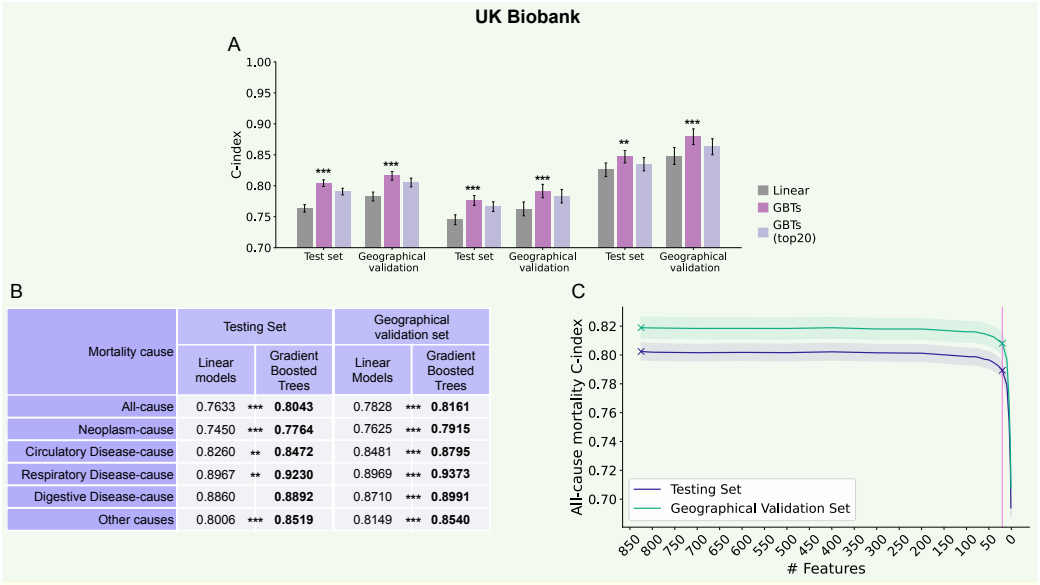
Supplementary Figure A.5.3: Sample size flow chart for the NHANES dataset. When predicting cause-specific mortality, the individuals deceased due to all other causes are excluded.

A Comparison of RMSEs for Exponential and Linear Curve Fitting

	UK Biobank		NHANES	
	Female	Male	Female	Male
Exponential curve	**0.7179**	**0.7993**	**0.6968**	**0.7601**
Linear curve	0.7209	0.8018	0.7568	0.8171



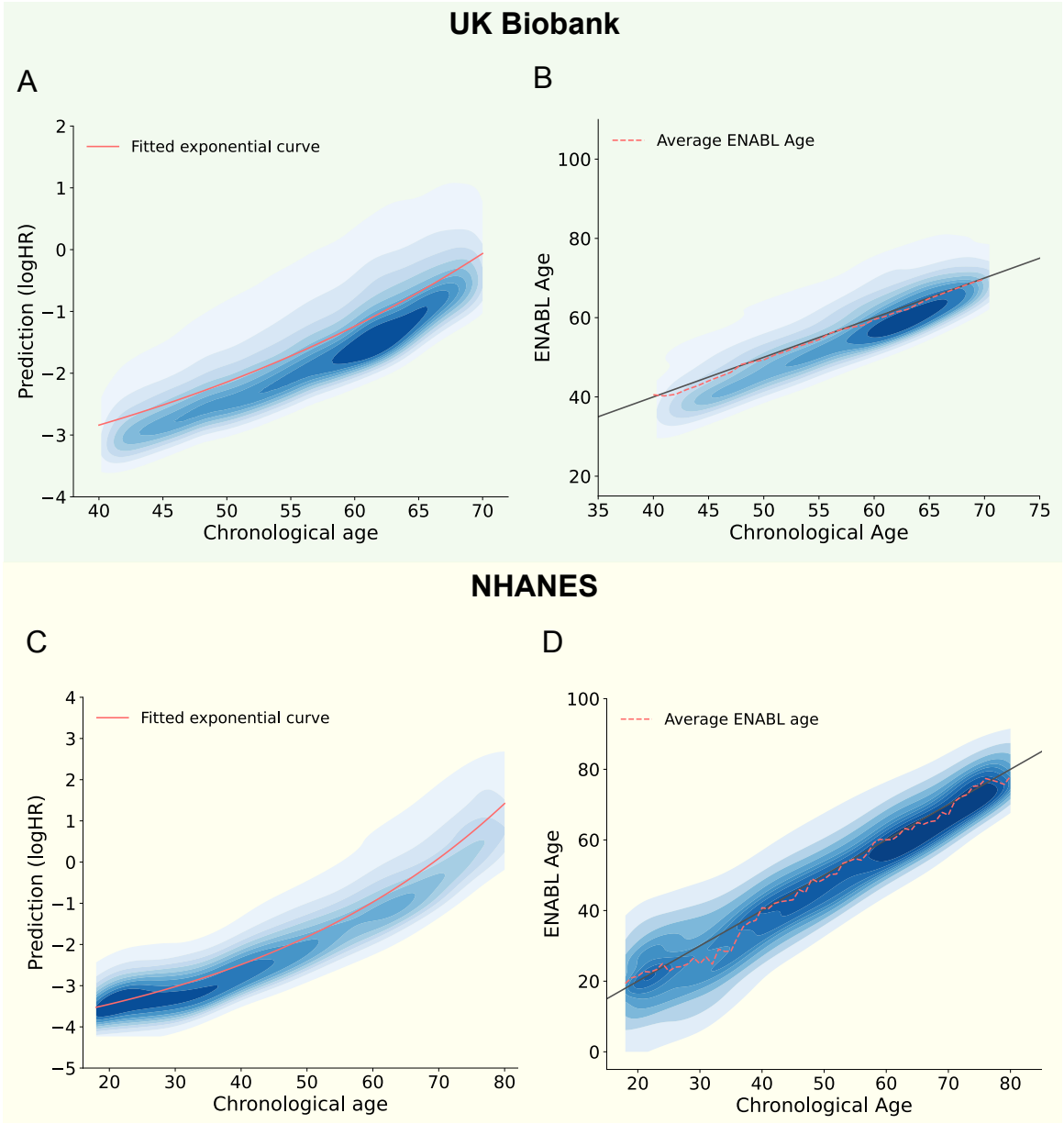
Supplementary Figure A.5.4: (A) Comparison of root mean squared errors (RMSEs) for exponential and linear curve fitting between chronological age and GBTs' predicted logHR, with lower values bolded. (B,C,E,F,H,I,K,L) Scatter plots of GBTs' predicted logHRs versus chronological ages and the fitted exponential/linear curve for female/male individuals in the UK Biobank/NHANES datasets. (D,G,J,M) Inverse exponential/-linear curves transforming GBTs' predicted logHR to ENABL Age in the UK Biobank/NHANES datasets.



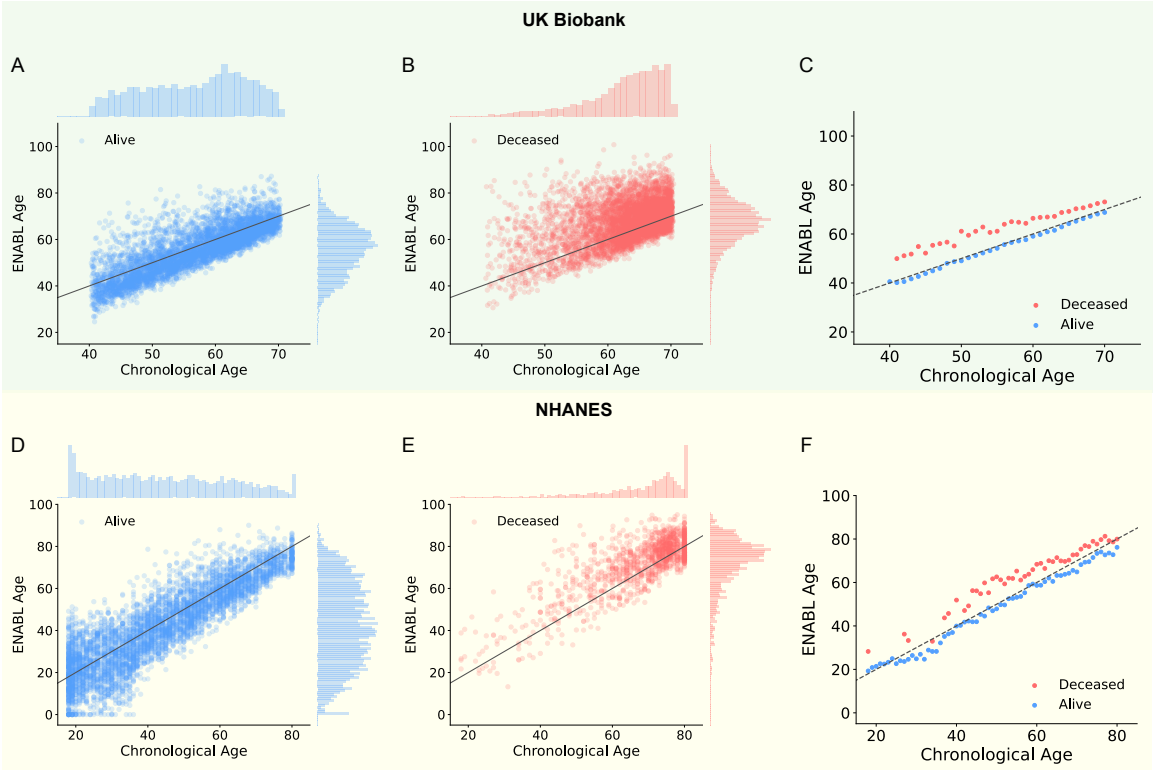
Mortality cause	Testing Set		Geographical validation set	
	Linear models	Gradient Boosted Trees	Linear Models	Gradient Boosted Trees
All-cause	0.7633 ***	**0.8043**	0.7828 ***	**0.8161**
Neoplasm-cause	0.7450 ***	**0.7764**	0.7625 ***	**0.7915**
Circulatory Disease-cause	0.8260 **	**0.8472**	0.8481 ***	**0.8795**
Respiratory Disease-cause	0.8967 **	**0.9230**	0.8969 ***	**0.9373**
Digestive Disease-cause	0.8860	**0.8892**	0.8710 ***	**0.8991**
Other causes	0.8006 ***	**0.8519**	0.8149 ***	**0.8540**

Mortality cause	Linear models	Gradient Boosted Trees	Mortality cause	Linear models	Gradient Boosted Trees
All-cause	0.8763 ***	**0.8834**	Alzheimer's Disease-cause	**0.9723**	0.9492
Diseases of Heart-cause	0.9316 ***	**0.9433**	Diabetes Mellitus-cause	0.9658	**0.9698**
Malignant Neoplasm-cause	0.8705	**0.8786**	Influenza and Pneumonia-cause	**0.9132**	0.8946
Chronic Lower Respiratory Disease-cause	0.9694	**0.9714**	Nephritis, Nephrotic Syndrome and Nephrosis-cause	0.9544	**0.9667**
Cerebrovascular Disease-cause	**0.8965**	0.8929	Other causes	0.8761 ***	**0.8912**

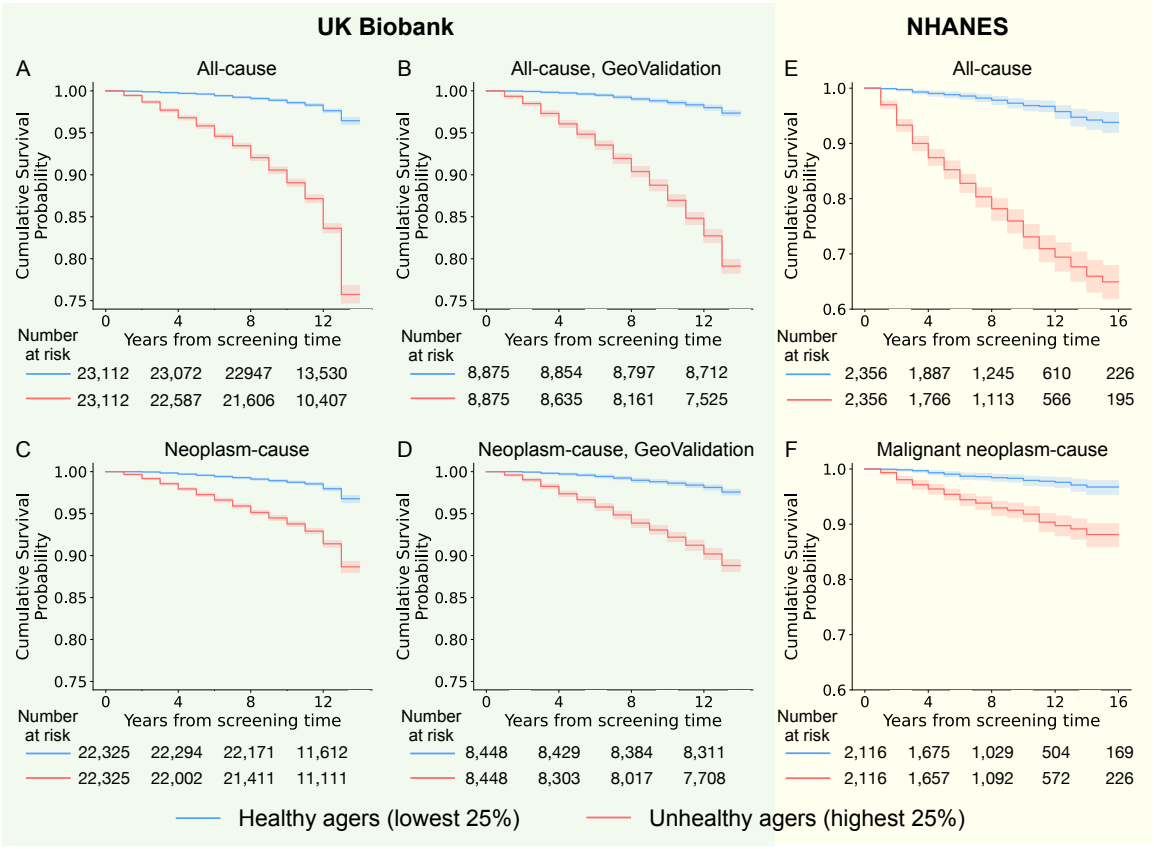
Supplementary Figure A.5.5: (A,D) The C-index of the all-cause mortality prediction and cause-specific mortality prediction on the UK Biobank test set, a left-out geographical validation set, and the NHANES test set using linear Cox regression models, GBTs Cox regression models, and GBTs Cox regression models with the top 20 important features selected by a recursive feature elimination method. Confidence intervals are computed using bootstrap re-sampling of the test set 1,000 times. P-values, derived from the bootstrap results, are used to assess the performance improvement of GBTs over linear models. (***) represents a p-value < 0.001, (**) represents a p-value < 0.01, and (*) represents a p-value < 0.05. (B,E) C-index comparison for all-cause and cause-specific mortality predictions on the UK Biobank and NHANES datasets using linear models and GBT models. More details are in Supplementary Methods **??**. (C,F) The C-index of the models using different feature sets after recursive feature elimination on UK Biobank and NHANES datasets. Lines are mean performance over 1000 random train/test splits, and shaded bands are 95 percent normal confidence intervals. The pink line is the vertical line of 20 features. "$\times$" indicates the C-index of the models using all features or the top 20 most important features. More details are in Supplementary Methods **??**.



Supplementary Figure A.5.6: (A,C) The contour plots of GBTs' predicted logHR versus chronological ages and the fitted exponential curve on the UK Biobank and NHANES datasets. (B,D) The contour plots of ENABL Ages versus chronological ages on the UK Biobank and NHANES datasets.



Supplementary Figure A.5.7: (A,B,D,E) The scatter plots and density plots of ENABL Age predictions versus chronological ages for the individuals that are alive or deceased at mortality data collection time in the UK Biobank and NHANES dataset. (C,F) Comparison of average ENABL Age between deceased and alive individuals across different age groups in the UK Biobank and NHANES datasets.

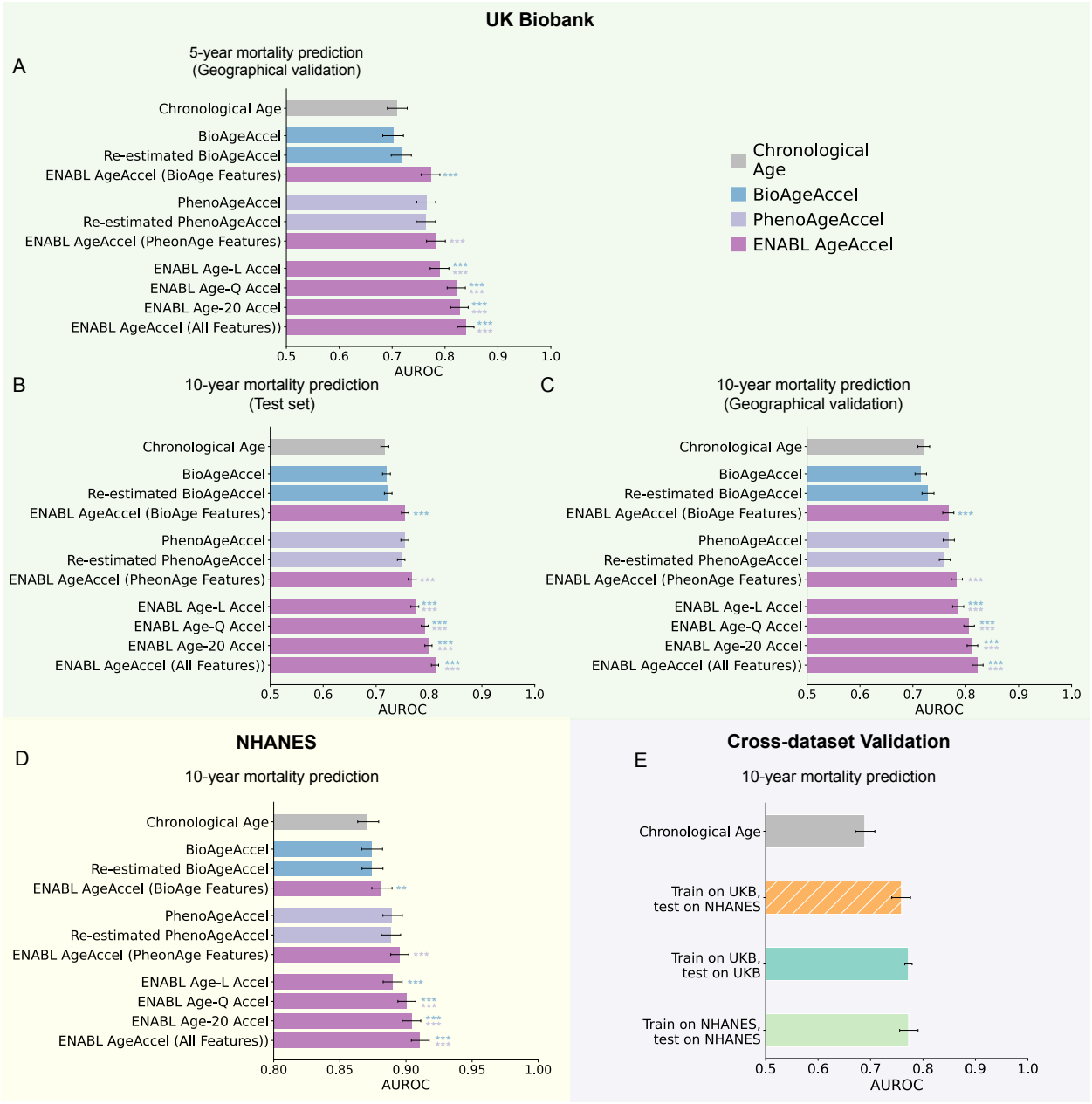


G

Average logHR

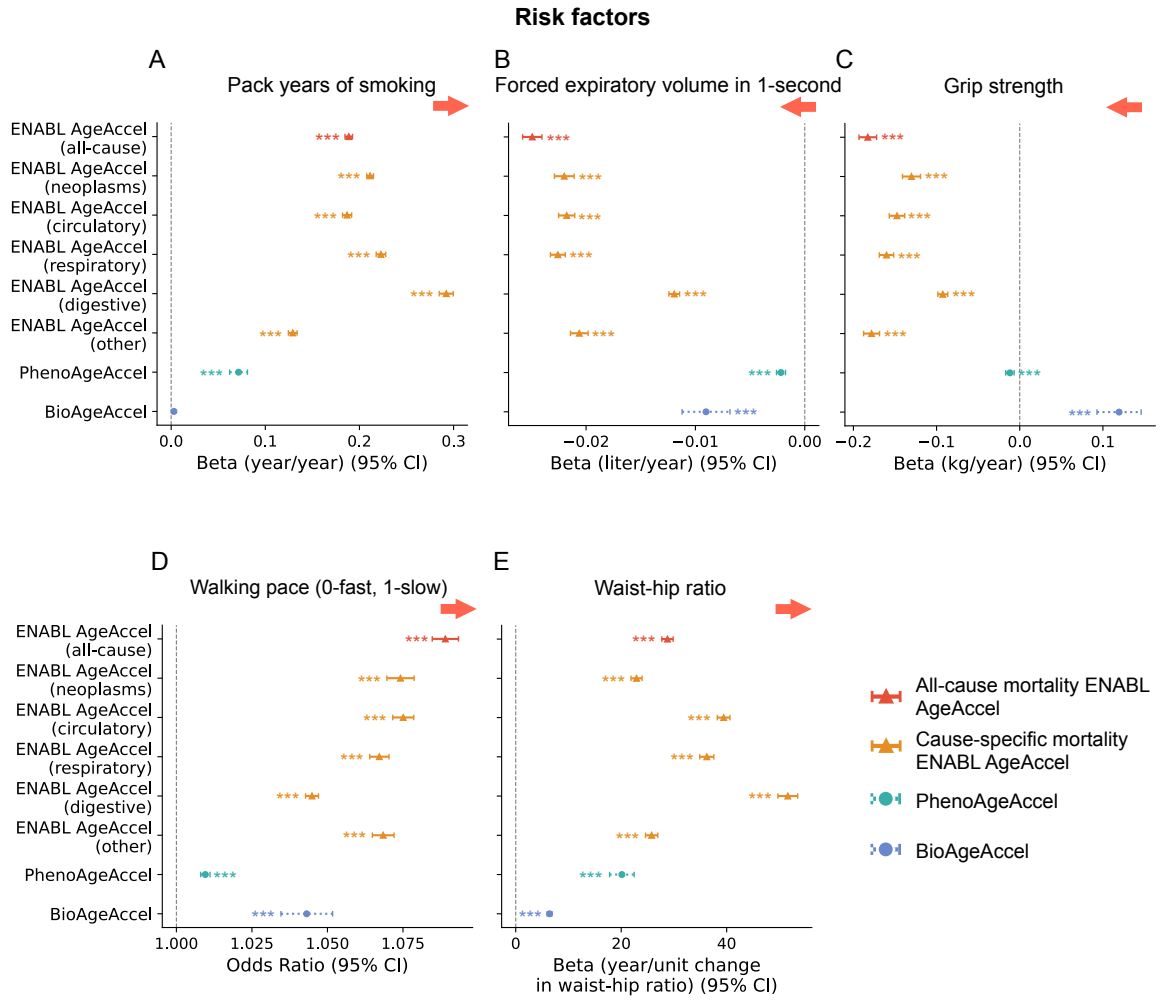
All-cause mortality			
	UK Biobank		NHANES
	Test set	GeoValidation	
Healthy agers	-2.0851	-2.1911	-2.7072
Unhealthy agers	-0.2950	-0.1787	-0.9008
Neoplasm-cause mortality			
	UK Biobank		NHANES
	Test set	GeoValidation	
Healthy agers	-1.5721	-1.5837	-2.0451
Unhealthy agers	-0.3336	-0.2634	-0.8423

Supplementary Figure A.5.8: **ENABL Age clocks successfully stratify unhealthy and healthy individuals.** (A-B) Kaplan–Meier curves for persons in the highest 25% (unhealthy agers) versus the lowest 25% (healthy agers) of all-cause mortality ENABL AgeAccel in the UK Biobank test set and geographical validation set. (C-D) Kaplan–Meier curves for persons in the highest 25% (unhealthy agers) versus the lowest 25% (healthy agers) of neoplasm-cause mortality ENABL AgeAccel in the UK Biobank test set and geographical validation set. (E-F) Kaplan–Meier curves for persons in the highest 25% (unhealthy agers) versus the lowest 25% (healthy agers) of all-cause mortality ENABL AgeAccel and malignant neoplasm-cause mortality ENABL AgeAccel in the NHANES test set. (G) The average logHR of the healthy and unhealthy age groups.

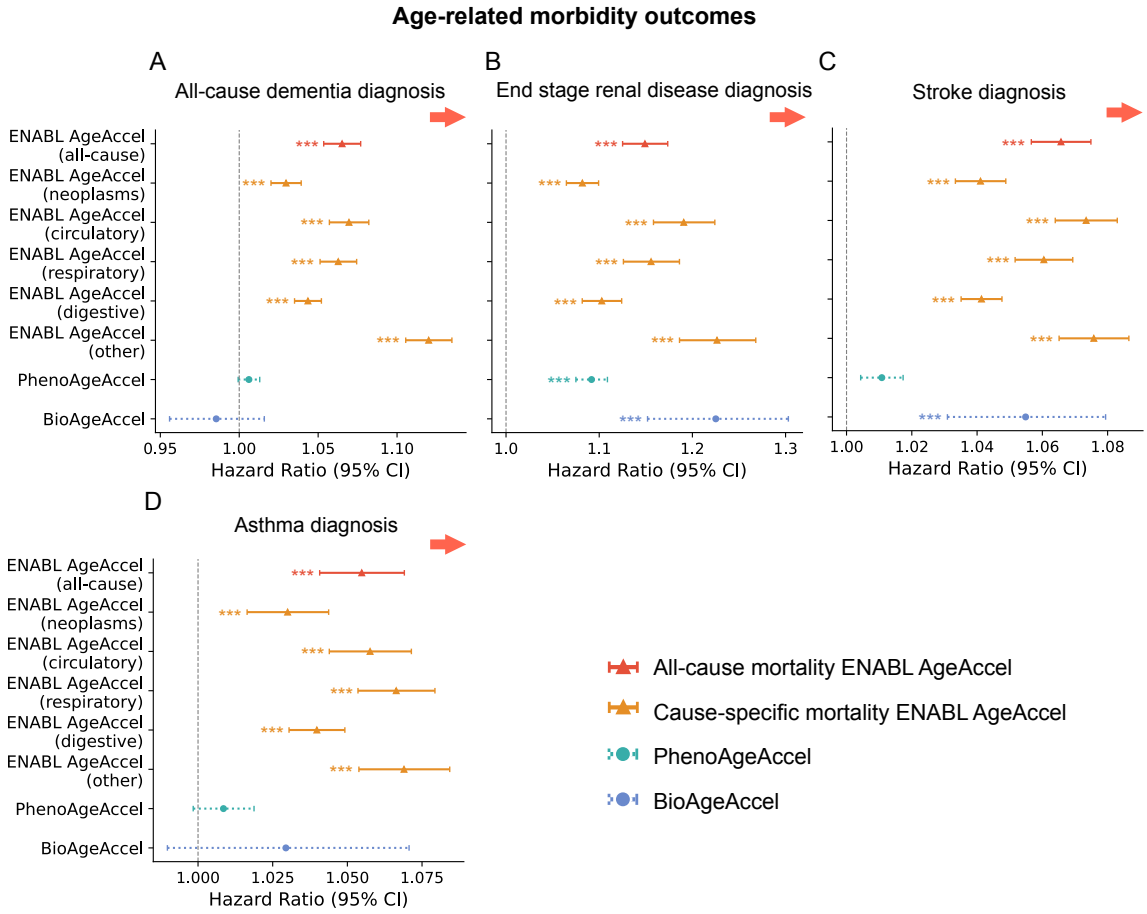


Supplementary Figure A.5.9

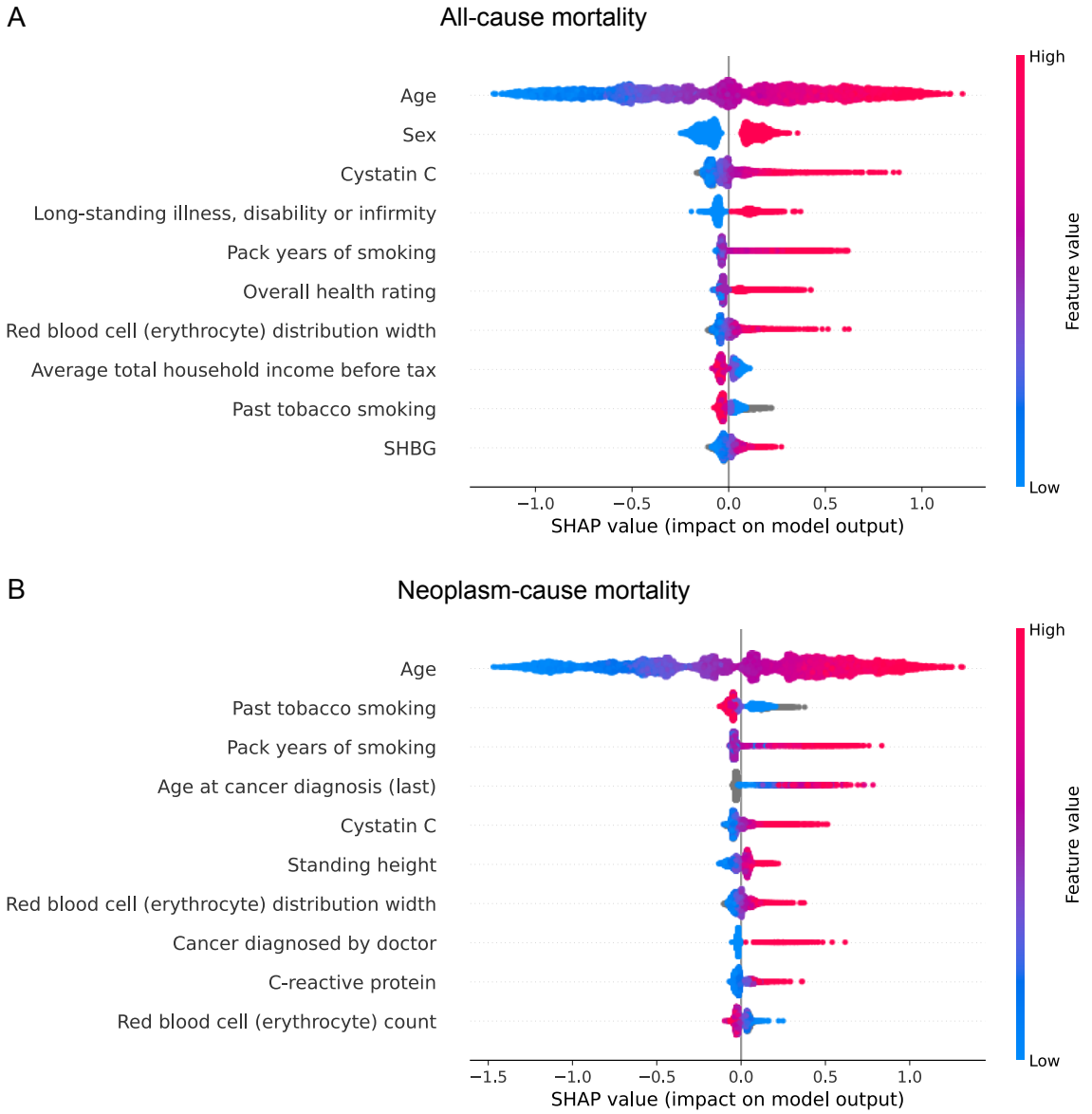
Supplementary Figure A.5.9: **ENABL AgeAccels have high mortality prediction power.** (A-D) The AUROCs of the 5-and 10- year mortality prediction GBT models of ENABL AgeAccels, PhenoAgeAccel, and BioAgeAccel adjusted by chronological age and sex trained and tested on the UK Biobank and NHANES datasets. "Chronological Age" refers to the GBT models trained by chronological age and sex. "BioAgeAccel" and "PhenoAgeAccel" are calculated using the formulae in their original papers. "Re-estimated BioAgeAccel" and "Re-estimated PhenoAgeAccel" are calculated using the re-estimated weights on the UK Biobank and NHANES datasets. "ENABL AgeAccel (BioAge Features)" and "ENABL AgeAccel (PhenoAge Features)" are calculated by the ENABL Age framework using only the features included in BioAge and PhenoAge, respectively. "ENABL Age-L Accel" refers to the ENABL Age clock that uses laboratory features in the four most popular blood panels (Supplementary Table **??**) – CBC, CMP, LP and WBC. "ENABL Age-Q Accel" refers to the ENABL Age clock that uses the top 20 most important questionnaire features selected by recursive feature elimination (Supplementary Table **??**). "ENABL Age-20 Accel" refers to the ENABL Age clock that uses the top 20 most important features out of all possible features selected by recursive feature elimination (Supplementary **??**; Supplementary Figures **??**B,D; Supplementary Table **??**). Colored squares indicate the feature types/laboratory panels (CBC – Complete Blood Count, CMP – Comprehensive Metabolic Panel, LP – Lipoprotein (a), and WBC – White Blood Cell Count) used to calculate biological ages. Confidence intervals are computed using bootstrap resampling of the test set 1,000 times. P-values, derived from the bootstrap results, are used to assess the performance improvement of ENABL AgeAccels over "Re-estimated BioAgeAccel" (blue asterisks) and "Re-estimated PhenoAgeAccel" (purple asterisks). (***) represents a p-value < 0.001, (**) represents a p-value < 0.01, and (*) represents a p-value < 0.05. (E) The cross-dataset validation of the 10-year mortality prediction model using ENABL Age-L. "Chronological Age" refers to the mortality prediction model using chronological age and sex trained and tested on NHANES samples aged 40-70. "Train on UKB, test on NHANES" refers to cross-dataset validation of the ENABL Age clock. We train ENABL Age-L on the UKB samples using the features in the four popular blood panels that overlap in the UKB and NHANES datasets and use it to evaluate the ENABL Age-L for NHANES samples. Then, we train and test the mortality prediction models using chronological age, sex, and ENABL Age-L acceleration (trained on UKB samples) on NHANES samples aged 40-70. "Train on UKB, test on UKB" and "Train on NHANES, test on NHANES" refer to the mortality prediction model using chronological age, sex, and ENABL Age-L acceleration trained and evaluated on UKB samples or NHANES samples aged 40-70, respectively.



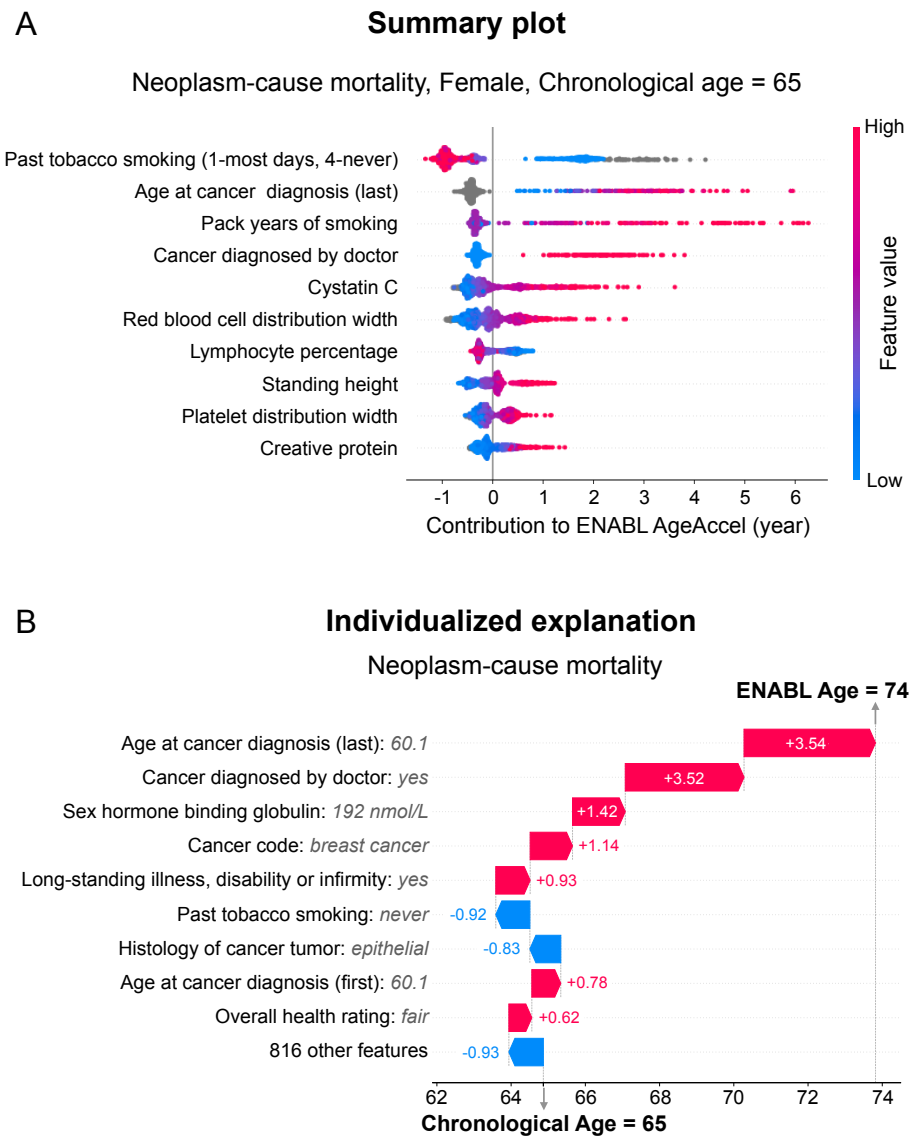
Supplementary Figure A.5.10: **ENABL AgeAccels are strongly associated with diverse risk factors.** (A-E) Associations of age accelerations (ENABL AgeAccel, PhenoAgeAccel, BioAgeAccel) with risk factors (i.e., pack years of smoking, forced expiratory volume in 1-second, grip strength, walking pace, and waist-hip ratio). The risk factors are regressed separately on each of the biological age accelerations, adjusting for chronological age and sex using ordinary least squares regression (pack years of smoking, forced expiratory volume in 1-second, grip strength, and waist-hip ratio), and logistic regression (walking pace), as appropriate. Height is included as a covariate when fitting our regression models for grip strength and walking pace. The biological age accelerations are in units of years. (***) represents a p-value $<$ 0.00001. (**) represents a p-value $<$ 0.00010. (*) represents a p-value $<$ 0.00052. The significance threshold is adjusted using the Bonferroni correction method, accounting for a total of 96 tests (8 biological ages and 12 traits). More details are in Supplementary Methods **??**.



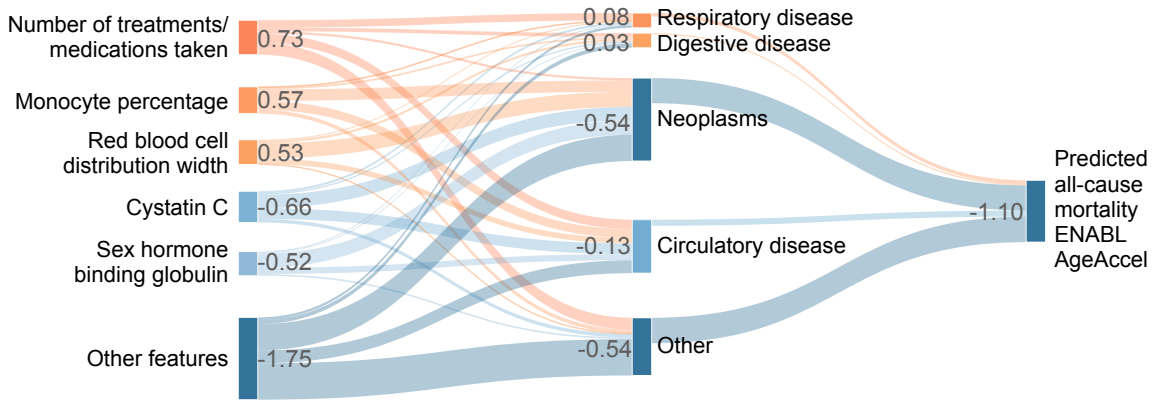
Supplementary Figure A.5.11: **ENABL AgeAccels are strongly associated with diverse age-related morbidity outcomes.** (A-D) Associations of age accelerations (ENABL AgeAccel, PhenoAgeAccel, BioAgeAccel) with age-related morbidity outcomes (i.e., all-cause dementia, end-stage renal disease, stroke, and asthma). The outcomes are regressed separately on each of the biological age accelerations, adjusting for chronological age and sex using Cox regression. The biological age accelerations are in units of years. (***) represents a p-value < 0.00001. (**) represents a p-value < 0.00010. (*) represents a p-value < 0.00052. The significance threshold is adjusted using the Bonferroni correction method, accounting for a total of 96 tests (8 biological ages and 12 traits). More details are in Supplementary Methods **??**.



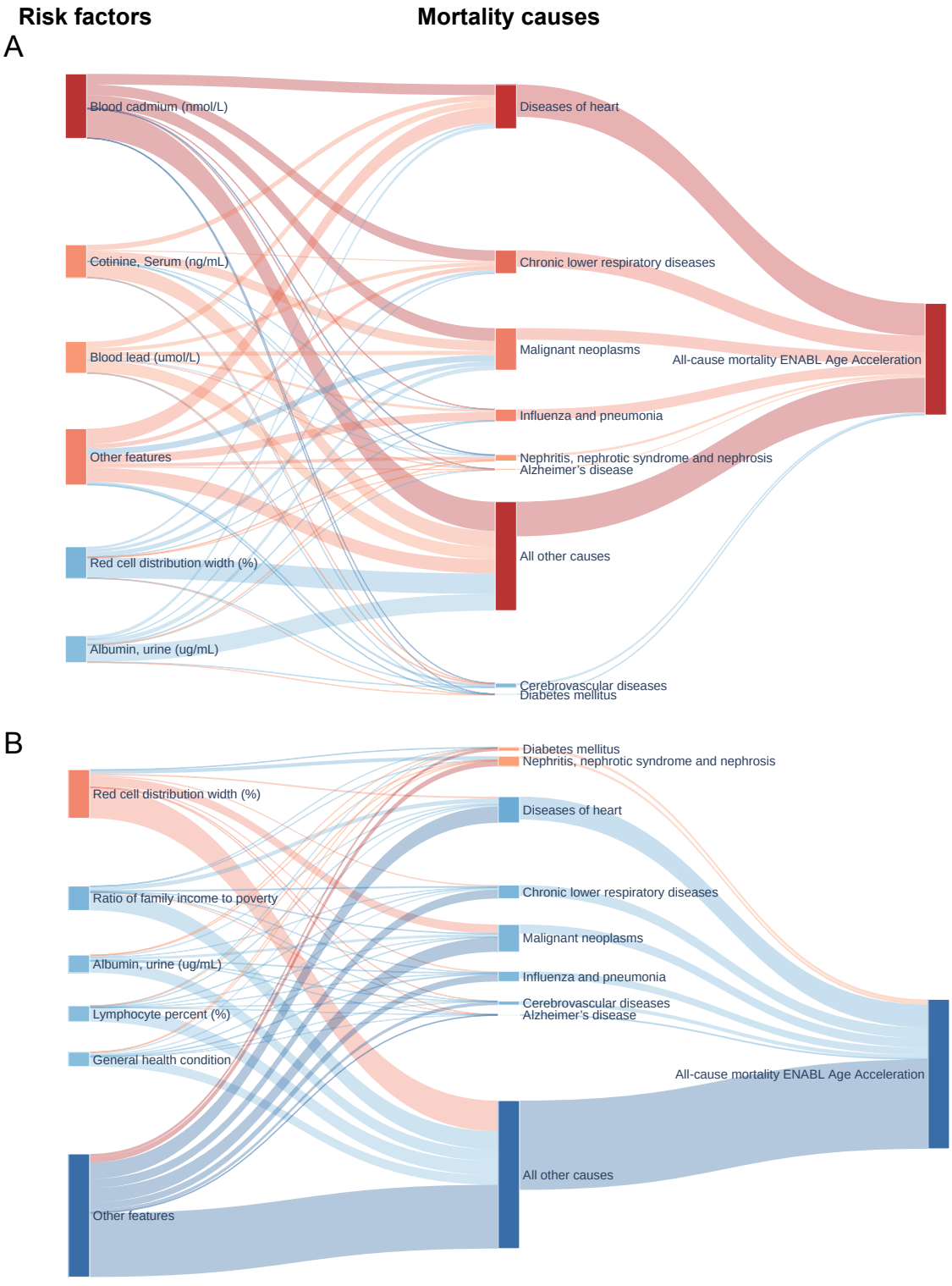
Supplementary Figure A.5.12: (A-B) SHAP summary plot for all-cause mortality and neoplasm-cause mortality prediction models using UKB datasets. SHAP values are consistent and accurate calculations of each feature's contribution to the model's prediction. The plot shows the most impactful features on prediction (ranked from most to least important) and the distribution of the impacts of each feature on model output, which includes a set of plots where each dot corresponds to an individual. The colors represent feature values for numeric features: red for larger values and blue for smaller ones. The thickness of the line comprised of individual dots is determined by the number of examples at a given value. A negative SHAP value (extending to the left) indicates a reduced mortality prediction value, while a positive one (extending to the right) indicates an increased mortality prediction value.



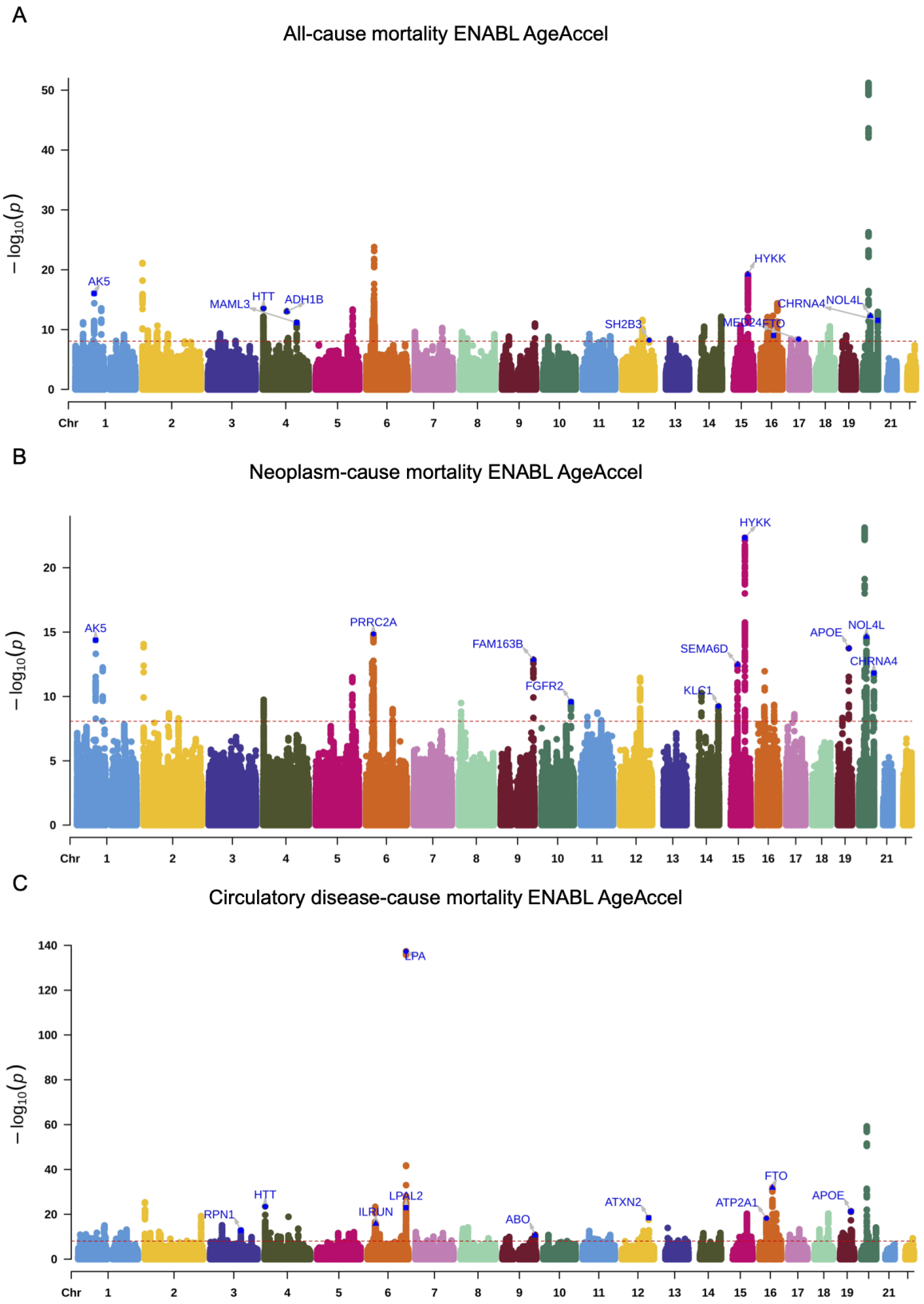
Supplementary Figure A.5.13: (A) SHAP summary plot for ENABL Age clocks trained on neoplasm-cause mortality using UKB datasets for female samples aged 65 years old. The plot shows the most impactful features on ENABL Age (ranked from most to least important). It also shows the distribution of the impacts of each feature on ENABL AgeAccel, which includes a set of plots where each dot corresponds to an individual. The colors represent feature values for numeric features: red for larger values, blue for smaller, and grey for not applicable (e.g., currently smoking on most or all days for past tobacco smoking). The thickness of the line (which actually comprises many individual dots) is determined by the number of examples at a given value, where dots spread out vertically if there are many examples. A negative rescaled SHAP value (extending to the left) indicates reduced ENABL AgeAccel, while a positive one (extending to the right) indicates increased ENABL AgeAccel. The SHAP values are calculated using explicands and baselines that have the same age and sex (i.e., females aged 65 years old). (B) The individualized explanation of neoplasm-cause mortality ENABL Age for a single female aged 65 years old. The output value (the gray dashed line with the number at the top of the plot) shows ENABL Age for that individual. The base value (the gray dashed line with the number at the bottom of the plot) approximates chronological age (i.e., 65). The features in red increase ENABL AgeAccel, and those in blue decrease it.



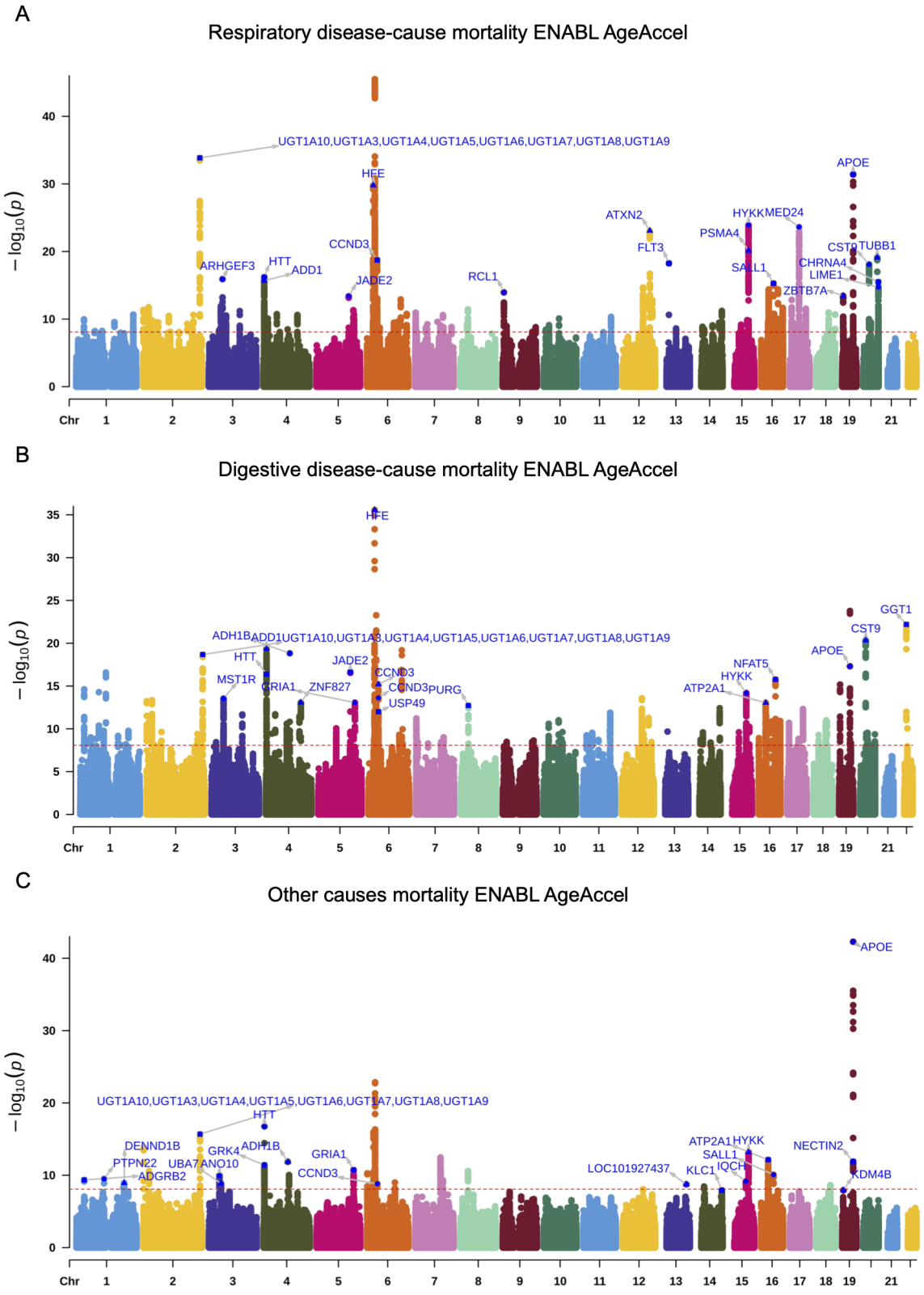
Supplementary Figure A.5.14: The two-layer all-cause mortality ENABL AgeAccel explanations depict the contributions of features to different mortality causes as well as mortality causes to all-cause mortality ENABL Age for a healthy individual from the UK Biobank dataset. Flows in red represent positive rescaled SHAP values, which increase the ENABL AgeAccel, while flows in blue indicate negative rescaled SHAP values, which decrease the ENABL AgeAccel. The width of each flow and the grey numbers correspond to the contribution of risk factors to mortality causes as well as mortality causes to all-cause mortality ENABL Age in units of years. The "Predicted all-cause mortality ENABL AgeAccel" refers to the predicted all-cause mortality ENABL Age using the cause-specific ENABL Ages minus chronological age.



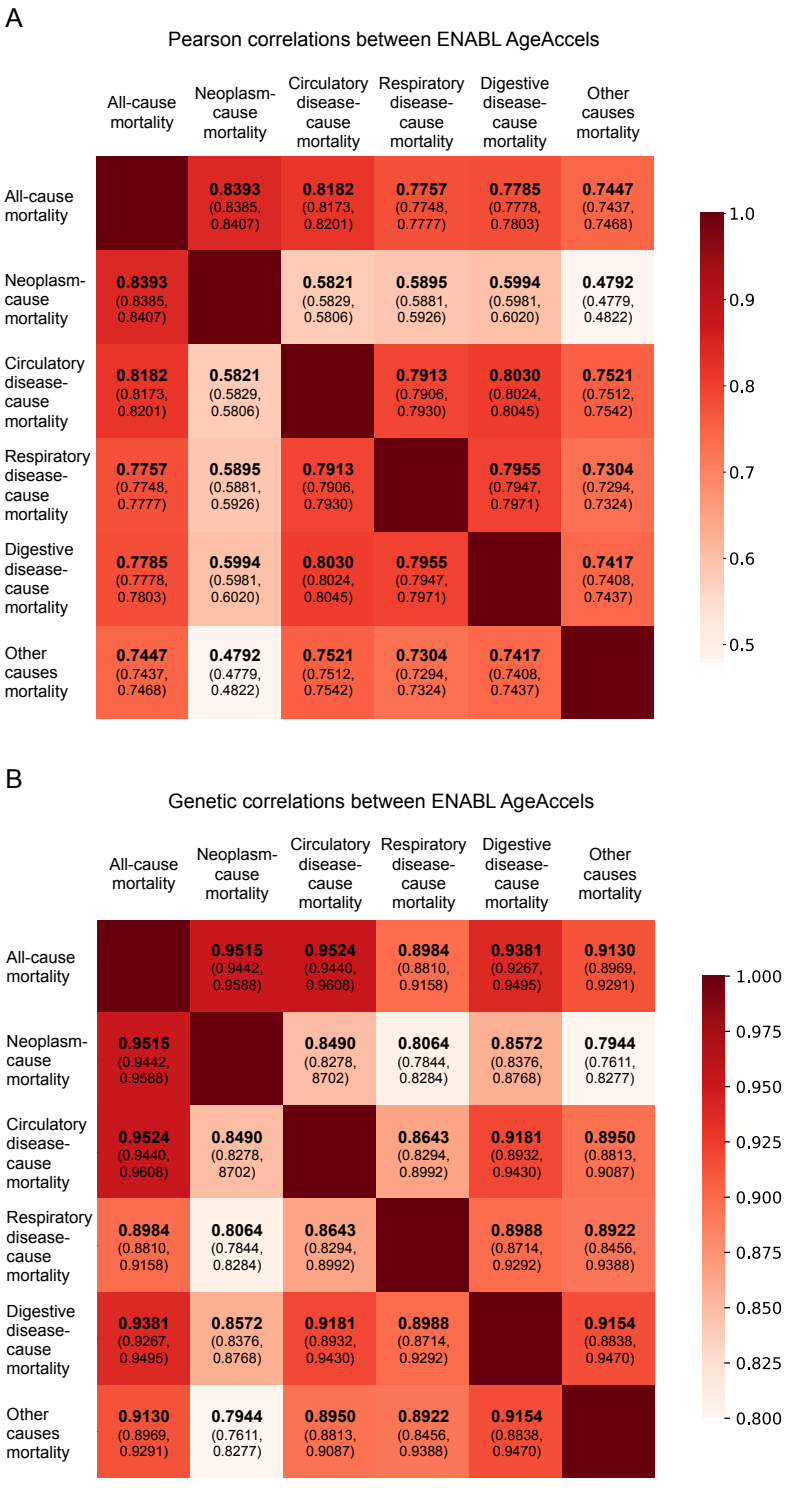
Supplementary Figure A.5.15: (A-B) The two-layer all-cause mortality ENABL AgeAccel explanations, i.e., the features of different mortality causes as well as mortality causes to all-cause mortality ENABL Age for two individuals from the NHANES dataset (Supplementary Methods **??**). The lines in red indicate positive rescaled SHAP values that increase the ENABL AgeAccel; those in blue indicate negative rescaled SHAP values that decrease the ENABL AgeAccel. The width of each flow and the grey numbers indicate the contribution of risk factors to mortality causes as well as mortality causes to all-cause mortality ENABL Age in units of years.



Supplementary Figure A.5.16: Manhattan plots for all-cause mortality ENABL AgeAccel, neoplasm-cause mortality ENABL AgeAccel, and circulatory disease-cause mortality ENABL AgeAccel (colors to separate adjacent chromosomes without other indications). SNP p-values smaller than 8.3×10^{-9} are deemed to be statistically significant.



Supplementary Figure A.5.17: Manhattan plots for respiratory disease-cause mortality ENABL AgeAccel, digestive-cause mortality ENABL AgeAccel, and other causes mortality ENABL AgeAccel (colors to separate adjacent chromosomes without other indications). SNP p-values smaller than 8.3×10^{-9} are deemed to be statistically significant.



Supplementary Figure A.5.18: The Pearson correlations and genetic correlations between different ENABL AgeAccels, with the numbers in brackets representing the corresponding 95% confidence intervals.

UK Biobank							
Training set + testing set							
Mortality cause		All (n=465,631)	Neoplasms (n=449,531)	Circulatory disease (n=439,292)	Respiratory disease (n=434,713)	Digestive disease (n=433,554)	Other (n=437,5485)
Number of deaths		33,380 (7.17%)	17,280 (3.84%)	7,041 (1.60%)	2,462 (0.57%)	1,303 (0.30%)	5,294 (1.21%)
Follow-up (years)		12.00 (11.31-12.65)	12.03 (11.36-12.67)	12.05 (11.39-12.68)	12.06 (11.41-12.68)	12.06 (11.41-12.68)	12.06 (11.40-12.68)
Age (years)		57.12 (8.10)	56.92 (8.08)	56.81 (8.08)	56.76 (8.08)	56.73 (8.08)	56.79 (8.09)
Sex	Male	212,568 (45.65%)	202,124 (44.96%)	197,767 (45.02%)	194,390 (44.72%)	193,646 (44.66%)	195,925 (44.78%)
	Female	253,063 (54.35%)	247,407 (55.04%)	241,525 (54.98%)	240,323 (55.28%)	239,908 (55.34%)	241,620 (55.22%)
Ethnicity	White	436,600 (93.77%)	421,334 (93.73%)	411,362 (93.64%)	407,062 (93.64%)	405,947 (93.63%)	409,691 (93.63%)
	Asian	111,12 (2.39%)	10,775 (2.40%)	10,747 (2.45%)	10,632 (2.45%)	10,606 (2.45%)	10,700 (2.45%)
	Black	7.987 (1.72%)	7,772 (1.73%)	7,695 (1.75%)	7,620 (1.75%)	7,617 (1.76%)	7,699 (1.76%)
	Mixed	2,861 (0.61%)	2,802 (0.62%)	2,758 (0.63%)	2,736 (0.63%)	2,734 (0.63%)	2,747 (0.63%)
	Other	4,419 (0.95%)	4,322 (0.96%)	4,251 (0.97%)	4,223 (0.97%)	4,221 (0.97%)	4,250 (0.97%)
	Unknown	2,652 (0.57%)	2,526 (0.56%)	2,479 (0.56%)	2,440 (0.56%)	2,429 (0.56%)	2,458 (0.56%)
Geographical validation set							
Mortality cause		All (n=35,735)	Neoplasms (n=33,986)	Circulatory disease (n=32,986)	Respiratory disease (n=32,443)	Digestive disease (n=32,399)	Other (n=32,849)
Number of deaths		3,504 (9.81%)	1,755 (5.16%)	755 (2.29%)	212 (0.65%)	168 (0.52%)	618 (1.88%)
Follow-up (years)		13.14 (12.96-13.34)	13.15 (12.98-13.34)	13.16 (12.99-13.35)	13.16 (12.99-13.35)	13.16 (12.99-13.35)	13.16 (12.99-13.35)
Age (years)		56.90 (8.07)	56.59 (8.03)	56.46 (8.04)	56.35 (8.02)	56.32 (8.02)	56.42 (8.04)
Sex	Male	15,830 (44.30%)	14,733 (43.35%)	14,305 (43.37%)	13,928 (42.93%)	13,900 (42.90%)	14,151 (43.08%)
	Female	19,905 (55.70%)	19,253 (56.65%)	18,681 (56.63%)	18,515 (57.07%)	18,499 (57.10%)	18,698 (56.92%)
Ethnicity	White	35,024 (98.01%)	33,309 (98.01%)	32,313 (97.96%)	31,780 (97.96%)	31,737 (97.96%)	32,181 (97.97%)
	Asian	322 (0.90%)	305 (0.90%)	303 (0.92%)	300 (0.92%)	297 (0.92%)	301 (0.92%)
	Black	57 (0.16%)	57 (0.17%)	55 (0.17%)	55 (0.17%)	55 (0.17%)	55 (0.17%)
	Mixed	90 (0.25%)	87 (0.26%)	85 (0.26%)	85 (0.26%)	85 (0.26%)	84 (0.26%)
	Other	129 (0.36%)	125 (0.37%)	125 (0.38%)	122 (0.38%)	122 (0.38%)	123 (0.37%)
	Unknown	113 (0.32%)	103 (0.30%)	105 (0.32%)	101 (0.31%)	103 (0.32%)	105 (0.32%)

Supplementary Table A.5.1: Population characteristics for the all-cause mortality and cause-specific mortality datasets of the UK Biobank study cohort. Data are mean (SD), median (IQR), or n/N (%). The sample size flow chart is shown in Supplementary Figure **??**.

NHANES						
Mortality cause		All (n=47,084)	Heart disease (n=42,077)	Malignant neoplasms (n=42,297)	Chronic lower respiratory disease (n=41,210)	Cerebrovascular disease (n=41,226)
Number of deaths		6,119 (13.00%)	1,112 (2.64%)	1,332 (3.15%)	245 (0.59%)	261 (0.63%)
Follow-up (years)		7.67 (4.25-11.83)	8.08 (4.42-12.17)	8.08 (4.42-12.17)	8.08 (4.58-12.33)	8.08 (4.50-12.33)
Age (years)		47.02 (19.07)	44.65 (18.24)	44.66 (18.19)	44.15 (18.01)	44.15 (18.02)
Sex	Male	22,677 (48.16%)	19,983 (47.49%)	20,108 (47.54%)	19,451 (47.20%)	19,450 (47.18%)
	Female	24,407 (51.84%)	22,904 (52.51%)	22,189 (52.46%)	21,759 (52.80%)	21,776 (52.82%)
Ethnicity	Non-Hispanic White	21,359 (45.36%)	18,413 (43.76%)	18,481 (43.69%)	17,925 (43.50%)	17,904 (43.43%)
	Non-Hispanic Black	10,021 (21.28%)	9,093 (21.61%)	9,173 (21.69%)	8,908 (21.62%)	8,917 (21.63%)
	Mexican American	8,869 (18.84%)	8,092 (19.23%)	8,133 (19.23%)	7,953 (19.30%)	7,982 (19.36%)
	Other Hispanic	3,443 (7.31%)	3,249 (7.72%)	3,264 (7.72%)	3,211 (7.79%)	3,212 (7.79%)
	Other	3,392 (7.20%)	3,230 (7.68%)	3,246 (7.67%)	3,213 (7.80%)	3,211 (7.79%)
Mortality cause		Alzheimer's disease (n=41,114)	Diabetes mellitus (n=41,126)	Influenza and pneumonia (n=41,082)	Nephritis, nephrotic syndrome and nephrosis (n=41,079)	Other (n=43,576)
Number of deaths		149 (0.36%)	161 (0.39%)	117 (0.28%)	114 (0.28%)	2,611 (5.99%)
Follow-up (years)		8.08 (4.58-12.33)	8.08 (4.58-12.33)	8.08 (4.58-12.33)	8.08 (4.58-12.33)	7.92 (4.42-12.08)
Age (years)		44.09 (18.00)	44.08 (17.98)	44.06 (17.98)	44.07 (17.98)	45.36 (18.58)
Sex	Male	19,379 (47.13%)	19,387 (47.14%)	19,377 (47.16%)	19,382 (47.18%)	20,671 (47.44%)
	Female	21,735 (52.87%)	21,739 (52.86%)	21,705 (52.83%)	21,697 (52.82%)	22,905 (52.56%)
Ethnicity	Non-Hispanic White	17,865 (43.45%)	17,806 (43.30%)	17,817 (43.37%)	17,811 (43.36%)	19,308 (44.31%)
	Non-Hispanic Black	8,894 (21.63%)	8,919 (21.69%)	8,897 (21.66%)	8,909 (21.69%)	9,337 (21.43%)
	Mexican American	7,942 (19.32%)	7,978 (19.40%)	7,950 (19.35%)	7,945 (19.34%)	8,329 (19.11%)
	Other Hispanic	3,206 (7.80%)	3,211 (7.81%)	3,211 (7.82%)	3,208 (7.81%)	3,311 (7.60%)
	Other	3,207 (7.80%)	3,212 (7.81%)	3,207 (7.81%)	3,206 (7.80%)	3,291 (7.55%)

Supplementary Table A.5.2: Population characteristics for the all-cause mortality and cause-specific mortality datasets of the NHANES study cohort. Data are mean (SD), median (IQR), or n/N (%). The sample size flow chart is shown in Supplementary Figure **??**.

		Original Weights			Re-estimated Weights on UKB			Re-estimated Weights on NHANES		
	Units	S	K	Q	S	K	Q	S	K	Q
Albumin	g/dL	0.3372	-0.0058	4.4377	1.6472	-0.0023	3.9805	0.3451	-0.0029	4.3950
Alkaline phosphatase	u/L	28.9126	0.4822	58.8800	32.8179	0.4851	50.2275	25.5826	0.1028	66.6535
Creatinine (serum)	mg/dL	0.2065	0.0029	0.9299	0.2905	0.0018	0.6620	0.4334	0.0044	0.6785
C-reactive protein	mg/dL	0.6015	0.0058	0.1414	0.4237	0.0029	0.0674	0.7537	0.0032	0.2817
Hba1c	%	0.9468	0.0197	4.4880	1.0613	0.0142	4.3985	0.9517	0.0167	4.8515
Systolic BP	mmHg	14.6464	0.6784	90.9866	18.5730	0.7875	94.7321	16.7040	0.5040	100.1305
Total cholesterol	mg/dL	39.9367	0.9721	163.2156	77.4899	0.3057	185.2020	40.3106	0.3803	177.0529
α		31.63			259.123			337.4045		

Supplementary Table A.5.3: Parameters for deriving BioAge: original weights from Kuo, Pilling, Liu, Atkins, and Levine [145] and re-estimated weights on UK Biobank and NHANES datasets. More details are in Supplementary Methods **??**.

		Units	Original Weights	Re-estimated Weights on UKB	Re-estimated Weights on NHANES
b_1	Albumin	g/L	-0.0336	-0.0137	-0.0291
	Creatinine	umol/L	0.0095	0.0032	0.0018
	Glucose, serum	mmol/L	0.1953	0.0943	0.0461
	C-reactive protein (log)	mg/dL	0.0954	0.1740	0.0427
	Lymphocyte percent	%	-0.0120	-0.0282	-0.0153
	Mean cell volumn	fL	0.0268	-0.0027	0.0332
	Red cell distribution width	%	0.3306	0.0844	0.1766
	Alkaline phosphatase	U/L	0.0019	0.0033	0.0034
	White blood cell count	1000 cells/uL	0.0554	0.0189	0.0226
	Age	years	0.0804	0.0946	0.0777
b_0			-19.9070	-14.7996	-19.9067
γ			0.0077	0.0141	0.0077
α			141.50	130.9765	139.7259
β			-0.0055	-0.0137	-0.0062
θ			0.0917	0.1028	0.0846

Supplementary Table A.5.4: Parameters for deriving PhenoAge: original weights from Levine, Lu, Quach, Chen, Assimes, Bandinelli, Hou, Baccarelli, Stewart, Li, et al. [153] and re-estimated weights on UKB and NHANES datasets. More details are in Supplementary Methods **??**.

Panel	ENABL Age-L	
	UK Biobank	NHANES
Complete blood count (CBC)	Hematocrit	Hematocrit
	Hemoglobin	Hemoglobin
	Red cell distribution width	Red cell distribution width
	Mean corpuscular volume	Mean corpuscular volume
	Red blood cell count	Red blood cell count
	Platelet count	Platelet count
	Mean platelet volume	Mean platelet volume
	Mean corpuscular hemoglobin	Mean corpuscular hemoglobin
	Mean corpuscular hemoglobin concentration	Mean corpuscular hemoglobin concentration
	Total white blood cell count	Total white blood cell count
Comprehensive metabolic panel (CMP)	Blood urea nitrogen	Sodium
	Creatinine	Potassium
	Glucose	Chloride
	Calcium	Bicarbonate
	Albumin	Blood urea nitrogen
	Total protein	Creatinine
	Alkaline phosphatase (ALP)	Glucose
	Alanine aminotransferase (ALT)	Calcium
	Aspartate aminotransferase (AST)	Albumin
	Bilirubin	Total protein
		Alkaline phosphatase (ALP)
		Alanine aminotransferase (ALT)
		Aspartate aminotransferase (AST)
		Bilirubin
Lipid panel (LP)	Total cholesterol level	Total cholesterol level
	Triglyceride level	Triglyceride level
	High-density lipoprotein (HDL)	High-density lipoprotein (HDL)
	Low-density lipoprotein (LDL)	Low-density lipoprotein (LDL)
White Blood Cell Count (WBC)	Neutrophill count	Neutrophill count
	Neutrophill percentage	Neutrophill percentage
	Lymphocyte count	Lymphocyte count
	Lymphocyte percentage	Lymphocyte percentage
	Monocyte count	Monocyte count
	Monocyte percentage	Monocyte percentage
	Eosinophill count	Eosinophill count
	Eosinophill percentage	Eosinophill percentage
	Basophill count	Basophill count
	Basophill percentage	Basophill percentage

Supplementary Table A.5.5: The features included in the ENABL Age-L of UK Biobank and NHANES datasets.

Importance Ranking	ENABL Age-Q	
	UK Biobank	NHANES
1	Age	Age
2	Sex	Systolic blood pressure
3	Long-standing illness, disability or infirmity	Arm Circumference
4	Waist circumference	Ratio of family income to poverty
5	Pack years of smoking	General health condition
6	Pulse rate	Number of months working in the main job
7	Overall health rating	Sex
8	Age at cancer diagnosis (last)	Education Level - Adults 20+
9	Past tobacco smoking	Require special healthcare equipment
10	Number of treatments/medications taken	Self-reported greatest weight
11	Systolic blood pressure	Avg # alcoholic drinks/day - past 12 months
12	Job involves heavy manual or physical work	Smoked at least 100 cigarettes in life
13	Cancer diagnosed by doctor	Shortness of breath on stairs/inclines
14	Mother still alive	Marital Status - Widowed
15	Average total household income before tax	Number of rooms in home
16	Serious illness, injury or assault to yourself in last 2 years	Diastolic blood pressure
17	Hip circumference	Self-reported weight-age 25
18	Taking other prescription medications	Duration of longest job
19	Usual walking pace	Race - Non-Hispanic White
20	Number of vehicles in household	Not a citizen of the US

Supplementary Table A.5.6: Selected top 20 questionnaire features of the UK Biobank and NHANES datasets.

Importance Ranking	ENABL Age-20	
	UK Biobank	NHANES
1	Age	Age
2	Sex	Blood lead
3	Cystatin C	Red cell distribution width
4	Long-standing illness, disability or infirmity	General health condition
5	Past tobacco smoking	Albumin, urine
6	Age at cancer diagnosis (last)	Ratio of family income to poverty
7	Red blood cell distribution width	Arm Circumference
8	Lymphocyte percentage	Blood cadmium
9	Creatinine	Albumin, serum
10	Alanine aminotransferase	Require special healthcare equipment
11	Pack years of smoking	Number of months working in the main job
12	Sex hormone binding globulin	Creatinine, serum
13	Overall health rating	Education Level - Adults 20+
14	Peak expiratory flow (PEF)	Mean cell volume
15	Average total household income before tax	Race - Non-Hispanic White
16	Cancer diagnosed by doctor	Shortness of breath on stairs/inclines
17	Pulse rate	Lymphocyte percent
18	Waist circumference	MIL: maximum inflation levels
19	Apolipoprotein A	Cotinine, serum
20	Gamma glutamyltransferase	Sex

Supplementary Table A.5.7: Selected top 20 features of the UK Biobank and NHANES datasets.

5-year Mortality Prediction				
	AUROC	P-value (Compare with Re-estimated BioAgeAccel Model)	P-value (Compare with Re-estimated PhenoAgeAccel Model)	ΔAUROC with Chronological Age Model
Chronological Age	0.6965	-	-	-
BioAgeAccel	0.7032	-	-	0.0067
Re-estimated BioAgeAccel	0.7077	-	-	0.0113
ENABL AgeAccel (BioAge Features)	0.7510	0	-	0.0545
PhenoAgeAccel	0.7493	-	-	0.0528
Re-estimated PhenoAgeAccel	0.7460	-	-	0.0495
ENABL AgeAccel (PheonAge Features)	0.7678	-	0	0.0713
ENABL Age-L Accel	0.7684	0	0	0.0720
ENABL Age-Q Accel	0.7968	0	0	0.1003
ENABL Age-20 Accel	0.8053	0	0	0.1089
ENABL AgeAccel (All Features))	0.8179	0	0	0.1214
10-year Mortality Prediction				
	AUROC	P-value (Compare with Re-estimated BioAgeAccel Model)	P-value (Compare with Re-estimated PhenoAgeAccel Model)	ΔAUROC with Chronological Age Model
Chronological Age	0.7163	-	-	-
BioAgeAccel	0.7196	-	-	0.0033
Re-estimated BioAgeAccel	0.7230	-	-	0.0067
ENABL AgeAccel (BioAge Features)	0.7549	0	-	0.0386
PhenoAgeAccel	0.7547	-	-	0.0384
Re-estimated PhenoAgeAccel	0.7478	-	-	0.0314
ENABL AgeAccel (PheonAge Features)	0.7683	-	0	0.0519
ENABL Age-L Accel	0.7737	0	0	0.0574
ENABL Age-Q Accel	0.7917	0	0	0.0754
ENABL Age-20 Accel	0.7991	0	0	0.0828
ENABL AgeAccel (All Features))	0.8115	0	0	0.0952

Supplementary Table A.5.8: Mortality prediction performance using ENABL Ages, BioAge, and PhenoAge on the UK Biobank test set. The "AUROC" column presents the AUROCs of the mortality prediction GBT models for ENABL AgeAccels, PhenoAgeAccel, and BioAgeAccel, adjusted for chronological age and sex. P-values are employed to evaluate the performance improvement of ENABL Ages over "Re-estimated BioAgeAccel" and "Re-estimated PhenoAgeAccel." The "ΔAUROC with Chronological Age Model" is calculated as the AUROC of the model using chronological age, sex, and biological ages minus the AUROC of the model using chronological age and sex alone.

5-year Mortality Prediction				
	AUROC	P-value (Compare with Re-estimated BioAgeAccel Model)	P-value (Compare with Re-estimated PhenoAgeAccel Model)	ΔAUROC with Chronological Age Model
Chronological Age	0.7093	-	-	-
BioAgeAccel	0.7014	-	-	-0.0078
Re-estimated BioAgeAccel	0.7177	-	-	0.0084
ENABL AgeAccel (BioAge Features)	0.7726	0	-	0.0633
PhenoAgeAccel	0.7646	-	-	0.0553
Re-estimated PhenoAgeAccel	0.7637	-	-	0.0545
ENABL AgeAccel (PheonAge Features)	0.7830	-	0	0.0738
ENABL Age-L Accel	0.7899	0	0	0.0806
ENABL Age-Q Accel	0.8213	0	0	0.1121
ENABL Age-20 Accel	0.8274	0	0	0.1181
ENABL AgeAccel (All Features))	0.8387	0	0	0.1294
10-year Mortality Prediction				
	AUROC	P-value (Compare with Re-estimated BioAgeAccel Model)	P-value (Compare with Re-estimated PhenoAgeAccel Model)	ΔAUROC with Chronological Age Model
Chronological Age	0.7211	-	-	-
BioAgeAccel	0.7151	-	-	-0.0060
Re-estimated BioAgeAccel	0.7290	-	-	0.0079
ENABL AgeAccel (BioAge Features)	0.7672	0	-	0.0461
PhenoAgeAccel	0.7682	-	-	0.0471
Re-estimated PhenoAgeAccel	0.7603	-	-	0.0392
ENABL AgeAccel (PheonAge Features)	0.7833	-	0	0.0622
ENABL Age-L Accel	0.7860	0	0	0.0649
ENABL Age-Q Accel	0.8066	0	0	0.0855
ENABL Age-20 Accel	0.8133	0	0	0.0922
ENABL AgeAccel (All Features))	0.8227	0	0	0.1017

Supplementary Table A.5.9: Mortality prediction performance using ENABL Ages, BioAge, and PhenoAge on the UK Biobank geographical validation set. The "AUROC" column presents the AUROCs of the mortality prediction GBT models for ENABL AgeAccels, PhenoAgeAccel, and BioAgeAccel, adjusted for chronological age and sex. P-values are employed to evaluate the performance improvement of ENABL Ages over "Re-estimated BioAgeAccel" and "Re-estimated PhenoAgeAccel." The "ΔAUROC with Chronological Age Model" is calculated as the AUROC of the model using chronological age, sex, and biological ages minus the AUROC of the model using chronological age and sex alone.

5-year Mortality Prediction				
	AUROC	P-value (Compare with Re-estimated BioAgeAccel Model)	P-value (Compare with Re-estimated PhenoAgeAccel Model)	Δ AUROC with Chronological Age Model
Chronological Age	0.8392	-	-	-
BioAgeAccel	0.8443	-	-	0.0051
Re-estimated BioAgeAccel	0.8429	-	-	0.0037
ENABL AgeAccel (BioAge Features)	0.8543	0.0080	-	0.0151
PhenoAgeAccel	0.8626	-	-	0.0234
Re-estimated PhenoAgeAccel	0.8669	-	-	0.0277
ENABL AgeAccel (PheonAge Features)	0.8765	-	0	0.0373
ENABL Age-L Accel	0.8721	0	0.082	0.0329
ENABL Age-Q Accel	0.8794	0	0.0010	0.0402
ENABL Age-20 Accel	0.8868	0	0	0.0476
ENABL AgeAccel (All Features))	0.8935	0	0	0.0543
10-year Mortality Prediction				
	AUROC	P-value (Compare with Re-estimated BioAgeAccel Model)	P-value (Compare with Re-estimated PhenoAgeAccel Model)	Δ AUROC with Chronological Age Model
Chronological Age	0.8741	-	-	-
BioAgeAccel	0.8741	-	-	0.0029
Re-estimated BioAgeAccel	0.8745	-	-	0.0033
ENABL AgeAccel (BioAge Features)	0.8813	0.0030	-	0.0101
PhenoAgeAccel	0.8896	-	-	0.0185
Re-estimated PhenoAgeAccel	0.8887	-	-	0.0176
ENABL AgeAccel (PheonAge Features)	0.8952	-	0	0.0240
ENABL Age-L Accel	0.8897	0	0.333	0.0185
ENABL Age-Q Accel	0.9008	0	0	0.0296
ENABL Age-20 Accel	0.9043	0	0	0.0331
ENABL AgeAccel (All Features))	0.9107	0	0	0.0395

Supplementary Table A.5.10: Mortality prediction performance using ENABL Ages, BioAge, and PhenoAge on the NHANES test set. The "AUROC" column presents the AUROCs of the mortality prediction GBT models for ENABL AgeAccels, PhenoAgeAccel, and BioAgeAccel, adjusted for chronological age and sex. P-values are employed to evaluate the performance improvement of ENABL Ages over "Re-estimated BioAgeAccel" and "Re-estimated PhenoAgeAccel." The "ΔAUROC with Chronological Age Model" is calculated as the AUROC of the model using chronological age, sex, and biological ages minus the AUROC of the model using chronological age and sex alone.

	Intercept of LD Score regression	SE
All-cause mortality ENABL AgeAccel	1.062	0.008
Neoplasm-cause mortality ENABL AgeAccel	1.050	0.007
Circulatory disease-cause mortality ENABL AgeAccel	1.088	0.009
Respiratory disease-cause mortality ENABL AgeAccel	1.081	0.009
Digestive disease-cause mortality ENABL AgeAccel	1.074	0.008
Other causes mortality ENABL AgeAccel	1.047	0.007

Supplementary Table A.5.11: The LD score regression intercepts and their standard errors of the GWAS on ENBAL AgeAccels.

SNP	Chr	bJ	pJ	Gene	Reported trait(s)	Other aging trait(s)
All-cause mortality ENABL AgeAccel						
rs71658797	1	0.18	1.29E-16	AK5	BMI, Estimated glomerular filtration rate, Lung cancer	
rs362307	4	0.18	3.63E-11	HTT	Type 2 diabetes, Mean platelet volume, Waist-hip ratio, Neutrophil count	Parental longevity, Multivariate aging
rs1229984	4	0.34	3.86E-13	ADH1B	Alcohol consumption, Protein quantitative trait loci (liver), Esophageal cancer, BMI	
rs809955	4	-0.10	5.41E-12	MAML3	BMI, Smoking initiation, Type 2 diabetes, C-reactive protein levels	
rs3184504	12	-0.08	6.59E-09	SH2B3	Platelet count, Blood pressure, Hemoglobin, Monocyte count, Colorectal cancer, Cystatin C levels, Coronary artery disease	Longevity, Parental longevity
rs9788721	15	-0.13	1.15E-19	HYKK	Post bronchodilator FEV1/FVC ratio, Lung cancer, Smoking behaviour, COPD	Parental longevity, Multivariate aging
rs7206629	16	0.09	6.90E-10	FTO	BMI, Type 2 diabetes, Obesity, Breast cancer, High density lipoprotein cholesterol levels	PhenoAgeAccel, Longevity
rs34003767	17	-0.09	4.40E-09	MED24	Asthma and cardiovascular disease, White blood cell count	GrimAgeAccel
rs159428	20	0.10	2.60E-12	NOL4L	Monocyte count, Smoking status, Red cell distribution width, Mean corpuscular hemoglobin	Multivariate aging
rs4809542	20	0.19	1.12E-11	CHRNA.	Smoking behaviour, COPD, Aerodigestive squamous cell cancer	Parental longevity
Neoplasm-cause mortality ENABL AgeAccel						
rs71658797	1	0.17	7.33E-15	AK5	BMI, Estimated glomerular filtration rate, Lung cancer	
rs10885	6	0.16	5.30E-17	PRRC2A	BMI, Body fat percentage, Smoking status, Cervical cancer	
rs3025316	9	0.17	2.45E-13	FAM163I	Smoking cessation, Serum alkaline phosphatase levels, White blood cell count	
rs2981575	10	-0.09	2.00E-10	FGFR2	Breast cancer, Cancer, Gamma glutamyl transpeptidase	
rs61637848	14	0.1	5.05E-10	KLC1	BMI, Urate levels, Neuropsychiatric disorders, Smoking behaviour, Breast cancer	
rs9788721	15	-0.15	4.67E-23	HYKK	Post bronchodilator FEV1/FVC ratio, Lung cancer, Smoking behaviour, COPD, Parental longevity	Parental longevity, Multivariate aging
rs11638216	15	-0.11	2.34E-13	SEMA6D	BMI, Alcohol consumption, Smoking status, Lung cancer	Parental longevity
rs429358	19	-0.15	1.64E-14	APOE	Blood protein levels, Triglycerides, Total cholesterol levels, Alzheimer's disease	PhenoAgeAccel, BioAgeAccel, Longevity, Parental longevity, Multivariate aging
rs6141319	20	0.12	3.20E-15	NOL4L	Monocyte count, Smoking status, Red cell distribution width, Non-melanoma skin cancer	Multivariate aging
rs4809542	20	0.20	5.19E-12	CHRNA.	Smoking behaviour, COPD, Aerodigestive squamous cell cancer	Parental longevity
Circulatory disease-cause ENABL AgeAccel						
rs55683935	3	-0.31	9.96E-14	RPN1	Monocyte count, Eosinophil counts, Basophil count, Neutrophil count	HorvathAgeAccel
rs362307	4	0.28	2.12E-19	HTT	Mean platelet volume, Waist-hip ratio, Neutrophil count	Parental longevity, Multivariate aging
rs11803927	6	0.78	1.46E-150	LPA	Lipoprotein levels, Coronary artery disease, Peripheral artery disease, Unstable angina pectoris, Myocardial infarction	Longevity, Parental longevity, Multivariate aging
rs14755559	6	0.90	6.10E-27	LPAL2	Lipoprotein levels, Triglyceride levels in VLDL, Coronary artery disease	Parental longevity
rs2744961	6	0.14	5.16E-16	ILRUN	BMI, HDL cholesterol levels, Coronary artery disease, Myocardial infarction	
rs550057	9	0.12	5.31E-11	ABO	Blood protein levels, Cholesterol, Venous thromboembolism, Coronary artery disease, Hemoglobin concentration	PhenoAgeAccel
rs10774625	12	-0.14	3.67E-19	ATXN2	Diastolic blood pressure, Platelet count, LDL cholesterol levels, Hemoglobin concentration, Coronary artery disease	Parental longevity, Multivariate aging
rs9923147	16	0.20	9.02E-33	FTO	BMI, Type 2 diabetes, Obesity, High density lipoprotein cholesterol levels, C-reactive protein	PhenoAgeAccel, Longevity
rs7189927	16	-0.15	5.34E-18	ATP2A1	BMI, Hip circumference, Type 2 diabetes, Mean corpuscular volume	
rs7412	19	-0.29	2.60E-22	APOE	Blood protein levels, Triglycerides, Total cholesterol levels, Alzheimer's disease	PhenoAgeAccel, BioAgeAccel, Longevity, Parental longevity, Multivariate aging

Supplementary Table A.5.12: Selected genetic loci that are significantly associated with ENABL AgeAccel for all-cause mortality, neoplasm-cause mortality, and circulatory disease-cause mortality. The loci are mapped to interesting genes associated with health-related traits.

6 DEEP PROFILING OF GENE EXPRESSION ACROSS 18 HUMAN CANCERS

6.1 INTRODUCTION

Gene expression profiles are the reflections of a complex network of underlying cellular and molecular processes. Unsupervised learning is a key step toward extracting meaningful biological information from expression profiles and reducing the dimensionality of the data for downstream tasks, such as prediction of phenotypes. Unsupervised learning projects high-dimensional input variables into a latent space consisting of a smaller set of latent variables, or factors, capable of explaining the variation in the original input space. Learned latent variables represent sources of genome-wide expression variation across samples, for example large-scale transcriptional programs that define intrinsic disease subtypes or reflect extrinsic stimuli such as hypoxia or treatment pressure. Each individual cancer has different characteristics and response to therapy, even cancers of the same type. Therefore, discovering and understanding biologically meaningful sources of expression variation is of considerable interest from a research and clinical perspective.

One key limitation of commonly used latent space learning approaches for expression data, such as principal component analysis (PCA), is that they can only extract latent variables that have linear relationships with gene expression levels, while gene interactions can be more complex. The artificial intelligence (AI) field has achieved notable success in unsupervised learning by using deep neural networks that can capture highly complex relationships between variables. It has been shown that the latent variables extracted by unsupervised deep learning approaches from image data represent high-level features that are intuitively important for the entire image in the training set, for example: skin color, age, and gender from face images [112], lighting and room geometry from scene images [102], and rotation and size of an object from 3D images[111]. These informative and complex image features cannot be captured by models limited to learning linear feature interactions [27].

The success of unsupervised deep learning in computer vision has motivated several recent applications of deep unsupervised learning methods to gene expression profiles. Prior approaches have used generative modeling to learn the latent factors underlying single cell sequencing data, separating technical artifacts from biological factors [168]. Furthermore, previous studies have conducted pan-cancer analyses with various approaches ranging from co-expression networks, to differential expression analysis, to deep unsupervised learning approaches [43, 47, 114, 139, 157, 242, 304, 307, 308, 326]. For example,

Kim, Kim, Choe, Lee, and Kang [139] introduced a deep learning architecture to enable transfer learning of unsupervised deep models to improve survival prediction and applied it to The Cancer Genome Atlas (TCGA) data. Way and Greene [307] pioneered the application of unsupervised deep learning to capture biologically relevant features from TCGA expression data.

However, three challenges still impede the successful application of deep unsupervised learning approaches to cancer expression data. First, deep learning has a high risk of overfitting when not provided with large sample numbers. Second, the non-deterministic nature of the learning process impairs the robustness of the learned latent spaces. Each run of neural network training, even using the same architecture, results in different models with different parameters, which makes it difficult to capture consistent signals [278]. Model consistency is of paramount importance in biology, where interpretation of the learned model is more important than obtaining high predictive accuracy. Third, neural networks with multiple hidden layers are "black boxes" by nature: since it is not clear how the model uses gene expression inputs to generate a latent variable, biological interpretation of latent variables is problematic.

In addressing the inherent non-determinism in training deep learning models, particularly for biological data analysis, model ensembles emerge as a potent solution. By aggregating outputs from multiple model runs, ensembles enhance the consistency and stability of predictions, crucial for biological applications [23, 42, 202]. Whereas prior techniques have suggested the use of model ensembles in unsupervised learning [278, 309], these methods have so far been limited to "shallow" models with a single hidden layer. Moreover, the application of Explainable AI (XAI) in life sciences [138, 232, 256], although widespread, often grapples with complex, multidimensional data. In this context, model ensembles offer a substantial advantage, improving the quality and reliability of feature attribution [121], thereby aligning with the growing emphasis on transparency and comprehension in AI models used for biological data analysis.

To resolve these challenges, we developed DeepProfile, a framework that enables a unique pan-cancer analysis by learning statistically robust and interpretable latent spaces from gene expression data (Figure **??**). To robustly train the neural networks, we incorporated expression datasets comprising 18 human cancers from 50,211 transcriptomes in the public gene expression data repository Gene Expression Omnibus (GEO) [73]. To address the non-deterministic nature of the deep learning process and capture robust latent spaces, we devised a unique ensemble approach to integrate the results of hundreds of deep unsupervised models generated from different random starting points and latent space sizes. While previous approaches have proposed using ensembles of models [278, 309], these methods have so far been limited to "shallow" unsupervised models with a single hidden layer. By incorporating state-of-the-art feature attribution methods that can provide gene importance values for each latent variable, DeepProfile is able to create ensembles of "deep" unsupervised models with multiple hidden layers. Finally, DeepProfile

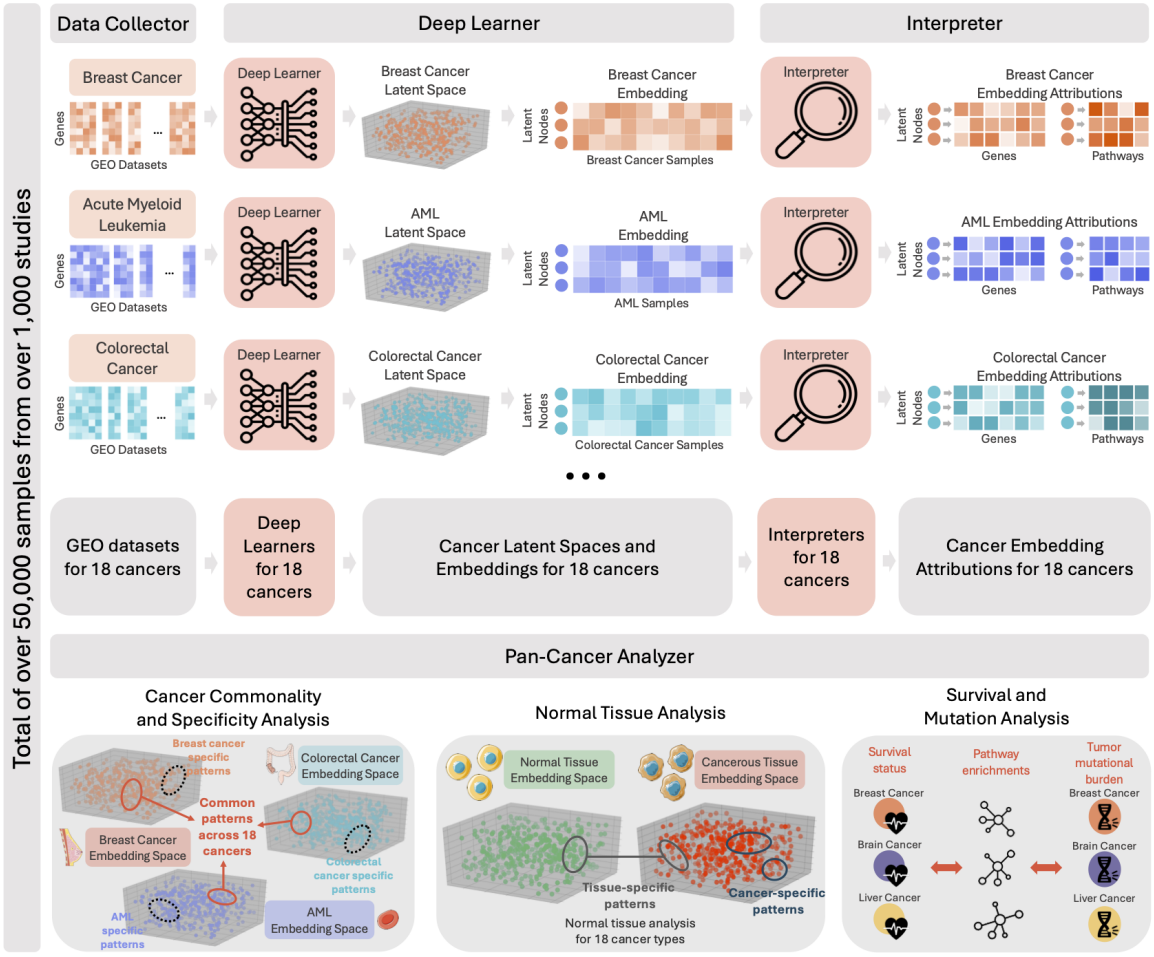


Figure 6.1: **DeepProfile pan-cancer framework.** Data Collector: We downloaded gene-expression datasets for 18 cancer types from the common microarray platforms, preprocessing and concatenating them into cancer-specific expression matrices. In total, we have over 50,000 samples from over 1,000 GEO datasets. Deep Learner: We passed the expression matrices to Deep Learner models to learn cancer-specific latent spaces. Deep Learner is an ensemble of VAEs that encodes the high-dimensional expression signals to a biologically informative 'latent space'. We then mapped the training samples to the learned latent spaces and defined cancer sample 'embeddings', where each DeepProfile latent variable encodes a certain source of variance across cancer samples. Interpreter: We passed the learned embeddings to Interpreter models to extract 'gene-level and pathway-level attributions' for each latent variable. Gene-level attributions denote how much each gene contributes to a latent variable. Similarly, pathway-level attributions denote the pathways significantly associated with the most important genes of each latent variable. Pan-Cancer Analyser: Using the cancer-specific embeddings and attributions; we carried out a detailed pan-cancer analysis including (1) analysing the latent spaces of 18 cancers to discover cancer-common and specific patterns, (2) differentiating cancer-specific patterns from tissue-specifying ones by contrasting cancer embeddings to normal tissue embeddings and (3) investigating survival and mutation related signals by integrating DeepProfile embeddings with survival and tumour mutational burden profiles.

extends previous studies by incorporating an extended set of gene expression profiles from GEO, The Cancer Genome Atlas (TCGA) [312], and the Genotype-Tissue Expression (GTEx) database [164], and by integrating different data modalities such as clinical and mutational features. This rich resource of robust cancer-specific deep embeddings, the values of the latent variables, and biological characterization of the latent variables enables us to examine cancer transcriptomic signals from a new angle and investigate their associations to various phenotypes.

Using the DeepProfile framework, we examine genes and pathways that capture major variation across all 18 cancer types. We find that universally important genes control aspects of the inflammatory response by modulating the transcriptional phenotypes of tumor-infiltrating immune cells. Cancer-type specific genes with large contributions to the latent spaces of only one particular cancer type, on the other hand, define molecular disease subtypes and reflect tissue-specific biology. We develop a methodology for linking DeepProfile embeddings to patient- and tumor-level characteristics and apply the method to study genes and pathways that - as seen through the lens of DeepProfile's latent spaces - correlate with tumor mutation burden and patient survival. We find that tumor mutation burden is significantly associated with the expression of cell cycle-related pathways across a large majority of cancers, while survival correlates with DNA mismatch repair and MHC class II antigen presentation pathway activity. Our methodologies to make deep neural network models biologically interpretable allow for complex, non-linear relationships to be learned while retaining stable models. Thus, DeepProfile's robustness and interpretability enables the discovery of unique biological patterns in large gene expression datasets.

6.2 RESULTS

6.2.1 *DeepProfile learns robust latent spaces for 18 cancer types*

Because highly expressive models such as deep neural networks tend to overfit when the sample size is small, we obtained all available expression datasets from the most common microarray platforms for 18 human cancers from GEO [73] (Figure **??**; Supplemental File 1) (see Methods), resulting in 50,211 samples from 1,098 datasets. DeepProfile projects the expression data into lower-dimensional latent space represented by a set of latent variables using an ensemble approach for the variational autoencoder (VAE) [141] (Extended Data Figure **??**). The VAE is a special type of deep neural network that compresses high-dimensional data (here, tens of thousands of genes) into low-dimensional embeddings with minimal information loss. More specifically, two neural networks - (i) the encoder that models the relationship between input variables and latent variables in the latent space and (ii) the decoder that models the relationship between the latent variables and

the reconstructed input variables - are trained such that the reconstructed input data are close to the input gene expression data (see Methods).

VAE is a unique model that can discover non-linear relations among genes to reflect the true nature of gene interactions. However, applying the model to expression data is not straightforward. Neural networks inherently suffer from learned model variability across different random initializations due to their intrinsic non-convex nature. This means that a conventional learning algorithm for VAE can result in a model that is different in every trial, an outcome that hinders the inference of robust biological signals. To improve robustness, we developed an ensemble of VAEs to combine the learned models from different random runs and latent dimension sizes (Extended Data Figure **??**) (see Methods). This approach integrates signals from hundreds of different latent spaces into one information-rich space. After learning these cancer-specific latent spaces, DeepProfile's 'interpreter' biologically characterizes each latent variable by mapping it to genes and pathways (Figure **??**). This process is based on the principled 'feature attribution' method, namely integrated gradients [272], to quantify how much each latent variable's value is attributed to input variables (Figure **??** and Extended Data Figure **??**). In particular, for each latent variable, DeepProfile produces a list of gene attribution scores, which indicate the relevance of each gene to that latent variable and uses the top-listed genes for pathway enrichment tests, which provide pathway-level attribution scores (see Methods).

The input gene expression datasets, their lower-dimensional embeddings, gene-level and pathway-level relevance, and the results of our pan-cancer analysis are publicly available at: `https://github.com/suinleelab/deepprofile-study` (code), and `https://doi.org/10.6084/m9.figshare.25414765.v2` (data).

The trained DeepProfile model explains the relevant factors of gene expression variation in each sample by encoding high-dimensional measurements of thousands of gene expression levels into 150 latent variables. The number of latent variables was determined using an algorithm that iteratively decides whether to add an additional latent variable using a statistical test of Gaussianity (see Methods). DeepProfile can be applied to any new cancer gene expression dataset to reduce its dimensionality (Extended Data Figure **??**; Methods). To demonstrate the consistency with independent RNA-Seq data, we used RNA-seq data from TCGA [312] containing 9,079 samples across 18 cancers which were not used for training DeepProfile (Extended Data Figure **??**) (see Methods). Our result also highlights that DeepProfile can be successfully applied to RNA-Seq expression profiles despite being trained on microarray data. This is further supported by the high correlation between DeepProfile embeddings generated from microarray and RNA-seq data (Extended Data Figure **??**).

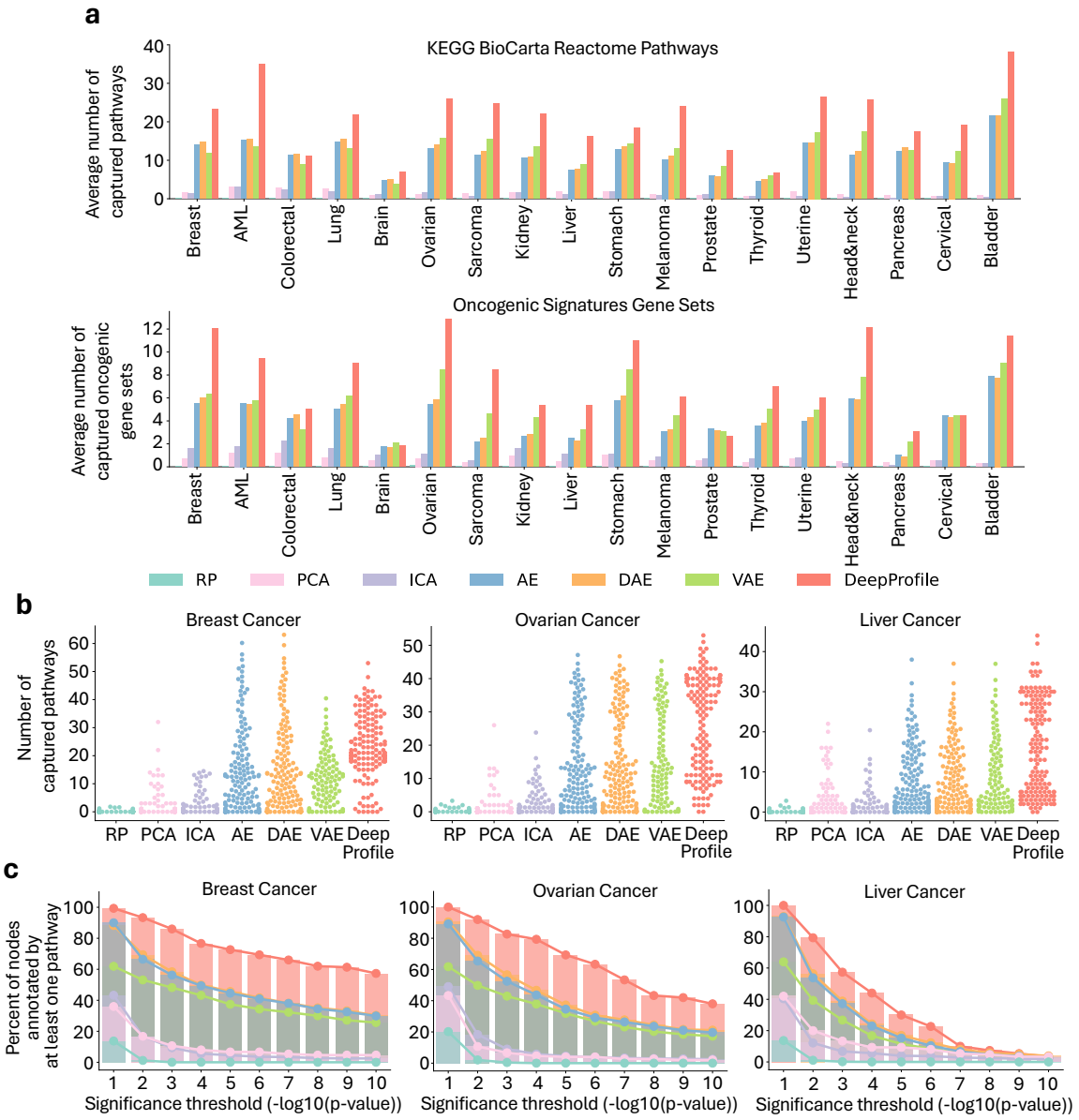


Figure 6.2: **Comparisons of pathway enrichment from DeepProfile with other dimensionality-reduction methods.** a, The average number of pathways significantly captured (FDR-corrected $P < 0.05$) by latent variables of latent embeddings of DeepProfile and other dimensionality-reduction methods are shown for KEGG, BioCarta and Reactome pathways (top), and Oncogenic Signatures gene sets (bottom). Each latent variable of each embedding is associated with each pathway with a P value and we count the number of pathways significantly captured by each latent variable. We then average these pathway counts over all latent variables to define the average number of pathways significantly captured by a method. b, Distribution plots of number of KEGG, BioCarta and Reactome pathways significantly captured (FDR-corrected $P < 0.05$) by each latent variable shown for 3 cancer types. c, Comparison of the percent of latent variables annotated by at least one pathway above the significance threshold. The percent of annotated latent variables are shown for multiple significance thresholds for DeepProfile and alternative dimensionality-reduction methods.

6.2.2 *DeepProfile can learn biologically interpretable latent variables enriched for a wide set of pathways*

It is desirable for latent variables to be biologically interpretable. DeepProfile provides gene attribution scores for each latent variable, thereby enabling a standard enrichment test to assess the overlap's statistical significance using the Fisher's exact test between the top-scoring genes and predefined gene sets, available through curated pathway databases such as KEGG, BioCarta, and Reactome. Pathway annotation dramatically facilitates the interpretation of a latent variable's biological meaning; ideally, latent variables will capture many known pathways. We compared the average number of pathways captured by DeepProfile's latent variables to results from other dimensionality reduction methods (see Methods). DeepProfile latent variables captured more pathways than alternative methods (106 test cases out of 108, proportions z-test $P = 1.62 \times 10^{-301}$) (Figure **??**a top). Further, when we focused on oncogenic pathways (as defined by MSigDb) specifically, DeepProfile outperformed the other methods in terms of total gene sets captured (102 tests cases out of 108, proportions z-test $P = 2.03 \times 10^{-90}$) (Figure **??**a bottom). This means DeepProfile not only captures more pathways but also identifies the pathways relevant to cancer.

A latent variable not associated with any known pathway is difficult to characterize biologically, thus decreasing overall interpretability. We found that DeepProfile produces fewer such pathways than other methods (Figure **??**b and Extended Data Figure **??**) (see Methods). Further, we show that, for varying p-value thresholds, a higher percentage of DeepProfile latent variables are biologically annotated compared to other methods (Figure **??**c and Extended Data Figure **??**) (see Methods). To validate DeepProfile's discriminatory power against random patterns, we explored its performance on Gaussian noise datasets, simulating conditions devoid of actual biological signals. The results highlight the model's precision in differentiating genuine biological signals from noise. These results demonstrate that DeepProfile's unique deep learning ensemble approach improves latent variables' biological interpretability. Using the robustly identified latent space and embeddings and the gene-level and pathway-level interpretation of each latent variable, we next proceeded to perform in-depth analyses of the biology revealed by DeepProfile.

6.2.3 *Universally important genes modulate inflammatory pathways*

We began by investigating genes with universally large gene attribution scores to DeepProfile latent variables across all cancer types (see Methods). These genes represent dominant gene expression programs that consistently explain considerable portions of the transcriptional variance across many different cancers. We found that universal genes with high average gene attribution scores were primarily involved in immune response regulation and antigen presentation (35 out of the top universal 100 genes, $P = 9.4 \times 10^{-6}$) (Figs. 3a-c). Given that solid tumors (which constitute most of our data) can be infiltrated by

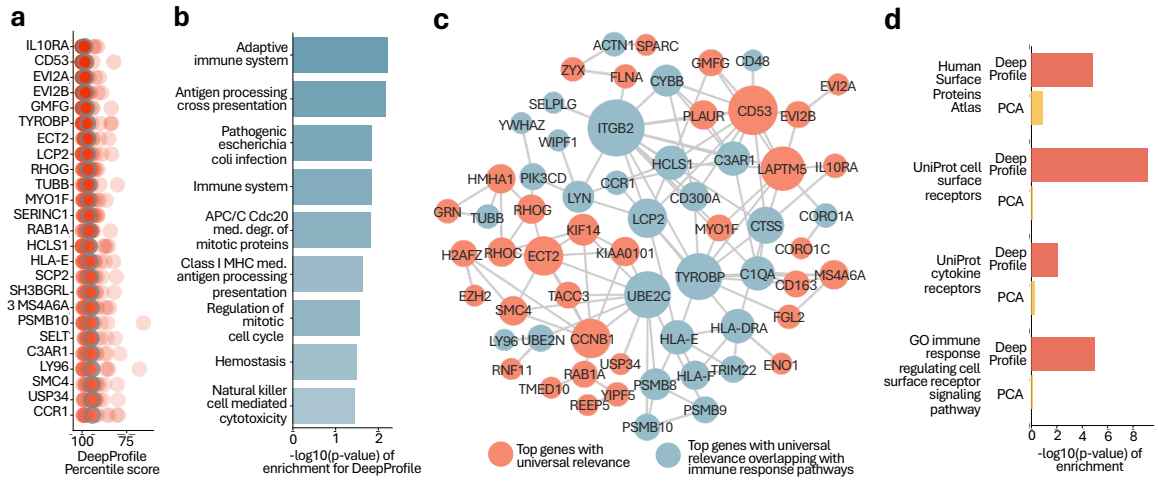


Figure 6.3: **DeepProfile cancer-commonality analysis.** a, List of top highest-scoring genes across 18 cancer types for DeepProfile. The percentile scores of the top- scoring genes are shown for all cancers and the average percentile scores across 18 cancers are highlighted. We calculated the average importance of a gene for DeepProfile embedding by calculating the average gene importance scores across all latent variables of the embedding, converting the average importance scores to percentile scores and averaging these percentile scores across all 18 cancers. The plot is zoomed in for clear comparison. b, The top enriched pathways (KEGG, BioCarta, Reactome) for the top 100 universally important DeepProfile genes and the corresponding FDR-corrected P values. c, Network of top 100 genes with universal importance. The network was generated with StringDB and disconnected latent variables are excluded. The size of a latent variable was determined by hubness, that is, the number of edges. Genes that are included in immune response-related pathways are coloured blue. d, The enrichment P values for cell surface and cytokine receptors for DeepProfile and PCA top 100 universally important genes.

immune cells to varying degrees, we hypothesized that universal genes may reflect the gene expression signatures of various admixing immune cell types. To test this hypothesis, we assessed the overlap between four signatures of major immune cell types (T cells, B cells, neutrophils, and macrophages; see Methods) [30] and genes with top DeepProfile attribution scores (see Methods). We found that there was a small overlap between top DeepProfile genes and the macrophage signature (2 out of the top universal 100 genes, $P = 2.5 \times 10^{-2}$), but not any of the other immune cell type signatures. To enhance our analysis, we utilized pre-computed immune cell fractions from the TCGA data [284]. We calculated Pearson's correlations between gene expression levels and the proportions of various immune cell types, identifying the top 100 genes most correlated with each cell type. Subsequent Fisher's exact tests showed minimal or no significant overlap between these correlation-based top genes and the top 100 DeepProfile genes (Methods), suggesting that the gene sets driving immune infiltration are distinct from those identified by DeepProfile's signatures.

Next, we hypothesized that DeepProfile prioritized genes whose expression was associated with recurrent transcriptional phenotypes in tumor-infiltrating immune cells, such as signatures linked with immune cell activation or suppression. To illustrate this concept, consider the gene with the highest average attribution, the alpha subunit of the interleukin 10 receptor (*IL10RA*). *IL10RA* scored among the top 1% of genes in 14 out of 18 cancers (78% of cancer types) and top 10% in all 18 cancer types, indicating that DeepProfile consistently ascribed high explanatory power to this gene, regardless of tissue context (**??**a). Upon encountering an inflammatory stimulus, a variety of immune cells upregulate *IL10RA*, which mediates the activation of a compensatory anti-inflammatory gene expression program; *IL10RA* has consequently been described as a "master switch" regulating the balance between pro- and anti-tumor inflammation [206]. Therefore, transcript levels of *IL10RA* do not only reflect the presence or absence of *IL10RA*-expressing immune cells, they also predict several thousand genes regulated by *IL10RA* [130], potentially explaining the large role this gene plays in DeepProfile's latent spaces.

To test the hypothesis that universally high-scoring DeepProfile genes were enriched for transcripts that, like *IL10RA*, modulate immune cells' transcriptional phenotypes, we quantified cell surface receptors among genes with top attribution scores. We reasoned that cell surface receptors are enriched for proteins that relay extra-cellular signals and thus have the potential to regulate immune cells' transcriptional phenotypes. We collected gene sets containing cell surface proteins and receptors from the Cell Surface Protein Atlas (CSPA) [22], the UniProt database [61], and the Gene Ontology database (GO) [60]. We found highly significant overlap between these gene sets and genes with top average DeepProfile attribution scores across all cancers (15, 32, and 12 out of the top universal 100 genes, respectively; p-values: 1.5×10^{-5}, 7.0×10^{-10}, 1.0×10^{-5}) (Figure **??**d) (see Methods). Importantly, PCA did not recover these cell surface proteins and receptors (Figure **??**d) (see Methods), indicating that DeepProfile's ability to identify non-linear relation-

ships is essential in capturing this source of variance, and that the functional relationship between receptor expression and gene expression modulation may itself have non-linear form.

In addition to *IL10RA*, DeepProfile's top attributions contained many lesser known but potentially important genes that are consistently involved in the latent spaces of most cancer types. These included *CD53*, an immune-cell specific tetraspanin [72]; *EVI2A* and *EVI2B*, genes that control granulocytic differentiation [341]; and *TYROBP*, an adaptor protein that in association with various receptors mediates immune cell activation [277] (Figure **??**a). As indicated above, none of these genes appear to signal the presence of a particular immune cell type in the tumor microenvironment, as they are broadly expressed by many different cells, but instead may be involved in modulating tumor-resident immune cells' transcriptional phenotypes.

6.2.4 *Universally important pathways include cell cycle, immune system, and oxidative phosphorylation*

Next, to investigate pathway-level information captured by DeepProfile, we studied the relationship between the embeddings and curated pathway gene sets available through the KEGG, BioCarta, and Reactome databases. We considered a pathway to be significantly enriched in a given cancer type if it overlapped with an FDR-corrected P value below 0.05 with at least one DeepProfile latent variable (see Methods). We then extracted the pathways captured in the largest number of cancer types, grouped these pathways by functional category, and sorted the categories by the average number of cancer types in which they were significantly detected.

As expected, cell cycle-related gene sets were almost universally important, confirming that differences in proliferative index are a major source of variation across cancer transcriptomes (Figure **??**). This observation is consistent with long-standing clinical experience - some cancers evidently have higher mitotic rates than others - and the cell cycle consequently is found to play a role in nearly every morphological or molecular characterization of cancer [93, 155, 175, 249, 318]. Two cancer types had notably less pronounced contributions from cell cycle-related gene sets: AML, whose latent space mainly captured pathways related to adaptive immune response, and thyroid cancer, for which the most important pathways were related to mitochondrial function. The two most common types of thyroid cancer (papillary and follicular) are exceptionally slow-growing neoplasms, which may explain this relative lack of contribution by cell cycle-related pathways. In AML, growth rates are more difficult to assess [34], but it may be that most patients experience uniformly high growth rates due to the disease's aggressiveness and its lack of spatial restraint. In both cases, a lack of variation in proliferative fractions across patients would explain why DeepProfile did not detect the cell cycle as an important contributor of variance in these cancers' transcriptomes.

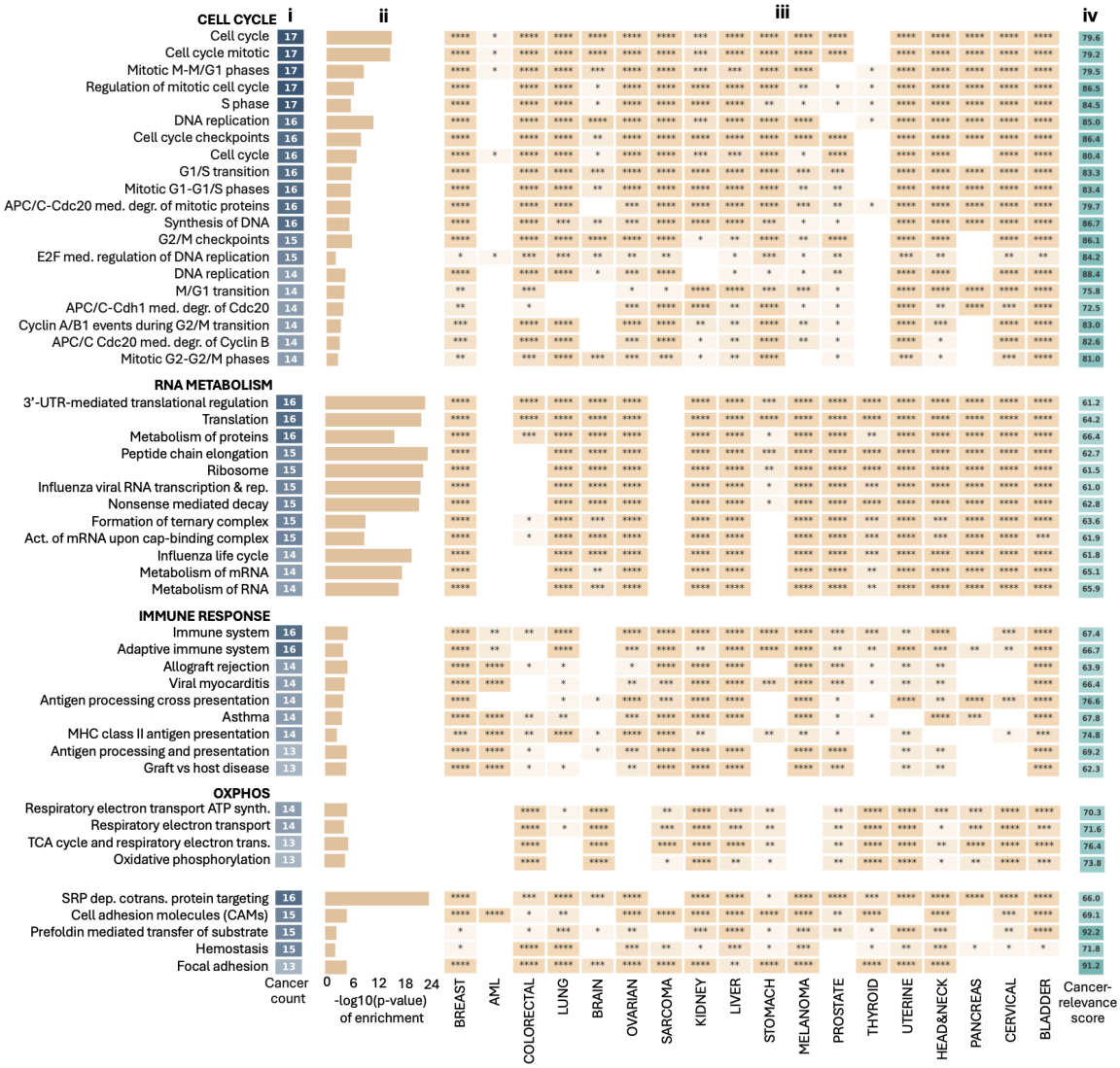


Figure 6.4: **List of top KEGG, BioCarta and Reactome pathways that are universally important.** The pathways are sorted on the basis of the number of cancer types significantly capturing the pathway. All the scores for all pathways are available in Supplementary Dataset 3. a, Number of cancer types (out of 18) significantly capturing (FDR-corrected $P < 0.05$) each pathway. b, $-\log_{10}$(P value of enrichment) averaged over all cancers significantly capturing the pathway. c, Heat map denoting the significance of enrichment P values for top pathways and all cancer types. The star annotations correspond to the significance of enrichment (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, ****$P < 0.0001$). d, Cancer character scores of pathways. The cancer character score denotes the relevance of each pathway to normal or cancerous tissue, where a higher score indicates that the pathway is specifically important for cancerous tissues. The pathways are grouped manually in terms of their functional relations. The order of the groups is determined by the average cancer character score of each pathway group.

Immune-related pathways, as discussed in detail above, were the third-most frequently captured category (Figure **??**) followed by gene sets related to oxidative phosphorylation (OXPHOS), indicating that individual tumors' position on the metabolic continuum between glycolysis and aerobic respiration explains global differences in their gene expression profiles [128]. Genes related to RNA metabolism and ribosome function also emerged as relevant across a large number of cancers; enrichment p-values were particularly significant in this category (Figure **??**). Consistent with prior pan-cancer analyses [18, 93, 155, 242], our study reinforces the significance of both immune-related and metabolism-related pathways across various cancer types, underlining their critical role in cancer biology. The identification of these well-established pathways initially validates the effectiveness of our approach, confirming that DeepProfile is capturing key biological processes known to be pivotal in cancer, and paving the way for uncovering more profound insights in subsequent sections of our analysis.

6.2.5 *DeepProfile latent variables capture both cancer and normal tissue-specific expression signatures*

We hypothesized that RNA metabolism/ribosomal gene sets were not necessarily identified by DeepProfile because they captured variance related to the presence of different disease subtypes within a tissue of origin, but rather because they contained genes that are constitutively expressed in a highly correlated manner. To test this hypothesis, we generated DeepProfile embeddings for normal tissue gene expression profiles from the GTEx database [164]. By fitting predictor models to differentiate between normal and cancer embeddings, we generated a score for each DeepProfile latent variable denoting how successfully it can separate cancer from normal tissue (see Methods). Using DeepProfile pathway-level latent variable attributions, we mapped these latent variable-level scores to pathways to define a cancer-relevance score for each pathway (see Methods). A high cancer-relevance score indicates that the pathway is specifically important for cancer because it shows stronger expression variance in cancer than in normal tissues (Figure **??** iv). We found that in comparison with cell cycle pathways, the ribosomal gene sets' cancer-specificity score was indeed lower (average cancer-specificity score of 82.39 for cell cycle compared to 63.19 for ribosomal pathways; $P = 1.6 \times 10^{-17}$, Welch's t-test), indicating that these genes also capture significant variance across normal tissue gene expression profiles. Nonetheless, we note that the degree of biosynthetic activity (as reflected by ribosomal protein expression) has recently been shown to be associated with differentiation state in colorectal cancer [196], raising the intriguing possibility that DeepProfile's capture of ribosomal genes reflects variance in differentiation states across tumor samples within a given tumor type. This may explain why some relatively narrowly defined (and therefore more homogeneous) cancer types such as AML did not show significantly enriched ribosome-related pathways. We further note that the two near-universally important path-

ways with the highest cancer-relevance scores were related to protein folding (prefoldin) and focal adhesions (Figure **??**). The latter result is consistent with DeepProfile capturing variation in epithelial-to-mesenchymal transition states that may exist among tumors but would not be expected to occur in healthy tissues.

6.2.6 *Cancer type-specific genes and pathways define molecular disease subtypes*

After studying genes and pathways that DeepProfile considered universally relevant, we aimed to identify genes that only capture variance in specific cancer types. We calculated a per-gene cancer type specificity score, defined as the difference between the gene percentile score for one cancer type and the highest gene percentile score across all other cancer types. High specificity scores indicate that a gene captures a large amount of variance in one cancer type but plays a more subordinate role in others (see Methods). We found that genes with high specificity scores generally defined dominant subtypes or grades of differentiation within a tissue category (Figure **??**a). For example, the top breast-specific transcripts were prolactin-induced protein (*PIP*), a gene predominantly expressed in well-differentiated estrogen receptor-positive tumors [57]; *FOXC1*, a gene expressed in basal-like breast cancer [154]; and *GFRA1*, which is specific to the luminal A subtype [29] (Figure **??**a).

To formally test the hypothesis that DeepProfile captured genes that are differentially expressed among breast cancer subtypes, we calculated the overlap between breast cancer-specific genes and PAM50, a gene set that effectively distinguishes between basal-like, normal-like, luminal A, luminal B, and HER2-enriched subtypes [211] and obtained significant results ($P = 3.8 \times 10^{-3}$) (see Methods). Importantly, a linear model (PCA) could not effectively select subtype-specific genes ($P = 1.0$, for PAM50 gene set enrichment), indicating that DeepProfile's ability to capture non-linear relationships is crucial for learning of biologically meaningful patterns. We further explored the abilities of DeepProfile and traditional linear models (PCA, ICA, RP) to distinguish cancer subtypes, leveraging the Metabric dataset renowned for its detailed subtype labels in breast cancer. The results demonstrated that DeepProfile excels in distinguishing cancer subtypes. However, it is noteworthy that our subsequent analysis also revealed that PCA, despite not efficiently selecting subtype-specific genes, could in fact distinguish between different cancer subtypes. This suggests that while DeepProfile is capable of identifying specific genes tied to cancer subtypes, PCA, with a broader analytical approach, also holds the capability to differentiate between cancer subtypes.

Similarly, AML-specific genes comprised transcripts that had previously been associated with AML subtypes (such as *HOXA7*, *TRH*, *MYL4*, *ANK1*) [8, 295] (Figure **??**a) and showed significant overlap with genes identifying AML subtypes ($P = 4.2 \times 10^{-5}$) [297], while PCA again failed ($P = 1.0$). In the brain, DeepProfile identified genes that distinguish oligodendrogliomas from astrocytomas (such as *CNP* [225]) or vary across glioblas-

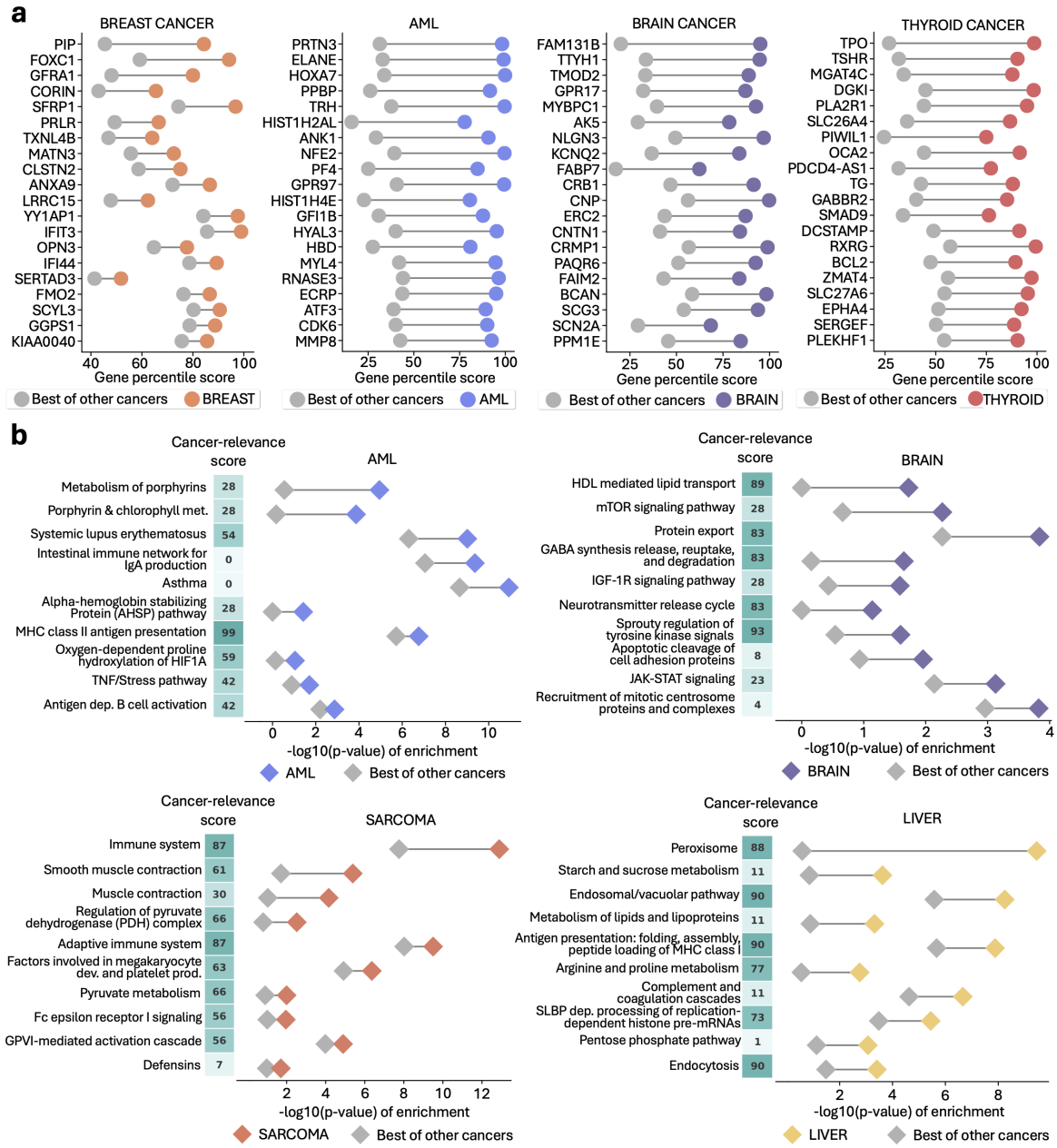


Figure 6.5: **DeepProfile cancer-specificity analysis.** a, The plots of cancer-specific genes shown for 4 cancer types. The difference between the percentile score for the specific cancer type and the highest percentile score among all the other 17 cancer types for the top 20 genes with the highest difference score are shown for each cancer type separately. The coloured dots show the percentile score of one gene for the specific cancer type and the grey dots show the highest percentile score the same gene has among all the other cancer types. The genes are sorted on the basis of the difference values. b, The plots of cancer-specific pathways along with cancer character scores for 4 cancer types. Pathways are sorted on the basis of the difference between the $-\log_{10}$(P value) for the specific cancer type and the highest $-\log_{10}$(P value) among all the other 17 cancer types. Each dot pair represents the $-\log_{10}$(P value) corresponding to one pathway for the specific cancer type and the highest $-\log_{10}$(P value) among all the other cancer types. The vector of cancer character scores shows the cancer character percentile score of the latent variable that is capturing the shown pathway. A higher cancer character score indicates that the given latent variable, therefore pathway, is specifically important in cancerous tissue.

toma subtypes (such as *BCAN* [300]). Top thyroid cancer-specific genes included thyroid peroxidase (*TPO*) and thyroid stimulating hormone receptor (*TSHR*), two transcripts that have critical functions in normal thyroid physiology. These genes may indicate the presence of well-differentiated thyroid cancers, which to some degree retain the expression profiles from their normal tissue of origin, versus highly undifferentiated cancers, which lose tissue-specific transcript expression to a larger degree. To support this hypothesis, we compared DeepProfile thyroid cancer-specific genes with genes associated with thyroid cancer subtypes [333]. We observed that the two gene groups significantly overlapped ($P = 4.4 \times 10^{-10}$) while the same analysis for the thyroid cancer-specific genes discovered by PCA showed no significance ($P = 1.0$). These case studies demonstrate how DeepProfile successfully detects genes that differentiate cancer subtypes, while a linear model fails to capture these patterns.

Next, we extracted curated pathway gene sets that DeepProfile recognized as cancer-specific (see Methods). Potentially more informative than a gene-level view, this approach can go beyond categorizing subtype 'marker genes' to reveal coherent pathways that vary dominantly among cancers from one tissue of origin. Thus, the analysis provides concrete information about the molecular mechanisms driving expression heterogeneity within cancer types. Indeed, DeepProfile assigned highly characteristic molecular processes to each cancer type.

Top AML-specific pathways were related to porphyrin metabolism and heme biosynthesis (Figure **??**b). That leukemic cells show increased heme biosynthesis has been known for more than half a century [303]; but little is known about the mechanistic relevance of the porphyrin production pathway in leukemogenesis. Importantly, recent evidence showing that MYC-overexpressing leukemic progenitors require porphyrin biosynthesis for self-renewal [87] demonstrates a role for this pathway in driving or facilitating leukemogenesis in a subset of these cancers. It is notable that DeepProfile identified this pathway as relevant to AML, as we are not aware of prior unsupervised analyses that have highlighted porphyrin production. As in our analysis of genes and pathways that were universally important across cancers, we also calculated 'cancer-relevance' scores (by comparing matched normal tissue embeddings from GTEx) that determine to what degree a pathway's importance was specific to malignancy. The AML-specific pathway with the highest cancer-relevance score was MHC class II antigen presentation, represented by *HLA-DMA*, *HLA-DRB1*, *HLA-DMB*, *HLA-DPA1*, and *HLA-DPB1* genes. Downregulated *HLA-DPA1*, *HLA-DPB*, and *HLA-DRB1* in AML has recently been reported during relapse after allogeneic bone marrow transplant and has been interpreted as evidence of graft pressure on leukemic cells [56]. However, DeepProfile's identification of the MHC class II antigen presentation pathway's prominence indicates that MHC class II protein expression heterogeneity may be a more general disease feature distinguishing AML subtypes, a concept that has not been described in the literature thus far to our knowledge.

In brain cancer (Figure **??**b), lipid transport scored as the most important pathway, with a high cancer-relevance score. Cholesterol is an essential component of myelin, and the brain contains approximately 20% of the body's total cholesterol [5]. Astrocytes normally produce most of the the brain's cholesterol, since it cannot be transported across the blood-brain-barrier. In glioblastoma, the brain's normal lipid metabolism is altered: glioblastoma cells limit cholesterol biosynthesis and depend on exogenous cholesterol uptake for survival [298], making DeepProfile's selection of this pathway a notable result. The Sprouty (SPRY) pathway obtained the highest cancer-relevance score, driven mainly by *SPRY1* and *SPRY4*. These two genes negatively regulate FGFR signaling, a pathway that is key to glioblastoma progression and is currently being targeted in clinical trials [123]. These and other examples - such as the identification of an important role for the peroxisome in liver cancer [38] (Figure **??**b) - illustrate DeepProfile's ability to extract cancer-specific and biologically meaningful expression patterns from large unstructured data depositories. While understanding expression subtypes and the pathways defining them is valuable from a basic science perspective, determining pathways connected to clinical variables is arguably even more important from a translational point of view. We therefore set out to develop a rigorous methodology for connecting DeepProfile embeddings to relevant patient and tumor-level characteristics.

6.2.7 *Detecting survival- and mutation burden-associated pathways via DeepProfile*

A pathway's contribution to DeepProfile latent variables reflects to what degree it captures variance in the primary gene expression data but does not reveal whether the pathway relates to variables of clinical interest. We developed a general methodology for connecting pathways to clinical characteristics via DeepProfile latent variables (Extended Data Figure **??** and Methods). We tested the approach by extracting pathways that are relevant to two important patient-level and tumor-level features: survival and tumor mutation burden (TMB). Specifically, we associated each DeepProfile latent variable with survival or TMB and generated p-values denoting the association significance between each latent variable and the phenotypes. Then, using the pathway-level attributions for DeepProfile latent variables, we mapped the latent variable-level phenotype associations to pathway-level associations, thereby obtaining survival and TMB association p-values for each pathway (see Extended Data Figure **??**, Methods). The same approach can readily be adapted to other variables of interest, for example tumor stage, tumor grade, or treatment response. There are two advantages of using DeepProfile latent variables (instead of genes or pathways themselves). First, as we demonstrated, DeepProfile embeddings encode robust sources of variation among cancer samples; thus, the association search space is reduced to potentially more biologically meaningful variables. These latent variables distill the comprehensive and intricate biological information from the data without relying on predefined features, enabling exploration of relationships with any biological and

clinical features. With these latent variables, DeepProfile allows researchers to uncover patterns and associations that might be obscured in the high-dimensional space of gene expression data. Second, since each DeepProfile latent variable is a non-linear combination of genes, it has the unique ability to capture complex interactions between genes and phenotypes of interest. This non-linear mapping allows for the integration of multifaceted biological information, going beyond simple additive effects to model the complex, often non-linear relationships inherent in gene regulation and cellular function. Although these latent variables derived from deep neural networks can offer a more nuanced view, the inherent complexity of these models often complicates interpretation. However, by utilizing XAI methods, we are able to clarify these models, providing interpretable insights that pave the way for the discovery of novel insights into cancer biology.

To test the effectiveness of this approach, we first investigated the curated pathway gene sets that DeepProfile recognized to be significantly related to arguably the most important patient-level trait - survival. As in our previous analyses, we initially focused on pathways associated with survival across all cancer types (Figure **??**a) (see Methods). Notably, in this pan-cancer analysis, the unifying theme of most survival-related pathways was adaptive immunity (Figure **??**a). High-scoring gene sets included adaptive immune system, MHC class I antigen presentation, antigen processing cross-presentation, B cell receptor signaling, the proteasome pathway, and activation of NF-κB (all significantly detected in five cancer types). Three pathways stood out for scoring in more than five cancer types. These included DNA mismatch repair (six cancers), a process that can give rise to large numbers of neoantigens when impaired, and MHC class II antigen presentation, which was the highest-scoring pathway overall (significantly detected in seven cancer types). These two pathways will be explored in more detail further below.

To provide a contrast and comparison for these results, we next studied pathways with significant connection to a tumor-level characteristic, TMB (**??**b) (see Methods). Unlike survival, TMB-relevant pathways were most consistently linked to the cell cycle (Figure **??**b) and included DNA replication, mitotic M-M/G1 phases, mitotic prometaphase, chromosome maintenance, and others. The top-scoring TMB-linked pathway was mitotic G2-G2/M phases, which was significantly detected in 11 out 18 cancers. These results establish a link between a tumor's proliferative activity and its mutation burden, consistent with DNA replication acting as a powerful mutagen. This connection carries interesting implications given the strong interest in TMB as a predictor of immunotherapy response [45].

Analogously to previous analyses, we also studied the pathways with the highest survival and TMB scores for each cancer type. Again, we found that DeepProfile identified distinct sets of pathways as being relevant to both traits. For example, survival-related pathways in brain cancer were dominated by interferon type I and II signaling and MHC class I-mediated immunity, while TMB-related pathways prominently featured cell-cell and cell-matrix interactions (Figure **??**c). In sarcoma, survival-related pathways almost

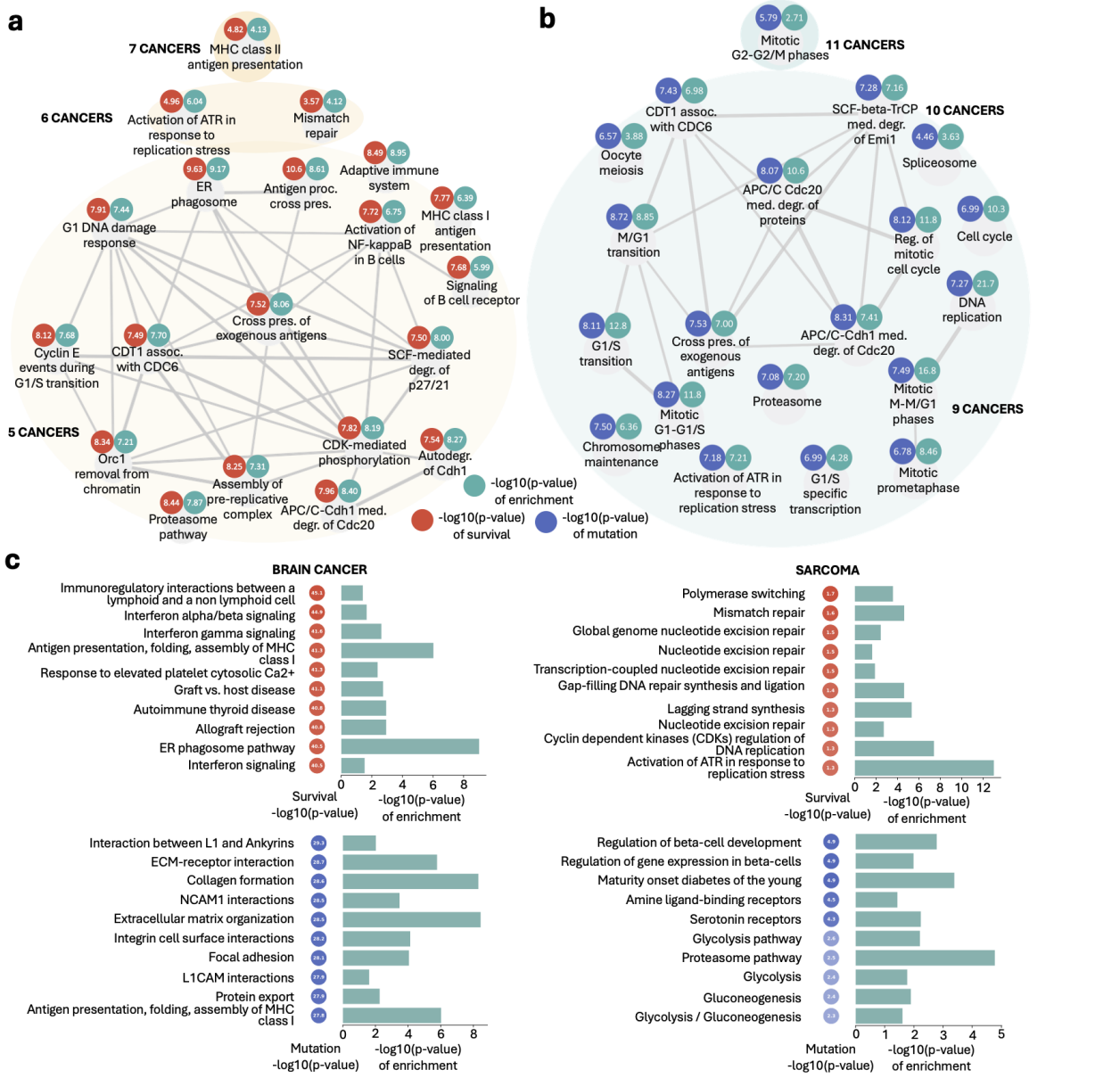


Figure 6.6: **DeepProfile survival and mutation analysis.** a,b, The network of top survival- (a) and TMB-related (b) pathways. For each pathway group, we show the number of cancers for which the pathway is significantly enriched and significantly associated with survival/TMB ($P < 0.05$). We further show the $-\log_{10}$(P value) of enrichment and $-\log_{10}$(P value) of survival/TMB association averaged across all cancers detecting the pathway to be relevant to survival/ mutation. The connections between pathways were determined on the basis of gene membership Jaccard similarities.

exclusively concerned DNA repair processes (mismatch repair, nucleotide excision repair) and replisome function, whereas TMB gene sets were strongly related to glucose metabolism (Figure **??**c).

6.2.8 *DNA mismatch repair and antigen presentation via MHC class II are common survival-related pathways*

We then explored the unexpected pan-cancer association between survival and DNA mismatch repair and MHC class II antigen presentation in more detail. DeepProfile detects robust correlations between pathways and survival; however, it does not reveal these associations' directions. Therefore, to define this direction, we fitted univariate Cox regression models on the genes in the pathways being investigated. This returned a survival z-score for each gene and cancer type pair (see Methods; a negative z-score means that lower expression leads to better chance of survival whereas a positive z-score means that higher expression leads to a better chance of survival).

Examining the z-scores of DNA mismatch repair genes across all cancers, we confirmed a strong correlation with survival (Figure **??**a), validating DeepProfile's findings at the primary gene expression level. The association direction tended to be negative (indicating that lower expression of DNA mismatch repair proteins associates with improved survival), particularly for the six cancers with statistically significant scores in the DeepProfile-based analysis (Figure **??**a). We confirmed this finding further using Kaplan-Meier analyses that yielded consistent results (Figure **??**b and Extended Data Figure **??**) (see Methods). The prognostic relevance of DNA mismatch repair gene expression across many cancers is particularly notable given DeepProfile's identification of the adaptive immune response as a central survival-related pathway hub. Anti-tumor immune responses are thought to depend substantially on the presence of neoantigens, whose abundance increases in cancers with deficient DNA mismatch repair [324]. Similarly, reduced expression of mismatch repair proteins can increase mutability and microsatellite instability [246]. Therefore, higher neoantigen levels in tumors with fewer mismatch repair proteins may make these tumors more visible to the immune system and thus contribute to the improved survival of patients with low DNA mismatch repair protein expression (Figure **??**c).

Next, we investigated the MHC class II antigen presentation pathway more thoroughly. We focused on *HLA-D* genes because they had top-level attribution scores and survival z-scores across all 18 cancer types among all genes in the MHC class II antigen presentation pathway. Unlike the DNA mismatch repair z-scores, which showed a negative correlation between expression and survival across most cancer types, the association for *HLA-D* expression was bifurcated (Figure **??**a). Pancreas, kidney, AML, and brain had a strong negative association between *HLA-D* gene expression and survival change, while the correlation was positive for most other cancers, especially melanoma and uterine cancer.

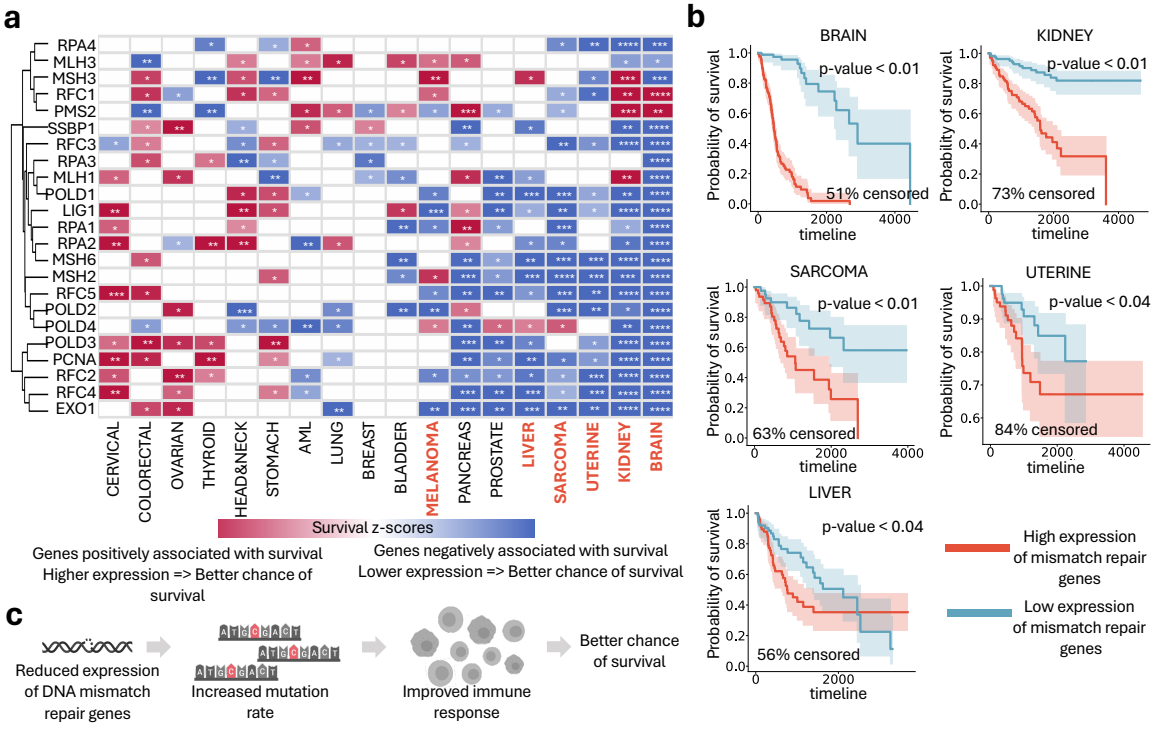


Figure 6.7: **Mismatch-repair-pathway survival analysis.** a, Heat map of survival z-scores of all genes included in the KEGG mismatch repair pathway (*magnitude of $z-score > 1$, **magnitude of $z-score > 2$, ***magnitude of $z-score > 3$, ****magnitude of $z-score > 4$). Six cancer types detected by DeepProfile are highlighted in red. b, Kaplan–Meier plots of average expression of mismatch repair pathway. The samples with an expression above the $mean + 1$s.d. are marked as highly expressed and below $-(mean + 1$s.d.$)$ are marked as lowly expressed. The shaded areas represent the confidence intervals. The log rank test P values and the percent of censored samples are reported for each cancer. Five cancer types with a log rank test P value below 0.05 are shown. c, Schematic of mismatch repair mechanism.

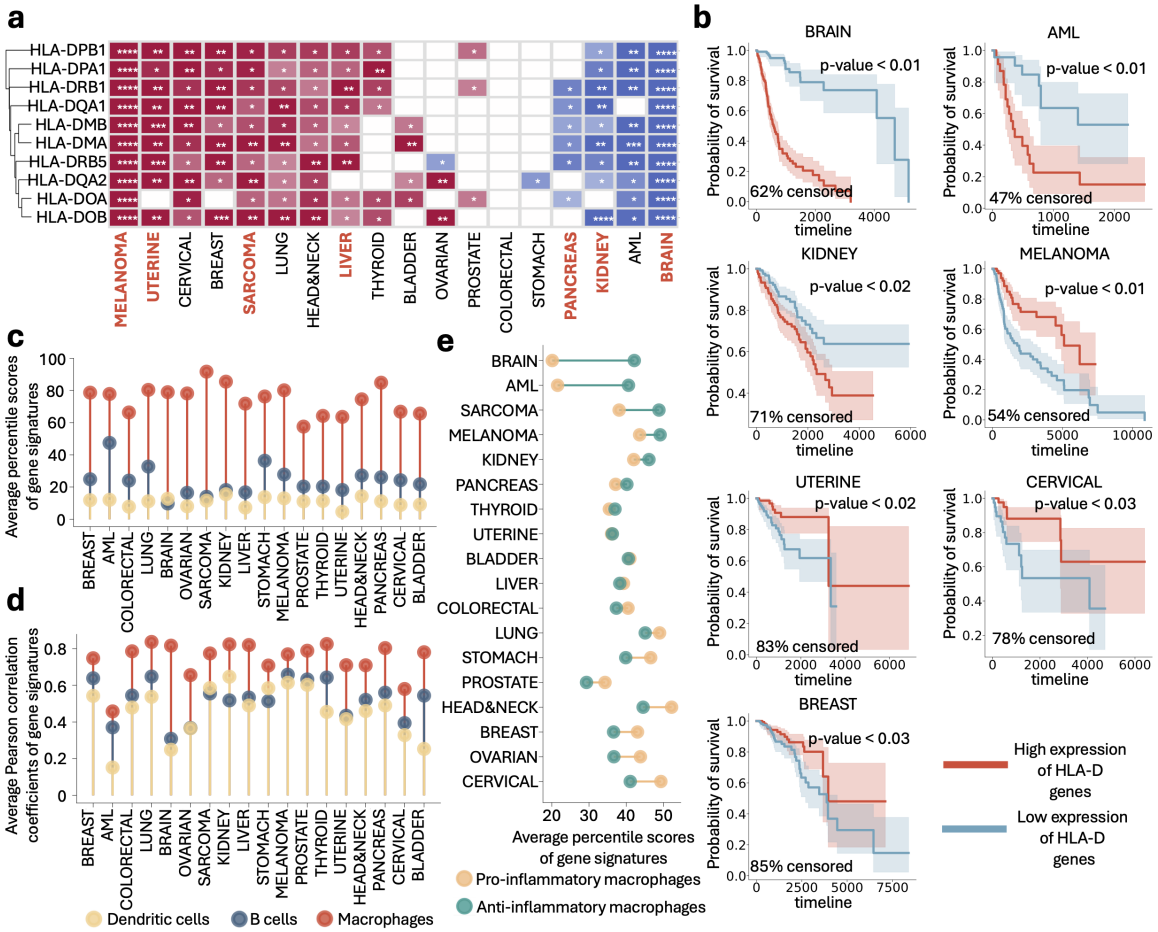


Figure 6.8: **MHC-class-II-pathway survival analysis.** a, Heat map of survival z-scores of all HLA-D genes included in the Reactome MHC class II antigen presentation pathway. Seven cancer types detected by DeepProfile are highlighted in red. b, Kaplan–Meier plots of average expression of HLA-D genes for cancer types with a log rank test P value below 0.05. The samples with an expression above the $\mathrm{mean} + 1\mathrm{s.d.}$ are marked as highly expressed and below $-(\mathrm{mean} + 1\mathrm{s.d.})$ are marked as lowly expressed. The shaded areas represent the confidence intervals. The log rank test P values and the percent of censored samples are reported for each cancer. c, Comparison of average percentile scores of gene dendritic cells, B cells and macrophages shown for 18 cancers. d, Comparison of average Pearson correlation between the expression of HLA-D genes and cell type signatures for the three cell types shown for 18 cancers. e, Comparison of average percentile scores of pro- and anti-inflammatory macrophages shown for 18 cancers (Supplementary Fig. 8).

Again, we confirmed these findings via Kaplan-Meier analyses (Figure **??**b and Extended Data Figure **??**). These results suggested that *HLA-D* gene expression in the tumor and/or its environment is beneficial in some cancer types (melanoma, uterine cancer, breast cancer) and detrimental in others (brain cancer, kidney cancer).

Since most cancers do not express MHC class II genes (with the exception of AML, in which *HLA-D* expression is associated with an inflamed phenotype and therapy relapse [56]), we wondered which cell type in the tumor microenvironment might be the primary source of the *HLA-D* transcripts and, by extension, linked to differential survival. Tumor-resident immune cell types that express MHC class II genes include macrophages, dendritic cells, and B cells. To gauge these cells' relative abundance in the tumor microenvironment, we measured the signature genes' average percentile score for each cell type, where the most highly expressed gene had a score of 100 (see Methods). We found that of the three cell types, macrophage-specific genes were by far the most abundant across all studied cancers, in line with the fact that macrophages can be highly abundant in many cancer types [74, 228, 235] (Figure **??**c). Also, we found that in all cancers, the macrophage signature correlated best with *HLA-D* expression (Figure **??**d; see Methods), further supporting the notion that macrophages are the largest contributors to *HLA-D* transcript abundance in bulk tumor samples. Considering that macrophages' divergent functions range from pro-tumor to anti-tumor activity [74, 228, 235], we wondered whether the phenotypes of tumor-associated, *HLA-D*-expressing macrophages might explain the observed bifurcation in the correlation between *HLA-D* expression and survival. To this end, we examined gene transcripts that may reflect macrophage function. Specifically, we assessed expression of *CD40*, *CXCL9*, *CXCL10*, *CXCL11*, *SLAMF1*, and *TNIP3*, which associate with anti-tumor activity, and of *CFP*, *HRH1*, *NPL*, *PDCD1LG2*, and *CFP*, which typically indicate immunosuppression and tumor promotion [69]. While these genes are not necessarily uniquely expressed by macrophages, the macrophages' abundance (Figure **??**c) makes them plausible main sources of these transcripts. Examining the relative prevalence of the gene transcripts mentioned above revealed that most tumor types expressed both signatures at similar levels (Figure **??**e) (see Methods). The only large gap, with a large preponderance of immunosuppressive transcripts, was observed in brain cancer and AML - the two cancer types with the most significant negative association between *HLA-D* expression and survival ($P = 3.4 \times 10^{-2}$ and $P = 1.6 \times 10^{-1}$, Welch's T test for brain cancer and AML, respectively). We repeated the same test with an extended list of pro- and anti-inflammatory macrophage signatures [185] and again observed a significantly stronger immunosuppressive macrophage abundance in brain cancer ($P = 5.0 \times 10^{-2}$, Welch's T test) (Extended Data Figure **??**d) (see Methods). The presence of macrophages that are polarized towards an immunosuppressive phenotype might therefore contribute to the negative correlation between *HLA-D* expression and survival in brain cancers and AML. In most other cancer types, *HLA-D* expression correlates with improved outcomes, which is consistent with a net positive effect of macrophages on patient survival.

6.3 DISCUSSION

DeepProfile represents a paradigm for applying unsupervised learning to the analysis of gene-expression data. Common unsupervised machine learning techniques in this area fall into three categories: clustering, network inference and representation learning. The mechanism by which statistical patterns are translated into concrete biological insights is important. DeepProfile represents a major departure from existing unsupervised learning paradigms. While the patterns learned by clustering and network inference algorithms have natural biological interpretations, with gene clusters corresponding to expression modules and network edges corresponding to potential regulatory interactions, representation learning largely lacks methods for such a translation. Linear methods such as PCA, ICA or 'shallow' autoencoders have been interpreted by examining the magnitude of their learned weights; however, the 'black box' nature of deep neural networks (DNNs) makes it difficult to understand how genes or biological processes are associated with each learned latent variable and how gene-expression levels are related to phenotypes. DeepProfile provides a language based on rigorous machine learning principles to 'read out' biologically meaningful information from deep representations, enabling discoveries not captured by existing unsupervised analysis paradigms. While DNNs have been successful mainly in tasks where a supervisory label is present [16, 23, 186, 339], DeepProfile opens the door for DNN-based approaches to be applied to unsupervised, comprehensive, exploratory analysis of accumulating published gene-expression data.

DeepProfile introduces a series of rigorous methodologies to 'interrogate' DNNs to generate biological hypotheses. First, one of our key innovations is in the way each latent variable is biologically annotated. We adopted the axiomatic feature attribution method, Integrated Gradients [272], a principled way of estimating the contribution of each input gene variable onto each latent variable. This enabled the computation of gene importance scores for each latent variable, which can be followed by curated pathway gene set enrichment analysis on top-scoring genes. Biological characterization of these latent variables is important, for example, in cancer, to understand the individual variation in clinical outcomes, response to therapy and coordinated transcriptional programmes underlying cancer progression. The overall gene importance scores computed across all latent variables in the entire model result in top-scoring genes whose expression variation across samples explains a large portion of the expression variation of genes. These genes can be interpreted as master regulators, analogous to 'hubs' that are considered important in traditional gene network learning approaches. In addition, DeepProfile introduces various generalizable methodologies to examine the biological characterization of sample-level phenotypes (such as clinical outcomes and tumour mutational burden) on the basis of the latent variables, the difference between samples with different labels (that is, cancer vs normal tissues) and differences between different models (that is, cancer types). We show-

case DeepProfile's ability to reveal biological insights through our pan-cancer analysis using these methodologies detailed below.

DeepProfile also introduces a way to ensemble the latent variables from many variational autoencoder models trained using varying numbers of latent dimensionalities and random initializations. The use of Integrated Gradients [272] allowed the latent variables of our deep model (Extended Data Figure **??**) to be directly ensembled, increasing model stability and consistency, while remaining interpretable. Our experimental results show that DeepProfile's ensembled latent variables encode general and transferrable information about the cancer transcriptome (Figure **??** and Extended Data Figure **??**). We also demonstrated that DeepProfile's ensemble approach can learn better embeddings than individual variational autoencoders trained using specific dimensionalities, consistent with the conclusion that models with different latent dimensionalities may learn different information [309]. The improvement in performance across a variety of tasks that DeepProfile attains suggests that further studies into ensemble methods for unsupervised gene-expression analysis may be fruitful. Furthermore, while DeepProfile was able to extract more underlying biological signal than other unsupervised approaches (Figure **??**), the high-dimensional and highly correlated nature of gene-expression data means that there may have been more biological signal that was not able to be uncovered. Feature attribution methods tend to split credit among correlated features, potentially 'washing out' the signal from large, correlated groups [305]. Future work will be necessary to scale methods for disentangling causal effects from observational data to high-dimensional cancer expression data at the level of either the models or the feature attributions [171, 305].

The application of DeepProfile to a pan-cancer gene-expression compendium exposed several intriguing biological patterns. These analyses were enabled by DeepProfile's integration of the learned model with independent biological databases, including normal tissue expression data, patient-level phenotype data and protein-protein interaction databases. First, we observed that DeepProfile tagged as universally important a very specific category of immune-related genes. Our analysis suggested that these genes did not merely reflect the admixture of different immune cell types in the tumour microenvironment. Instead, they were enriched for cell surface receptors that transduce external signals and thus influence downstream gene expression in a variety of immune cells. Why do these genes capture variance so efficiently? The simplest explanation is that they are representative of recurring transcriptional phenotypes of common immune cells. Depending on the level of immune cell admixture, and thus the magnitude of the immune cell contribution to the overall expression profile, this may be sufficient to propel these genes to such a prominent position. However, an even more powerful explanation is that transcriptional states of malignant cells and infiltrating immune cells are correlated to some degree. For example, cancers with high expression of genes indicative of epithelial-to-mesenchymal transition exhibit a distinct, suppressed immune landscape [44]. Single-cell sequencing studies have shown that transcriptional profiles of immune and cancer cells

can co-vary, and suggest the existence of recurring 'hubs' of interacting cells [51]. Genes that are characteristic of such hubs would be expected to capture particularly high levels of variance, as they would be predictive of both immune and tumour cell transcriptomes. Identification of such genes may be of particular interest from a therapeutic perspective. Careful investigation of top universal DeepProfile genes in single-cell gene-expression data across different cancers will undoubtedly shed more light on this question in the future.

In our cancer-specificity analysis, DeepProfile excelled at extracting disease subtype-specific signatures from the data in an unsupervised manner. We consider this impressive, given that the input datasets were not curated and carefully standardized, such as the ones that were used for the initial discovery of these signatures, but unstructured and variable data deposited in a public database by hundreds of different research groups. DeepProfile's excellent performance in this setting shows that it can robustly identify relevant biological signals in challenging situations in which other methods (such as PCA) do not perform adequately. Analysis of cancer-specific DeepProfile pathways identified disease-specific processes, such as porphyrin metabolism in AML or lipid transport in brain cancer. By further annotating these pathways by their specificity to malignancy, highlighting those that play a comparatively minor role in normal tissue gene expression (via embeddings of GTEx profiles), DeepProfile has generated a list of prime candidate pathways that can be explored for therapeutic intervention opportunities.

Perhaps the most interesting aspect of our analysis was the establishment of a quantitatively rigorous connection between DeepProfile embeddings and patient survival characteristics. The results were unexpected and surprising. Low expression of DNA-mismatch repair transcripts was significantly associated with improved survival in this large cohort of varied cancer types, most of which are expected to be mismatch repair proficient. These results suggest that capacity for DNA-mismatch repair may exist on a transcriptionally driven spectrum and that a tumour's exact position on this continuum may be therapeutically relevant. Microsatellite unstable tumours across all tissues respond well to immune checkpoint therapy and are thus universally approved for treatment with pembrolizumab [226]. Our results raise the question of whether cancers with low DNA-mismatch repair gene expression might also benefit from immune checkpoint inhibition.

Finally, analysis based on DeepProfile's latent spaces showed that adaptive immunity pathways, particularly those related to MHC class II antigen presentation, were the most consistently survival related among 1,077 tested functional gene sets, the latter surpassing even DNA-mismatch repair. This surprising result was highly specific to patient survival, as demonstrated by a comparative analysis for TMB, in which the adaptive immune system did not play a significant role. Focusing on the top-scoring genes from the MHC class II antigen presentation gene set, we found that *HLA-D* transcripts were largely responsible for the strong outcome association. Given that a limited number of immune cells express *HLA-D* genes, we were able to nominate macrophages as the 'prime suspect' source of

these survival-associated transcripts in the tumour microenvironment. The effect of *HLA-D* expression, however, was bifurcated across tumour types. Brain cancer and AML patients had a worse outcome if *HLA-D* expression was high, while melanoma and uterine cancer patients benefitted. We speculate that the transcriptional phenotype of tumour-resident macrophages (pro- or anti-inflammatory) determines whether the presence of these cells has a net beneficial or harmful effect. We found that in glioblastoma, expression of transcripts characteristic of anti-inflammatory macrophages, which are thought to drive tumour progression [165], was predominant, potentially explaining the negative correlation between *HLA-D* expression and outcome. Pro- and anti-inflammatory macrophage transcripts were more balanced in other tumour types, including melanoma and uterine cancer. In these cases, the net effect of the total macrophage population appears to be positive. Importantly, these results are in line with a recent meta-analysis which suggested that expression of anti-inflammatory macrophage markers was correlated with worse prognosis across multiple cancer types, while expression of pro-inflammatory markers was associated with improved survival [165]. Again, it will be important to follow up on these observations in single-cell datasets, once their size has grown sufficiently to conduct robust survival analyses, or in more extensive immunohistochemical studies of macrophage polarization across large patient cohorts.

In summary, we have devised and implemented a deep learning framework to extract robust biological signals from large-scale cancer gene-expression data. DeepProfile is designed to be a resource for the cancer research community. Using our framework, researchers can create robust and interpretable embeddings of new expression data (Extended Data Figure **??**), improving performance on downstream tasks and increasing insight into relevant transcriptional programmes in their samples. The demonstrated compatibility between microarray data and bulk RNA-seq data (Extended Data Figure **??**) suggests that the learned model can be used for bulk RNA-seq data as well. Beyond the computational advance represented by this approach, DeepProfile provides hundreds of biological insights gleaned from existing compendia that can be mined by researchers to advance our understanding of different human malignancies.

6.4 METHODS

6.4.1 *Data processing*

We downloaded publicly available gene expression datasets generated by either of the two microarray platforms: Affymetrix GeneChip Human Genome U133 Plus 2.0 (Affy HG-U133 Plus 2.0) and Affymetrix GeneChip Human Genome U133A 2.0 (Affy HG-U133A 2.0). These datasets were available from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database [73] for 18 cancer types and we

used the 'GEOparse' Python library (https://github.com/guma44/GEOparse) for downloading the datasets.

While GEO searching filters results according to supplied keywords, the returned results may still include gene expression samples from healthy tissues or patients with cancer types other than the queried cancer type. To eliminate these irrelevant samples, we removed the samples that do not contain the search keywords in their 'titles', 'characteristics', or 'descriptions'. To further clean our data without unnecessarily eliminating relevant samples, we manually curated it. Using these steps, we aimed to minimize the number of incorrectly included and incorrectly excluded samples. We also excluded cell line expression samples and used only patient samples because the same cell line's low expression variance across datasets might prevent deep neural networks from learning a reliable model. Despite our automated and manual curation to eliminate samples from cell lines, other cancer types, and healthy tissue, it is still possible that some outlier samples are included in our GEO data collection.

To integrate data from various platforms, we converted platform-specific probe IDs to gene symbols using the probe ID to gene symbol conversion lists for each platform available in GEO. For each cancer, we took the genes present in all data series we have available. A study might have different sample batches submitted on different dates indicated in the 'submission_date' field. We corrected for these potential batch effects within each study using the Python ComBat [126] library's 'combat' function with the default parameters (https://github.com/brentp/combat.py), where different batches correspond to data subsets submitted at different dates. We log transformed the expression measurements, standardized (i.e., zero-mean and unit variance) each gene in each dataset to ensure that different input features (i.e., gene expression levels) are on the same scale, and applied mean imputation to impute missing gene-level measurements. We also excluded duplicate samples with the same GEO IDs. We concatenated all datasets and applied batch effect correction, once again using ComBat with the same parameters, considering each study to be a separate batch in order to minimize the effect of potential study-specific confounders.

6.4.2 *Training variational autoencoder models*

An 'autoencoder' is a type of neural network that consists of an encoder and a decoder network with an information bottleneck layer with D latent variables (i.e., $D \ll M$) in the middle [113]. It generates an embedding Z such that the information present in the original space is preserved in this lower dimensional space as well. Specifically, the encoder network, defined as $f : X \longrightarrow Z$, maps from the input space $X \in \mathbb{R}^M$ to latent embedding $Z \in \mathbb{R}^D$. Similarly, the decoder network, defined as $g_\varphi : Z \longrightarrow X$, maps the embedding Z back to input space. We optimize over the both networks to minimize the squared 2-norm distance between our input X and the reconstructed input as follows:

$\min_{\phi,\varphi} \mathbb{E}\| x - g_\varphi(f(x)) \|_2^2$.

A 'variational autoencoder' (VAE) is an extension of a standard autoencoder that takes as input an $N \times M$ matrix X, where N denotes the number of samples, M denotes the number of features and X_{ij} denotes the feature j of sample i. It also consists of encoder and decoder networks but adopts a regularized training such that the model is robust to overfitting[29]. To perform regularization, VAE learns a distribution of the latent space rather than learning the encoding directly and samples from the learned distribution to generate an embedding. VAE trains the model to bring the distribution of the latent space as close to a standard Gaussian distribution (i.e., $N(0, 1)$) as possible, which ensures that the learned distribution is regularized.

We define the encoder network as $f : X \longrightarrow \mu_x, \sigma_x$, which maps from the input space $X \in \mathbb{R}^M$ to latent space distribution mean $\mu_x \in \mathbb{R}^D$ and distribution variance $\sigma_x \in \mathbb{R}^D$. We then sample from the distribution to define the low-dimensional embedding $Z \in \mathbb{R}^D$:

$Z \sim N(\mu_x, \sigma_x)$.

The decoder is defined the same way as it is in a standard autoencoder. To regularize the distribution over the latent space, VAE adds a regularization term to the model's loss function, i.e., Kullback-Leibler divergence between the learned distribution and a normal distribution [144]. The network is trained to be optimized as follows:

$\min_{\phi,\varphi} \mathbb{E}\| x - g_\psi(f(x)) \|_2^2 + KL[(\mu_x, \sigma_x), N(0, 1)]$,

where $KL[(\mu_x, \sigma_x), N(0, 1)]$ denotes the Kullback-Leibler divergence between the distributions. This regularization component forces the encoder and decoder networks to learn a generalizable, smooth latent space that embeds similar samples close to each other.

Before training different VAE models for a cancer type, we extracted the principal components [319] of the expression matrix; we trained the VAEs using these components as inputs, a commonly used approach for training deep neural networks to prevent overfitting [296]. We chose the number of principal components based on the number of samples and their ability to explain a significant portion of the variance in the data. Specifically, we selected 1,000 components for cancer types with more than 1,000 samples, 500 components for those with 500 to 1,000 samples, and 250 components for those with fewer than 500 samples. Our criteria ensure that the selected components account for approximately 80% of the variance in almost all cancer types and 90% for most.

We trained VAE models using the principal components of the cancer-specific gene expression matrix as inputs; the encoder and decoder networks both include 3 fully connected layers, and the two networks mirror each other in structure. The minibatch size is set to 50, and we trained the models using the Adam optimizer [140] with a learning rate of 0.0005. We initialized each VAE model with a different random set of weights using 'Glorot_uniform' weight initialization. We built the entire model in Python using 'Keras' with 'Tensorflow' backend (https://github.com/keras-team/keras).

In determining the size of the latent space for our VAE models, we specifically selected a set of sizes - 5, 10, 25, 50, 75, and 100. This deliberate selection was made to give our

models a broad scope to capture a comprehensive range of information from the data. We established these sizes to provide a structured approach to encompass the variety and complexity of the data patterns we are analyzing. All layers use rectified linear unit activation except the last layers of both networks, where we applied linear activation and batch normalization on all encoder layers. Additionally, we fine-tuned the VAE models' hyperparameters, including the dropout rate and the number of neurons per layer, utilizing 5-fold cross-validation and gauging the fine-tuning by the metric of validation reconstruction error. Our options for dropout rate included 0, 0.2, 0.4, and 0.6. Regarding the number of latent variables in the intermediate layers, we considered configurations such as (50, 5), (100, 25), (250, 50), (250, 100), and (300, 150).

Initially, we calibrated the dropout rate by averaging the validation reconstruction errors across all models with different configurations of intermediate layers, particularly highlighting the findings for breast cancer (which has the largest sample size), sarcoma (with a sample size that represents the average), and bladder cancer (with the smallest sample size). Models set with a dropout rate of 0 showed the lowest average reconstruction errors. Fixing the dropout rate at zero, we then optimized the count of latent variables in the intermediate layers. The results show that models with an increased number of neurons exhibit improved performance. Nevertheless, to ensure a balance between the efficiency of model training and the precision of feature importance assessments, we settled on 250 and 100 latent variables for the first two layers for latent space sizes of 25, 50, 75, and 100. For a latent space size of 10, the numbers were 250 and 50 latent variables, and for the size of 5, 100 and 25 latent variables were selected.

The maximum GPU memory usage documented across 18 different cancers is 475MB, demonstrating the efficiency of DeepProfile framework in handling large-scale genomic data.

6.4.3 *Learning DeepProfile latent variables*

DeepProfile combines all embeddings generated by VAE models to learn a single, robust latent space that can preserve both high- and low-level features. We trained a total of $|D||R|$ models, where D is a set of possible latent space sizes for individual VAE models and R is a set of random seeds used to initialize model weights. We trained a VAE model for each latent space size $d \in D$ and for each random seed $r \in R$ for the initial weights, which we denote as $VAE_{d,r}$. For our experiments, we used $D = \{5, 10, 25, 50, 75, 100\}$ and $R = \{0, \ldots, 99\}$, which corresponds to 100 random models for each of the 6 latent space sizes, for a total of 600 VAE models. Each VAE model takes the expression matrix $X \in \mathbb{R}^M$ as input and outputs an embedding $Z \in \mathbb{R}^d$. We assessed the robustness of the DeepProfile model by analyzing gene ranking consistency across multiple VAE model ensembles. This analysis showed a substantial increase in gene ranking overlap as more models were added to each ensemble, demonstrating the model's robustness and stability.

Across all |D||R| models, we have |D||R| embeddings and $\sum_{d \in D} d * |R|$ latent variables in total (600 embeddings and 26,500 latent variables for our setting). To group similar data encodings, we applied k-means clustering to cluster all latent variables from all models. We used the Python 'sklearn' library's 'KMeans' model with k-means++ initialization and 10 different starting points [14]. k-means assigns each of $\sum_{d \in D} d * |R|$ latent variables to one of the L clusters, where L is the number of DeepProfile latent variables. Note that we disregard the information about which latent variable came from which model: we simply applied clustering to all latent variables by treating them as independent and identically distributed (i.i.d.). As a result, different latent variables of the same VAE model might be in different clusters as well as in the same cluster. Also, one cluster may include latent variables from different models with the same latent space size (i.e., different runs), or it can also include latent variables from models with different latent space size. After k-means groups the latent variables that are similar across runs and dimensions, we created one ensemble latent variable per cluster by averaging the values of all latent variables in that cluster to obtain a final embedding, $Z \in \mathbb{R}^L$ (Extended Data Figure **??** and **??**a).

To select the latent embedding size for DeepProfile, we applied 'G-means clustering', an extension of k-means clustering that determines the optimal number of clusters k [107]. We used Python's 'gmeans' package and trained with strictness criteria 3, maximum depth 10, and minimum observation count 1 (`https://github.com/flylo/g-means`). For each cancer type, we fitted G-means clustering before training the k-means models to select the optimal k value. We averaged the optimal number of clusters across 18 cancers to set $L = 150$ as the final latent embedding size after rounding down the exact average, which was 157. We selected the same latent size for each cancer type to enable direct comparison between cancer-specific embeddings. To address the inherent variability in k-means clustering, particularly in initial centroid selection, we performed stability analyses. These analyses, involving Normalized Mutual Information (NMI) scores and gene/pathway comparisons across runs, consistently affirmed the stability of our model in identifying key genetic elements.

Our DeepProfile framework can encode user cancer expression samples. When user expression samples are passed to the DeepProfile model, we first apply the same preprocessing procedure we applied to our training samples after eliminating the genes not available for the training samples. We pass the preprocessed expression matrices to our trained VAE models to generate embeddings. In other words, we use the learned weights for VAE models to encode the user samples and generate an embedding from each VAE model. We then use the learned ensemble assignments to cluster VAE latent variables and take the average value in each cluster to define the final DeepProfile embedding for user samples (Extended Data Figure **??**b). Users can select the number of latent dimensions, in which case ensemble label assignments will be calculated again to define the new ensemble latent variables for the user-selected latent dimension size.

We estimate training and testing times for all cancer types. On average, the average training time for each cancer type is approximately 1.36 hours, and the average testing time is notably efficient at just 0.10 hours. This indicates that while the training phase of the DeepProfile models requires a reasonable amount of time, the testing phase is exceptionally efficient, which is advantageous for practical applications.

6.4.4 *Gene- and pathway-level attributions of DeepProfile latent variables*

To calculate gene-level attributions of DeepProfile latent variables, which denote how much each gene contributes to the learned latent variables, we used Python's 'Keras' implementation of Integrated Gradients (`https://github.com/hiranumn/IntegratedGradients`), a gradient-based feature attribution method for neural networks [272] (Extended Data Figure **??**c). When applied to a neural network model, Integrated Gradients learns the sample-level importance values of each input feature for each output variable.

To compute the gene importance values for each latent variable in our finalized DeepProfile model, we follow a two-step approach. Firstly, we calculate the Integrated Gradient (IG) values for each principal component relating to every Variational Autoencoder (VAE) latent variable. Subsequently, these IG values are multiplied by the respective principal component weights, also known as eigenvectors. The process of multiplying the IG values by the eigenvectors provides a mechanism for scaling the importance values according to the influence of each principal component on the original genes. Thus, we are able to obtain the gene importance values linked to each VAE latent variable. As the DeepProfile model is an ensemble of VAE models, the DeepProfile latent variables include numerous VAE latent variables. Therefore, importance values for each DeepProfile latent variable are calculated by averaging the attributions of the corresponding VAE latent variables. To determine the global importance of each gene for a latent variable, we calculate the absolute valued average of attribution scores across all training samples for each cancer type. Since DeepProfile is an ensemble of VAE models, where each DeepProfile latent variable combines multiple VAE latent variables, feature attributions for each DeepProfile latent variable are calculated by averaging the attributions of the VAE latent variables defining that ensemble latent variable.

To calculate pathway-level attributions, we used gene-level attributions and ran pathway enrichment tests using a total of 1,077 functional pathways from Reactome [78], BioCarta [243], and KEGG [132] from the C2 collection of the version 6.2 of MSigDB [159, 270] . For enrichment tests, we used Fisher's Exact Test's (FET) [239] 'fisher_exact' method from Python's 'scipy.stats' module. From the gene list for each pathway, we removed the genes that are not present in our input expression matrix and passed the top G genes with the highest importance values for a DeepProfile latent variable to FET, where G is the average pathway length across all 1,077 functional pathways from Reactome, BioCarta, and KEGG. For multiple hypothesis correction, we applied Benjamini-Hochberg FDR correction [28]

across all latent variables, using the 'multipletests' function from Python's 'statsmodels' library.

6.4.5 *Comparing DeepProfile to alternative dimensionality reduction methods*

We compared DeepProfile to alternative dimensionality reduction algorithms, including the commonly used linear methods as well as other deep learning approaches. We trained these algorithms using the same preprocessed gene expression levels that we used as input to DeepProfile VAE models.

Gaussian random projection maps the original input to a lower dimensional space, where each component is randomly drawn from a normal distribution. From Python's 'sklearn' library, we used 'GaussianRandomProjection' and repeated training 10 times with different random seeds to output 10 different embeddings.

Principal Component Analysis (PCA) [319] is a linear dimensionality reduction method that generates orthogonal components to encode variation in the original input space. We used the 'PCA' module from Python's 'sklearn' library and used the top 150 principal components when comparing it to the DeepProfile embedding, which has 150 latent variables.

Independent Component Analysis (ICA) [59] is also a linear dimensionality reduction method that learns independent components from the original space. We trained ICA using Python's 'sklearn FastICA' with 100,000 iterations; we repeated the training 10 times with different random seeds to output 10 different embeddings.

Autoencoder (AE) [113] is a deep unsupervised neural network consisting of an encoder and decoder network trained to learn a latent space that can reconstruct the original space as successfully as possible. For autoencoder trainings, we used the same top principal components of the preprocessed gene expression levels as we did for training DeepProfile to enable a fair comparison between models. We tuned the hyperparameters of AE models, the number of layers, number of latent variables, dropout rate, and batch size using 5-fold cross validation with reconstruction error as the metric. In the final AE model, we have 1 hidden layer each in encoder and decoder networks with 750 latent variables, 0.1 dropout rate, and batch size of 100. The model was trained with the Adam optimizer using a learning rate of 0.0005. Since each different random initialization of the model can output a different representation, we repeated the autoencoder training 10 times with different random weight initializations. The models were implemented using the 'Keras' with 'Tensorflow' backend.

Denoising Autoencoder (DAE) [299] is a regularized autoencoder model that adds noise to the input data in order to generate more robust embeddings. We applied the same procedure to denoising autoencoder models as autoencoders: we passed the same preprocessed gene expression levels as input to DAE models and selected the hyperparameters with 5-fold cross validation. The final tuned model has 1 hidden layer each in

encoder and decoder networks, 750 latent variables and 0.1 dropout rate. We optimized the model using the Adam optimizer with a learning rate of 0.0005 and batch size of 100. We again repeated the training of DAE models 10 times with different random weight initializations. The models were implemented using the 'Keras' with 'Tensorflow' backend.

Variational Autoencoder (VAE). We included the single VAE models with 100 latent variables, the most powerful configuration among all models within our DeepProfile ensemble, as a baseline for comparison. The VAE models have two hidden layers, featuring 250 and 100 latent variables respectively. The dropout rate is set to 0. Optimization is achieved using the Adam optimizer at a learning rate of 0.0005 and a minibatch size of 50.

6.4.6 *Creating TCGA RNA-Seq embeddings*

We downloaded TCGA RSEM normalized log2 transformed RNA-Seq expression matrices for all cancer types from Broad Institute data version 2016_01_28 (https://gdac.broadinstitute.org/) and generated by TCGA Research Network (https://www.cancer.gov/tcga/). We preprocessed the TCGA expressions with the same pipeline we used for preprocessing GEO expression datasets: we selected the genes available only in the training data, zero imputed the genes missing in the TCGA dataset, and standardized each gene to zero-mean univariance.

Since we take the top principal components of the training data to train DeepProfile, we applied the same processing step for generating TCGA embeddings. We encoded the TCGA samples using the PCA model trained on the training data. To generate the DeepProfile embeddings, we loaded all trained VAE models, encoded TCGA PCA transformed input features with each of the models, and used the pre-learned ensemble labels to cluster the latent variables of our VAE embeddings and define a 150-dimensional DeepProfile embedding for TCGA RNA-Seq samples. We repeated this procedure for each cancer type. To assess the generalizability of the DeepProfile model to TCGA data, we measured the Mean Square Error (MSE) on both GEO and TCGA datasets across different latent dimensions. Significantly lower MSE values on TCGA data underscore DeepProfile's proficiency in effectively reconstructing and adapting to unseen data, demonstrating its robust generalization capabilities.

Similarly, for all the alterative dimensionality reduction approaches, we used the trained models to encode TCGA RNA-Seq samples.

6.4.7 *Comparison of DeepProfile microarray and RNA-Seq embeddings*

To demonstrate that DeepProfile can learn informative latent spaces from both microarray and RNA-Seq test data, we used the TCGA cancer samples for which we have both RNA-Seq and microarray expression available. We downloaded TCGA $\log_2 LOWESSnormalizedmicroarrayex$

generated by the TCGA Research Network (https://www.cancer.gov/tcga/). We selected the genes present in both microarray and RNA-Seq datasets to enable a fair comparison and preprocessed microarray expression profiles following the same preprocessing steps applied to GEO samples. We then measured the Pearson correlation coefficient between the gene expression matrices generated with the two technologies using the Python 'scipy.stats' library's 'pearsonr' method. Thus, we obtained a correlation coefficient for each TCGA sample, which denotes the similarity between the two expression profiles.

Following the same procedure for creating DeepProfile TCGA RNA-Seq embeddings, we created DeepProfile embeddings from TCGA microarray profiles. In this way, using the DeepProfile framework, we obtained two separate embeddings for a cancer type: (1) embeddings generated from microarray expression, and (2) embeddings generated from RNA-Seq expression. We then measured the Pearson correlation between DeepProfile RNA-Seq and microarray embeddings for each TCGA sample using the Python 'scipy.stats' library's 'pearsonr' method. Again, we obtained a correlation coefficient for each TCGA sample, which denotes the similarity between two expression embeddings.

6.4.8 *Comparing DeepProfile pathway coverage to alternative dimensionality-reduction methods*

When comparing DeepProfile to other dimension reduction methods in terms of pathway coverage, which we used as a metric for evaluating the biological relevance of the learned latent space, we followed the same procedure as we used for DeepProfile. We applied Fisher's Exact Test (FET) [239] 'fisher_exact' method from Python's 'scipy.stats' module and obtained a P value for each latent variable-pathway pair, denoting the significance of enrichment.

To run pathway enrichment tests, we first obtained the gene-level attributions for each dimensionality reduction method. For PCA, we obtained the component matrix, which denotes the contribution of each gene to each principal component, and we took the absolute values of the component matrix for use in enrichment tests. Similarly, for ICA and RP, we obtained the absolute valued component matrices. Since we trained each model 10 times with different random initializations, we repeated the FET for each of the 10 models and averaged the pathway enrichment results over 10 runs. For autoencoder and denoising autoencoder models, we used Integrated Gradients [272] to obtain gene-level attributions for the embedding latent variables, following the same procedure we applied for VAE models. Again, we obtained gene-level attributions for each of the 10 random trainings, conducted FET enrichment tests for each run, and reported average pathway enrichments over 10 models.

We compared DeepProfile's pathway coverages to other dimension reduction methods using 3 different metrics:

1. We compared the 'average pathway coverages'. The enrichment tests we conducted provided us with an enrichment p-value for each latent variable-pathway pair. Af-

ter FDR correction, we marked the latent variable-pathway pairs with a $P < 0.05$ as significant and calculated the total number of significant enrichments for each latent variable. We defined the number of pathways significantly captured by each latent variable as the pathway coverage of that latent variable. Then, we averaged these latent variable-level pathway coverages across all latent variables to calculate the average final coverage of an embedding. This metric let us define an average pathway coverage score per model and per cancer type.

2. We compared the 'distributions of latent variable-level pathway coverages' across models. Again, using the same pathway-level attribution p-values, we counted the number of pathways significantly captured (FDR $P < 0.05$) by each latent variable of each embedding. We compared distributions for each method and each cancer type.

3. We compared the 'percent of latent variables annotated by at least one pathway'. For various significance threshold values that range from a P value of 1×10^{-1} to 1×10^{-10}, we counted the number of pathways with a P value below the threshold for each latent variable. This again returned a pathway coverage value for each latent variable of the embedding. We then calculated the percent of latent variables with a pathway coverage above one, which is effectively the percent of latent variables annotated by at least one pathway with a P value below the threshold. We again repeated the calculations for each method and cancer type.

6.4.9 *Comparing DeepProfile pathway coverage to VAE models*

When comparing pathway enrichment of DeepProfile to VAE models, we used the gene-level attributions for each different dimensional VAE model and applied FET to obtain a p-value for each latent variable of each 600 different models. We again considered a latent variable to be significantly capturing a pathway if the FDR corrected P value is below 0.05. DeepProfile is an ensemble model that combines 600 VAE models to define an ensemble embedding, and our aim was to show that the DeepProfile model can preserve the pathways captured by the individual VAE models. Accordingly, we used two different metrics to compare the pathway coverages of VAE models to DeepProfile:

1. We compared DeepProfile pathway coverages to the 'average pathway coverages' of all 600 different VAE models. For each pathway, we calculated the percent of VAE models that captured this pathway significantly (i.e., with at least one latent variable of the embedding with an FDR corrected $P < 0.05$). We then compared the pathways captured by the threshold percent of the VAE models, where the threshold ranges from 50 to 90, to DeepProfile to investigate whether the pathways captured by VAE models could also be captured by DeepProfile.

2. We compared DeepProfile model to VAE models with 'different dimension sizes'. For each different dimensional VAE model, e.g., a 5-dimensional VAE model, we marked a pathway to be captured if the majority of the VAE models (at least 51 of 100 models) significantly captured the pathway (FDR-corrected $P < 0.05$). We repeated the same procedure for each of the 6 different dimensional VAE models to mark the pathways captured by different dimensional VAE models. We then compared the pathways captured by a threshold number of different dimensional models, where the threshold ranges from 1 to 6, to the DeepProfile model in order to investigate whether the pathways captured by these different VAE models could be detected by DeepProfile as well.

6.4.10 *Detecting universally important genes*

To detect the highest-scoring genes across all cancer types, we used the gene-level attributions of DeepProfile latent variables. We calculated the average attribution score across all latent variables to define an overall importance score for each gene, and we converted these scores to percentile scores, where the highest scored gene takes value of 100 and the lowest scored gene takes value of 0. Once we separately obtained these percentile scores for each gene for each cancer type, we calculated the average percentile score across 18 cancer types as the universal percentile score of a gene. We could then sort the genes by their universal percentile scores to detect the top universally important ones. We generated a network of the top 100 universally important genes using STRING [276] with a medium confidence level of 0.4, using all interaction sources and eliminating disconnected latent variables. We visualized the network using Cytoscape [259].

To detect the pathways enriched for the top universally important genes, we used Fisher's Exact Test [239] 'fisher_exact' method from Python's 'scipy.stats' module on the top 100 universally important genes using Reactome [78], BioCarta [243], KEGG [132], and GO Biological Process (BP) [60] gene sets from MSigDB (v.6.2) [159, 270]. We applied FDR correction across all pathways.

Furthermore, to detect whether DeepProfile's universally important genes are enriched for various immune cell type signatures, we collected gene signatures of T cells, B cells, neutrophils, and macrophages [30] and obtained a total of 108 genes available in our training expression dataset. We again used FET for the top 100 universally important genes using these markers to calculate enrichment score. Additionally, we incorporated the pre-computed immune cell fractions from the PanCan Immunity paper in our analysis [284]. We mapped these immune cell fractions to the TCGA gene expression data used in our study. To assess whether the top 100 DeepProfile genes could readily be explained by the genes identified as immune cell related in the PanCan study, we calculated Pearson's correlations between the fractions of four major immune cell types–T cells, B cells, neutrophils, and macrophages, as well as their subtypes–and all genes across all

TCGA samples. In this analysis, the top correlated genes are likely to be specifically expressed by those immune cells. Then, for each immune cell type, we conducted FETs to assess the overlap between its top 100 correlated genes and the top 100 genes identified by our DeepProfile analysis, and applied a Bonferroni correction to adjust for multiple comparisons.

To determine whether DeepProfile's universally important genes are enriched for cell surface and cytok ine receptors, we first collected gene sets from the Cell Surface Protein Atlas (CSPA) [21], UniProt database [61], and Gene Ontology (GO). From CSPA, we downloaded the list of human surfaceome proteins and their annotations, selected the proteins with a 'high confidence' CSPA category and protein probability of 1.0, and obtained a list of 555 human surface proteins. From the UniProt database, we downloaded human cell surface receptors using the keyword 'cell surface receptor,' selected the reviewed proteins, and obtained a list of 1,307 genes. Similarly, we downloaded human cytokine receptors using the keyword 'cytokine receptor', selected the reviewed proteins, and obtained a list of 773 genes. From GO, we used the gene set 'Immune response regulating cell surface receptor signaling pathway', which has 346 genes. Note that for each gene set, we used the genes available only in DeepProfile's training expression matrix and reported the intersecting gene counts. We again ran FET for the top 100 universally important genes using these 4 distinct gene lists to calculate enrichment scores.

To compare the top universally important genes detected by DeepProfile to the universally important genes for PCA, we ran the same analysis that we applied to DeepProfile to PCA. Using the attribution scores (the absolute valued component matrices) of each gene for the PCA model, we calculated the average attribution score across all top 150 principal components, converted the scores to percentile scores, and calculated the average across 18 cancers to define a universal importance score for each gene for PCA. Similarly, we ran FET for the top 100 universally important PCA genes using KEGG, BioCarta, Reactome pathways and GO BP gene sets. We also repeated FET enrichment tests for the same 4 receptor gene lists used for DeepProfile, again using the top 100 universally important PCA genes.

6.4.11 *Detecting universally important pathways*

To detect the highest-scoring pathways across all cancer types, we used the pathway-level attributions we have for each DeepProfile latent variable, which contains the P value of enrichment for each latent variable-pathway pair. To define an overall pathway enrichment score for an embedding, we selected the maximum $-\log_{10}$(P value) across all latent variables for each pathway and obtained an enrichment score for each cancer type/pathway pair. We marked a pathway as being significantly captured by a cancer type if the FDR corrected P value is below 0.05. To determine the universally important pathways, we counted the number of cancer types that significantly captured each pathway. We also

recorded the average $-\log_{10}$(P value) of enrichment for each pathway by taking the mean across all cancer types that significantly captured the pathway. After obtaining the number of cancers and average enrichment score for each pathway, we sorted the pathways first by the number of cancers and then by enrichment scores to get the list of universally important pathways.

6.4.12 *Calculating cancer character scores for pathways*

To calculate a cancer character score for each pathway, we conducted what we called a 'normal tissue analysis.' First, from the GTEX portal (https://gtexportal.org/home/datasets), we downloaded RNA-Seq expression (gene TPMs) with accession number phs000424.v7.p2 [164] and selected the tissues corresponding to the 18 cancer types we have. We preprocessed and encoded GTEX expression profiles following the same pipeline used for TCGA RNA-Seq expression and passed the expression values to our already trained DeepProfile models to generate normal tissue embeddings.

To detect how successfully each latent variable can differentiate cancer vs. normal tissue, we trained logistic regression classifiers by passing the cancer and normal tissue DeepProfile embeddings as input and predicted cancer vs. normal tissue labels. We used the Python 'sklearn' library's 'LogisticRegression' with 'liblinear' solver and 'l2' regularization, repeated the training 500 times with different random samplings, and recorded the mean of absolute value of classifier weights from all models. We defined the cancer character score of each latent variable as the absolute valued classifier weight, denoting the importance of each DeepProfile latent variable in differentiating cancer from normal tissue, where a high cancer character score would indicate that the latent variable is quite important for differentiating the tissue type.

We then mapped these latent variable-level cancer character scores to pathways to determine the cancer-tissue specificity of each pathway for each cancer type. For each pathway, we calculated the weighted average of cancer character scores using the $-\log_{10}$(P value) of enrichment scores for that pathway as the weights and obtained an average cancer character score for each pathway-cancer type pair. Note that if a pathway is not enriched for any of the latent variables, we assigned it a cancer character score of 0. To define a universal cancer character score for each pathway, we calculated the average across the cancer character scores of 18 cancers, excluding the cancers with a score of 0.

6.4.13 *Detecting cancer-specific genes and pathways*

To identify the genes that are high scoring specifically for a certain cancer type, we used the importance scores we calculated for each gene-cancer type pair. For each gene, we calculated the difference between the percentile score of a cancer and the maximum per-

centile score across all other 17 cancers. These cancer-specific difference scores allowed us to detect the top cancer-specific genes for each cancer.

To calculate the enrichment score for the PAM50 genes [211], we ran FET for the top N highest scoring genes for breast cancer where N ranges from 1 to 1000. We then applied Bonferroni correction over all thresholds to report the final P value of enrichment.

To detect cancer-specific pathways, we used the $-\log_{10}$(P value) of enrichment scores we calculated for each pathway/cancer type pair and calculated the difference between the enrichment score for one cancer type and the maximum enrichment score across all other 17 cancers. Thus, we obtained a cancer-specific difference score for each cancer type and for each pathway, allowing us to detect the pathways that are cancer type-specific. We also assigned a cancer character score to each of these cancer-specific pathways using the absolute-valued classifier weights from the normal tissue analysis after converting them to percentile scores.

6.4.14 *Pan-cancer survival and mutation analysis*

With the goal of associating each pathway with patient survival, we used the TCGA RNA-Seq DeepProfile embeddings we learned for 18 cancers along with the survival status. We separately fitted univariate Cox regression models [64] to each DeepProfile latent variable. We used the R 'survival' library's 'coxph' method, recorded P values of model coefficients for each latent variable, and applied FDR correction over all latent variables for each pathway. We repeated the model trainings for each cancer type.

To detect the pathways determining patient survival, we mapped these latent variable-level survival scores to pathways. For each pathway, we calculated the weighted average of $-\log_{10}$(survival P value) across all latent variables using the $-\log_{10}$(enrichment P value) as the weights (Extended Data Figure **??**). When none of the latent variables could significantly capture a pathway (FDR-corrected $P < 0.05$), we assigned a survival score of 0 to that pathway. We also masked the pathway enrichment P values with a z-score below 0.25 to prevent the involvement of lowly ranked pathways in the average score calculation. We calculated the average survival $-\log_{10}$(P value) for each pathway/cancer type pair. Since the survival analysis using enrichment scores from FET did not provide a rich enrichment for survival, we repeated the FET using a broader set of top genes ($4 \times$ average pathway size) to run an informative survival analysis. This pipeline let us define a survival p-value for each pathway and for each cancer type. We then marked a pathway to be relevant to survival if both the FDR corrected enrichment P value and the survival P value were below 0.05.

To detect universally survival-related pathways, for each pathway, we counted the number of cancers (of the 18 cancers) with significant survival scores. We also calculated the average enrichment $-\log_{10}$(P value) and the average survival $-\log_{10}$(P value) across all the cancer types detected to be significantly associated with survival. Sorting the path-

ways by the number of cancer types and then by the average survival score provided us with the top universally survival-associated pathways. We visualized the top 20 universally survival-associated pathways using Cytoscape [259] EnrichmentMap tool [193] with latent variable cutoff value of 1.0 and edge cutoff similarity value of 0.5, where the connections between pathways were determined by the Jaccard similarity of gene memberships of pathways.

To conduct mutation analysis, we downloaded TCGA mutation profiles for all cancer types from Broad Institute data version 2016_01_28 (https://gdac.broadinstitute.org/) generated by the TCGA Research Network (https://www.cancer.gov/tcga/). We selected the samples for which we have both expression measurements and tumor mutational burden (TMB) data available and calculated the total number of mutations for each cancer sample by summing the number of mutations for each gene. If k different mutations occurred for one gene, we increased the total mutation count by k.

To assign a mutation association score to each DeepProfile latent variable, we calculated the Pearson correlation, using the Python 'scipy.stats' library's 'pearsonr' method, between each latent variable of DeepProfile embedding and the log of total mutation count after eliminating outlier mutation scores beyond the 95% confidence level (z-score > 1.96). We repeated these experiments for each of the 18 cancer types and obtained latent variable-level TMB correlation P values.

To map the latent variable-level P values to pathways, we repeated the same procedure we followed for the survival analysis: we calculated the weighted average $-\log_{10}$(TMB P value) across all latent variables for each pathway, where the weights were defined as $-\log_{10}$(enrichment P value) of the latent variables. We again calculated the average enrichment and TMB $-\log_{10}$(P value) across all cancer types with significant scores (FDR-corrected $P < 0.05$) and detected the number of cancers significantly associated with TMB for each pathway. We visualized the top 20 mutation-associated pathways using Cytoscape's [259] EnrichmentMap tool [193] with the same setting as the survival-associated pathways network.

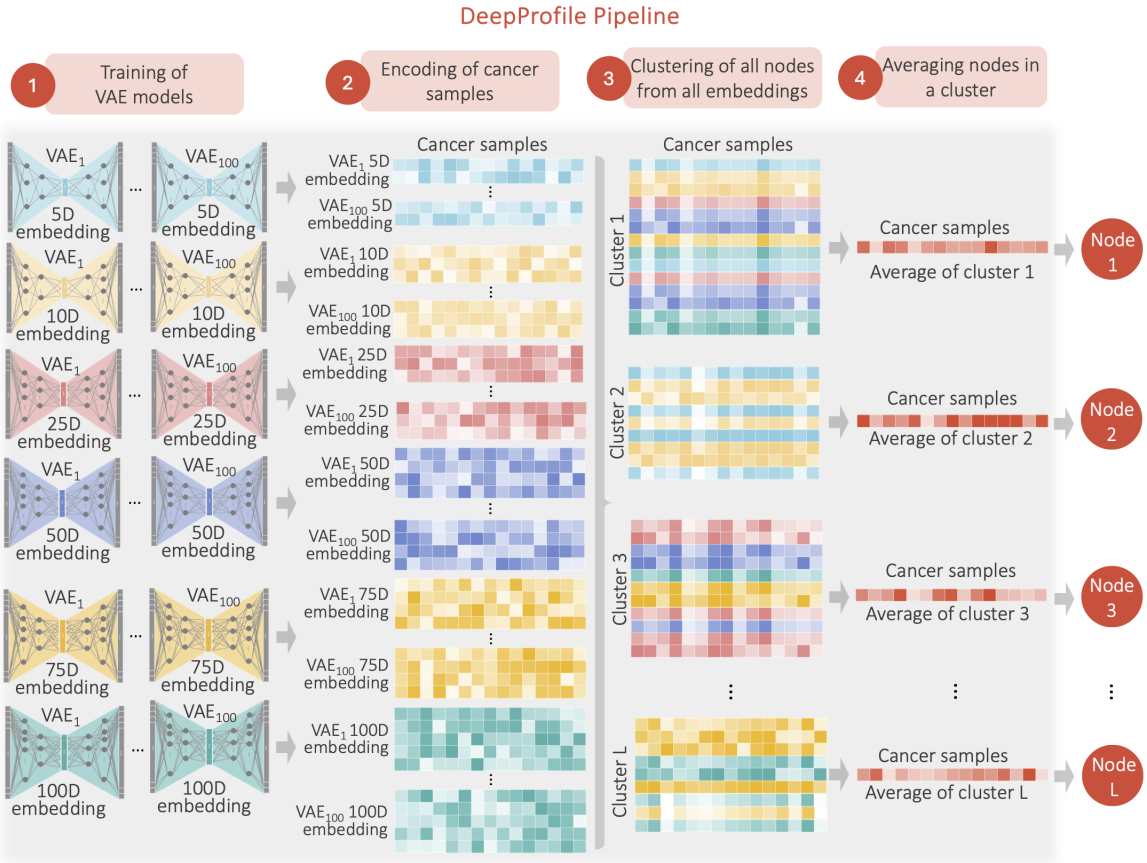
6.4.15 *Downstream survival analysis*

For pathways detected to be relevant to patient prognosis, we conducted a downstream survival analysis independent from DeepProfile model. We first fitted univariate Cox regression models to each of the 23 genes included in the KEGG mismatch repair pathway and 91 genes included in the Reactome MHC class II antigen presentation pathway using TCGA RNA-Seq profiles. We used the R 'survival' library's 'coxph' method to predict survival, trained the models separately for each cancer type, and recorded z-scores to get the direction of association. After obtaining survival z-scores from Cox models, we created heatmaps of gene-level survival scores for 18 cancers by clustering the genes with hierarchical clustering.

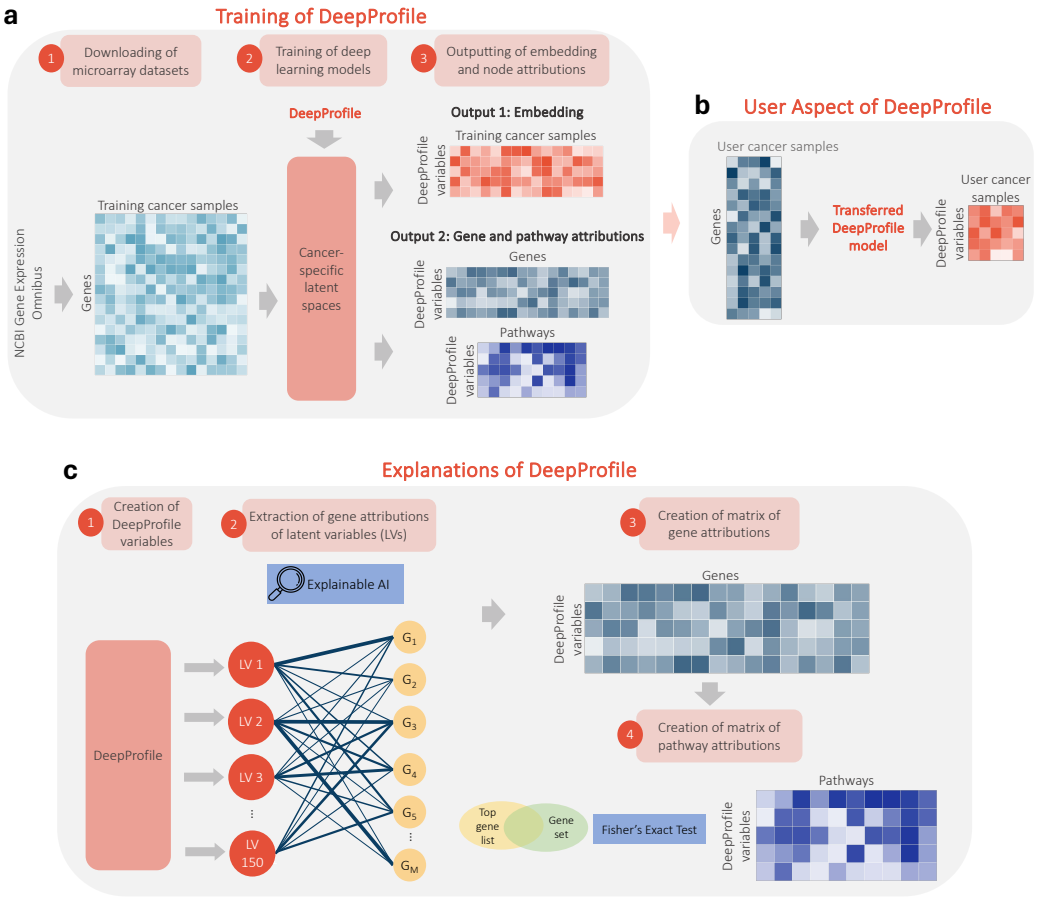
To investigate the association of average expression of the selected pathways with survival, we first calculated the average expression of genes from the KEGG mismatch repair pathway and the average expression of *HLA-D* genes (*HLA-DMA*, *HLA-DMB*, *HLA-DOA*, *HLA-DOB*, *HLA-DPA1*, *HLA-DPB1*, *HLA-DQA1*, *HLA-DQA2*, *HLA-DRB1*, *HLA-DRB5*) from the Reactome MHC class II antigen presentation pathway across all TCGA samples with survival record. We then created Kaplan-Meier plots [133] for each pathway using the calculated average expression values. We used the Python 'lifelines' library's 'KaplanMeierFitter' class for generating Kaplan-Meier survival plots. When generating the plots, we separated the patients into two groups based on their average expressions: one group with expression above mean + s.d. and one group with expression below -(mean + s.d.). We then fitted Kaplan-Meier models to these 2 groups. We also recorded the P values using the 'lifelines logrank_test', which tests how significantly the two curves are separated from each other.

To detect the immune cell type responsible from expression *HLA-D* genes, we first calculated the average expression of each gene across all TCGA samples for each cancer. Note that we performed the mean operation prior to preprocessing the expression matrices. We sorted the genes by their average expression and converted the rankings to percentile scores. To order the importance of different immune cell types, we calculated the average gene percentile score of the gene signatures for each immune cell type, which we defined as *XCR1* and *CLEC9A* for dendritic cells; *MS4A1*, *CD79A*, and *PAX5* for B cells; and *CD163*, *CD68*, *CSF1R* for macrophages. To measure the association between the immune cells and *HLA-D* expression, we measured the Pearson correlation between average expression of the cell type signatures listed above and the average expression of *HLA-D* genes using the Python 'scipy.stats' library's 'pearsonr' method.

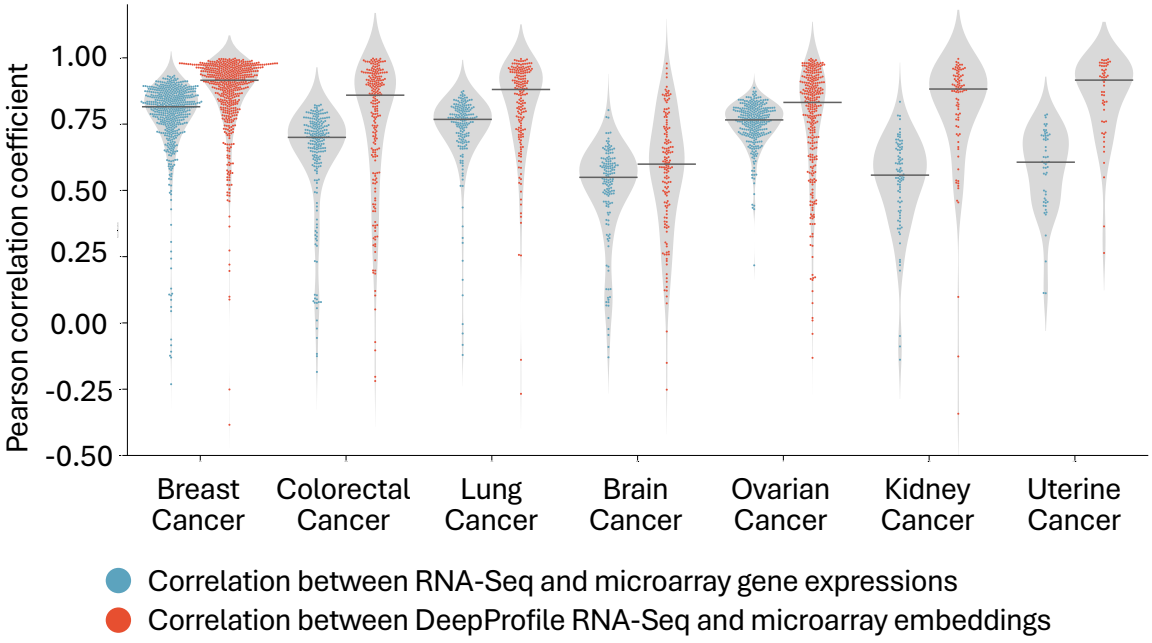
For the macrophage analysis, we downloaded the gene signatures for pro-inflammatory and immunosuppressive macrophages [69] and used the list of unique genes for each macrophage group (*CD40*, *CXCL9*, *CXCL10*, *CXCL11*, *SLAMF1*, and *TNIP3* for pro-inflammatory macrophages; *CFP*, *HRH1*, *NPL*, *PDCD1LG2*, and *RENBP* for immunosuppressive macrophages) to again measure gene ranking percentile scores. We conducted the same analysis as we did for immune cell types: we measured the average expression for genes, ranked the genes, and calculated the average percentile scores for the genes included in pro-inflammatory or immunosuppressive macrophage signatures. We also repeated this pro-inflammatory/immunosuppressive macrophage analysis using the extensive list of pro-inflammatory or immunosuppressive macrophage signatures [185].



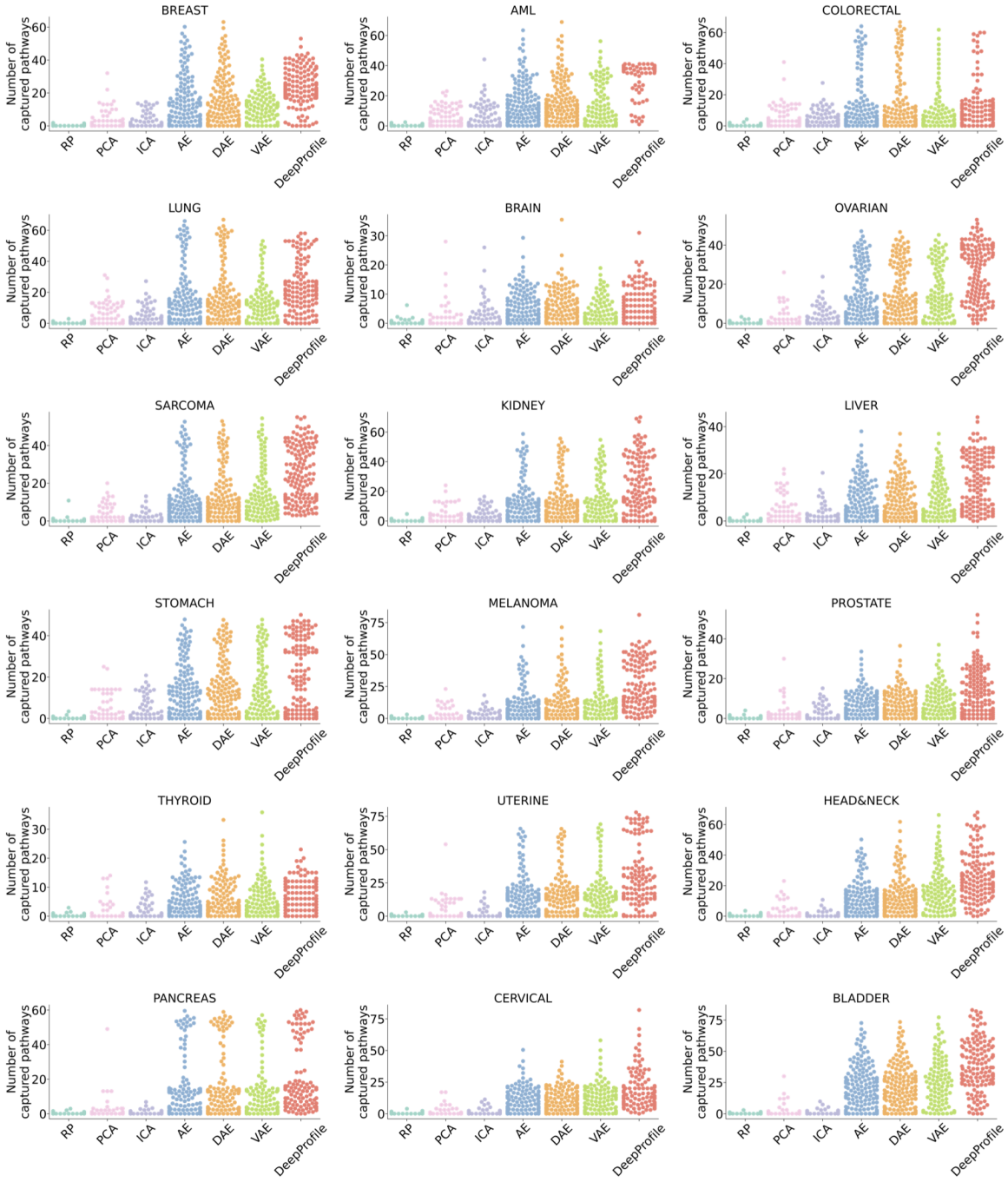
Extended Data Figure 6.9: **DeepProfile Pipeline (related to Figure ??)**. 1.Using the training cancer samples, we train hundreds of different Variational Autoencoder (VAE) models. We train each VAE model with a different latent dimension size 100 times with different random weight initializations. 2. We encode the training cancer samples using each of these VAE models to generate an embedding for each VAE model. Since VAE models have varying number of latent dimension sizes, the generated embeddings also have varying number of latent variables. 3. We cluster all latent variables of all VAE embeddings to group together latent variables that have similar patterns for the training samples. We cluster all VAE latent variables without any constraints which means that the latent variables from one VAE embedding can be placed in the same cluster or separate clusters. Thus, each cluster might contain latent variables from the same VAE embedding, VAE embeddings with the same latent space size from different random runs, or embeddings with different latent space sizes. 4. To define the final latent variable values, we average the latent variable values in a cluster and combine them to define the final DeepProfile embedding and latent variables. DeepProfile learns L latent variables where each latent variable is an ensemble of VAE latent variables from different models.



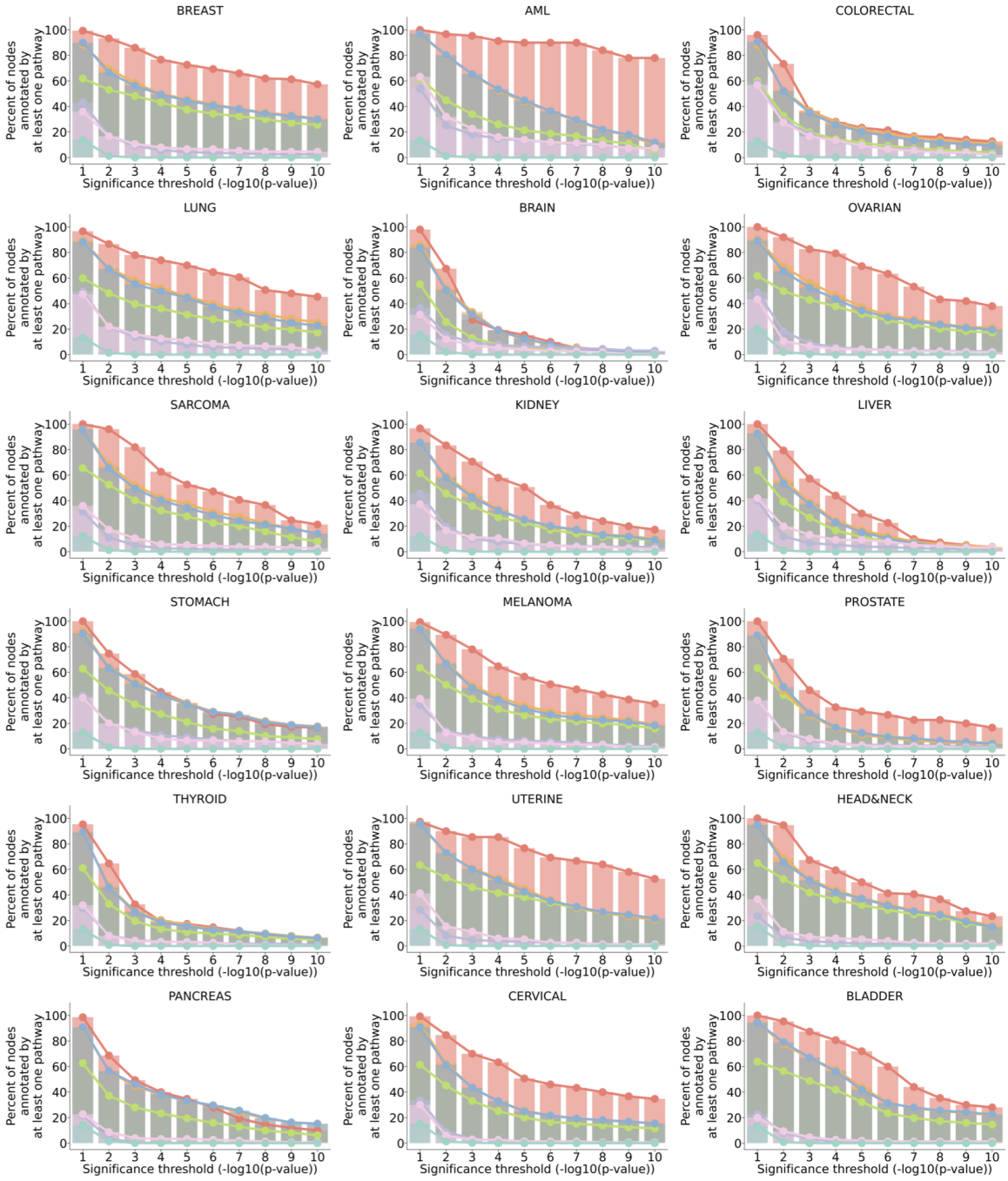
Extended Data Figure 6.10: **Overall framework of DeepProfile training, user aspect, and explanations (related to Figure ??). a. The training of DeepProfile.** 1. DeepProfile first collects publicly available expression dataset for each cancer type from NCBI GEO and defines an expression matrix, containing the expression measurements for all genes and samples. 2. DeepProfile is a deep learning model that uses all the cancer samples to learn cancer-specific latent spaces. 3. DeepProfile generates two outputs. Output 1 is an embedding of all the cancer samples where the number of DeepProfile latent variables is much smaller than the original number of genes passed to the model. Output 2 is the gene-level and pathway-level attributions for each DeepProfile latent variable. **b. The user aspect of DeepProfile.** DeepProfile model trained from thousands of cancer samples can be used to generate embeddings for new expression profiles. When an expression matrix of user cancer samples is passed to DeepProfile, it uses the already trained model and maps the user samples to the learned cancer-specific latent space and outputs an embedding for the user samples. **c. Explanations of DeepProfile latent variables.** 1. DeepProfile pipeline shown in a generates L latent variables. 2. For each latent variable we obtain gene attributions. Each DeepProfile latent variable is connected to genes through a set of fully or densely connected multilayer perceptron layers. We use Interpreter model (Figure **??**) to simplify the multilayer connections and define the weights connecting each gene to each DeepProfile latent variable. 3. The graph of gene-level attributions of DeepProfile latent variables in 2 is converted to a gene attribution matrix, containing the contribution of each DeepProfile latent variable to each gene. 4. From gene-level attributions, using the pathway memberships of genes, pathway-level attributions are obtained, containing the p-value indicating the significance of the overlap between gene sets and the most important genes of DeepProfile latent variables.



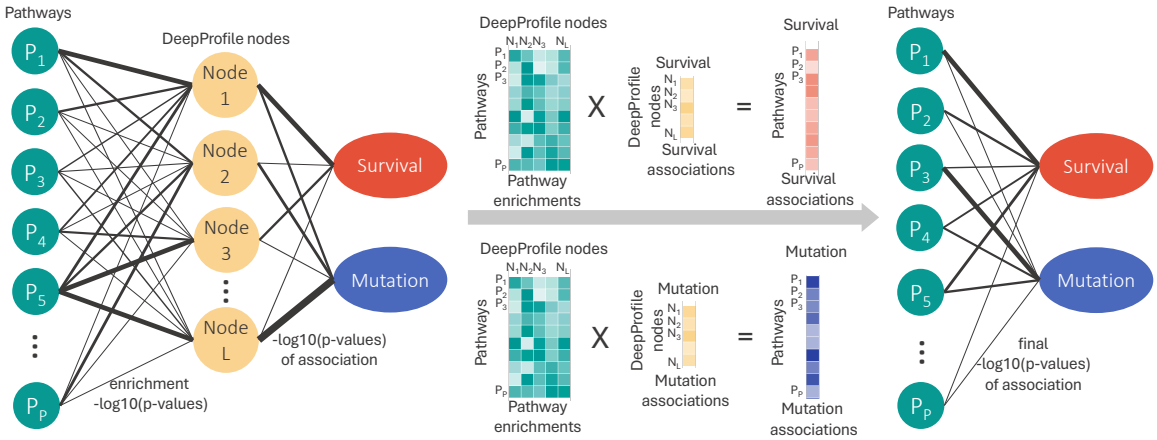
Extended Data Figure 6.11: **DeepProfile increases concordance between RNA-seq and microarray expression profiles.** Distribution of correlation coefficients between TCGA RNA-Seq and microarray expression profiles and correlation coefficients between TCGA RNA-Seq and microarray DeepProfile embeddings. For TCGA cancer samples with both RNA-Seq and microarray expression available, we generated DeepProfile embeddings separately for RNA-Seq and microarray data. We measured the correlation between the DeepProfile RNA-Seq and microarray embeddings on per-sample basis. We also measured the correlation between original expression levels from RNA-Seq and microarray expression matrices. We generated distribution plots of Pearson correlation coefficients for DeepProfile embeddings (red dots) and original expression measurements (blue dots). Each dot represents the correlation coefficient for one cancer sample and the distributions are plotted for 7 cancer types. The median correlation coefficient for each sub distribution is shown with the dark colored bars. DeepProfile embedding scores are significantly higher than the original expression scores (Wilcoxon sign-ranked test $P < 0.005$ for all 7 cancer types).



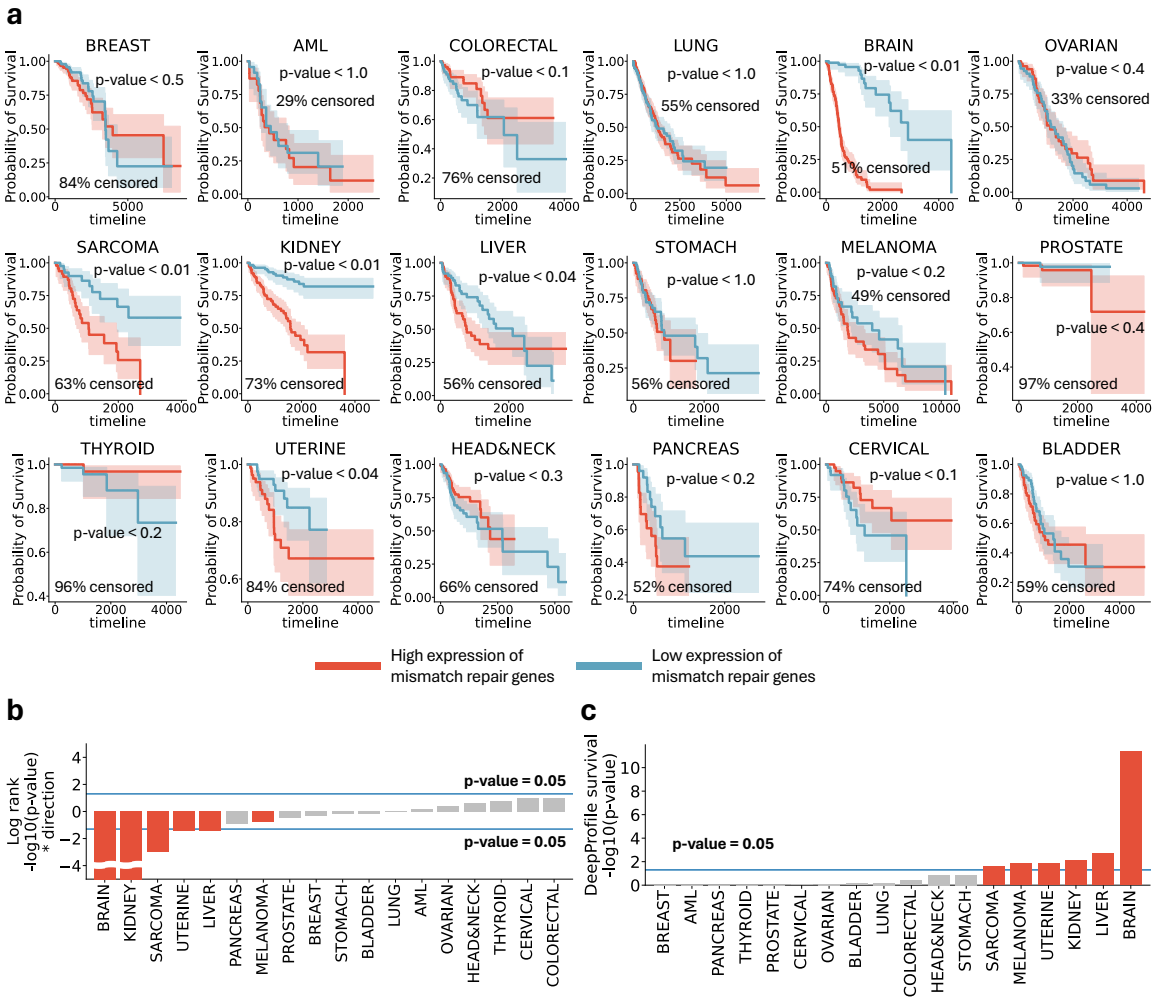
Extended Data Figure 6.12: **Distribution plots of pathway coverage (related to Figure ??).** Distribution plots of number of KEGG, BioCarta, Reactome pathways significantly captured (FDR-corrected $P < 0.05$) by each latent variable of embeddings generated by DeepProfile and other methods are shown for all 18 cancer types.



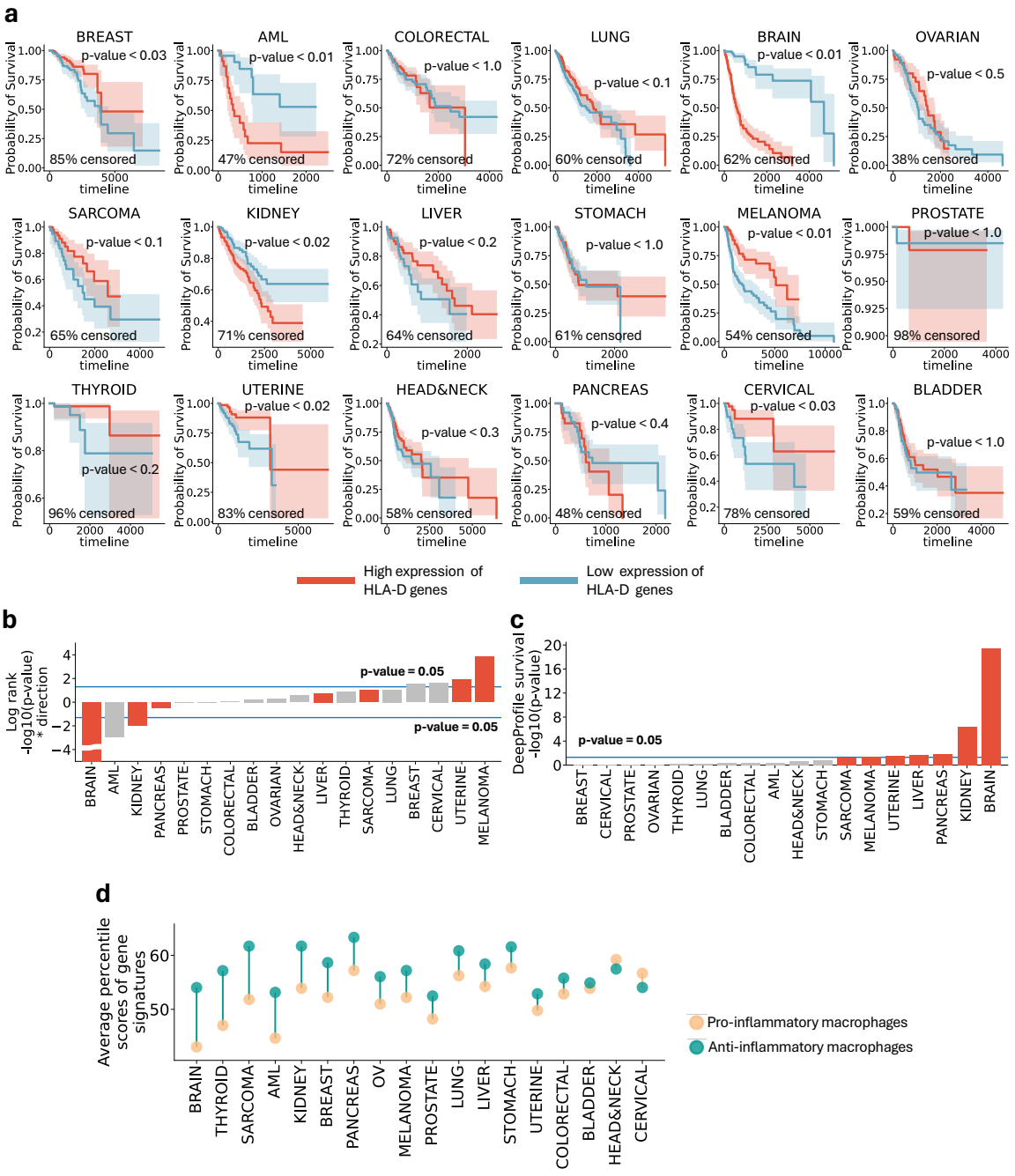
Extended Data Figure 6.13: **Comparison of biologically annotated latent variables (related to Figure ??).** Comparison of the percent of latent variables annotated by at least one pathway above the significance threshold. The percent of annotated latent variables are shown for multiple significance thresholds for DeepProfile and alternative dimensionality reduction methods. The plots are shown for all 18 cancer types.



Extended Data Figure 6.14: **The schematic of DeepProfile survival and mutation analysis at pathway level (related to Figure ??).** Each pathway is connected to each DeepProfile latent variable with a certain enrichment score ($-\log_{10}$(P value)) as we extracted pathway-level explanations for DeepProfile latent variables before. We then fit univariate Cox survival regression models to each DeepProfile latent variable and obtain a p-value denoting the significance of association of a latent variable with survival. We also measure the Pearson correlation between each DeepProfile latent variable and tumor mutational burden (TMB) and obtain a P value denoting the significance of association of a latent variable with TMB. In order to calculate the overall pathway-level survival and mutation association scores, we take the inner product of enrichment and latent variable- level association matrices and normalize the matrix. This way, we obtain the final ($-\log_{10}$(P value)) of survival and mutation association for each pathway. We repeated this process for each cancer type which allows us to carry cancer common and specific survival and mutation analyses.



Extended Data Figure 6.15: **Mismatch-repair-pathway survival analysis (related to Figure ??).** a. Kaplan-Meier plots of average expression of KEGG mismatch repair pathway. The samples with an expression above (mean + 1 s.d.) are marked as highly expressed and below -(mean + 1 s.d.) are marked as lowly expressed. The shaded areas represent the confidence intervals. The log rank test P values and the percent of censored samples are reported for each cancer. b. Plot of log rank test P values and the direction from Kaplan-Meier plots. The cancer types are sorted by the direction of association and the $-\log_{10}$(P value). c. Plot of DeepProfile survival p-values for KEGG mismatch repair pathway. The cancer types are sorted by the $-\log_{10}$(P value).



Extended Data Figure 6.16: **MHC-class-II-pathway survival analysis (related to Figure ??).** a. Kaplan-Meier plots of average expression of HLA-D genes in Reactome MHC class II antigen presentation pathway. The samples with an expression above (mean + 1 s.d.) are marked as highly expressed and below -(mean + 1 s.d.) are marked as lowly expressed. The shaded areas represent the confidence intervals. The log rank test P values and the percent of censored samples are reported for each cancer. b. Plot of log rank test P values and the direction from Kaplan-Meier plots. The cancer types are sorted by the direction of association and the $-\log_{10}$(P value). c. Plot of DeepProfile survival P values for Reactome MHC class II antigen presentation pathway. The cancer types are sorted by the $-\log_{10}$(P value). d. Comparison of average percentile scores of pro- and anti- inflammatory macrophages shown for 18 cancers.

7

AN EXPLAINABLE AI FRAMEWORK FOR IDENTIFYING UNIVERSAL AGING SIGNATURES IN CELL EMBEDDINGS

7.1 INTRODUCTION

Aging is a complex, multifactorial process characterized by the progressive loss of physiological integrity, which leads to functional decline and increased vulnerability to disease [166]. It is the primary risk factor for a wide range of chronic diseases, including cancer, diabetes, cardiovascular disorders, and neurodegenerative conditions [137, 200]. Unraveling the biological mechanisms that drive aging is essential for identifying therapeutic targets and biomarkers that can promote healthy longevity and delay aging-related diseases [120, 191]. Although numerous studies have uncovered molecular and cellular changes associated with aging [262, 306], identifying the key drivers remains one of the field's most pressing challenges [282].

The advent of single-cell transcriptomic technologies has enabled unprecedented resolution in exploring the aging process. By capturing gene expression profiles at the individual cell level, single-cell RNA sequencing (scRNA-seq) facilitates the discovery of cell-type-specific and context-dependent aging signatures [3, 173, 244]. However, accurately characterizing aging signatures from scRNA-seq data remains challenging due to the other non-aging-related sources of variation such as tissue, cell type, sex, and batch effects. These background factors often obscure the more subtle transcriptomic changes associated with aging, hindering efforts to isolate true aging-related signatures. To address this complexity, previous studies have developed separate linear models for different tissues and cell types [35, 183, 336], an approach that requires training many tissue- and cell-specific models and risks overlooking globally conserved aging genes.

Recent single-cell representation learning methods, including probabilistic latent variable models and contrastive approaches, capture broad cellular heterogeneity but generally do not disentangle subtle aging-related signatures from other dominant biological factors [4, 75, 92, 168, 170, 310, 311]. These frameworks often rely on a single latent space or on binary case–control contrasts, which do not reflect the continuous nature of aging. In recent disentanglement frameworks [222], all cells belonging to the same category, such as a specific cell type or condition, share a single embedding, making it difficult to capture gradual, continuous processes such as aging. In parallel, Patches [25] was developed to disentangle discrete condition-specific variation without explicitly supervising background factors. Together, these limitations motivate an approach that explicitly models aging-related and background variation.

To address these challenges, we introduce ACE (Aging Cell Embeddings), a representation learning framework designed to disentangle aging-related signatures from background biological variation in scRNA-seq data (Figure **??**a). ACE models gene expression using two distinct sets of latent variables: *age variables*, which capture variations related to age, and *background variables*, which capture other non-aging-related signatures (Figure **??**b). The framework supports both global (*i.e.*, shared across tissues and cell types) and local (*i.e.*, tissue-cell-type-specific) aging analyses, enabling the identification of both shared and context-specific aging trajectories.

We apply ACE to large-scale single-cell aging atlases, including datasets from mouse [3], fly [173], and human [236]. Our results show that ACE effectively isolates aging-related signatures and enables several key applications (Figure **??**c): (1) identifying global and local aging genes using explainable artificial intelligence (XAI) methods; (2) predicting biological age at both the cell and subject levels across diverse contexts; and (3) analyzing conserved aging signatures across species, including mouse, fly, and human, by aligning aging trajectories and identifying conserved aging genes. We further validate ACE's robustness through cross-cell-type generalization, highlighting its utility in uncovering context-independent aging mechanisms. Finally, we experimentally confirmed ACE's findings through RNAi knockdown experiments in *C. elegans*, where multiple prioritized genes significantly impacted lifespan. Notably, ACE identified *Uba52* as an aging-associated gene across species, and its role was experimentally validated, demonstrating ACE's ability to uncover biologically conserved aging mechanisms. The implementation of the ACE model is publicly available online.[1]

7.2 RESULTS

7.2.1 *ACE effectively disentangles aging signatures from single-cell RNA-seq data*

We designed ACE (Aging Cell Embeddings), an explainable variational autoencoder-based model, to isolate aging-related gene expression variations in single-cell RNA sequencing datasets (Figure **??**a). scRNA-seq datasets are inherently complex, with gene expression signatures reflecting diverse biological factors including species, tissue, cell type, sex, and technical artifacts. ACE is a representation learning framework [169] that explicitly models and disentangles aging-related variations from other non-aging-related factors in scRNA-seq data. It enables the extraction of conserved aging signatures from heterogeneous biological contexts.

The architecture of ACE (Figure **??**b; **??**) comprises two parallel encoder networks. The age encoder extracts *age embeddings* that specifically capture aging-related signatures. Simultaneously, the background encoder captures embeddings that represent non–aging-related variation, such as species, tissue, cell type, and sex (*i.e.*, background factors). To

1 https://github.com/suinleelab/ACE

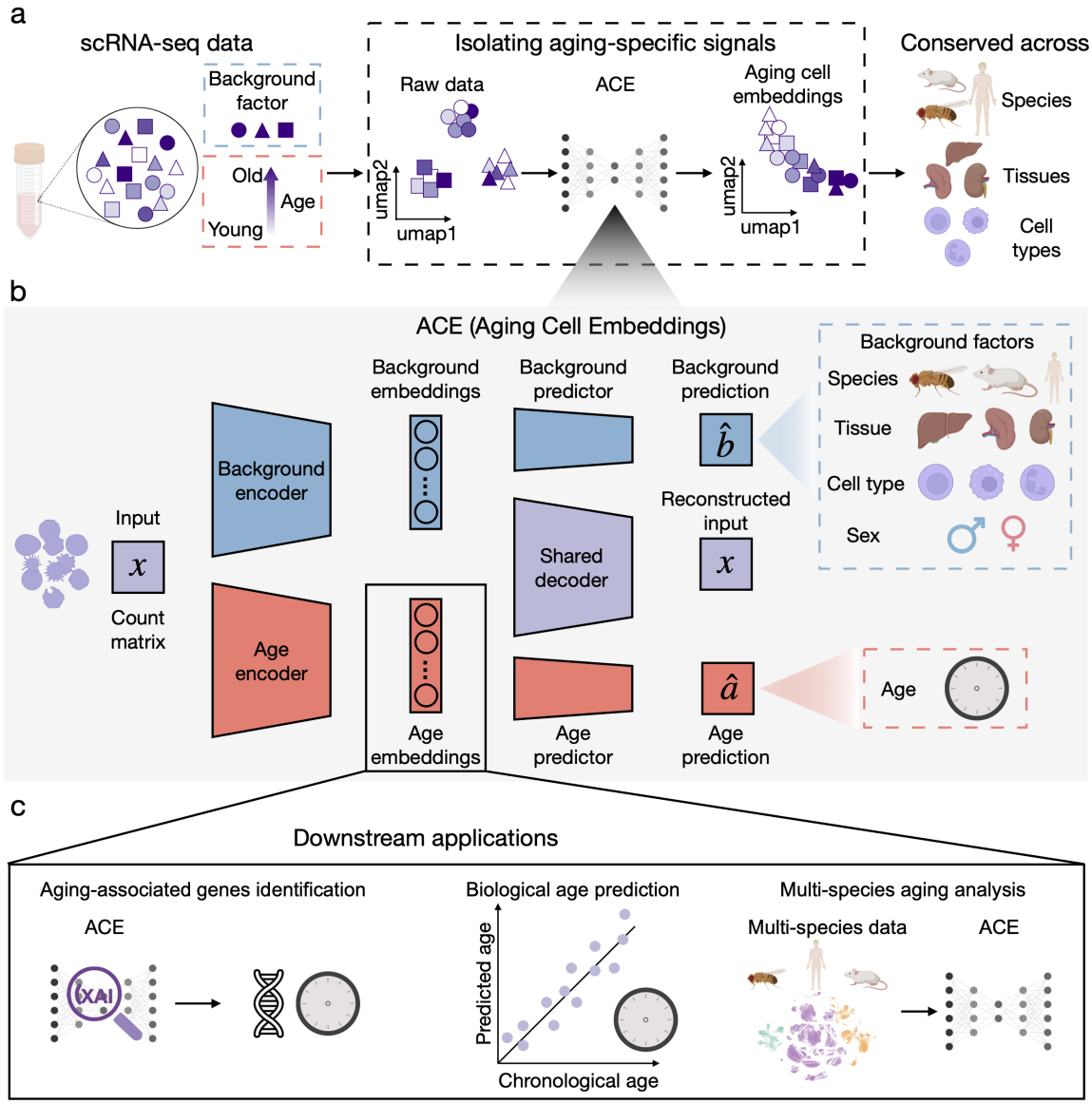


Figure 7.1: **The ACE Framework. (a)** scRNA-seq datasets contain information about both aging signatures and biological factors like tissue, cell type, and sex. The goal of ACE is to disentangle aging from other background factors. **(b)** ACE model structure. Each cell is represented by two latent variables: age variables and background variables, which encode aging signatures and background factors, respectively. Age variables predict the subjects' chronological ages, while background variables predict background factors such as tissue, cell type, and sex. The Hilbert-Schmidt independence criterion (HSIC) enforces independence between the two embeddings. A decoder reconstructs the gene expression counts. **(c)** Downstream applications of the ACE model. The age variables enable a variety of downstream analyses, including explainable aging gene identification, biological age prediction, and conserved aging signature analysis across species.

ensure effective separation of these embeddings, ACE incorporates prediction networks: age embeddings are used to predict the subjects' chronological age, while background embeddings predict background factors. The two sets of embeddings are constrained to remain independent using the Hilbert-Schmidt Independence Criterion (HSIC) [99, 220, 275], ensuring that the learned biological signatures remain distinct. To reconstruct gene expression while accounting for technical variation, ACE includes a decoder that models the observed RNA measurements using distributions conditioned on the background and age embeddings. These distributions account for technical factors such as sequencing depth, batch effects, and dropouts by following the probabilistic modeling approach used in scVI [168].

The age variables learned by ACE enable a variety of downstream applications essential to aging research (Figure **??**c). Specifically, the age embeddings support: (1) identification of global (*i.e.*, shared across tissues and cell types) and local (*i.e.*, tissue-cell-type-specific) aging genes using explainable AI methods, (2) accurate estimation of biological age, defined as the prediction of aging based on gene-expression profiles, at both the cell and subject levels, and (3) analysis of conserved aging signatures across different species by aligning their aging trajectories. These applications underscore the model's utility in systematically disentangling aging signatures from complex scRNA-seq data.

7.2.2 *ACE effectively captures global aging trajectories*

Aging manifests through both global and local molecular programs: some transcriptional signatures are shared across diverse tissues and cell types, whereas others are restricted to particular cellular contexts. ACE supports global aging analysis by explicitly modeling aging-related variation that is consistent across tissues and cell types (Figure **??**a). For this analysis, ACE was trained separately on each species (*e.g.*, mouse and fly), with tissue, cell type, and sex as background factors, allowing the age embeddings to capture variation specifically associated with aging while controlling for other biological sources of heterogeneity. We evaluated ACE's ability to disentangle global aging signatures using two cross-tissue, multi-cell-type aging atlases — the Tabula Muris Senis (TMS) mouse dataset [3] and the Aging Fly Cell Atlas [173]. We applied Expected Gradients [76] to the global ACE model to identify global aging genes that contribute significantly to aging across tissues and cell types.

We first trained ACE on the TMS dataset generated using droplet-based scRNA-seq, including cells from mice across different ages (1m, 3m, 18m, 21m, 24m, 30m) (Figure **??**b). We focus the analysis on the 20 most commonly occurring cell types. In total, the data we use contains 135,420 cells taken from 13 different tissues across 23 subjects (**??**). The raw data's dominant biological variations include cell type, tissue, and sex (Supplementary Figure **??**). An effective global aging model should learn the age embeddings that capture aging-related variations while retaining other biological variations unrelated to

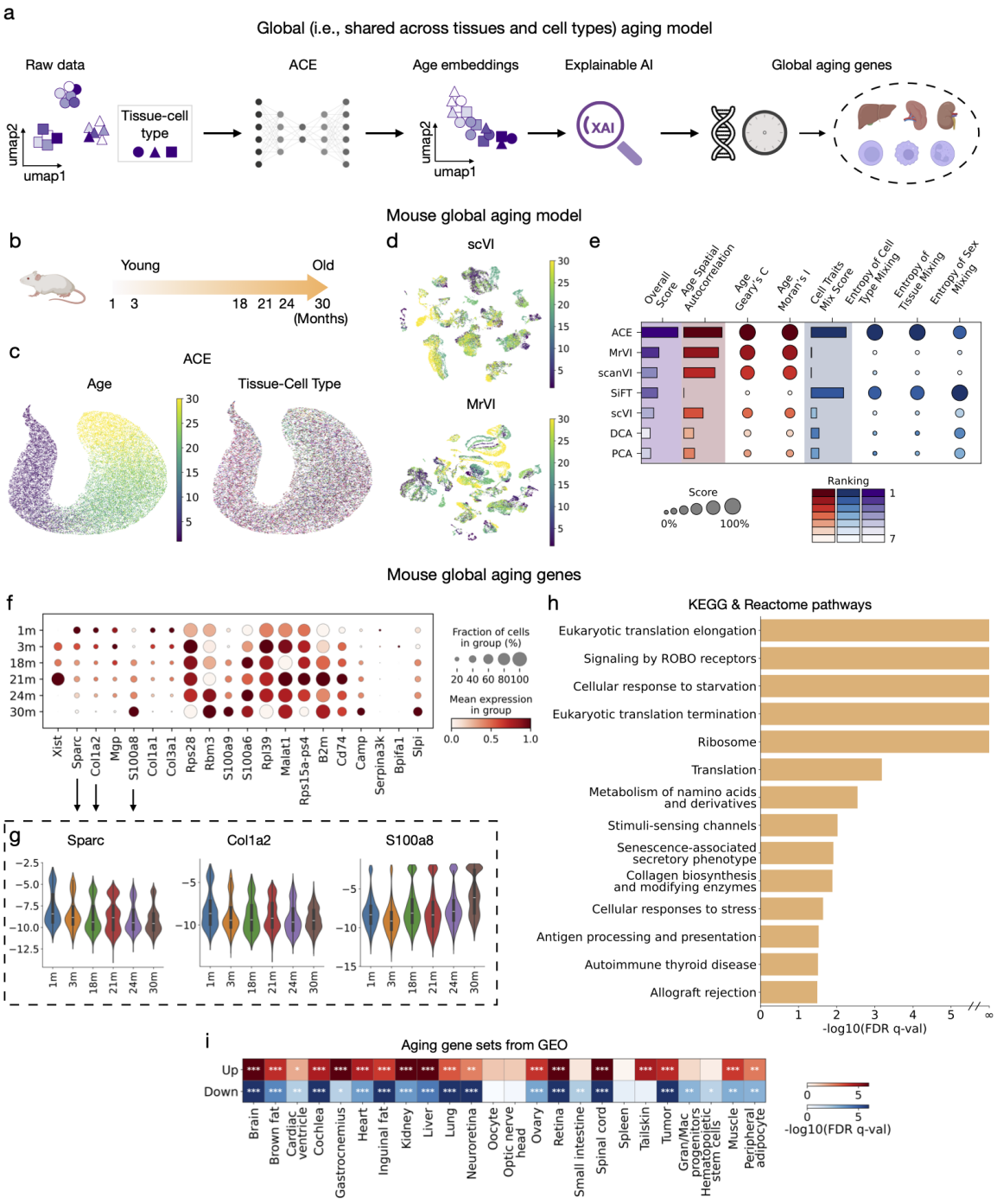

a
Global (i.e., shared across tissues and cell types) aging model
Raw data
Tissue-cell type
ACE
Age embeddings
Explainable AI
Global aging genes
umap2
umap1
XAI
Mouse global aging model
b
Young
Old
1 3 18 21 24 30
(Months)
c
ACE
Age
Tissue-Cell Type
d
scVI
MrVI
e
Overall Score
Age Spatial Autocorrelation
Age Geary's C
Age Moran's I
Cell Traits Mix Score
Entropy of Cell Type Mixing
Entropy of Tissue Mixing
Entropy of Sex Mixing
ACE
MrVI
scanVI
SiFT
scVI
DCA
PCA
Score
0%
100%
Ranking
Mouse global aging genes
f
1m
3m
18m
21m
24m
30m
Xist
Sparc
Col1a2
Mgp
S100a8
Col1a1
Col3a1
Rps28
Rbm3
S100a9
S100a6
Rpl39
Malat1
Rps15a-ps4
B2m
Cd74
Camp
Serpina3k
Bpifa1
Slpi
Fraction of cells in group (%)
20 40 60 80100
Mean expression in group
0.0 0.5 1.0
g
Sparc
Col1a2
S100a8
h
KEGG & Reactome pathways
Eukaryotic translation elongation
Signaling by ROBO receptors
Cellular response to starvation
Eukaryotic translation termination
Ribosome
Translation
Metabolism of namino acids and derivatives
Stimuli-sensing channels
Senescence-associated secretory phenotype
Collagen biosynthesis and modifying enzymes
Cellular responses to stress
Antigen processing and presentation
Autoimmune thyroid disease
Allograft rejection
-log10(FDR q-val)
i
Aging gene sets from GEO
Up
Down
Brain
Brown fat
Cardiac ventricle
Cochlea
Gastrocnemius
Heart
Inguinal fat
Kidney
Liver
Lung
Neuroretina
Oocyte
Optic nerve head
Ovary
Retina
Small intestine
Spinal cord
Spleen
Tailskin
Tumor
Gran/Mac progenitors
Hematopoietic stem cells
Muscle
Peripheral adipocyte
-log10(FDR q-val)

Figure 7.2

Figure 7.2: **ACE enables global (*i.e.*, shared across tissues and cell types) modeling of aging trajectories. a.** Overview of the global aging model pipeline. ACE is applied to scRNA-seq data to learn aging-related cell embeddings shared across tissues and cell types. XAI techniques are then used to identify aging-associated genes across tissues and cell types. **b.** Visualization of aging timepoints available in the TMS Droplet mouse dataset, from 1 to 30 months, capturing the life span of a mouse. **c.** UMAPs of the age embeddings learned by ACE, colored by age (left) and tissue-cell-type identity (right). The age embeddings capture aging-related signatures while being disentangled from background factors. **d.** UMAPs of latent embeddings learned by alternative methods: scVI (top) and MrVI (bottom), colored by age. **e.** Quantitative comparison of ACE and baseline methods (MrVI, scanVI, SiFT, scVI, DCA, and PCA) across multiple evaluation metrics. Circles represent individual metric scores, with circle size indicating the normalized score (scaled between 0 and 1) and color reflecting the relative ranking. For consistency, all metrics are scaled such that higher values indicate better performance. Bars represent composite scores: Age Spatial Autocorrelation is the average of Geary's C and Moran's I; Cell Traits Mix Score is the average of cell type, tissue, and sex mixing metrics; and Overall Score is the average of Age Spatial Autocorrelation and Cell Traits Mix Score. Bar color indicates the method's relative ranking. **f.** Dot plot showing expression patterns of top 20 global aging genes identified by the ACE global aging model in the TMS Droplet dataset. Dot size indicates the fraction of cells expressing the gene in each age group; color represents mean expression. **g.** Violin plots showing expression patterns of representative global aging genes (*Sparc*, *Col1a2*, *S100a8*). Violin plots display gene expression distributions across different age groups, aggregated over all tissues and cell types. These visualizations highlight consistent aging-associated expression changes for the selected genes. **h.** Gene set enrichment analysis of the full ranked list of global aging genes reveals KEGG and selected Reactome pathways that are significantly enriched and play key roles in aging-related biological processes. Significance was assessed at FDR $q < 0.05$ using the Benjamini-Hochberg correction. **i.** Gene set enrichment analysis using the "Aging Perturbations from GEO Up" and "Aging Perturbations from GEO Down" gene set collections. These databases consist of gene sets curated from GEO studies comparing aged versus young samples, capturing genes consistently upregulated or downregulated with age across various tissues and cell types. ACE-derived ranked list of global aging genes shows significant enrichment in many of these aging-associated gene sets. Red boxes indicate enrichment in *upregulated* gene sets; blue boxes indicate enrichment in *downregulated* gene sets. Color intensity reflects $-\log_{10}$(FDR q-value), with significance determined by FDR-adjusted q-values using the Benjamini-Hochberg correction. Asterisks denote significance thresholds (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$). Gran/Mac progenitors represent granulocyte/-macrophage progenitors.

aging in the background embeddings. We compared ACE with existing baseline models, including MrVI [32], scANVI [325], which utilize the age labels during training, SiFT [223], which uses the background labels during training, and scVI [168], DCA [75], and PCA, which are unsupervised single-cell representation learning methods. UMAP of ACE's age embeddings reveals a smooth and continuous aging trajectory (Figure **??**c, left), while tissue-cell-types are well mixed in the same space (Figure **??**c, right), indicating that ACE effectively disentangles aging signatures from background factors and captures global aging signatures shared across tissues and cell types (see Supplementary Figure **??** for UMAPs of both age and background embeddings colored by age and background factors).

By contrast, embeddings from baseline methods, such as scVI and MrVI, exhibited significant confounding due to tissue and cell type variations, obscuring aging trajectories (Figure **??**d). A quantitative comparison further shows that ACE outperforms baseline methods across multiple metrics by a large margin, including spatial autocorrelation between the age embeddings and chronological age, as well as mixing scores for cell type, tissue, and sex (Figure **??**e; **??**). Overall, ACE consistently ranks highest in capturing aging signatures while minimizing confounding from background factors.

We applied Expected Gradients to the global ACE model to identify genes contributing most strongly to aging-related variation (**??**). Among the top 20 global aging genes (Figure **??**f), 16 have prior reports of aging-associated regulation: *Xist* [100], *Sparc* [248], *Col1a2* [263], *Mgp* [20], *S100a8* [197, 334], *Col1a1* [187], *Col3a1* [187], *Rps28* [122], *Rbm3* [110], *S100a9* [334], *S100a6* [71], *Malat1* [245], *Rps15a-ps4* [238] *B2m* [264], *Cd74* [124], and *Serpina3k* [161]. The remaining genes are enriched for pathways related to innate immune activation (*Camp* [322], *Bpifa1* [7], and *Slpi* [180]), suggesting plausible, but less systematically characterized, links to aging.

Several of these (*e.g.*, *Sparc*, *Col1a2*, and *S100a8*) display clear, monotonic expression shifts across the lifespan (Figure **??**g). *Sparc* functions as an immunometabolic checkpoint that promotes inflammation and interferon signaling, and its inhibition has been shown to reduce inflammation and extend health span during aging [248]. The *Col1a2* gene encodes for a component of Type I Collagen that is present in the extracellular matrix (ECM) of various tissues. ECM dynamics have been increasingly identified as important hallmarks of aging [266]. *Col1a2* has been identified as a key player in skin, bone, and cardiac health in various mouse disease models [146, 261, 273], suggesting that aging-associated changes in expression of *Col1a2* could lead to disruption of the health and homeostasis of these organ systems. *S100a8* is a pro-inflammatory gene upregulated with age in multiple tissues, where it promotes oxidative stress, inflammation, and cellular senescence, making it a potential driver and biomarker of aging [334]. The identification of these genes by ACE supports its ability to uncover biologically meaningful and well-conserved markers of aging.

We then performed gene set enrichment analysis using the full ranked list of global aging genes and identified multiple significantly enriched KEGG and Reactome pathways (Figure **??**h; see Supplementary Figure **??** for the full list of enriched pathways; **??**). These include translation-related processes (*e.g.*, ribosome, eukaryotic translation elongation and termination), cellular stress responses (*e.g.*, response to starvation), and immune-related pathways (*e.g.*, antigen processing and presentation, senescence-associated secretory phenotype), extracellular matrix pathways (*e.g.*, collagen biosynthesis and modifying enzymes) and axonal pathfinding (*e.g.*, signaling by ROBO receptors). These findings underscore the relevance of proteostasis, stress adaptation, and immune regulation, cell-ECM interactions, and the nervous system in the aging process. Furthermore, the ranked gene list of global aging genes show significant enrichment across aging-associated gene sets curated from GEO, which are derived from comparisons between aged and young samples (Figure **??**i). The broad enrichment across diverse tissues and cell types highlights ACE's ability to capture conserved aging signatures shared across multiple biological contexts.

To externally validate the mouse global aging model, we applied ACE to the independent TMS FACS dataset (**??**). Similar to the Droplet data, the raw FACS data are dominated by tissue, cell type, and sex signatures, with age effects being comparatively subtle (Supplementary Figure **??**). ACE effectively disentangles these factors, producing age embeddings that capture a smooth aging trajectory (Supplementary Figure **??**b, Supplementary Figure **??**). The global aging gene rankings from the FACS model are highly concordant with those from the Droplet model (Supplementary Figure **??**c,d). Gene set enrichment analysis recapitulates key aging-associated pathways, including translation, stress response, and ROBO signaling (Supplementary Figure **??**e, Supplementary Figure **??**), and shows strong overlap with external GEO aging signatures (Supplementary Figure **??**f), demonstrating the robustness and reproducibility of ACE across datasets.

We next evaluated ACE's ability to extract global aging signatures on the Aging Fly Cell dataset [173], which includes cells spanning ages from 5 to 70 days (Figure **??**a; **??**). The data we use contains 424,863 cells from two tissues (head and body) and 14 cell types across 30 subjects (see Supplementary Figure **??** for the UMAPs of the raw data). Consistent with mouse data, the UMAP of ACE's age embeddings exhibits a continuous aging trajectory that is largely independent of tissue and cell type differences (Figure **??**b; see Supplementary Figure **??** for UMAPs of both age and background embeddings colored by age and background factors). ACE outperforms baseline models by best capturing aging-related variations while minimizing confounding effects from tissue, cell type, and sex (Figure **??**c,d). ACE also identifies global aging genes with consistent aging-associated expression patterns (Figure **??**e,f), including known aging-associated markers such as *lovit*, *Pzl*, and *Hsp26*. *lovit* is a synaptic protein that plays a critical role in transporting the neurotransmitter, histamine, in photoreceptor cells in insects and other arthropods. RNAi-mediated downregulation of *lovit* in flies resulted in disrupted visual transmission

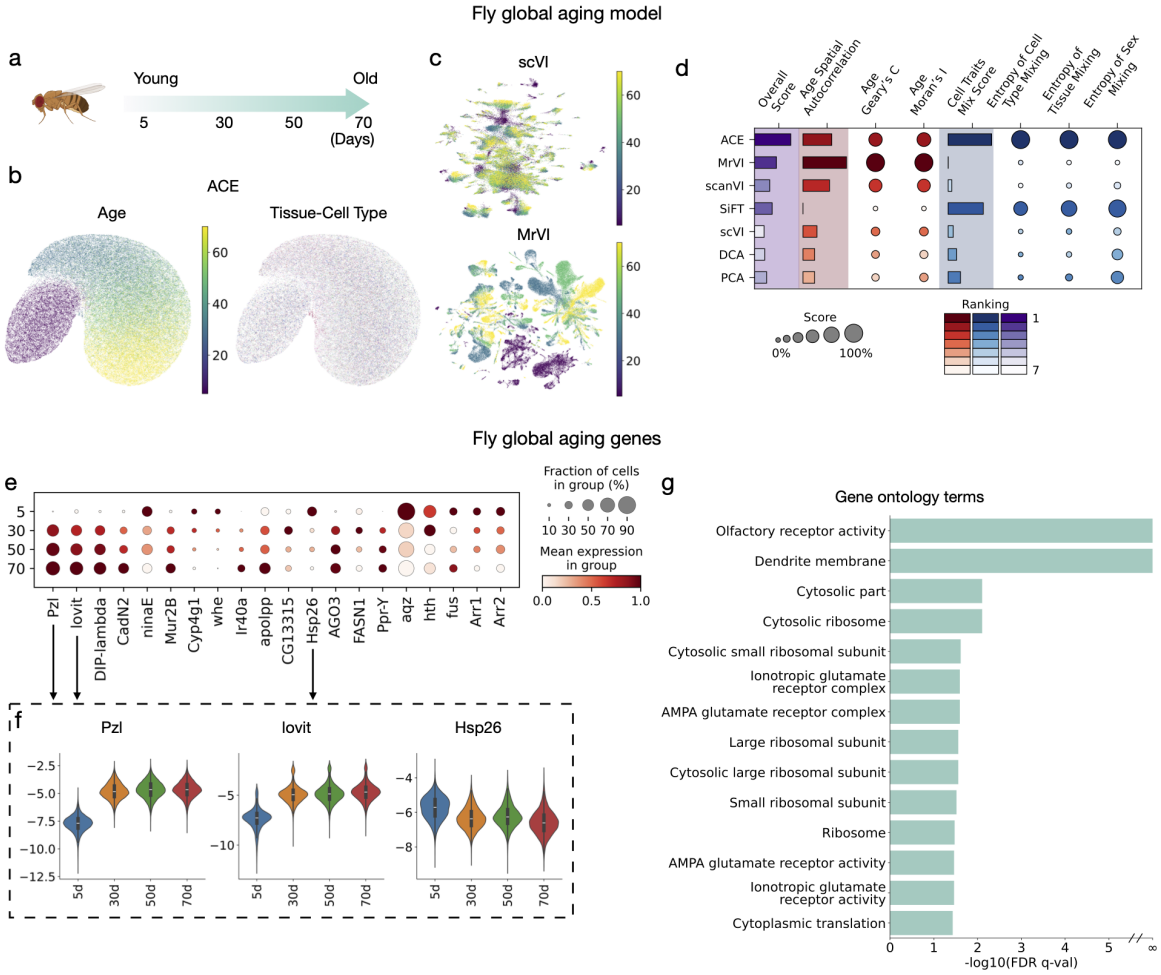


Figure 7.3: **ACE enables global (*i.e.*, shared across tissues and cell types) modeling of aging trajectories in fly. a.** Visualization of the aging timepoints available in the Aging Fly Cell Atlas (AFCA) dataset, ranging from 5 to 70 days, covering the lifespan of a fly. **b.** UMAPs of the age embeddings learned by ACE, colored by age (left) and tissue-cell-type identity (right). The age variables capture aging-related signatures while being disentangled from background factors. **c.** UMAPs of latent variables learned by alternative methods: scVI (top) and MrVI (bottom), colored by age. **d.** Quantitative comparison of ACE and baseline methods (MrVI, scanVI, SiFT, scVI, DCA, and PCA) across multiple evaluation metrics. Circles represent individual metric scores, with circle size indicating the normalized score (scaled between 0 and 1) and color reflecting the relative ranking. For consistency, all metrics are scaled such that higher values indicate better performance. Bars represent composite scores: Age Spatial Autocorrelation is the average of Geary's C and Moran's I; Cell Traits Mix Score is the average of cell type, tissue, and sex mixing metrics; and Overall Score is the average of Age Spatial Autocorrelation and Cell Traits Mix Score. Bar color indicates the method's relative ranking. **e.** Dot plot showing expression patterns of top 20 global aging genes identified by the ACE global aging model in the AFCA. Dot size indicates the fraction of cells expressing the gene in each age group; color represents mean expression. **f.** Violin plots showing expression patterns of representative global aging genes (*Iovit*, *Pzl*, and *Hsp26*) in AFCA data. Violin plots display gene expression distributions across different age groups, aggregated over all tissues and cell types. These visualizations highlight consistent aging-associated expression changes for the selected genes. **g.** Gene set enrichment analysis using the full ranked list of global aging genes from AFCA data identified significantly enriched Gene Ontology terms that play key roles in aging-related biological processes. Significance was assessed at FDR $q < 0.05$ using the Benjamini-Hochberg correction.

[327]. *Pzl* is expressed in specific locomotor neurons in flies, where mutant variants of *Pzl* result in disrupted locomotor function in fly larva [119]. *Hsp26*, which encodes a small heat shock protein, has been linked to increased resistance to oxidative stress and extended lifespan in flies [158]. Gene Ontology (GO) enrichment analysis of these genes highlights aging-associated pathways (Figure **??**g). Similar to the mouse global model, the fly global model identifies several pathways involved in protein translation (*e.g.*, Cytosolic ribosome, large ribosomal subunit, cytoplasmic translation) and synaptic function (*e.g.*, Dendrite membrane, ionotropic glutamate receptor complex, and AMPA glutamate receptor complex). The agreement between these two analyses highlights the importance of proteostasis and nervous system function in aging.

Altogether, the results demonstrate that ACE effectively disentangles global aging signatures from dominant non-aging background factors. Using the XAI method, ACE not only identifies key global aging genes with conserved expression patterns but also uncovers biologically meaningful pathways and gene sets associated with aging. Consistent results across both mouse and fly datasets highlight ACE's robustness and its ability to reveal aging pathways that are consistent across species at single-cell resolution.

7.2.3 *ACE effectively captures local aging trajectories*

While global aging analysis reveals shared, conserved aging signatures, many aging processes are tissue- and cell-type-specific, reflecting the distinct functions and environments of different biological systems. Local (*i.e.*, tissue–cell-type–specific) analysis at single-cell resolution is therefore critical for uncovering these context-dependent aging signatures that global models may overlook. As shown in Figure **??**a, ACE can also be configured as a local aging model by excluding tissue and cell type from the background factors and using only sex (if available) in the background network. For training the age embedding, we continue to use only the age variable as the predictor. This setup allows the age embeddings to encode both aging-related variation and tissue-cell-type-specific structure, enabling ACE to learn fine-grained, local aging signatures. We applied this configuration to the TMS Droplet dataset, where the resulting UMAPs of ACE's local age embeddings reveal distinct, coherent trajectories within each tissue–cell-type cluster (Figure **??**b,c; Supplementary Figure **??**). These results indicate that ACE effectively models tissue-cell-type-specific aging trajectories within a unified framework, without requiring explicit supervision on tissue or cell type

We apply the Expected Gradients (EG) to identify local aging genes for four representative tissue-cell-type pairs, including limb muscle mesenchymal stem cells, lung classical monocytes, kidney proximal convoluted tubule epithelial cells, and heart endothelial cell of coronary artery (Figure **??**d, top). For each pair, we calculate the difference between a gene's percentile score (*i.e.*, the relative importance of a gene for a tissue-cell-type pair as estimated by EG) in the target tissue-cell-type and its highest percentile score across

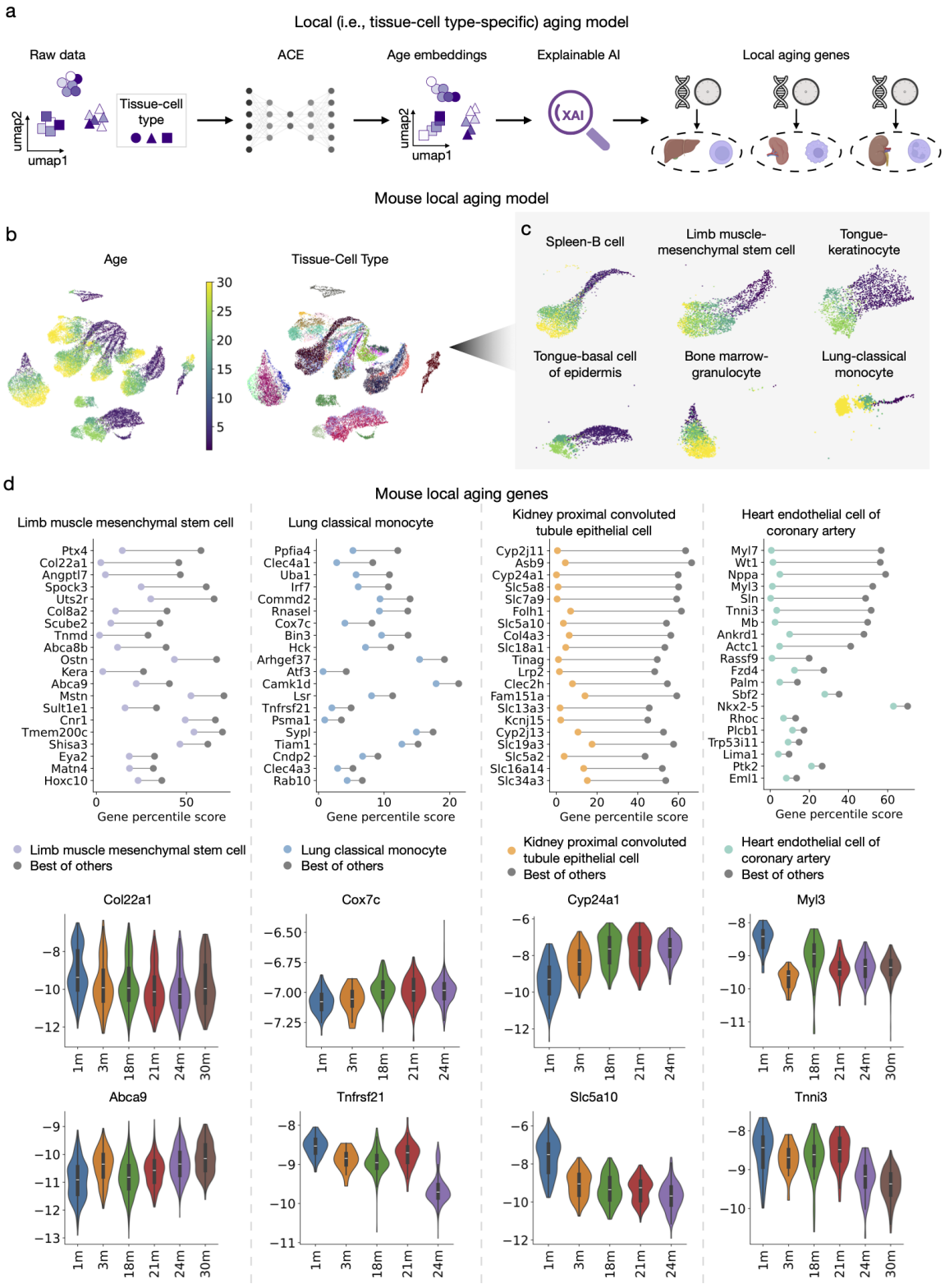

a
Local (i.e., tissue-cell type-specific) aging model
Raw data
ACE
Age embeddings
Explainable AI
Local aging genes
Tissue-cell type
umap2
umap1
XAI
Mouse local aging model
b
Age
Tissue-Cell Type
c
Spleen-B cell
Limb muscle-mesenchymal stem cell
Tongue-keratinocyte
Tongue-basal cell of epidermis
Bone marrow-granulocyte
Lung-classical monocyte
d
Mouse local aging genes
Limb muscle mesenchymal stem cell
Lung classical monocyte
Kidney proximal convoluted tubule epithelial cell
Heart endothelial cell of coronary artery
Gene percentile score
Best of others
Col22a1
Cox7c
Cyp24a1
Myl3
Abca9
Tnfrsf21
Slc5a10
Tnni3

Figure 7.4

Figure 7.4: **ACE enables local (*i.e.*, tissue-cell-type-specific) modeling of aging trajectories. a.** Overview of the local aging model pipeline. To enable local aging analysis, we train an ACE model while excluding tissue and cell type information from the background factors. This design allows the age embedding to capture both tissue-cell-type-specific clustering and aging-related variation. XAI method, Expected Gradients, is then applied to the ACE model to identify local aging-associated genes specific to tissue-cell-type contexts. **b.** UMAPs of the learned age embeddings, colored by age (left) and tissue-cell-type identity (right). The embedding shows clear clustering by tissue-cell-type while preserving aging trajectories within each cluster. **c.** UMAPs of the age embeddings for examples of tissue-cell-type pairs, illustrating clear aging patterns within individual cell populations. **d.** Local aging genes identified in specific tissue-cell-type pairs in the TMS dataset. **top:** Local aging genes for limb muscle mesenchymal stem cell, lung classical monocyte, kidney proximal convoluted tubule epithelial cell, and heart endothelial cell of coronary artery. The genes are ranked by the difference of their percentile scores within each tissue-cell-type context (colored dots) and their best scores in all other cell types (gray dots). Percentile scores are derived from the full ranked gene list, with lower percentiles corresponding to higher importance. **bottom:** Violin plots showing expression of selected top-ranked local aging genes across age groups, illustrating aging-associated expression trends specific to each tissue-cell-type.

all other pairs, and rank the genes by this difference. A high difference indicates that a gene contributes more strongly to aging in one specific tissue-cell-type context while playing a lesser role elsewhere. Many of the top-ranked genes display clear and progressive expression changes across age groups within their respective tissue-cell-type (Figure **??**d, bottom). For example, *Col22a1* and *Abca9* in mesenchymal stem cells, *Cox7c* and *Tnfrsf21* in lung classical monocytes, *Cyp24a1* and *Slc5a10* in kidney epithelial cells, and *Myl3* and *Tnni3* in heart endothelial cells of coronary artery, show consistent aging-associated expression patterns. These trends suggest potential roles for these genes in mediating tissue-cell-type-specific aging processes.

Specifically, several of the above-identified genes have been implicated as modulators of tissue-specific pathogenesis. *Col22a1*, a collagen type XXII component expressed in the myotendinous junction, has been identified as a candidate gene for human myopathies [181]. *Abca9* encodes a cholesterol transporter where changes in cholesterol metabolism have been implicated in aging for several tissues [205]. *Cox7c* is a component of the mitochondrial electron transport chain complex IV, where mitochondrial dysfunction is well known to be perturbed in aging [167]. *Tnfrsf21* is a tumor necrosis factor (TNF) receptor superfamily member contributing to inflammatory responses. TNF receptors have been shown to change expression in humans across various ages, indicating aging-dependent remodeling of immune function throughout life [10]. *Cyp24a1* is a major component of vitamin D homeostasis, catabolizing the active form of vitamin D [1,25(OH)2D] into the inactive 1,24,25(OH)2D molecule. Mutations in this gene has been shown to lead to dysregulated vitamin D levels leading to a number of pathological conditions [219, 252]. *Slc5a10*

is a kidney-specific SLC transporter that is involved in various carbohydrate reabsorption and has been indicated in diabetic kidney disease [260]. *Myl3* encodes the Myosin light chain 3 protein that has been shown to negatively regulate osteoarthritis in a mouse model, while genetic mutations in this gene have been associated with hypertrophic cardiomyopathy in humans [11, 41]. *Tnni3* encodes troponin I, a component of the myocardial sarcomere structure, where mutations in this structure have been shown to play a role in several cardiomyopathies [265]. The implication of the identified genes in tissue-specific pathologies highlights their potential roles in modulating tissue function across various age stages in humans, suggesting that these genes may be potential targets for therapeutic interventions. We also compared ACE with a previously published linear model Zhang, Pisco, Darmanis, and Zou [336] to assess overlap and differences in global and tissue-cell-type-specific aging genes (Supplementary Figure **??**).

Overall, ACE not only captures global aging signatures shared across tissues and cell types, but also enables fine-grained analysis of local aging signatures unique to individual tissue-cell-type pairs. This dual capability is essential for uncovering aging mechanisms that are specific to particular biological environments and may be overlooked in global analyses, offering a more comprehensive understanding of the aging process at single-cell resolution.

7.2.4 *ACE enables accurate biological age clocks at both cell and subject levels*

ACE's ability to disentangle aging-related variation from background biological factors enables it to isolate aging-related transcriptomic signatures that reflect the fundamental biology of aging rather than tissue- or cell-type-specific effects. By capturing smooth and continuous aging trajectories, ACE provides a powerful foundation for constructing biological age clocks.

We first assess cell-level biological age prediction using the TMS Droplet mouse dataset. We train a multilayer perceptron (MLP) model that takes ACE-derived age embeddings as input to predict the chronological age of the subject from which each cell was collected (Figure **??**a). The model is trained on a randomly selected subset of cells and tested on the remaining held-out cells. The predicted age is defined as the cell's biological age, reflecting the transcriptomic state of aging. Predicting chronological age from molecular measurements is a widely used approach for estimating biological age [117, 207, 218]. The biological ages closely match the chronological ages (Pearson's $r \approx 0.9$; Figure **??**b), with ACE achieving slightly higher correlations than all baseline methods (Supplementary Figure **??**a). When evaluated across individual tissue-cell-type pairs, this model achieves strong and statistically significant positive Pearson correlations in nearly all pairs (Figure **??**c), with over half of the pairs exceeding a correlation of 0.8. These results demonstrate that ACE effectively captures continuous and generalizable aging signatures, enabling accurate biological age prediction at single-cell resolution.

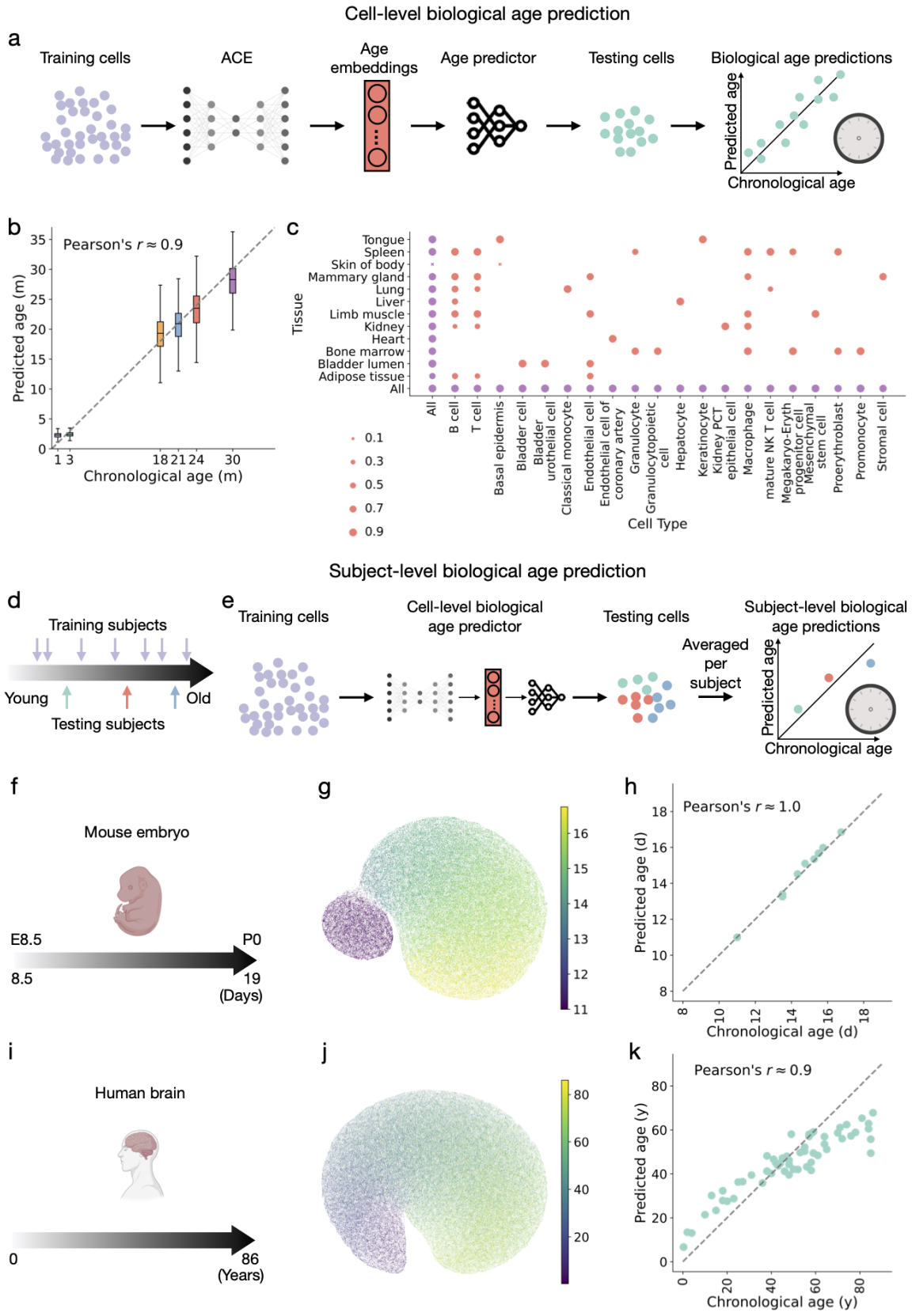

Cell-level biological age prediction
a
Training cells
ACE
Age embeddings
Age predictor
Testing cells
Biological age predictions
Predicted age
Chronological age
b
Pearson's r ≈ 0.9
Predicted age (m)
Chronological age (m)
c
Tissue
Tongue
Spleen
Skin of body
Mammary gland
Lung
Liver
Limb muscle
Kidney
Heart
Bone marrow
Bladder lumen
Adipose tissue
All
Cell Type
Subject-level biological age prediction
d
Training subjects
Young
Old
Testing subjects
e
Training cells
Cell-level biological age predictor
Testing cells
Averaged per subject
Subject-level biological age predictions
Predicted age
Chronological age
f
Mouse embryo
E8.5
P0
8.5
19
(Days)
g
h
Pearson's r ≈ 1.0
Predicted age (d)
Chronological age (d)
i
Human brain
0
86
(Years)
j
k
Pearson's r ≈ 0.9
Predicted age (y)
Chronological age (y)

Figure 7.5

Figure 7.5: **Biological age prediction using age embeddings learned by ACE. (a-c)** Cell-level aging clock based on ACE embeddings. **a.** Overview of the cell-level aging clock model. ACE learns aging-related latent embeddings from training cells, which are used as input to a multilayer perceptron (MLP) layer for single-cell level age prediction. **b.** Box plots comparing predicted and chronological age across timepoints in TMS Droplet dataset, showing that ACE enables accurate and progressive biological age estimation at the single-cell level. **c.** Dot plot of Pearson correlations between predicted and chronological age across tissue-cell-type pairs. Each dot represents the correlation coefficient for a specific pair, with dot size indicating the strength of the correlation. Cross marks ($\times$) denote negative correlations. All displayed correlations are statistically significant ($p < 0.05$, two-tailed t-test), confirming that the model reliably captures aging-related variation across a broad range of cell type and tissue contexts. **(d-k)** Subject-level aging clock based on ACE embeddings. **d.** Illustration of the subject-level biological age prediction setup. A subset of subjects spanning different ages is held out for testing, while the remaining subjects are used for training. **e.** Overview of the subject-level aging clock model. ACE is trained on cells from the training subjects to learn age embeddings. These embeddings are used to train a cell-level age predictor, which is then applied to cells from the held-out subjects. Predicted cell-level ages are averaged per subject to estimate subject-level biological age. **(f-h)** Developmental age prediction using ACE on a mouse embryo scRNA-seq dataset [229]. The dataset includes 74 subjects sampled between embryonic day 8.5 (E8.5) and postnatal day 0 (P0). Eight developmental time points (11.0, 13.5, 14.333, 14.75, 15.25, 15.5, 15.75, and 16.75 days), each represented by a single subject, are held out entirely from training. **f.** Developmental timeline in the mouse embryo dataset. **g.** UMAP visualization of ACE-derived age embeddings colored by developmental day, illustrating that ACE effectively captures continuous developmental progression. **h.** Predicted versus chronological age for the held-out subjects, demonstrating accurate generalization to unseen developmental stages. **(i-k)** Human brain biological age prediction using ACE. The dataset consists of scRNA-seq profiles from the prefrontal cortex of 286 individuals spanning 0.3 to 86 years of age [328]. A subset of 57 individuals is held out for testing. **i.** Lifespan timeline of the dataset. **j.** UMAP visualization of ACE-derived age embeddings, colored by age at death, illustrating ACE's ability to learn a continuous aging representation across the human lifespan. **k.** Scatter plot showing predicted versus chronological age at the subject level, indicating strong predictive accuracy across unseen individuals.

We next extend this framework to the subject-level biological age clock, enabling estimation of an individual's biological age from single-cell profiles. In this setup, we hold out entire test set subjects for evaluation while training ACE and the downstream age predictor on cells from the remaining individuals (Figure **??**d,e). For each held-out subject, we apply the trained model to cells from that subject and average the predicted cell-level biological ages to obtain a subject-level biological age estimate. We validate this approach using two distinct datasets: a mouse embryonic development dataset [229] and a human brain aging dataset [328].

The mouse embryo scRNA-seq dataset profiles the transcriptional states of embryos precisely staged at 2- to 6-hour intervals, spanning late gastrulation to birth [229]. After preprocessing, the data we use contains 416,315 cells from 74 embryos spanning embryonic day 8.5 (E8.5) to postnatal day 0 (P0) (Figure **??**f; **??**). We train the ACE model using cell type and sex as background factors. Eight randomly selected developmental time points (11.0, 13.5, 14.333, 14.75, 15.25, 15.5, 15.75, and 16.75 days), each corresponding to a single embryo, are completely excluded from training and reserved for testing. ACE captures a smooth developmental trajectory across the held-out time points, generating age embeddings that reflect the known progression of developmental stages (Figure **??**g). The predicted ages for the held-out embryos show near-perfect correlation with their true developmental days (Pearson's $r \approx 1.0$; Figure **??**h). ACE achieves performance comparable to other deep learning models and clearly exceeds simpler baselines (Supplementary Figure **??**b), demonstrating ACE's ability to generalize across unobserved developmental stages and highlighting the robustness of the ACE-based biological age clock on unseen subjects.

The human brain aging dataset includes scRNA-seq profiles from the dorsolateral prefrontal cortex of postmortem samples spanning the full human lifespan [328]. After preprocessing, the data we use comprises 1,303,449 cells from 286 individuals aged 0.3 to 86 years (Figure **??**i; **??**). ACE is trained using cell type and sex as background factors. A randomly selected subset of 57 individuals is entirely held out during training and used for testing. After training, ACE captures a smooth aging trajectory across the human lifespan in the held-out individuals (Figure **??**j). Predicted subject-level biological ages show a strong correlation with chronological age (Pearson's $r \approx 0.9$; Figure **??**k), with ACE outperforming baseline models (Supplementary Figure **??**c), demonstrating its effectiveness for lifespan-wide age prediction from single-cell data and highlighting the generalizability of the ACE-based biological age clock to unseen individuals.

Overall, these findings demonstrate the broad applicability of ACE for biological age prediction at both the single-cell and subject levels. By learning continuous and generalizable age embeddings, ACE supports accurate age estimation across diverse biological scales from individual cells to whole organisms, and performs robustly across various biological contexts. This versatility underscores its potential as a general-purpose framework for biological age estimation across species, tissues, and life stages.

7.2.5 *ACE enables multi-species aging analysis and identification of aging genes conserved across species*

Understanding aging across species is crucial for identifying evolutionarily conserved mechanisms that underlie fundamental aspects of the aging process. While many studies focus on organism-specific patterns, learning aging signatures that are shared across species provides an opportunity for discovering *universal* biomarkers and interventions. However, such analyses are challenging due to biological differences and variability in dataset characteristics across species. ACE is well-suited for this task, as its disentangled representation framework enables it to isolate aging signatures from background factors, such as species identity, in multi-species contexts. This allows ACE to extract consistent aging signatures across diverse organisms, positioning it as a unique tool for the analysis of cross-species aging analysis.

To assess whether ACE can uncover conserved aging signatures across species, we extend our analysis to a multi-species setting by integrating three large-scale scRNA-seq datasets: the Tabula Sapiens (TS) human cell atlas [236], the TMS mouse aging cell atlas [3], and the Aging Fly Cell Atlas [173], spanning diverse tissues and life stages (Figure **??**a). To enable biologically meaningful comparisons, we map mouse and fly ages to human-equivalent ages using phase equivalencies from established mouse-human mappings [83] and lifespan curve alignment for fly-human mapping (Figure **??**b). For humans and mice, we retain the 30 common and most abundant cell types; for flies, we include the 10 most abundant cell types. After preprocessing, the integrated dataset comprises 520,139 human cells, 171,924 mouse cells, and 101,684 fly cells across 19 tissues and 40 cell types (see Supplementary Figure **??** for the UMAPs of the raw data; **??**). The inclusion of mouse and fly data substantially broadens the age distribution compared to human data alone (Figure **??**b). ACE is then trained on the multi-species dataset using 2,515 one-to-one orthologous genes shared across all species, with species, tissue, cell type, and sex specified as background factors.

The learned age embeddings reveal a smooth and continuous aging trajectory shared across all species, with cells from different organisms, tissues, and cell types well mixed in the same embedding space (Figure **??**c; see Supplementary Figure **??** for UMAPs of both age and background embeddings colored by age and background factors). We then apply the Expected Gradients to identify aging-associated genes that are consistently important across species. Several conserved genes, such as *H3f3b*, *Lars2*, *Cdc42*, and *Pdcd4*, exhibit progressive and consistent expression changes across age groups in all species (Figure **??**d). The conserved genes across species all play a role in known aging-related processes. *H3f3b* encodes a histone protein involved in chromatin structure and epigenetic regulation of gene expression, which is a known pathway involved in aging [306]. *Lars2* encodes a mitochondrial leucyl-tRNA synthetase protein. A study utilizing *C. elegans* demonstrated that downregulation of this gene was associated with longer lifespan [148].

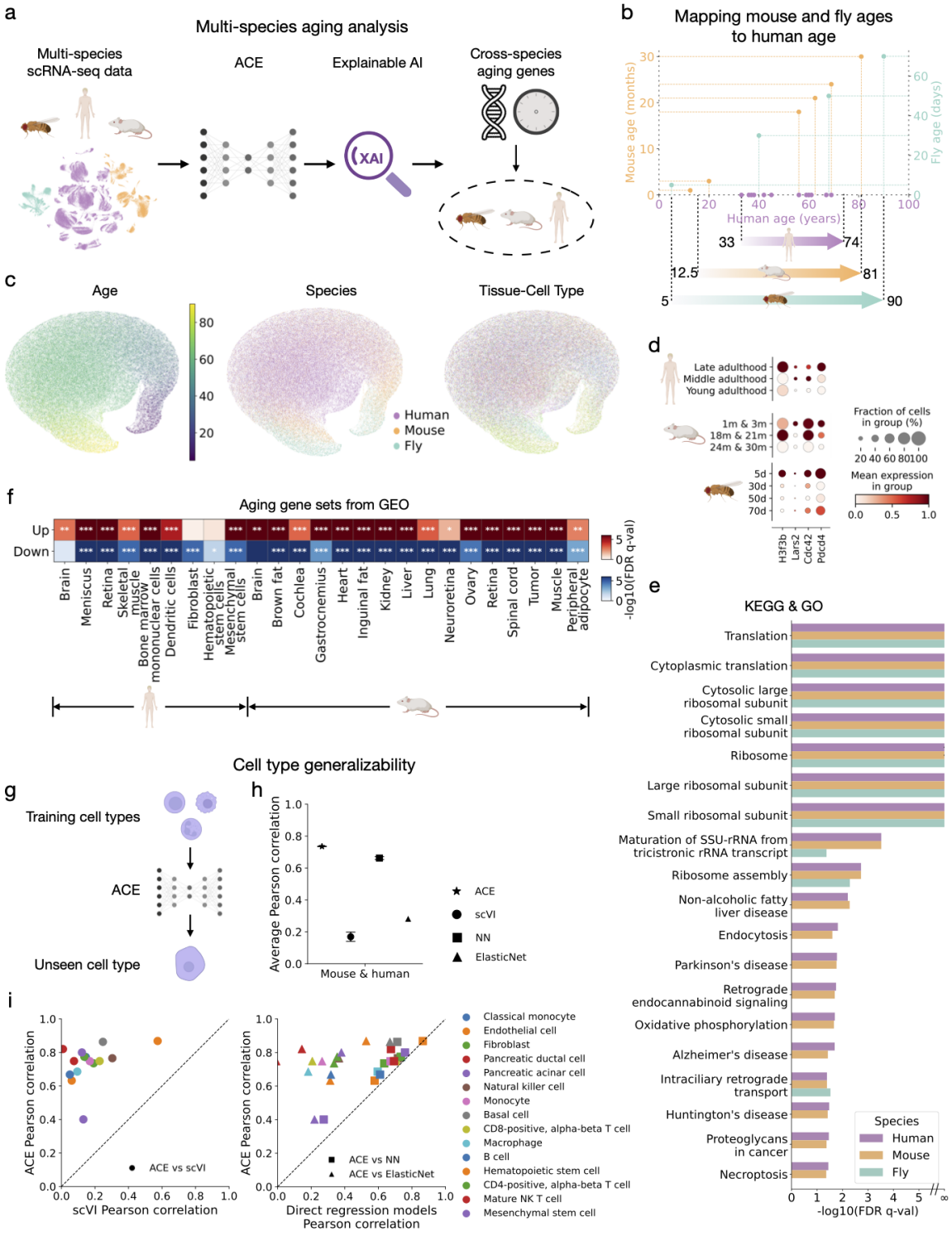

a
Multi-species aging analysis
Multi-species scRNA-seq data
ACE
Explainable AI
Cross-species aging genes
XAI
b
Mapping mouse and fly ages to human age
Mouse age (months)
Fly age (days)
Human age (years)
33
74
12.5
81
5
90
c
Age
Species
Tissue-Cell Type
Human
Mouse
Fly
d
Late adulthood
Middle adulthood
Young adulthood
1m & 3m
18m & 21m
24m & 30m
5d
30d
50d
70d
H3f3b
Lars2
Cdc42
Pdcd4
Fraction of cells in group (%)
Mean expression in group
e
KEGG & GO
Translation
Cytoplasmic translation
Cytosolic large ribosomal subunit
Cytosolic small ribosomal subunit
Ribosome
Large ribosomal subunit
Small ribosomal subunit
Maturation of SSU-rRNA from tricistronic rRNA transcript
Ribosome assembly
Non-alcoholic fatty liver disease
Endocytosis
Parkinson's disease
Retrograde endocannabinoid signaling
Oxidative phosphorylation
Alzheimer's disease
Intraciliary retrograde transport
Huntington's disease
Proteoglycans in cancer
Necroptosis
Species
Human
Mouse
Fly
-log10(FDR q-val)
f
Aging gene sets from GEO
Up
Down
-log10(FDR q-val)
Brain
Meniscus
Retina
Skeletal muscle
Bone marrow mononuclear cells
Dendritic cells
Fibroblast
Hematopoietic stem cells
Mesenchymal stem cells
Brain
Brown fat
Cochlea
Gastrocnemius
Heart
Inguinal fat
Kidney
Liver
Lung
Neuroretina
Ovary
Retina
Spinal cord
Tumor
Muscle
Peripheral adipocyte
Cell type generalizability
g
Training cell types
ACE
Unseen cell type
h
Average Pearson correlation
Mouse & human
ACE
scVI
NN
ElasticNet
i
ACE Pearson correlation
scVI Pearson correlation
ACE vs scVI
ACE vs NN
ACE vs ElasticNet
Direct regression models Pearson correlation
Classical monocyte
Endothelial cell
Fibroblast
Pancreatic ductal cell
Pancreatic acinar cell
Natural killer cell
Monocyte
Basal cell
CD8-positive, alpha-beta T cell
Macrophage
B cell
Hematopoietic stem cell
CD4-positive, alpha-beta T cell
Mature NK T cell
Mesenchymal stem cell

Figure 7.6

Figure 7.6: **Multi-species aging analysis and cell type generalizability using ACE. a.** Overview of the multi-species aging analysis pipeline. ACE is applied to multi-species scRNA-seq data from human, mouse, and fly to disentangle aging-related signatures shared across species. XAI method, Expected Gradients, is then used to identify conserved aging genes shared across species. **b.** Mapping of mouse and fly ages to human-equivalent ages using lifespan alignment, enabling biologically meaningful cross-species comparisons. **c.** UMAPs of the age embeddings learned by ACE from integrated human, mouse, and fly data. Embeddings are colored by age (left), species (middle), and tissue-cell-type labels (right), showing that the model captures a smooth aging trajectory across species, with cells from different organisms well mixed in the age latent space. **d.** Expression patterns of selected cross-species aging-associated genes identified by ACE. Shown are normalized expression levels of orthologous genes across different age groups in human, mouse, and fly, labeled using mouse gene symbols (*H3f3b*, *Lars2*, *Cdc42*, and *Pdcd4*). Dot size indicates the fraction of cells expressing the gene; color represents mean expression. **e.** KEGG and GO pathway enrichment analysis of conserved aging genes across human, mouse, and fly identified conserved pathways that are significantly enriched and involved in key aging-related biological processes. Significance was assessed at FDR $q < 0.05$ using the Benjamini-Hochberg correction. **f.** Gene set enrichment analysis using the "Aging Perturbations from GEO Up" and "Aging Perturbations from GEO Down" gene set collections. These databases consist of gene sets curated from studies in GEO that compare aged versus young samples, capturing genes consistently upregulated or downregulated with age for various tissues and cell types. The results demonstrate that multi-species ACE-derived aging gene rankings are significantly enriched in multiple known human and mouse aging-associated gene sets across various tissues and cell types. Red boxes indicate enrichment in *upregulated* gene sets; blue boxes indicate enrichment in *downregulated* gene sets. Color intensity reflects $-\log_{10}$(FDR q-value), with significance determined by FDR-adjusted q-values using the Benjamini-Hochberg correction. Asterisks denote significance thresholds (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$). **g.** Schematic illustrating the evaluation of cell type generalizability. To assess how well age embeddings learnt by ACE generalize to new biological contexts, the model is trained on mouse and human data excluding one cell type. Age embeddings are then extracted for the held-out (unseen) cell type, and used to train a multilayer perceptron (MLP) for age prediction. This setup evaluates how effectively the learned age embeddings capture aging signatures that generalize across cell types. **h.** Average Pearson correlation between predicted and true age across unseen cell types in mouse and human, comparing ACE to baseline models including scVI, a neural network (NN), and ElasticNet. Age prediction is performed using different inputs and models: MLPs are trained on age embeddings from ACE, latent variables from scVI, and raw gene expression for NN; ElasticNet is trained directly on raw gene expression. Each dot represents the mean Pearson correlation across 10 runs with different random seeds (for ACE, scVI, and NN); error bars indicate the standard deviation. ACE consistently outperforms baseline models. **i.** Scatter plots comparing ACE to baseline models across individual cell types, demonstrating that ACE provides stronger generalization to unseen cell types.

Cdc42 encodes a GTPase of the Rho-superfamily that regulates the actin cytoskeleton, cell cycle, and proliferation of many cell types and has been implicated in numerous studies as a factor in aging [294]. *Pdcd4* is involved in regulating programmed cell death and the knockdown of this gene in hepatoma cancer cells has been shown to induce cellular senescence [105].

Pathway enrichment analysis using KEGG and GO shows that these conserved aging genes are significantly enriched for translation-related functions (*e.g.*, ribosome, cytoplasmic translation) across all 3 species, highlighting the importance of proteostasis as a conserved aging mechanism across species. A greater number of conserved pathways are present between the human and mouse datasets, including immune signaling (*e.g.*, endocytosis), metabolic processes (*e.g.*, oxidative phosphorylation), and neurodegenerative pathways (*e.g.*, Alzheimer's and Parkinson's diseases) (Figure **??**e; see Supplementary Figure **??**,**??**,**??** for the full list of enriched pathways). Furthermore, ACE's ranked gene list demonstrates strong and consistent enrichment across nearly all aging-associated gene sets curated from GEO for both human and mouse, spanning multiple tissues and cell types (Figure **??**f). These results support its ability to identify aging-related genes that are shared across species, tissues, and cell types.

Because ACE captures aging signatures shared across diverse biological contexts, it should also generalize to new, unseen contexts. To evaluate this, we test ACE's ability to generalize to previously unseen cell types. Using a combined dataset of mouse and human cells across 40 common cell types, we select 15 with sufficient cell numbers and age diversity to serve as held-out cell types. For each held-out evaluation, we train ACE on the 39 cell types and assess its age prediction performance on the excluded one (Figure **??**g). An MLP is trained on the ACE-derived age embeddings to predict age. We compare ACE to baseline methods including scVI, which uses its learned embeddings with an MLP for age prediction, as well as a neural network (NN) and an ElasticNet model trained directly on raw gene expression for age prediction. Across the 15 held-out cell types, ACE achieves an average Pearson correlation of 0.73, with 11 cell types exceeding 0.7 (Figure **??**h), representing a 10.6% improvement over the best alternative baseline, NN. ACE consistently outperforms all baselines, demonstrating its ability to generalize aging predictions to previously unseen cell types and supporting its utility in modeling shared aging signatures across heterogeneous biological contexts (Figure **??**h,i).

Together, these findings demonstrate ACE's ability to model aging trajectories across diverse organisms and uncover conserved aging signatures, while also generalizing to previously unseen cell types. By learning shared aging representations that disentangle background biological factors, ACE enables robust aging analysis across species, tissues, and cell types. This highlights its versatility as a unified framework for cross-species and cross-context aging studies at the single-cell resolution.

7.2.6 *Experimental Validation of ACE-Identified Aging Genes*

To assess the biological relevance of aging-related genes identified by ACE, we performed RNAi knockdown experiments in *C. elegans* and measured their effects on lifespan (Figure **??**; **??**). We focused on genes identified by the mouse global aging model, as the mouse aging dataset provides a comprehensive, aging-focused single-cell resource that yields less confounded aging signatures. Specifically, we selected 8 mouse genes from the top 150 most important global aging genes identified by the mouse global aging model: *Ppia*, *Ftl1*, *Gpx3*, *Rpl37*, *Serpina3k*, *Tpt1*, *Rps27a*, and *Uba52*. Mouse genes were mapped to their *C. elegans* orthologs, resulting in a total of 15 tested *C. elegans* genes. These genes were selected based on their potential involvement in proteostasis, their evolutionary conservation across species, and the feasibility of functional testing in *C. elegans* via RNAi, as they do not produce larval arrest or lethality phenotypes.

For each gene, *C. elegans* orthologs were knocked down using RNAi, and survival curves were compared to those of empty vector (EV) controls. As shown in Figure **??**b, four knockdowns, *cyn-1*, *tct-1*, *rpl-37*, and *ubl-1*, led to significant lifespan alterations, with p-values ranging from 3.1×10^{-4} to 3.8×10^{-15}, confirming that ACE effectively prioritizes functionally relevant aging genes. Importantly, three of these four genes (*cyn-1*, *rpl-37*, *ubl-1*) are involved in protein translation. These results corroborate the cross-species pathway analysis (Figure **??**e), further emphasizing the importance of proteostasis as a conserved aging mechanism across species. Of the 8 mouse genes tested, 6 had at least one *C. elegans* ortholog whose knockdown significantly impacted lifespan. Among the 15 *C. elegans* genes tested, 8 showed significant effects (see Supplementary Figure **??** for full survival curves).

We also observed that one of the tested genes, *Uba52*, ranked highly in the cross-species ACE model (ranked 33), suggesting its conserved relevance across mouse, human, and fly. As shown in Figure **??**c, knockdown of *ubq-2* (the *C. elegans* ortholog of *Uba52*) led to significantly reduced lifespan, further supporting the biological validity of ACE's prioritizations across species. These results highlight ACE's ability to uncover evolutionarily conserved regulators of aging that can be experimentally validated in model organisms.

These experimental validations demonstrate ACE's power in prioritizing biologically meaningful aging-associated genes, with a strong emphasis on proteostasis. The observed lifespan effects in *C. elegans* following knockdown of ACE identified genes highlight the model's functional relevance and cross-species generalizability. Notably, several of the validated genes play key roles in ribosomal function and ubiquitin-mediated protein turnover – hallmarks of proteostasis that are increasingly recognized as central to aging. These findings underscore ACE's potential to uncover evolutionarily conserved regulators of aging and provide mechanistic insights with broad implications for aging biology and translational geroscience.

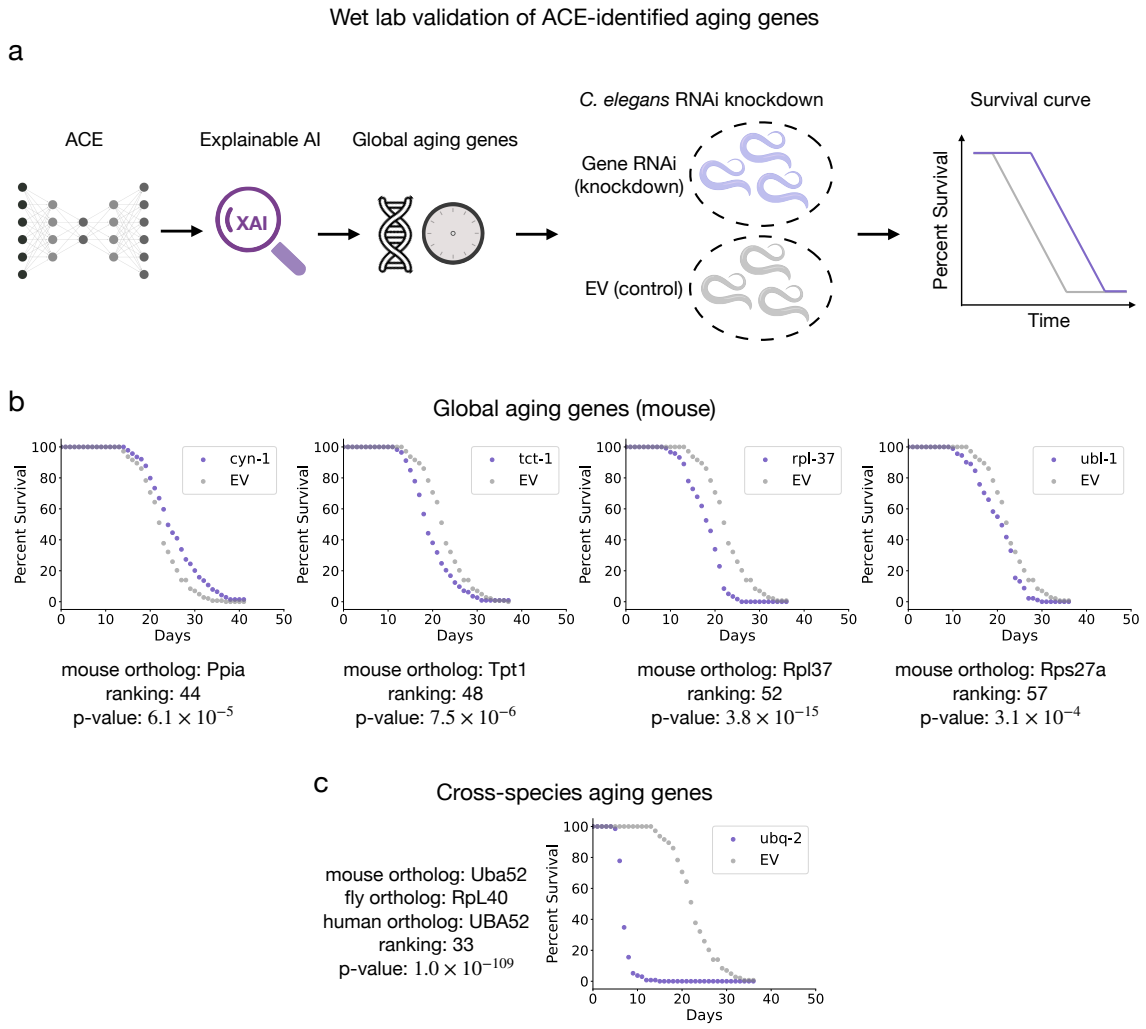


Figure 7.7: **Wet lab validation of ACE-identified aging genes using *C. elegans* RNAi lifespan assays. a.** Overview of the experimental validation pipeline. ACE identifies global aging genes using XAI techniques. Selected genes are knocked down in *C. elegans* via RNA interference (RNAi), and survival is compared to empty vector (EV) controls. **b.** Survival curves for knockdown of *C. elegans* orthologs corresponding to selected global aging genes identified by the mouse global ACE model. Each plot shows the percent survival of animals treated with gene-specific RNAi versus empty vector control. Mouse gene names, gene rankings from the mouse global ACE model, and p-values from t-tests (adjusted using the Benjamini-Hochberg correction) are shown below each plot. **c.** Survival curves for knockdown of *C. elegans* genes corresponding to a top-ranked aging-associated gene from the cross-species ACE model. Orthologs are selected based on mouse gene mappings. The plot shows percent survival for animals treated with gene-specific RNAi versus EV. Mouse gene name, its one-to-one orthologs in human and fly, gene rankings from the cross-species ACE model, and adjusted p-value from t-tests (adjusted using the Benjamini-Hochberg correction) are shown below each plot.

7.3 DISCUSSION

ACE provides an interpretable and generalizable framework for disentangling aging-related signatures from background biological variation in single-cell transcriptomic data. By modeling aging trajectories across tissues, cell types, and species, ACE enables the identification of both global and local aging signatures, revealing how aging manifests in conserved and context-dependent ways. Applied to large-scale mouse, fly, and human atlases, ACE uncovered robust, biologically meaningful aging-associated genes and pathways, including well-studied aging genes such as *Sparc* [94], *S100a8* [101], and *Hsp26* [258, 289]. ACE's learned embeddings also enabled accurate subject-level biological age prediction in mouse embryo and human brain datasets, highlighting their utility for biological age estimation.

Across species, our results overwhelmingly implicated protein homeostasis in aging. Protein homeostasis, or proteostasis, incorporates several processes, including protein synthesis, maintenance, and destruction/turnover in cells and has long been implicated in the aging process [19]. ACE has identified critical conserved components of proteostasis that affect aging across several species, including several genes involved in ribosome and ubiquitination, which are cellular machinery involved in protein synthesis and protein degradation, respectively. Importantly, several of these ribosomal and protein ubiquitination genes identified in the mouse model and multi-species model were validated in a *C. elegans* model and were found to significantly impact lifespan. The validation of these genes in an animal model unique from the animal models used in the ACE training data highlights the importance of these processes as conserved aging mechanisms across species. Additionally, this highlights the generalizability of ACE towards other model organisms and its broad biological utility. The identification of proteostasis genes is consistent with other experimental results showing that long-lived clams and long-lived naked mole rats have exceptional proteostasis compared to other similar species (shorter-lived clams and rats), further implicating proteostasis as foundational in the aging process. Indeed, it seems logical that cells need to maintain the function of the proteins that produce, repair and maintain all aspects of cellular physiology. Together, these results support the utility of ACE to ascertain genes and pathways involved in aging from publicly available scRNA-seq datasets.

The potential for ACE to identify previously unidentified driver genes involved in cellular aging is an important use-case for the model. With an increasingly aging population, the identification of aging driver genes can allow for the identification of druggable targets that can help ease the burden of aging and improve healthy lifespan. The application of ACE to identify tissue- and cell type-specific aging genes can assist with the therapeutic intervention for tissue-specific aging-associated diseases, such as aging-associated neurodegeneration, cardiovascular disease, and chronic inflammatory diseases. Genes identified using ACE can be studied in commonly used animal models, as performed here,

or they can be studied in human-specific models such as primary cell lines or human induced pluripotent stem cell (hiPSC) derived cell lines and organoids. The utilization of an AI model like ACE in combination with human-derived cell models is in direct line with the 2022 FDA Modernization Act 2.0, which allows for the use of new approach methodologies (NAMs) as an alternative to animal models for use in safety and efficacy assessments in the drug development process [342]. Models like ACE could prove to be a powerful tool in the drug discovery phase during preclinical drug development.

From a technical perspective, ACE introduces an interpretable and modular generative framework that disentangles aging-related variation from complex background factors in scRNA-seq data. Unlike conventional models that compress all variation into a single latent space [32, 168, 325], ACE learns two embeddings to isolate aging signatures, improving both interpretability and generalization. Its flexible design extends naturally to other biological settings, enabling disentanglement of molecular signatures linked to diverse phenotypes such as neurodegeneration, tumor progression, and immune activation. ACE also lays the groundwork for developing single-cell aging foundation models that generalize across cell types, tissues, species, and experimental conditions. Unlike existing foundation models trained on heterogeneous datasets that may conflate aging with other biological effects [53, 66, 283], ACE explicitly disentangles aging signatures from background variation, offering a targeted and interpretable method for building aging-specific foundation models. This disentanglement principle further positions ACE as a framework for developing phenotype-specific foundation models in single-cell biology.

The limitation of ACE is that it currently relies on accurate annotations for background factors such as tissue and cell type, which may be incomplete or inconsistent across datasets. This dependence could affect model performance on less curated or cross-study data. In addition, ACE currently operates on transcriptomic data only. Incorporating additional omics layers, such as DNA methylation [54], chromatin accessibility [12], or proteomics [134], could provide a more comprehensive understanding of aging. Advances in multi-modal generative modeling [15, 92] present promising opportunities to extend ACE toward multi-omics integration, thereby enhancing its biological scope.

In summary, ACE enables interpretable and disentangled modeling of aging signatures at single-cell resolution across species, tissues, and cell types. Its scalability and generalizability make ACE a powerful framework for uncovering biologically meaningful signatures from high-dimensional single-cell data and for advancing phenotype-specific modeling in aging and disease research.

7.4 METHODS

7.4.1 *The ACE model*

Here, we present the ACE model in more detail. We begin by describing the model's generative process and then the model's inference procedure.

7.4.1.1 *The ACE generative process*

For a data point x_n, we assume that each expression value x_{ng} for sample n and gene g is generated through the following process:

$$
\begin{aligned}
u_n &\sim \text{Normal}(0, I) \\
z_n &\sim \text{Normal}(0, I) \\
\ell_n &\sim \text{LogNormal}(\ell_\mu^\top s_n, (\ell_\sigma^2)^\top s_n) \\
\rho_n &= f_w(u_n, z_n, s_n) \\
w_{ng} &\sim \text{Gamma}(\rho_{ng}, \theta_g) \\
y_{ng} &\sim \text{Poisson}(\ell_n w_{ng}) \\
h_{ng} &\sim \text{Bernoulli}\big(f_h^g(u_n, z_n, s_n)\big) \\
x_{ng} &= \begin{cases} y_{ng} & \text{if } h_{ng} = 0 \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

Here, u_n and z_n denote two sets of latent variables that account for variations in scRNA-seq expression data. Specifically, u_n is designed to capture variation primarily driven by age (a_n), aiming to isolate aging-related biological signatures. In contrast, z_n models background variation associated with covariates b_n (e.g., cell type, tissue, sex, species), which are often dominant sources of variation and may overshadow subtle aging signatures. ACE's design enables the disentanglement of these two factors, helping to recover aging-related signatures across complex biological contexts. We place standard multivariate Gaussian priors on both u_n and z_n to support efficient inference within the variational autoencoder (VAE) framework [141].

To ensure that the latent variables u_n and z_n accurately capture distinct sources of biological variation, specifically aging-related signatures and background-related signatures, we adopt two complementary strategies. First, during training, we add supervised prediction losses that encourage each latent variable to retain information about its corresponding covariates. Specifically, a age prediction network is trained to predict the chronological age of the donors (a_n) from u_n, while a background prediction network is trained to predict background covariates b_n (e.g., cell type, tissue, sex, species) from z_n. These

covariates can be binary, categorical, or continuous; accordingly, we apply cross-entropy loss for classification tasks and mean squared error for regression. Second, we apply the d-variable Hilbert-Schmidt Independence Criterion (dHSIC) [99, 169, 220, 274] to enforce statistical independence between u_n and z_n. Since the dHSIC equals zero if and only if the variables are independent, minimizing this quantity ensures that the two latent spaces encode non-overlapping biological signatures, thereby disentangling aging-related variation from dominant background effects.

Here ℓ_μ and $\ell_\sigma^2 \in \mathbb{R}_+^B$, where B denotes the cardinality of an optional label denoting experimental batch, parameterize the prior for a latent RNA library size scaling factor on a log scale, and s_n is a B-dimensional one-hot vector encoding the batch label for each cell. For each batch, ℓ_μ and ℓ_σ^2 are set to the empirical mean and variance of the log library size. $\rho_{ng} \in \mathbb{R}_+$ and shape $\theta_g \in \mathbb{R}_+$ parameterize our Gamma distribution with a mean-shape parameterization. Moreover, we note that θ_g can be viewed as a gene-specific inverse dispersion parameter for a negative binomial distribution, and we learn $\theta \in \mathbb{R}_+^G$ through variational inference. f_w and f_g are neural networks that transform the latent space and batch annotations to the original gene space, i.e. f_w: $\mathbb{R}^d \times \{0,1\}^B \to \mathbb{R}^G$ and similarly for f_g, where d is the combined dimensionality of the concatenated age and background latent spaces. The outputs of the network f_w represent the mean proportion of transcripts expressed across all genes, and this constraint is enforced by using a softmax activation function in the last layer. That is, letting $f_w^g(u_n, z_n, s_n)$ denote the entry in the output of f_w corresponding to gene g, we have $\sum_g f_w^g(u_n, z_n, s_n) = 1$. The output of the neural network f_h denotes whether a dropout event has occurred (i.e., a gene's expression is read as zero due to technical factors rather than meaningful biological phenomena).

Our generative process closely follows that of contrastiveVI [311], but with the notable additions of prediction networks and the dHSIC penalty. While contrastiveVI's modeling approach excels situations when explicit case and control groups of cells are available, it is not appropriate for scenarios without a clear control group. By incorporating age covariate a_n and background covariates b_n, and introducing prediction networks for both, ACE can effectively isolate the aging variations even without an explicit group of corresponding control cells.

7.4.1.2 *Inference with ACE*

We cannot compute the ACE posterior distribution using Bayes' rule as the integrals required to compute the model evidence $p(x_n|s_n)$ are analytically intractable. As such, we instead approximate our posterior distribution using variational inference [31]. We approximate our posterior with a distribution factorized as follows:

$$q_{\phi_x}(u_n, z_n, \ell_n|x_n, s_n) = q_{\phi_u}(u_n|x_n, s_n)q_{\phi_z}(z_n|x_n, s_n)q_{\phi_\ell}(\ell_n|x_n, s_n). \tag{7.1}$$

Here ϕ_x denotes a set of learned weights used to infer the parameters of our approximate posterior. Based on our factorization of the posterior in Eq. **??**, we can divide our

full set of parameters ϕ_x into disjoint subsets ϕ_u, ϕ_z and ϕ_ℓ for inferring the parameters of the distributions of u, z and ℓ respectively. As in the VAE framework [141], we approximate the posterior for each set of latent variables via a deep neural network that takes in expression levels as input and returns the parameters of its corresponding approximate posterior distribution. Moreover, we note that each factor in the posterior approximation shares the same family as its respective prior distribution (e.g. $q(u_n|x_n, s_n)$ follows a normal distribution). By marginalizing out w_{ng}, h_{ng}, and y_{ng}, we can simplify our likelihood yielding $p_\nu(x_{ng}|u_n, z_n, s_n, \ell_n)$, which has a closed form of a zero-inflated negative binomial (ZINB) distribution and where ν denotes the parameters of our generative model. We implement our generative model with deep neural networks as done for our approximate posterior distributions. For Eq. **??**, We derive the following Evidence Lower Bound (ELBO) for the marginal log-likelihood:

$$\begin{aligned}\log p(x|s) \geqslant & \mathbb{E}_{q(u,z,\ell|x,s)} \log p(x|u,z,\ell,s) - D_{KL}(q(u|x,s)||p(u)) \\ & - D_{KL}(q(z|x,s)||p(z)) - D_{KL}(q(\ell|x,s)||p(\ell|s))\end{aligned} \tag{7.2}$$

We optimize the parameters of the generative model, inference networks, and prediction networks jointly using stochastic gradient descent. The optimization objective is a weighted composite loss that combines the ELBO, prediction losses for the age covariate a_n and background covariate b_n, and a penalty term based on the dHSIC, which encourages disentanglement between u_n and z_n. The neural networks used for both the variational and generative distributions are standard feedforward architectures with typical activation functions. Formally, the total loss is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ELBO}} + \lambda_a \cdot \mathcal{L}^{(a)}_{\text{pred}} + \lambda_b \cdot \mathcal{L}^{(b)}_{\text{pred}} + \lambda_{\text{dHSIC}} \cdot \text{dHSIC}(u_n, z_n) \tag{7.3}$$

where, $\mathcal{L}_{\text{ELBO}}$ denotes the negative ELBO as defined in Eq. **??**; $\mathcal{L}^{(a)}_{\text{pred}}$ is the prediction loss for the chronological age a_n, for which we use mean squared error (MSE) since age is a continuous variable; $\mathcal{L}^{(b)}_{\text{pred}}$ represents the prediction loss for the background covariates b_n, such as cell type, tissue, sex, and species, and we apply cross-entropy loss due to their categorical nature; $\text{dHSIC}(u_n, z_n)$ quantifies the statistical dependence between the latent variables u_n and z_n. The hyperparameters $\lambda_a, \lambda_b, \lambda_{\text{dHSIC}}$ are used to scale the auxiliary loss components so that their magnitudes are comparable to that of $\mathcal{L}_{\text{ELBO}}$. This normalization facilitates stable joint optimization by preventing any single auxiliary loss from dominating the training dynamics.

7.4.2 *Model optimization details*

For all datasets, ACE models were trained using 64% of cells, validated on 16% to determine the optimal number of training epochs, and evaluated on the remaining 20% for

both quantitative and qualitative analyses. Early stopping was applied when the validation ELBO failed to improve for 45 consecutive epochs. All models were optimized using the Adam optimizer with a learning rate of 0.0001, $\epsilon = 0.01$, and a weight decay of 10^{-6}. The hyperparameters $\lambda_a, \lambda_b, \lambda_{\text{dHSIC}}$ were chosen separately for each model to ensure that auxiliary loss terms were on a comparable scale to the ELBO.

Each ACE model was trained with 20 latent dimensions: 3 reserved for aging-related variation and 17 for background variation, unless otherwise noted. For the mouse and fly global aging models, and the mouse embryo model, background covariates included cell type, tissue, and sex. The human brain aging model used cell type and sex as background covariates. For the mouse local aging model, background included only sex, and 13 age dimensions and 7 background dimensions were used to encourage learning tissue- and cell type-specific aging signatures. For the multi-species aging model, background covariates included cell type, tissue, sex, and species.

We compared ACE with the following baseline methods: MrVI [32], scANVI [325], SiFT [223], scVI [168], DCA [75], and PCA. Each baseline was trained with 20 latent dimensions for consistency. For MrVI, chronological age was used as the "sample key" to estimate age effects. For scANVI, we treated age as a categorical "cell type label" for conditional modeling. For SiFT, we treated cell type, tissue, and sex as unwanted variation sources and removed them accordingly during training.

7.4.3 *Datasets and preprocessing*

We now briefly describe all datasets used in this work along with any corresponding preprocessing steps. All preprocessing steps were performed using the `Scanpy` Python package.

7.4.3.1 *Tabula Muris Senis mouse dataset*

The Tabula Muris Senis (TMS) dataset [3] is a large-scale single-cell RNA-sequencing (scRNA-seq) atlas characterizing aging-associated transcriptomic changes across multiple tissues in the mouse. It consists of two branches based on sequencing protocols: TMS Droplet, generated using microfluidic droplet-based platforms, and TMS FACS, produced via fluorescence-activated cell sorting (FACS) followed by Smart-seq2 library preparation. In our study, we used the TMS Droplet dataset as the primary resource due to its larger number of cells and used the TMS FACS dataset to validate key biological findings.

PREPROCESSING FOR TMS DROPLET: We first filtered out genes expressed in fewer than 5 cells and removed cells expressing fewer than 500 genes. Cells with fewer than 3,000 total counts were also excluded. To ensure sufficient representation, we retained only the top 20 most abundant cell types and included tissues with more than 1,000 cells. We then selected the top 2,000 highly variable genes using the Seurat v3 method with batch

correction applied via the `batch` key, based on raw counts. In addition to these HVGs, we incorporated aging-associated genes identified in the study by Zhang et al. [336].

After preprocessing, the TMS Droplet dataset comprised 135,420 cells and 4,327 genes, derived from 17 male and 6 female mice across six age groups (1m, 3m, 18m, 21m, 24m, 30m), 13 tissues, and 20 cell types.

PREPROCESSING FOR TMS FACS: We applied similar preprocessing steps to the TMS FACS dataset. Specifically, we removed cells expressing fewer than 500 genes and excluded those having fewer than 3,000 total counts. We then filtered to retain only tissues and cell types that were also present in the TMS Droplet data. Furthermore, we removed cell types, tissues, and donors with fewer than 100 cells, and limited the dataset to age groups with more than 1,000 cells. We restricted the gene set to those retained in the processed TMS Droplet data for consistency.

The final TMS FACS dataset comprised 24,546 cells and 4,120 genes across three age groups (3m, 18m, and 24m), including samples from 12 male and 6 female mice.

7.4.3.2 *Aging Fly Cell Atlas*

The Aging Fly Cell Atlas [173] is a large-scale single-cell RNA-sequencing resource that profiles aging in *Drosophila melanogaster* across multiple tissues, time points, and sexes. It provides comprehensive coverage of aging-associated transcriptional changes in both neural and peripheral tissues, offering insights into conserved and fly-specific aging processes.

We performed the following preprocessing steps. First, we removed genes expressed in fewer than 5 cells and filtered out cells expressing fewer than 300 genes. Cells with fewer than 500 total counts were also excluded. To ensure sufficient representation, we retained only cell types with more than 1,000 cells and excluded cells with ambiguous sex annotations labeled as "mix." We then selected the top 2,000 highly variable genes using the Seurat v3 method on the raw count layer with batch correction.

After preprocessing, the final dataset consisted of 424,863 cells and 2,000 genes, covering 14 cell types across two tissues (head and body), four age groups (5d, 30d, 50d, and 70d), and two sexes.

7.4.3.3 *Mouse embryo dataset*

To evaluate ACE for biological age prediction, we used the mouse embryo single-cell RNA-sequencing dataset [229]. This dataset provides high-resolution temporal profiling of developing mouse embryos, covering time points from embryonic day (E) 8.5 to postnatal day 0 (P0), sampled every 2 to 6 hours. It enables modeling of developmental trajectories as a continuous biological aging process.

We first filtered out genes expressed in fewer than 5 cells and removed cells expressing fewer than 500 genes. Cells with fewer than 3,000 total counts were also excluded. To ensure sufficient representation, we retained only the top 20 most abundant cell types. We then selected the top 2,000 highly variable genes using the Seurat v3 method, based on raw count values.

After preprocessing, the final dataset consisted of 416,315 cells from 74 embryos, spanning the developmental window from E8.5 to P0. For training, we modeled developmental stage as the aging covariate and treated cell type and sex as background covariates. To assess the model's ability to generalize to unseen time points, we excluded eight developmental stages, E11.0, E13.5, E14.333, E14.75, E15.25, E15.5, E15.75, and E16.75, from training. Each excluded time point corresponds to a single embryo and was reserved for testing.

7.4.3.4 *Human brain aging Dataset*

To evaluate biological age prediction in the human brain, we utilized the dorsolateral prefrontal cortex (dlPFC) single-cell RNA-sequencing dataset [328]. This dataset provides a comprehensive transcriptomic atlas of the human lifespan in a critical cognitive brain region.

We selected the top 2,000 highly variable genes using the Seurat v3 method, based on raw count values. After preprocessing, the final dataset consisted of 1,303,449 cells from 286 individuals aged between 0.3 and 86 years, spanning eight major cell types. To enable unbiased evaluation, a randomly selected subset of 57 individuals was entirely held out from training and used exclusively for testing.

7.4.3.5 *Multi-species dataset (Mouse + Fly + Human)*

To enable comparative analysis of aging across species, we constructed a multi-species dataset by combining the TMS mouse data, the Aging Fly Cell Atlas, and the Tabula Sapiens (TS) human cell atlas [236]. Tabula Sapiens is a comprehensive human single-cell transcriptomic atlas spanning multiple organs and systems from healthy adult donors, serving as a reference for human cellular diversity.

To align gene expression features across species, we first identified one-to-one orthologs between mouse and human, and between fly and human, retaining only genes with one-to-one orthology in both comparisons. We then merged the datasets across species based on the shared orthologous gene set.

We performed standard filtering to ensure quality and comparability: genes expressed in fewer than 5 cells and cells expressing fewer than 50 genes were removed. Additionally, we excluded cells with fewer than 300 total counts.

To ensure balanced representation of cell types and tissues, we retained the 30 most abundant cell types from the human and mouse datasets and the top 10 cell types from

the fly dataset. Additionally, for the mouse and human datasets, we restricted the data to the 17 overlapping tissues shared between the two species. We also filtered out age groups with fewer than 1,000 cells to maintain statistical robustness. Only genes present in the final filtered datasets were retained.

To enable biologically meaningful comparisons across species, we mapped mouse and fly ages to human-equivalent ages. Mouse-human equivalences followed established developmental phase mappings from Flurkey, Currer, and Harrison [83], while fly-human mappings were derived by aligning lifespan curves. The specific mappings used were:

- Mouse → Human: 1m → 12.5y, 3m → 20y, 18m → 56y, 21m → 62.5y, 24m → 69y, 30m → 81y.
- Fly → Human: 5d → 5y, 30d → 40y, 50d → 68y, 70d → 90y.

After preprocessing, the final multi-species dataset contained 793,747 cells and 2,515 genes. The human subset included 520,139 cells, the mouse subset included 171,924 cells, and the fly subset included 101,684 cells.

7.4.3.6 *Multi-species dataset (Mouse + Human)*

To evaluate ACE's ability to generalize to previously unseen cellular contexts, we constructed a multi-species dataset by combining the TMS mouse dataset [3] with the TS human dataset [236].

We began by identifying one-to-one orthologous genes between mouse and human. To ensure high-quality input data, we excluded genes expressed in fewer than 5 cells and filtered out cells with fewer than 500 detected genes or fewer than 3,000 total counts.

To maintain consistency and comparability across species, we retained only cell types shared between mouse and human with at least 100 cells in each species. Likewise, we filtered tissues to include only those represented by at least 100 cells in both species. This filtering yielded 40 common cell types and 15 shared tissues.

We selected the top 2,000 highly variable genes using the Seurat v3 method, applying batch correction via the `batch` key on the raw count layer. In addition to these highly variable genes, we included aging-associated genes identified by Zhang et al. [336] to augment aging-related signatures.

To enable age-aligned comparisons across species, we mapped mouse chronological ages to their human-equivalent values using established cross-species age conversion tables [83].

The resulting dataset consists of 513,924 cells and 3,991 genes across 40 cell types and 15 tissues. Among these, 15 cell types with sufficient cell numbers and age coverage were designated as held-out sets for evaluating ACE's cross-cell-type generalization capability.

7.4.4 *Evaluation metrics*

We used quantitative metrics implemented in the `scikit-learn` Python package to assess model performance. To enable visual comparisons across models and metrics, we generated overview tables following the format of contrastiveVI [311], where individual scores are shown as circles and aggregate scores as bars. All metrics were min-max scaled to allow cross-metric comparisons. These scaled values were then averaged into two composite scores: *Age Spatial Autocorrelation*, which quantifies the coherence of learned aging signatures, and *Cell Trait Mix Scores*, which evaluate how well cells from different background traits (e.g., cell types, tissues, sexes) are intermixed in the learned age space. A final overall score was computed by averaging these two aggregates.

For *Age Spatial Autocorrelation*, we computed both Geary's C and Moran's I to quantify the spatial autocorrelation of the aging variables learned by the global ACE model. While Moran's I captures broader global trends in spatial structure, Geary's C is more sensitive to local differences among neighboring cells.

For *Cell Trait Mix Scores*, we computed the Entropy of Mixing, which measures the diversity of background groups among the nearest neighbors of each cell in the learned age space. High entropy indicates that cells from different groups are well mixed, suggesting that the learned representation successfully removes background-specific variation. This is desirable for isolating aging signatures that generalize across different cellular contexts.

GEARY'S C. To quantify local spatial autocorrelation in the learned aging latent space, we compute Geary's C:

$$C = \frac{(N-1)\sum_{i,j} w_{ij}(u_i - u_j)^2}{2W\sum_i (u_i - \bar{u})^2} \tag{7.4}$$

Here, u_i denotes the aging variable for cell i, $\bar{u}$ is its global mean, w_{ij} is the spatial weight between cells i and j, and $W = \sum_{i,j} w_{ij}$. We define $w_{ij} = 1$ if cell j is among the 50 nearest neighbors of cell i, and 0 otherwise. Since lower values of C indicate stronger spatial autocorrelation, we report $1 - C$ to maintain consistency with other metrics where higher values indicate better performance.

MORAN'S I. To assess global spatial autocorrelation in the aging latent space, we compute Moran's I:

$$I = \frac{N\sum_{i,j} w_{ij}(u_i - \bar{u})(u_j - \bar{u})}{W\sum_i (u_i - \bar{u})^2} \tag{7.5}$$

where u_i is the aging variable for cell i, $\bar{u}$ is the global mean, and w_{ij} is defined as above. The denominator normalizes the spatial covariance between neighboring cells by the overall variance. Higher values of I indicate stronger spatial structure across the entire dataset.

ENTROPY OF MIXING. To quantify how well cells from different background covariates (e.g., cell types, tissues, sexes) are mixed in the learned embedding space, we compute the entropy of mixing as described in prior works [106, 168]. Specifically, for each cell u, we calculate the empirical distribution of background groups among its 50 nearest neighbors, denoted by B_u. Let p_i represent the proportion of neighbors from group i among the c possible groups, such that $\sum_{i=1}^{c} p_i = 1$. The local entropy of mixing for cell u is computed as:

$$H(u) = -\sum_{i=1}^{c} p_i \log p_i$$

We repeat this process over 100 randomly selected cells and report the average entropy. Higher values of this metric indicate stronger mixing of background groups, reflecting better removal of unwanted variation in the embedding.

7.4.5 *Interpretability of the ACE Model*

To interpret gene-level contributions to predicted biological age, we applied the *Expected Gradients* (EG) method [76]. Expected Gradients extends Integrated Gradients by averaging over a distribution of background samples, producing more stable and representative feature attributions for deep neural networks.

DEFINITION. Given a model f that maps an input gene expression vector $x \in \mathbb{R}^G$ to a scalar output (in our case, predicted biological age $\hat{a}$), the Expected Gradient attribution for gene g is defined as:

$$EG_g(x) = \mathbb{E}_{x' \sim \mathcal{D}_{bg}} \left[(x_g - x'_g) \cdot \int_{\alpha=0}^{1} \frac{\partial f(x' + \alpha(x - x'))}{\partial x_g} \, d\alpha \right] \tag{7.6}$$

Here: - x is a foreground (target) sample, - x' is drawn from a background distribution $\mathcal{D}_{bg}$, - g indexes the genes, - and the gradient is taken with respect to input gene g.

FOREGROUND AND BACKGROUND SAMPLES. To compute EG values in our framework: - **Foreground samples** are 200 cells randomly sampled from each tissue-cell-type pair. These represent the cells for which we wish to interpret the predicted biological age. - **Background samples** are 800 cells drawn from the same tissue-cell-type pair, representing the reference distribution $\mathcal{D}_{bg}$.

APPLICATION TO GLOBAL AND LOCAL MODELS. For the *global aging model*, all cells from multiple tissue-cell-type pairs were pooled, and EG values were computed using aggregated foreground and background samples. Gene attributions were then obtained by averaging the absolute EG values across all foreground samples.

For the *local aging model*, we computed EG values separately within each tissue-cell-type pair using the respective foreground and background sets. The absolute EG values were averaged across the foreground cells in each pair. Genes were then ranked according to their mean attributions, reflecting their local importance in age prediction.

7.4.6 *Pathway enrichment analysis*

Pathway enrichment analysis is a computational procedure used to determine whether a predefined set of genes (i.e., a biological pathway) shows statistically significant expression changes between different biological states. We used the open-source `GSEAPY` Python package for this analysis, applying the `prerank` method. Pathways with a false discovery rate (FDR) q-value below 0.05, adjusted using the Benjamini-Hochberg procedure, were considered significantly enriched and are reported in this study.

The `prerank` method is a variant of Gene Set Enrichment Analysis (GSEA) that operates on a pre-ordered list of genes instead of raw expression data. In our study, genes are ranked based on Expected Gradient values, which reflect their biological relevance. GSEA then evaluates whether genes from a particular pathway are disproportionately located near the top or bottom of the ranked list, indicating significant enrichment beyond random chance.

We conducted enrichment analysis across the following gene set libraries:

- **Aging Perturbations from GEO up/down:** Curated gene sets capturing consistent aging-associated upregulation or downregulation across multiple studies in the Gene Expression Omnibus (GEO), enabling cross-dataset evaluation of aging signatures.
- **KEGG (Kyoto Encyclopedia of Genes and Genomes):** A structured database of biological pathways encompassing metabolism, signal transduction, and disease mechanisms.
- **Reactome:** A manually curated knowledgebase detailing molecular events in signal transduction, immune function, gene regulation, and other cellular processes.
- **GO (Gene Ontology):** Hierarchical gene annotation system that organizes gene functions into three broad domains: Biological Process (BP), Cellular Component (CC), and Molecular Function (MF).

7.4.7 *Wet lab validation in C. elegans*

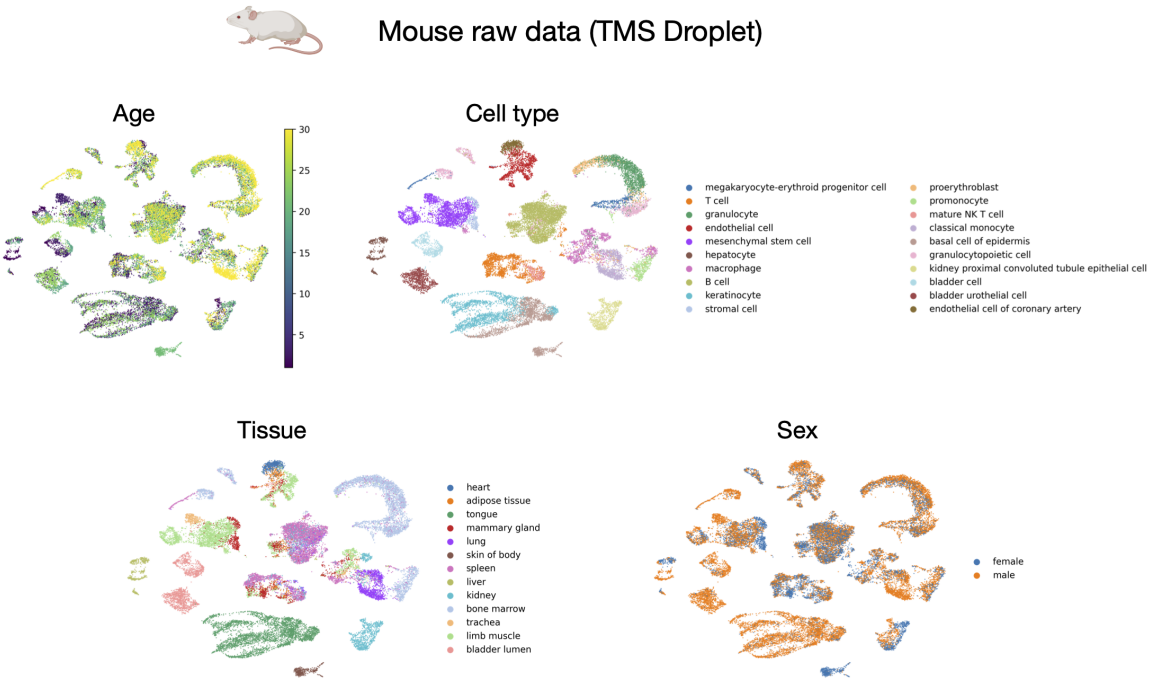
To assess the biological relevance of aging-associated genes identified by ACE, we performed RNA interference (RNAi) knockdown experiments in *Caenorhabditis elegans* and evaluated their effects on lifespan. Specifically, we selected 8 mouse genes from among the top 150 most important global aging genes identified by the ACE global aging model:

Ppia, *Ftl1*, *Gpx3*, *Rpl37*, *Serpina3k*, *Tpt1*, *Rps27a*, and *Uba52*. These genes were prioritized based on their relevance to proteostasis, evolutionary conservation across species, and feasibility for functional testing in *C. elegans*, avoiding those known to cause larval arrest or lethality upon knockdown.

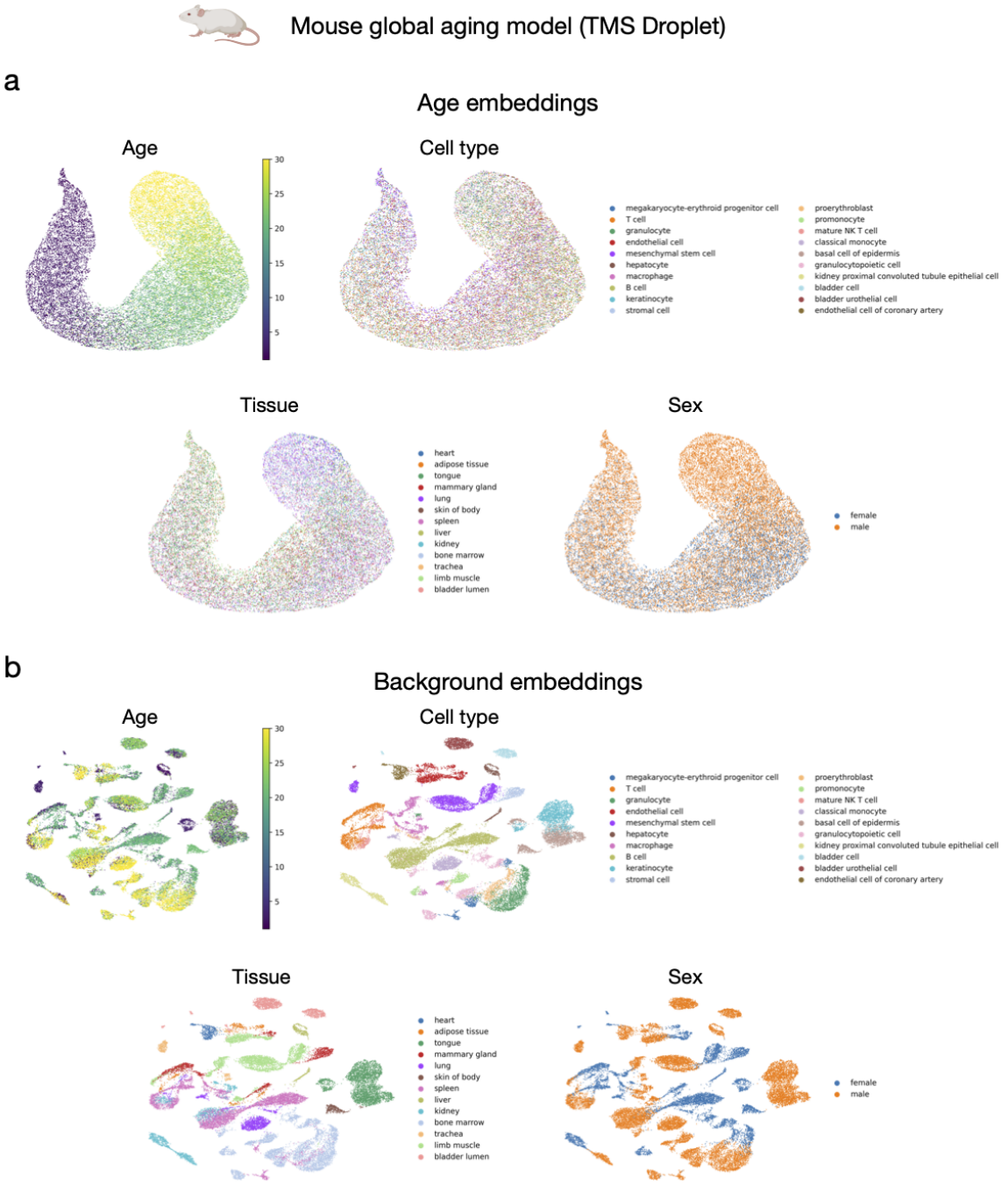
Each mouse gene was mapped to its corresponding *C. elegans* ortholog(s), resulting in a total of 15 *C. elegans* genes tested. RNAi knockdown was carried out individually for each ortholog, and survival curves were compared to those of animals fed with an empty vector (EV) control.

ANIMAL HUSBANDRY AND EXPERIMENTAL SETUP. Wild-type hermaphrodite *C. elegans* (strain N2 CGCb) were maintained under standard culture conditions at 20°C on nematode growth medium (NGM) plates seeded with *E. coli* OP50 [33]. Animals were synchronized using hypochlorite treatment and grown on OP50-seeded NGM plates until the late L4/young adult (YA) molt, approximately 50 hours post-feeding at 20°C. At this stage, animals were transferred to RNAi food on 12-well plates to initiate the aging experiments.

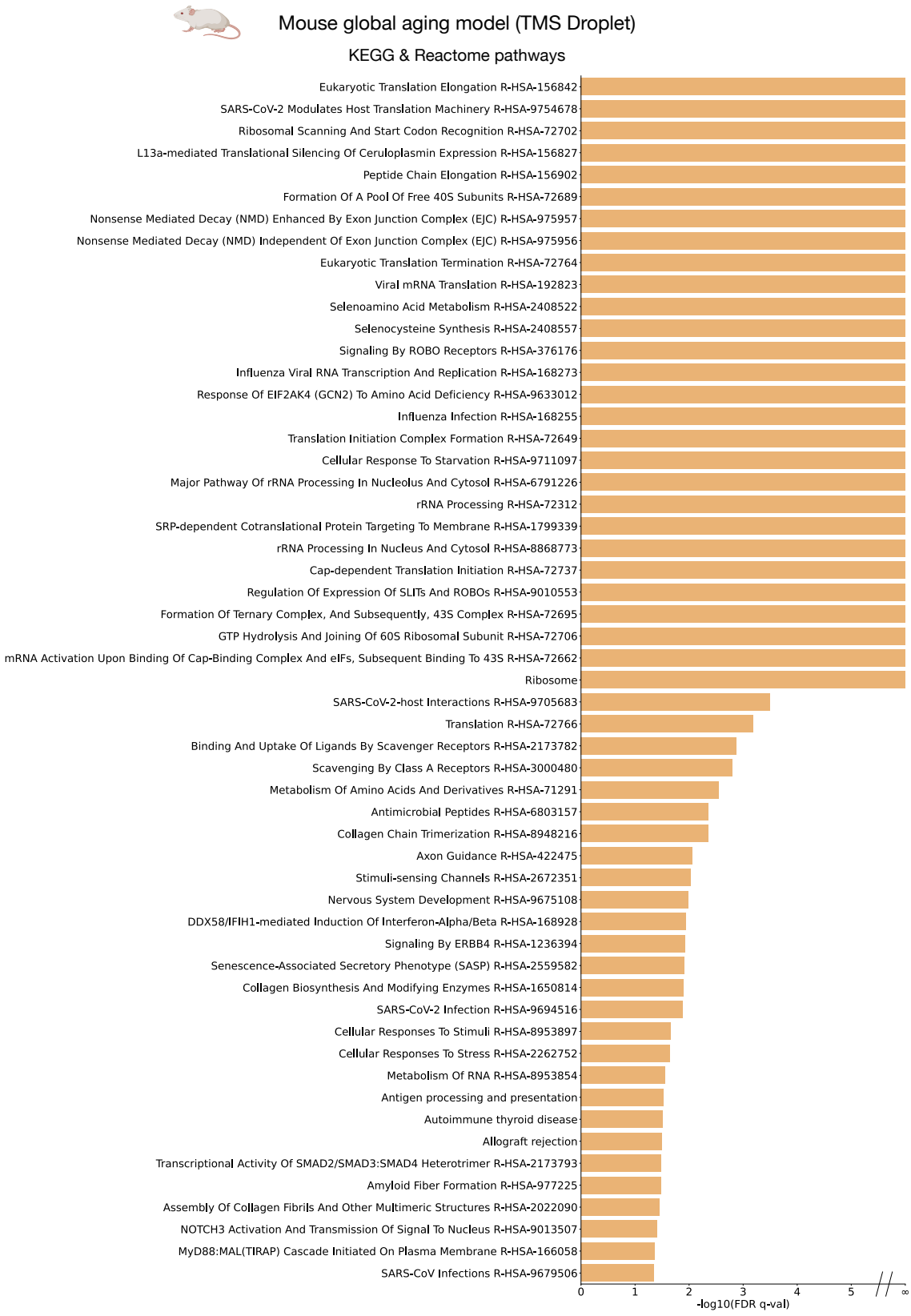
LIFESPAN MEASUREMENTS. Lifespan assays were conducted using the WormBot imaging platform [224], which quantifies the time of death (defined as cessation of movement) for individual worms based on time-lapse video recordings. Each RNAi knockdown experiment was performed using sequence-validated RNAi clones in HT115 *E. coli*, grown on NGM plates supplemented with 5-fluoro-2′-deoxyuridine (FUDR) to inhibit progeny production. The FUDR and RNAi clones may also inhibit mitotic development, so worms were plated at the L4/YA stage to ensure that any observed lifespan changes were attributable to gene knockdown effects during aging rather than development. Each well contained 20-30 animals.



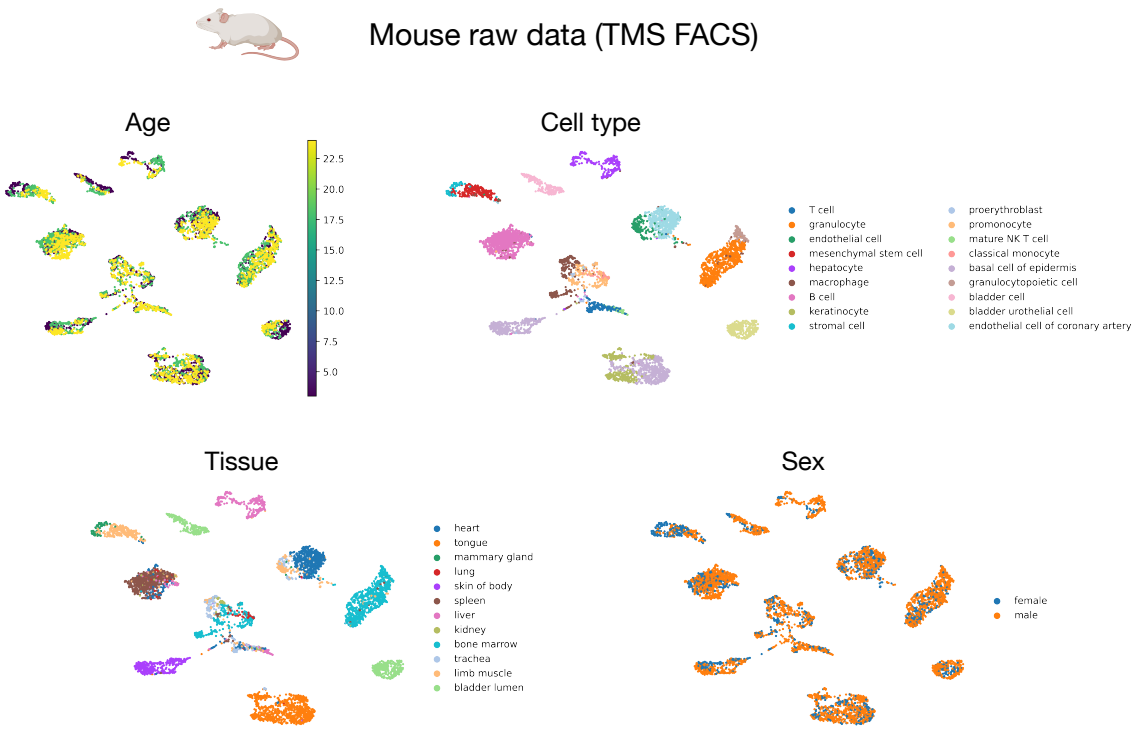
Supplementary Figure A.7.8: **Visualization of the TMS Droplet dataset using UMAP applied to normalized count data.** Plots are colored by age (top left), cell type (top right), tissue (bottom left), and sex (bottom right).



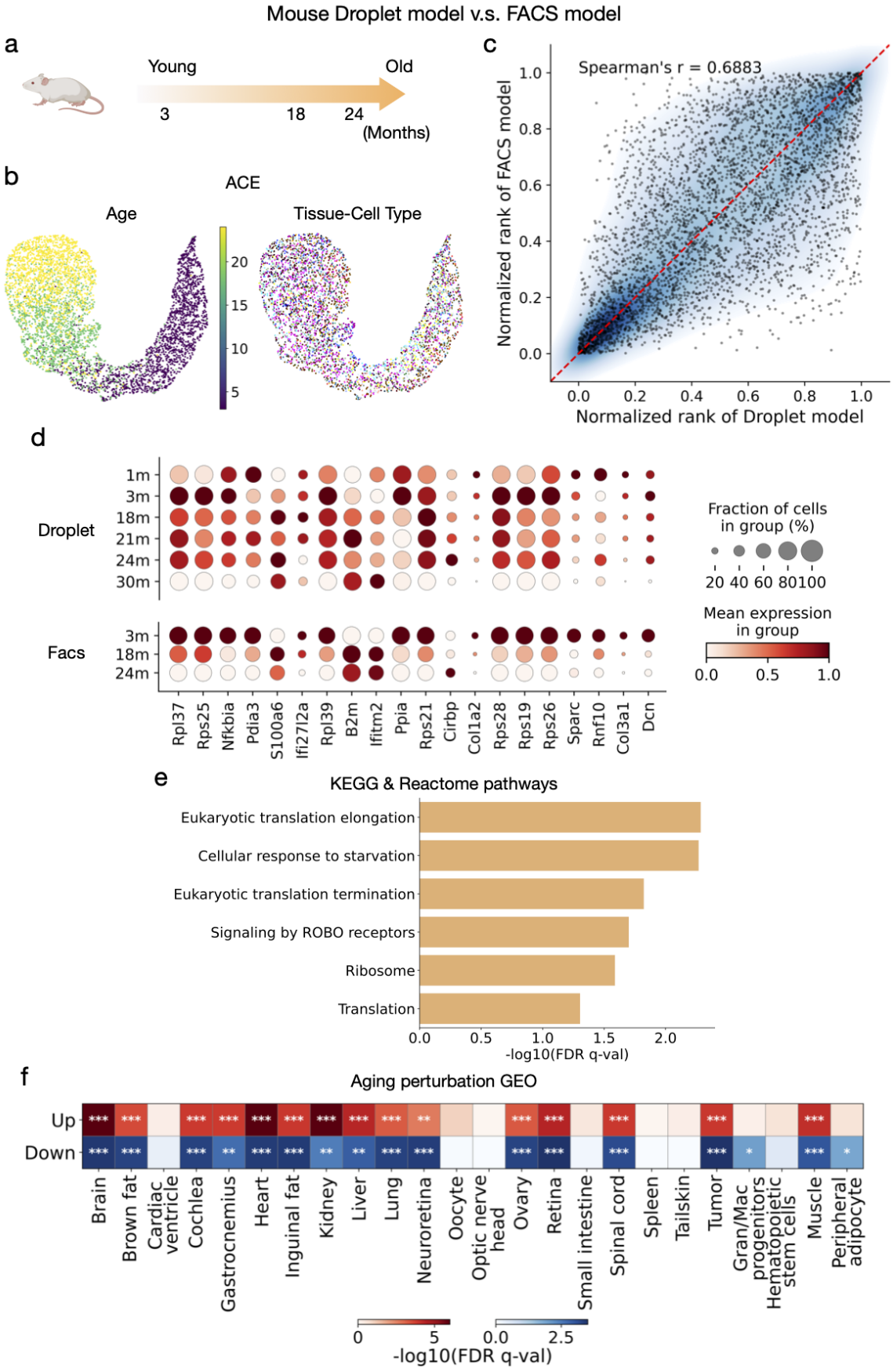
Supplementary Figure A.7.9: **Visualization of age and background embeddings from the global aging model learned by ACE using the TMS Droplet dataset. a.** UMAP of age embeddings colored by age, cell type, tissue, and sex. **b.** UMAP of background embeddings colored by the same attributes.



Supplementary Figure A.7.10: **Full list of KEGG and Reactome pathways significantly enriched by the mouse global aging model on the TMS Droplet dataset.** Significance was assessed at FDR q < 0.05 using the Benjamini-Hochberg correction.

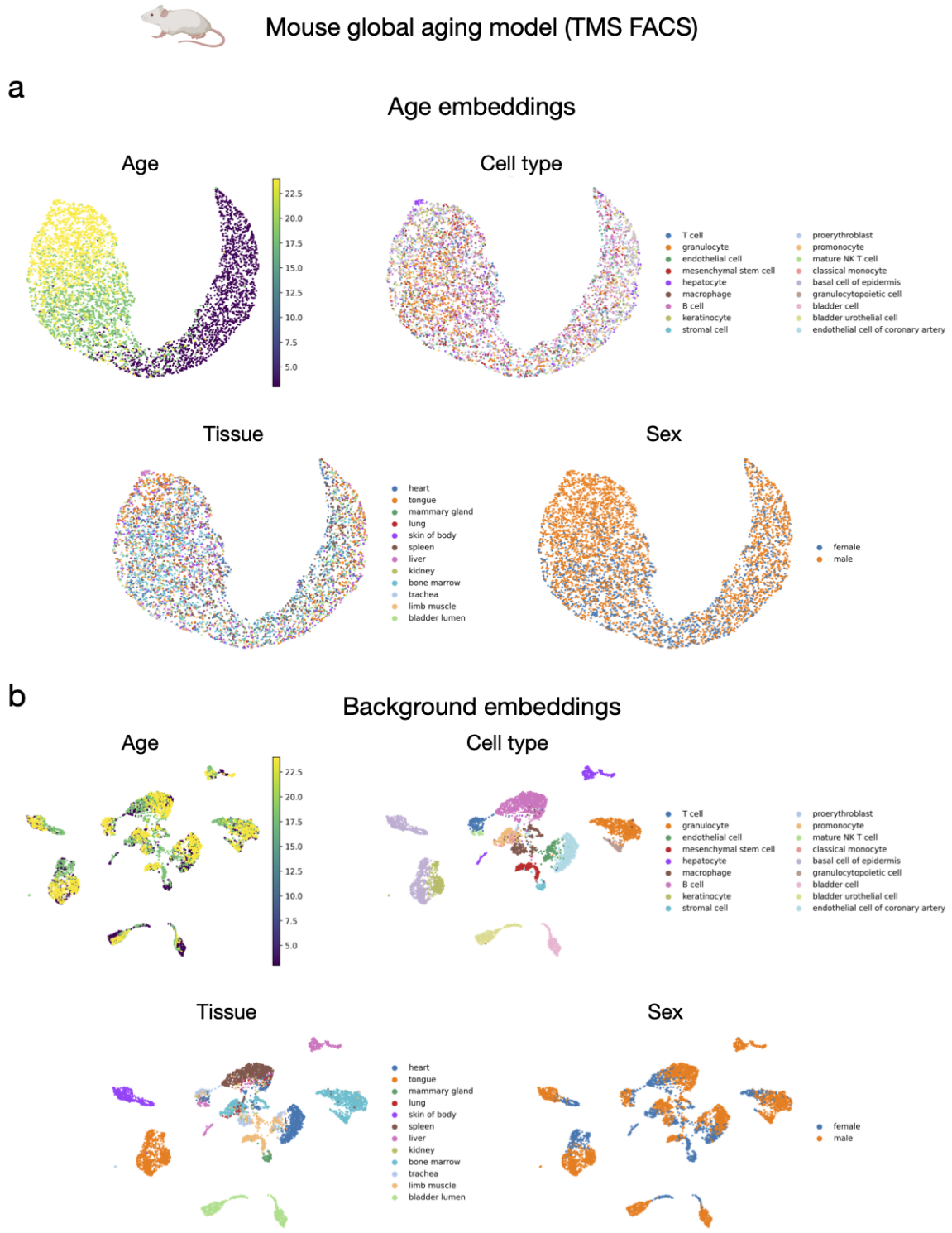


Supplementary Figure A.7.11: **Visualization of the TMS FACS dataset using UMAP applied to normalized count data.** Plots are colored by age (top left), cell type (top right), tissue (bottom left), and sex (bottom right).

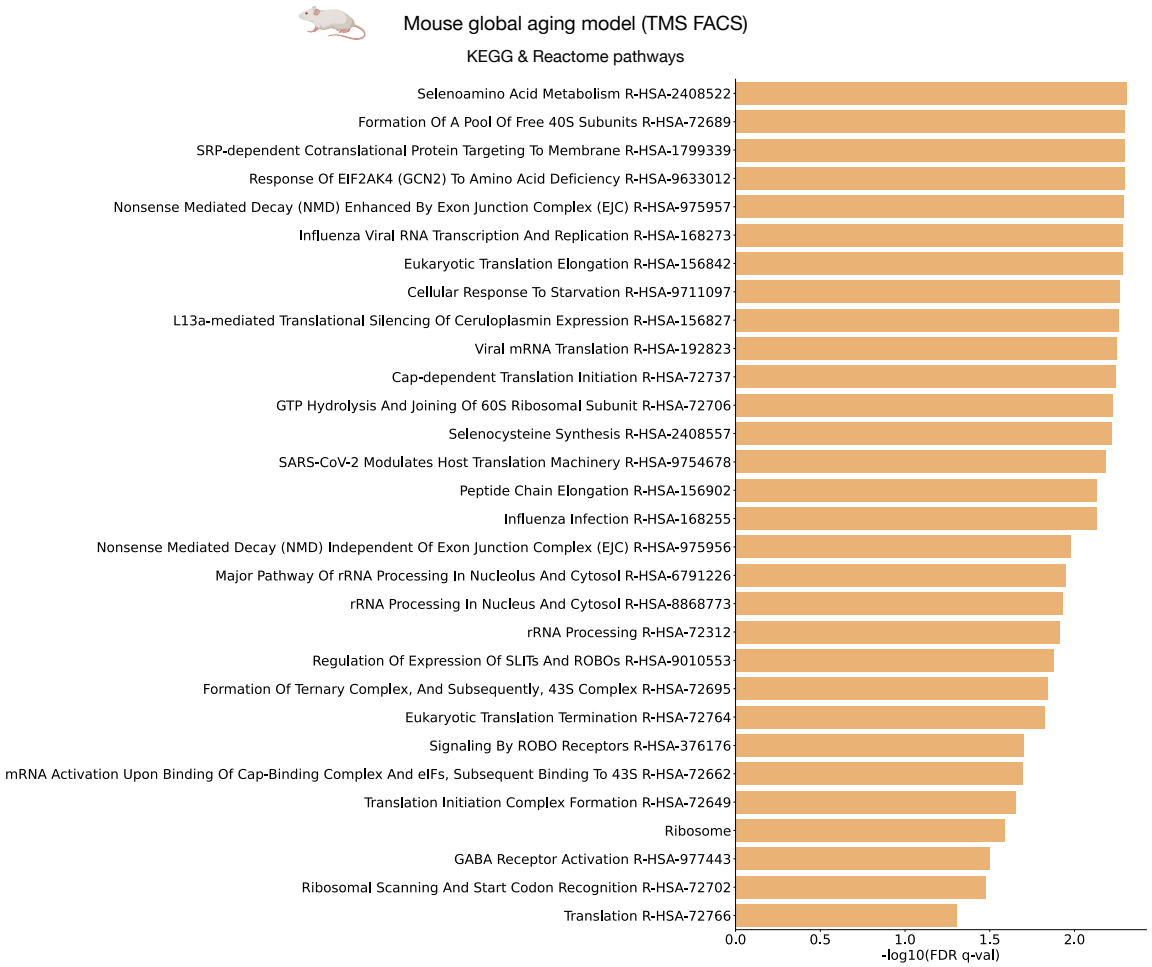


Supplementary Figure A.7.12

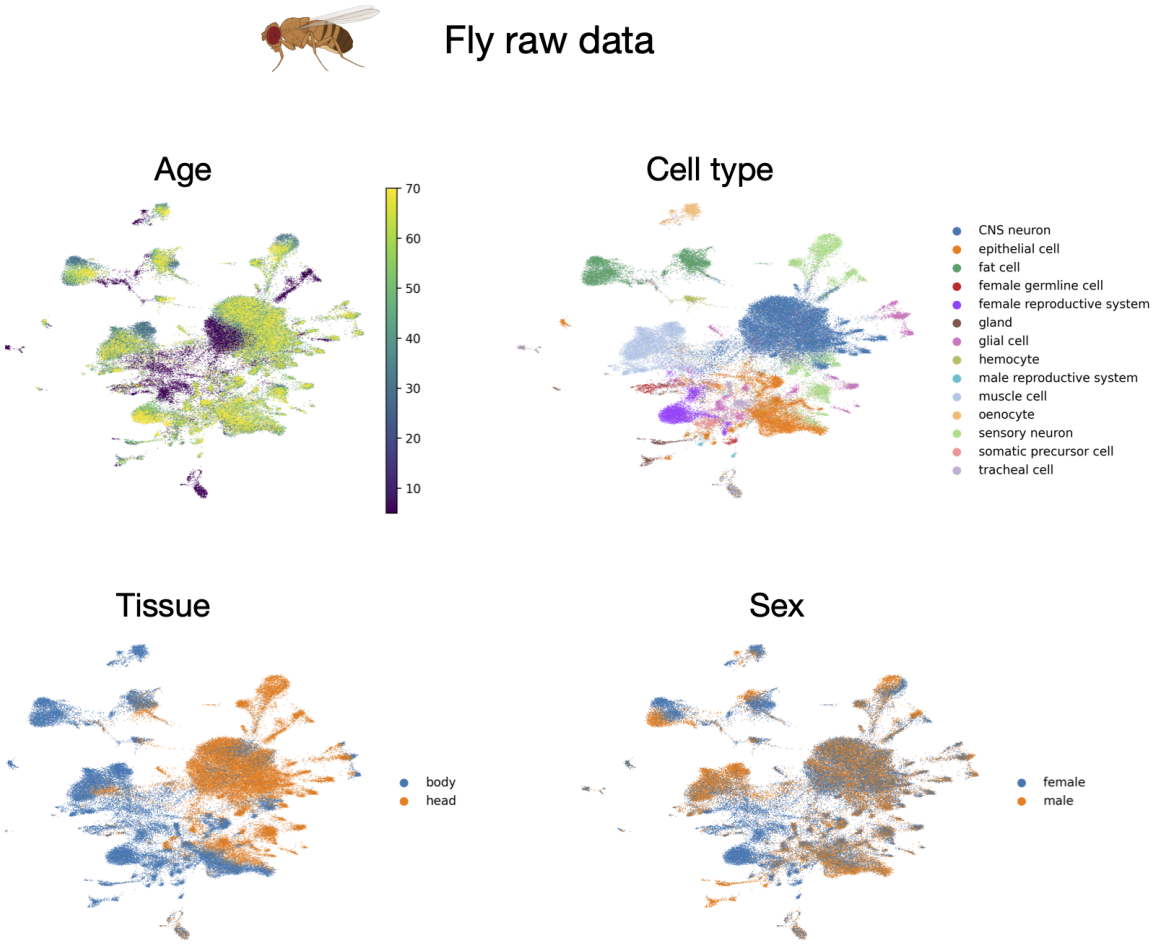
Supplementary Figure A.7.12: **Comparison of mouse Droplet and FACS models to validate global aging signatures. a.** Aging time points represented in the TMS FACS dataset, spanning from 3 to 24 months. The FACS dataset was processed using consistent filtering steps with the Droplet dataset, including the removal of low-quality cells, rare tissues/cell types, and donors with insufficient representation, resulting in 24,546 cells and 4,120 genes across three age groups (3m, 18m, and 24m). **b.** UMAP visualization of age embeddings learned by ACE using the TMS FACS dataset, showing that the learned embeddings capture a strong age-related gradient (left) while being disentangled from tissue and cell-type identity (right). **c.** Spearman correlation between the normalized gene rankings of the Droplet and FACS models (Spearman's $r = 0.6883$), demonstrating strong concordance and suggesting that the Droplet-derived global aging signatures are robust and reproducible in an independent dataset. **d.** Expression dynamics of selected overlapping genes from the top 150 ranked genes in both Droplet and FACS models. These genes exhibit consistent age-associated changes, with similar patterns of upregulation or downregulation across matched time points, reinforcing their roles as core aging-related genes. The size of each dot represents the fraction of cells in a group expressing the gene, while color intensity reflects the mean expression within that group. **e.** Selected overlapping KEGG and Reactome pathways identified by both models, including translation-related processes (e.g., eukaryotic translation elongation and termination), cellular response to starvation, and ROBO signaling. These representative pathways correspond to those in Figure **??**h and highlight biological processes consistently implicated in aging across datasets. **f.** Gene set enrichment analysis results from the FACS model, using the "Aging Perturbations from GEO Up" and "Aging Perturbations from GEO Down" collections. These databases aggregate results from GEO studies comparing aged versus young tissues and cell types. The FACS-derived global aging gene rankings show significant enrichment in many of these curated gene sets. Red boxes indicate enrichment in *upregulated* gene sets with age, whereas blue boxes indicate enrichment in *downregulated* gene sets. Color intensity represents $-\log_{10}$(FDR q-value), and significance was assessed using Benjamini-Hochberg correction. Asterisks indicate significance thresholds ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$). Notably, genes upregulated with aging are strongly enriched in multiple tissues (e.g., brain, heart, muscle, adipose), while downregulated genes are observed in select tissues, confirming that the FACS model captures both shared and tissue-specific aging perturbations. Gran/Mac progenitors represent granulocyte/macrophage progenitors.



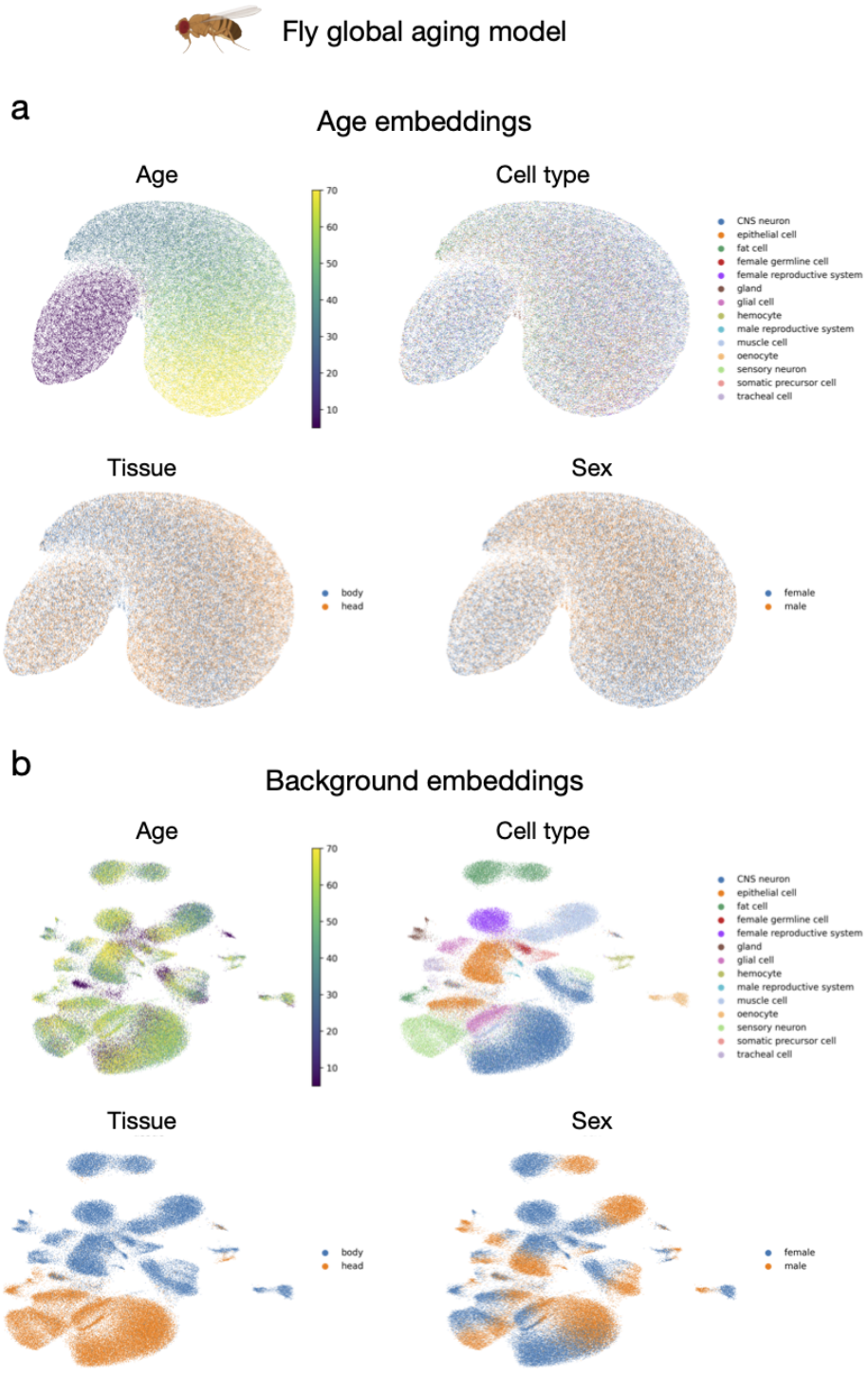
Supplementary Figure A.7.13: **Visualization of age and background embeddings from the global aging model learned by ACE using the TMS FACS dataset. a.** UMAP of age embeddings colored by age, cell type, tissue, and sex. **b.** UMAP of background embeddings colored by the same attributes.



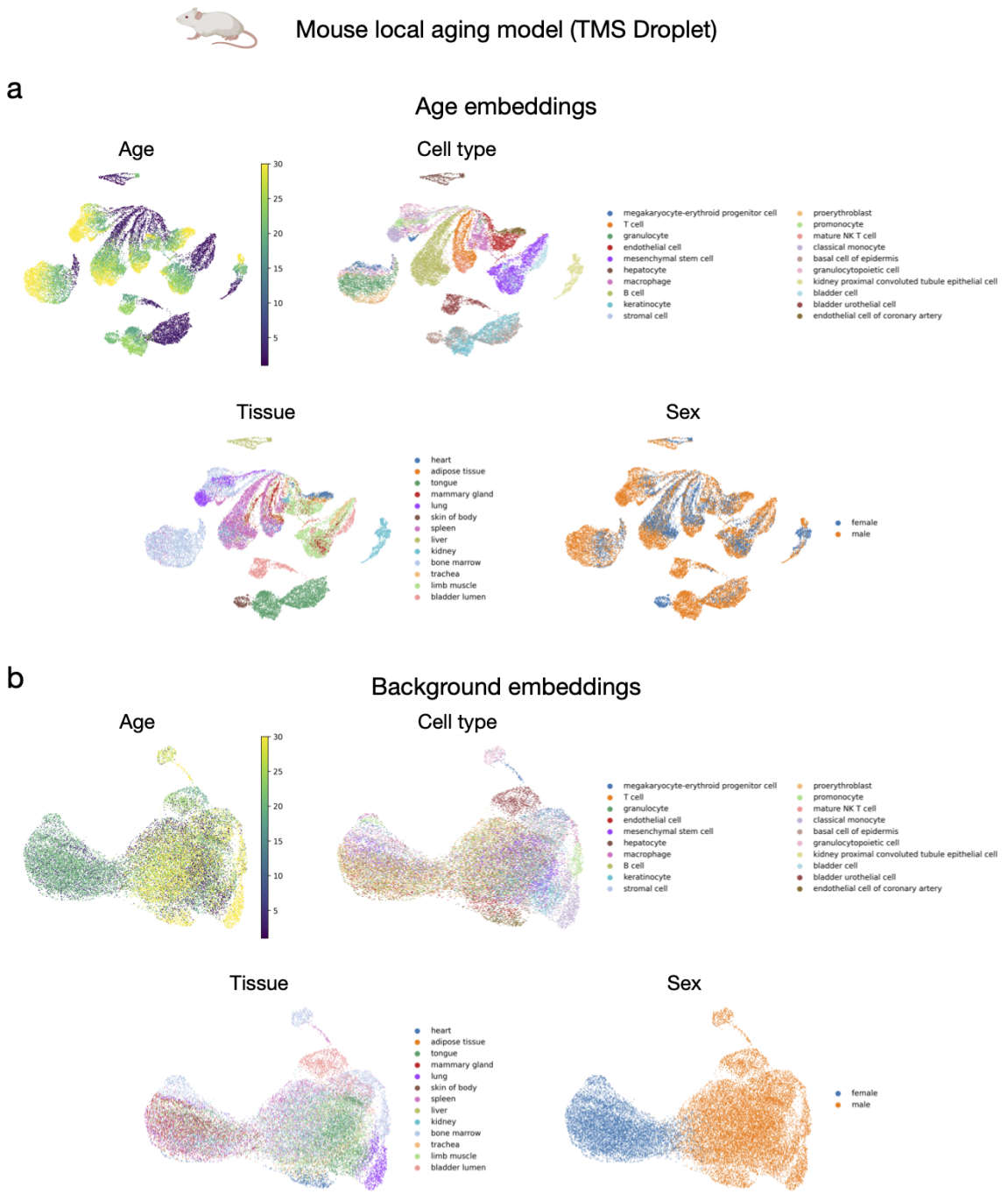
Supplementary Figure A.7.14: **Full list of KEGG and Reactome pathways significantly enriched by the mouse global aging model on the TMS FACS dataset.** Significance was assessed at FDR q < 0.05 using the Benjamini-Hochberg correction. A total of 29 pathways identified in the FACS model overlapped with those discovered by the Droplet model (Supplementary Figure **??**), highlighting robust and reproducible aging-associated biological processes across datasets.



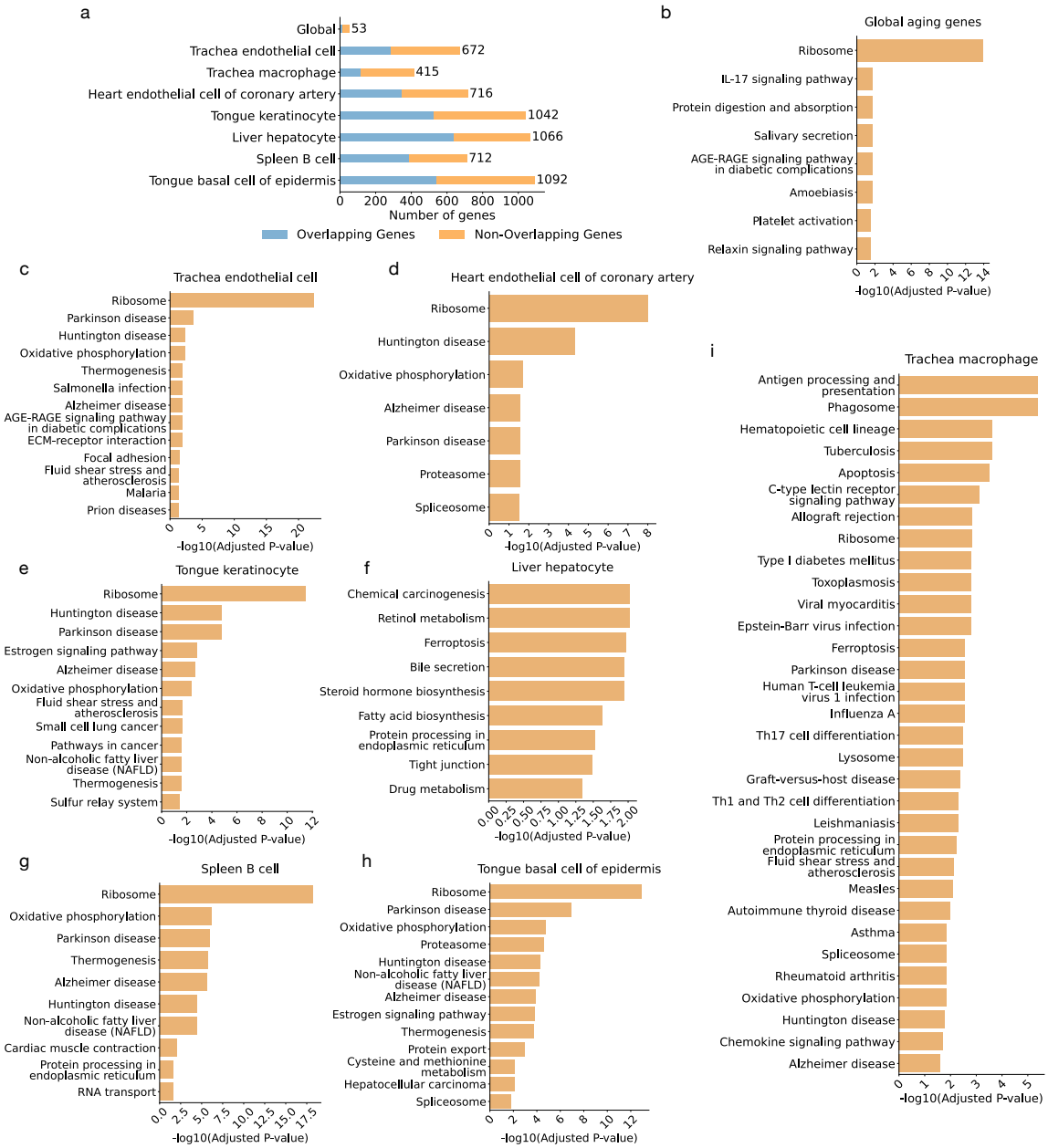
Supplementary Figure A.7.15: **Visualization of the Aging Fly Cell Atlas using UMAP applied to normalized count data.** Plots are colored by age (top left), cell type (top right), tissue (bottom left), and sex (bottom right).



Supplementary Figure A.7.16: **Visualization of age and background embeddings from the global aging model learned by ACE using the Aging Fly Cell Atlas. a.** UMAP of age embeddings colored by age, cell type, tissue, and sex. **b.** UMAP of background embeddings colored by the same attributes.

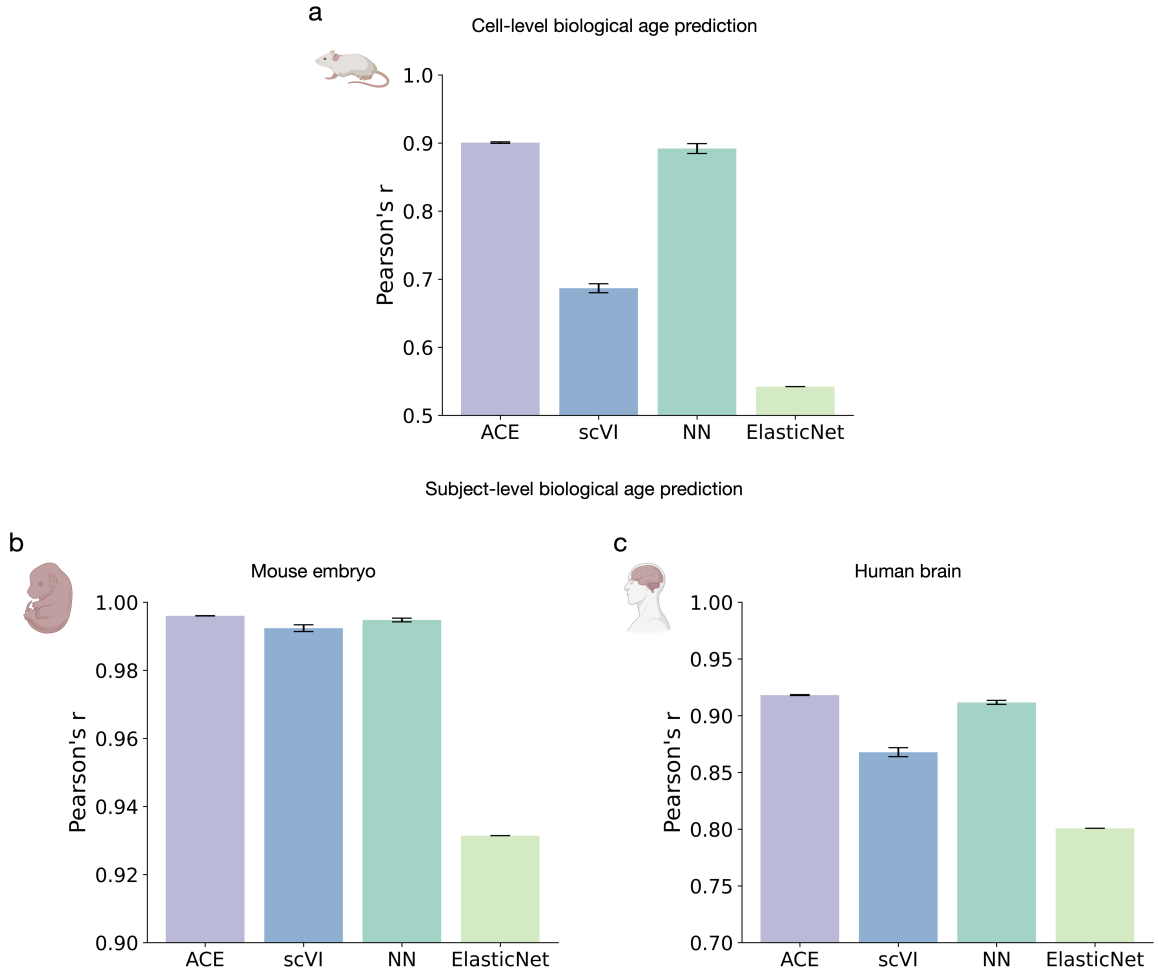


Supplementary Figure A.7.17: **Visualization of age and background embeddings from the local aging model learned by ACE using the TMS Droplet dataset. a.** UMAP of age embeddings colored by age, cell type, tissue, and sex. **b.** UMAP of background embeddings colored by the same attributes.

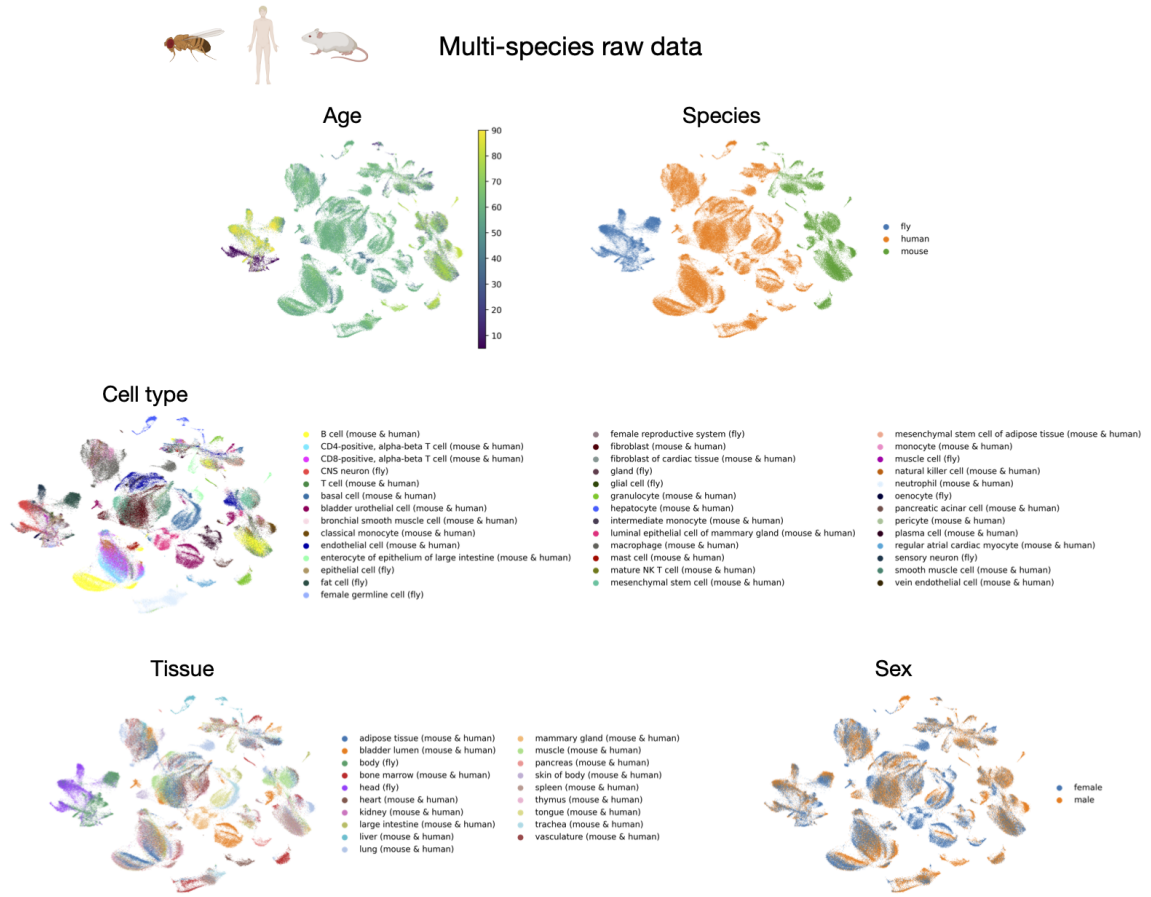


Supplementary Figure A.7.18

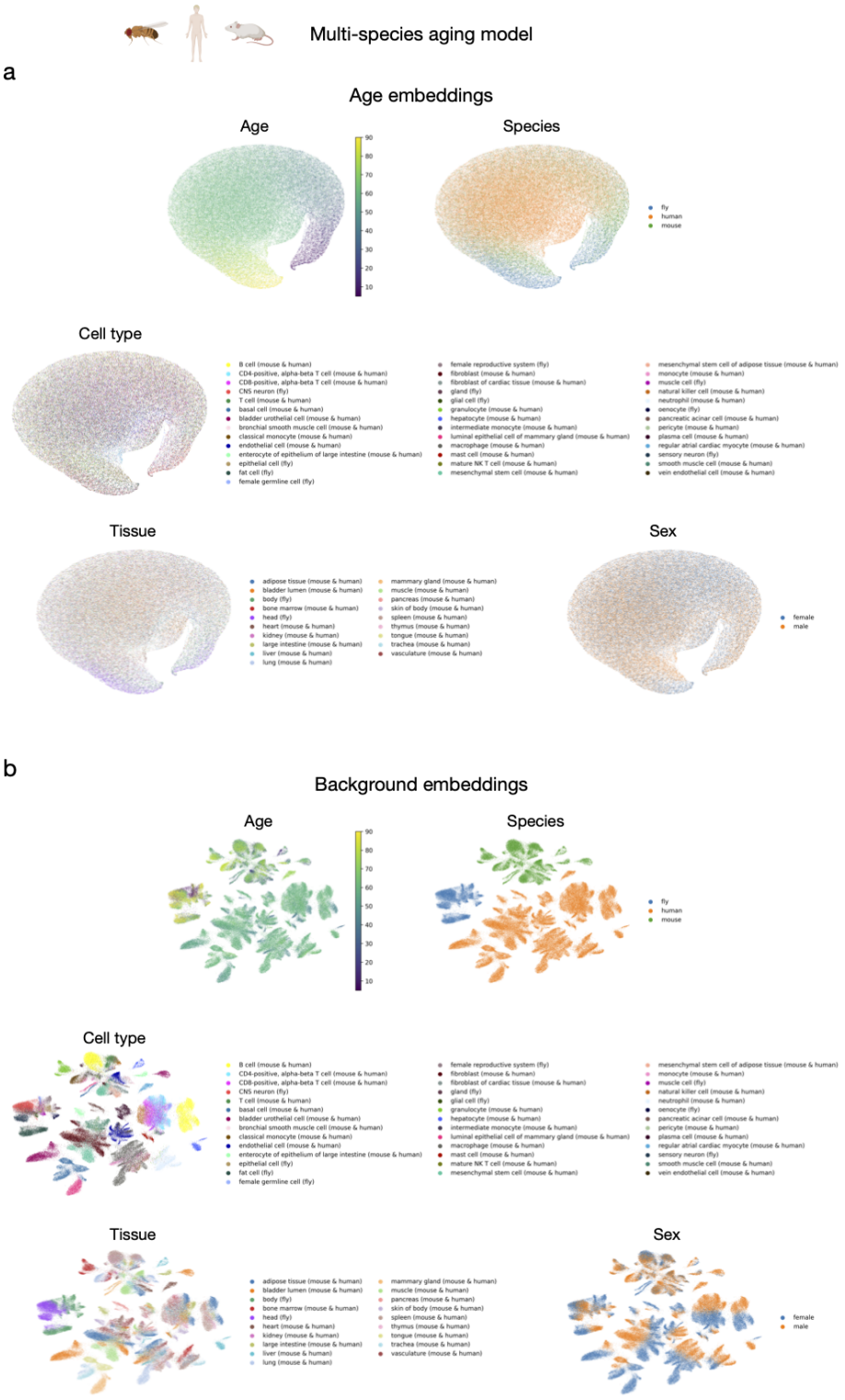
Supplementary Figure A.7.18: **Comparison of ACE-derived global and tissue-cell-type-specific aging genes with those identified by the linear model from Zhang, Pisco, Darmanis, and Zou [336]. a.** Bar plot showing the number of overlapping and non-overlapping aging genes identified by the ACE model and the linear model. For each group, we matched the number of top-ranked ACE genes to the number of aging genes reported by the linear model, allowing for a direct and fair comparison between the two approaches. This comparison shows that ACE identifies a substantial number of additional aging-related genes beyond those captured by the linear model, demonstrating its improved sensitivity for detecting comprehensive aging signatures. **b-i.** Gene set enrichment analysis of KEGG pathways performed on the non-overlapping ACE genes for the global group and each tissue-cell-type-specific group. The enriched pathways are strongly associated with known hallmarks of aging. These analyses demonstrate that ACE not only captures key aging pathways identified by the linear model but also uncovers additional aging signatures missed by the linear approach. By identifying both shared global aging pathways and tissue-cell-type-specific pathways, ACE provides a deeper understanding of the complex and heterogeneous nature of the aging process. Significance was assessed at FDR $q < 0.05$ using the Benjamini-Hochberg correction.



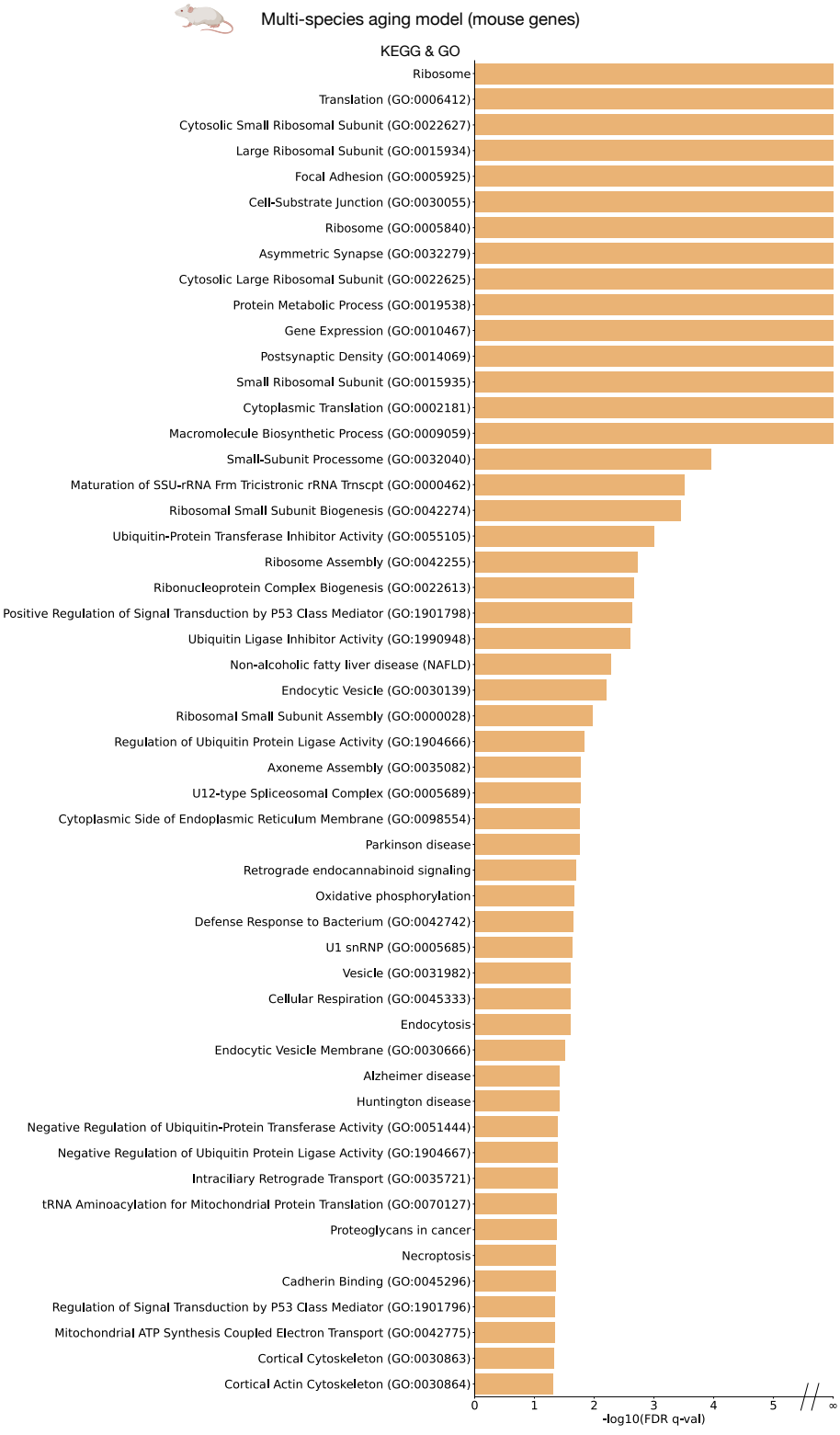
Supplementary Figure A.7.19: **Biological age prediction performance comparison.** (a) Cell-level biological age prediction across methods. ACE achieves high accuracy compared to baseline models, including scVI, NN, and ElasticNet. (b, c) Subject-level biological age prediction for mouse embryo (b) and human brain (c) datasets. ACE consistently performs competitively or better than other methods, demonstrating its effectiveness in capturing aging-related variation at both the cell and subject levels. Error bars represent standard errors of Pearson's correlation coefficients calculated across 10 replicates.



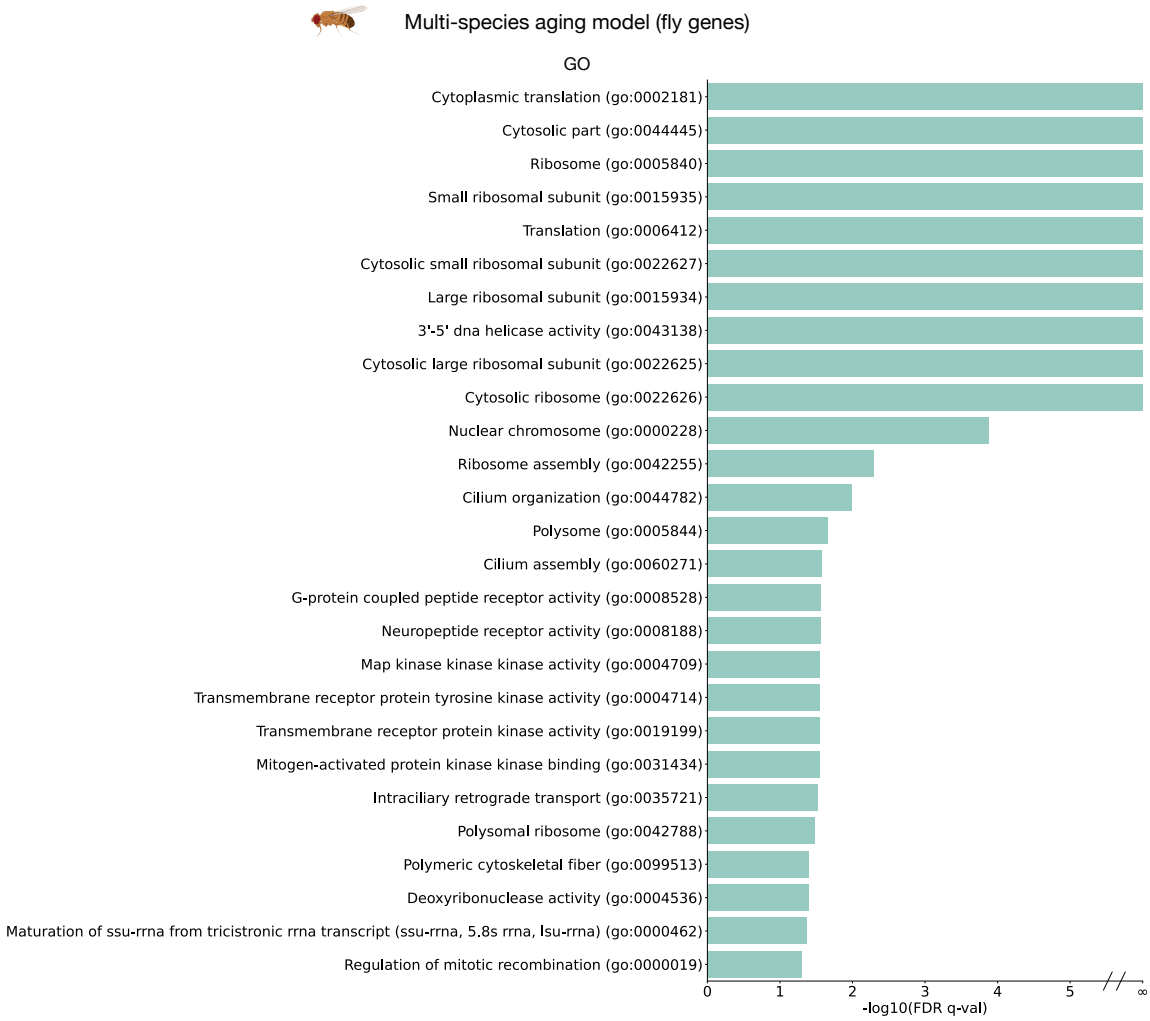
Supplementary Figure A.7.20: **Visualization of the multi-species dataset using UMAP applied to normalized count data.** Plots are colored by age (top left), species (top right), cell type (middle), tissue (bottom left), and sex (bottom right).



Supplementary Figure A.7.21: **Visualization of age and background embeddings from the multi-species aging model learned by ACE. a.** UMAP of age embeddings colored by age, species, cell type, tissue, and sex. **b.** UMAP of background embeddings colored by the same attributes.

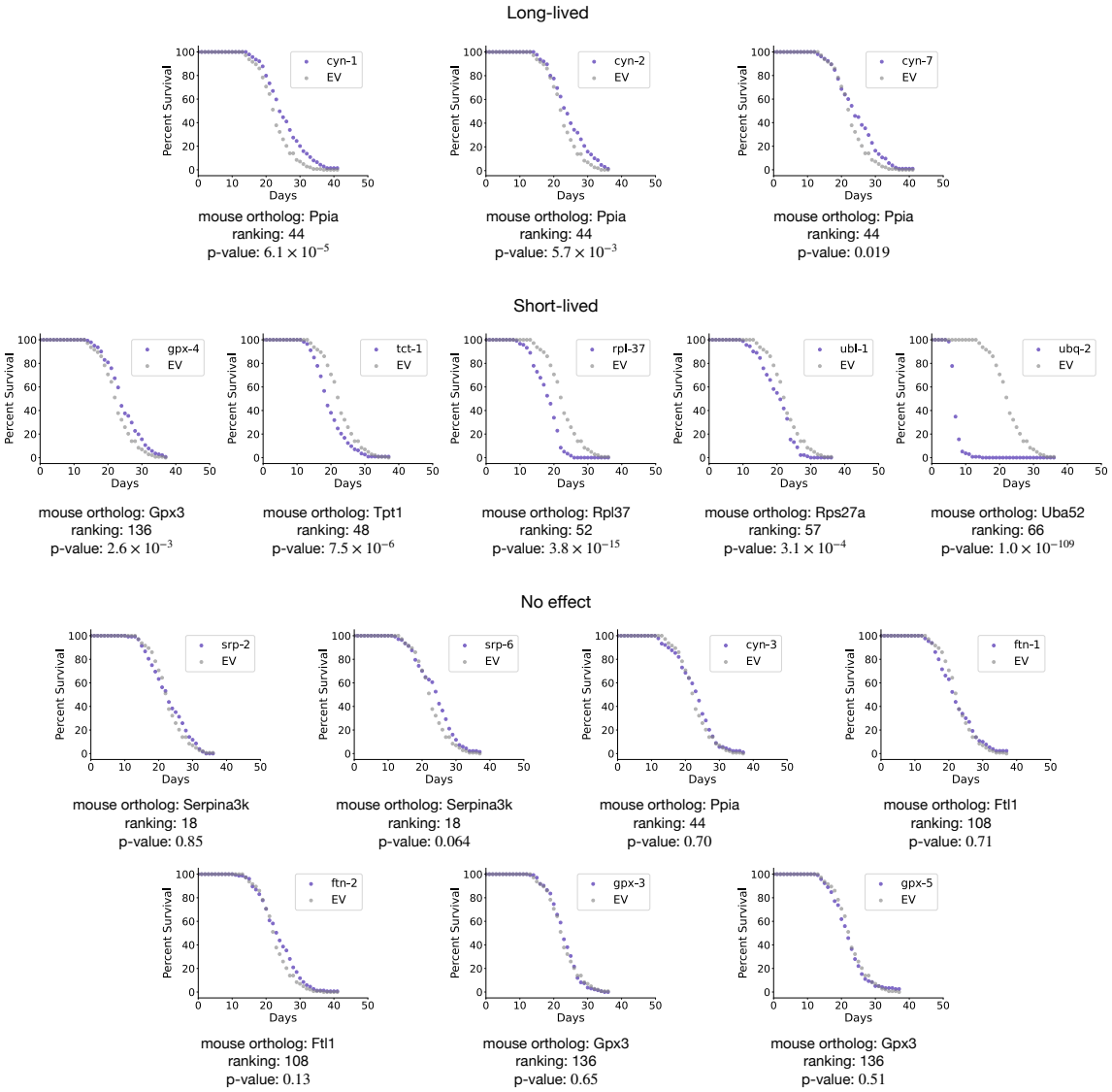


Supplementary Figure A.7.22: **Full list of KEGG and GO pathways significantly enriched by the multi-species aging model using mouse gene sets.** Significance was assessed at FDR q < 0.05 using the Benjamini-Hochberg correction.



Supplementary Figure A.7.23: **Full list of GO pathways significantly enriched by the multi-species aging model using fly gene sets.** Significance was assessed at FDR $q < 0.05$ using the Benjamini-Hochberg correction.

Supplementary Figure A.7.24: **Full list of KEGG and GO pathways significantly enriched by the multi-species aging model using human gene sets.** Significance was assessed at FDR q < 0.05 using the Benjamini-Hochberg correction.



Supplementary Figure A.7.25: **Wet lab validation of ACE-identified aging genes using C. elegans RNAi lifespan assays.** Survival curves for knockdown of *C. elegans* orthologs corresponding to selected global aging genes from the mouse global ACE model. Each plot shows the percent survival of animals treated with gene-specific RNAi (purple) compared to empty vector (EV) control (gray). Mouse gene names, their rankings from the ACE global aging model, and p-values from t-tests (adjusted using the Benjamini-Hochberg correction) are displayed below each plot. Genes are grouped into three categories based on their effects on lifespan: **Long-lived** (knockdown extends lifespan), **Short-lived** (knockdown shortens lifespan), and **No effect** (no significant impact on lifespan).

BIBLIOGRAPHY

[1] David A Bennett, Julie A Schneider, Zoe Arvanitakis, and Robert S Wilson. "Overview and findings from the religious orders study." In: *Current Alzheimer Research* 9.6 (2012), pp. 628–645.

[2] David A Bennett, Julie A Schneider, Aron S Buchman, Lisa L Barnes, Patricia A Boyle, and Robert S Wilson. "Overview and findings from the rush Memory and Aging Project." In: *Current Alzheimer Research* 9.6 (2012), pp. 646–663.

[3] "A single-cell transcriptomic atlas characterizes ageing tissues in the mouse." In: *Nature* 583.7817 (2020), pp. 590–595.

[4] Abubakar Abid and James Zou. "Contrastive variational autoencoder enhances salient features." In: *arXiv preprint arXiv:1902.04601* (2019).

[5] Fahim Ahmad, Qian Sun, Deven Patel, and Jayne M Stommel. "Cholesterol metabolism: A potential therapeutic target in glioblastoma." In: *Cancers* 11.2 (2019), p. 146.

[6] Ali Ahmed, Faiez Zannad, Thomas E Love, Jose Tallaj, Mihai Gheorghiade, Olaniyi James Ekundayo, and Bertram Pitt. "A propensity-matched study of the association of low serum potassium levels and mortality in chronic heart failure." In: *European heart journal* 28.11 (2007), pp. 1334–1343. ISSN: 1522-9645.

[7] Khondoker M Akram, Nathifa A Moyo, Gail H Leeming, Lynne Bingle, Seema Jasim, Saira Hussain, Anita Schorlemmer, Anja Kipar, Paul Digard, Ralph A Tripp, et al. "An innate defense peptide BPIFA1/SPLUNC1 restricts influenza A virus infection." In: *Mucosal immunology* 11.1 (2018), pp. 71–81.

[8] Raed A Alharbi, Ruth Pettengell, Hardev S Pandha, and Richard Morgan. "The role of HOX genes in normal hematopoiesis and acute leukemia." In: *Leukemia* 27.5 (2013), pp. 1000–1008.

[9] David B Allison, SK Zhu, Michael Plankey, Myles S Faith, and Moonseong Heo. "Differential associations of body mass index and adiposity with all-cause mortality among men in the first and second National Health and Nutrition Examination Surveys (NHANES I and NHANES II) follow-up studies." In: *International journal of obesity* 26.3 (2002), pp. 410–416.

[10] Alina Alshevskaya, Julia Zhukova, Julia Lopatnikova, Filipp Vasilyev, Ivan Khutornoy, Elena Golikova, Fedor Kireev, and Sergey Sennikov. "Nonlinear Dynamics of TNFR1 and TNFR2 Expression on Immune Cells: Genetic and Age-Related Aspects of Inflamm-Aging Mechanisms." In: *Biomedicines* 13.4 (2025), p. 852.

[11] Paal Skytt Andersen, Paula Louise Hedley, Stephen P Page, Petros Syrris, Johanna Catharina Moolman-Smook, William John McKenna, Perry Michael Elliott, and Michael Christiansen. "A novel Myosin essential light chain mutation causes hypertrophic cardiomyopathy with late onset and low expressivity." In: *Biochemistry research international* 2012.1 (2012), p. 685108.

[12] Brittany L Angarola, Siddhartha Sharma, Neerja Katiyar, Hyeon Gu Kang, Djamel Nehar-Belaid, SungHee Park, Rachel Gott, Giray N Eryilmaz, Mark A LaBarge, Karolina Palucka, et al. "Comprehensive single-cell aging atlas of healthy mammary tissues reveals shared epigenomic and transcriptomic signatures of aging and cancer." In: *Nature Aging* 5.1 (2025), pp. 122–143.

[13] M Austin Argentieri, Sihao Xiao, Derrick Bennett, Laura Winchester, Alejo J Nevado-Holgado, Upamanyu Ghose, Ashwag Albukhari, Pang Yao, Mohsen Mazidi, Jun Lv, et al. "Proteomic aging clock predicts mortality and risk of common age-related diseases in diverse populations." In: *Nature medicine* 30.9 (2024), pp. 2450–2460.

[14] David Arthur and Sergei Vassilvitskii. "k-means++ the advantages of careful seeding." In: *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. 2007, pp. 1027–1035.

[15] Tal Ashuach, Mariano I Gabitto, Rohan V Koodli, Giuseppe-Antonio Saldi, Michael I Jordan, and Nir Yosef. "MultiVI: deep generative model for the integration of multimodal data." In: *Nature Methods* 20.8 (2023), pp. 1222–1231.

[16] Žiga Avsec, Melanie Weilert, Avanti Shrikumar, Sabrina Krueger, Amr Alexandari, Khyati Dalal, Robin Fropf, Charles McAnany, Julien Gagneur, Anshul Kundaje, et al. "Base-resolution models of transcription-factor binding reveal soft motif syntax." In: *Nature genetics* 53.3 (2021), pp. 354–366.

[17] Chul-Young Bae, Yoori Im, Jonghoon Lee, Choong-Shik Park, Miyoung Kim, Hojeong Kwon, Boseon Kim, Chun-Koo Lee, Inhee Kim, JeongHoon Kim, et al. "Comparison of biological age prediction models using clinical biomarkers commonly measured in clinical practice settings: AI techniques vs. traditional statistical methods." In: *Frontiers in Analytical Science* (2021), p. 8.

[18] Matthew H Bailey, Collin Tokheim, Eduard Porta-Pardo, Sohini Sengupta, Denis Bertrand, Amila Weerasinghe, Antonio Colaprico, Michael C Wendl, Jaegil Kim, Brendan Reardon, et al. "Comprehensive characterization of cancer driver genes and mutations." In: *Cell* 173.2 (2018), pp. 371–385.

[19] William E Balch, Richard I Morimoto, Andrew Dillin, and Jeffery W Kelly. "Adapting proteostasis for disease intervention." In: *science* 319.5865 (2008), pp. 916–919.

[20] Hilary Barrett, Mary O'Keeffe, Eamon Kavanagh, Michael Walsh, and Eibhlís M O'Connor. "Is matrix Gla protein associated with vascular calcification? A systematic review." In: *Nutrients* 10.4 (2018), p. 415.

[21] Damaris Bausch-Fluck, Ulrich Goldmann, Sebastian Müller, Marc van Oostrum, Maik Müller, Olga T Schubert, and Bernd Wollscheid. "The in silico human surfaceome." In: *Proceedings of the National Academy of Sciences* 115.46 (2018), E10988–E10997.

[22] Damaris Bausch-Fluck, Andreas Hofmann, Thomas Bock, Andreas P Frei, Ferdinando Cerciello, Andrea Jacobs, Hansjoerg Moest, Ulrich Omasits, Rebekah L Gundry, Charles Yoon, et al. "A mass spectrometric-derived cell surface protein atlas." In: *PloS one* 10.4 (2015), e0121314.

[23] Nicasia Beebe-Wang, Safiye Celik, Ethan Weinberger, Pascal Sturmfels, Philip L De Jager, Sara Mostafavi, and Su-In Lee. "Unified AI framework to uncover deep interrelationships between gene expression and Alzheimer's disease neuropathologies." In: *Nature Communications* 12.1 (2021), p. 5369.

[24] Nicasia Beebe-Wang, Alex Okeson, Tim Althoff, and Su-In Lee. "Efficient and explainable risk assessments for imminent dementia in an aging cohort study." In: *IEEE Journal of Biomedical and Health Informatics* 25.7 (2021), pp. 2409–2420.

[25] Ozgur Beker, Dreyton Amador, Jose Francisco Pomarino Nima, Simon Van Deursen, Yvon Woappi, and Bianca Dumitrascu. "Patches: A Representation Learning framework for Decoding Shared and Condition-Specific Transcriptional Programs in Wound Healing." In: *bioRxiv* (2024), pp. 2024–12.

[26] Daniel W Belsky, Terrie E Moffitt, Alan A Cohen, David L Corcoran, Morgan E Levine, Joseph A Prinz, Jonathan Schaefer, Karen Sugden, Benjamin Williams, Richie Poulton, et al. "Eleven telomere, epigenetic clock, and biomarker-composite quantifications of biological aging: do they measure the same thing?" In: *American journal of epidemiology* 187.6 (2018), pp. 1220–1230.

[27] Yoshua Bengio, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives." In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), pp. 1798–1828.

[28] Yoav Benjamini and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300.

[29] Sunil Bhakta, Lisa M Crocker, Yvonne Chen, Meredith Hazen, Melissa M Schutten, Dongwei Li, Coenraad Kuijl, Rachana Ohri, Fiona Zhong, Kirsten A Poon, et al. "An anti-GDNF family receptor alpha 1 (GFRA1) antibody–drug conjugate for the treatment of hormone receptor–positive breast cancer." In: *Molecular cancer therapeutics* 17.3 (2018), pp. 638–649.

[30] Gabriela Bindea, Bernhard Mlecnik, Marie Tosolini, Amos Kirilovsky, Maximilian Waldner, Anna C Obenauf, Helen Angell, Tessa Fredriksen, Lucie Lafontaine, Anne Berger, et al. "Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer." In: *Immunity* 39.4 (2013), pp. 782–795.

[31] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. "Variational inference: A review for statisticians." In: *Journal of the American statistical Association* 112.518 (2017), pp. 859–877.

[32] Pierre Boyeau, Justin Hong, Adam Gayoso, Martin Kim, José L McFaline-Figueroa, Michael I Jordan, Elham Azizi, Can Ergen, and Nir Yosef. "Deep generative modeling of sample-level heterogeneity in single-cell genomics." In: *BioRxiv* (2022), pp. 2022–10.

[33] Sydney Brenner. "The genetics of Caenorhabditis elegans." In: *Genetics* 77.1 (1974), pp. 71–94.

[34] Paul PT Brons, Clemens Haanen, JBM Boezeman, Petra Muus, Rob SG Holdrinet, Arie HM Pennings, HM C Wessels, and Theo de Witte. "Proliferation patterns in acute myeloid leukemia: leukemic clonogenic growth and in vivo cell cycle kinetics." In: *Annals of hematology* 66.5 (1993), pp. 225–233.

[35] Matthew T Buckley, Eric D Sun, Benson M George, Ling Liu, Nicholas Schaum, Lucy Xu, Jaime M Reyes, Margaret A Goodell, Irving L Weissman, Tony Wyss-Coray, et al. "Cell-type-specific aging clocks to quantify aging and rejuvenation in neurogenic regions of the brain." In: *Nature Aging* 3.1 (2023), pp. 121–137.

[36] Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J Daly, Alkes L Price, and Benjamin M Neale. "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies." In: *Nature genetics* 47.3 (2015), pp. 291–295.

[37] Annalisa Buniello, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Sollis, et al. "The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019." In: *Nucleic acids research* 47.D1 (2019), pp. D1005–D1012.

[38] Mengjiao Cai, Xiao Sun, Wenchao Wang, Zhusheng Lian, Ping Wu, Suxia Han, Huan Chen, and Pumin Zhang. "Disruption of peroxisome function leads to metabolic stress, mTOR inhibition, and lethality in liver cancer cells." In: *Cancer letters* 421 (2018), pp. 82–93.

[39] Judith Campisi. "Aging, cellular senescence, and cancer." In: *Annual review of physiology* 75.1 (2013), pp. 685–705.

[40] Judith Campisi, Pankaj Kapahi, Gordon J Lithgow, Simon Melov, John C Newman, and Eric Verdin. "From discoveries in ageing research to therapeutics for healthy ageing." In: *Nature* 571.7764 (2019), pp. 183–192.

[41] He Cao, Panpan Yang, Jia Liu, Yan Shao, Honghao Li, Pinglin Lai, Hong Wang, Anling Liu, Bin Guo, Yujin Tang, et al. "MYL3 protects chondrocytes from senescence by inhibiting clathrin-mediated endocytosis and activating of Notch signaling." In: *Nature Communications* 14.1 (2023), p. 6190.

[42] Yue Cao, Thomas Andrew Geddes, Jean Yee Hwa Yang, and Pengyi Yang. "Ensemble deep learning in bioinformatics." In: *Nature Machine Intelligence* 2.9 (2020), pp. 500–508.

[43] Claudia Cava, Gloria Bertoli, Antonio Colaprico, Catharina Olsen, Gianluca Bontempi, and Isabella Castiglioni. "Integration of multiple networks and pathways identifies cancer driver genes in pan-cancer analysis." In: *BMC genomics* 19.1 (2018), pp. 1–16.

[44] Young Kwang Chae, Sangmin Chang, Taeyeong Ko, Jonathan Anker, Sarita Agte, Wade Iams, Wooyoung M Choi, Kyoungmin Lee, and Marcelo Cruz. "Epithelial-mesenchymal transition (EMT) signature is inversely associated with T-cell infiltration in non-small cell lung cancer (NSCLC)." In: *Scientific reports* 8.1 (2018), p. 2918.

[45] Timothy A Chan, Mark Yarchoan, Elizabeth Jaffee, Charles Swanton, Sergio A Quezada, Albrecht Stenzinger, and Solange Peters. "Development of tumor mutation burden as an immunotherapy biomarker: utility for the oncology clinic." In: *Annals of Oncology* 30.1 (2019), pp. 44–56.

[46] Josephine Y Chau, Anne C Grunseit, Tien Chey, Emmanuel Stamatakis, Wendy J Brown, Charles E Matthews, Adrian E Bauman, and Hidde P van der Ploeg. "Daily sitting time and all-cause mortality: a meta-analysis." In: *PloS one* 8.11 (2013), e80000.

[47] Fengju Chen, Yiqun Zhang, Sooryanarayana Varambally, and Chad J Creighton. "Molecular correlates of metastasis by systematic pan-cancer analysis across the cancer genome atlas." In: *Molecular Cancer Research* 17.2 (2019), pp. 476–487.

[48] Hugh Chen, Ian C Covert, Scott M Lundberg, and Su-In Lee. "Algorithms to estimate Shapley value feature attributions." In: *arXiv preprint arXiv:2207.07605* (2022).

[49] Hugh Chen, Scott M Lundberg, and Su-In Lee. "Explaining a Series of Models by Propagating Shapley Values." In: *arXiv preprint arXiv:2105.00108* (2021).

[50] Hugh Chen, Scott M Lundberg, and Su-In Lee. "Explaining a series of models by propagating Shapley values." In: *Nature communications* 13.1 (2022), p. 4512.

[51] Jonathan H Chen, Karin Pelka, Matan Hofree, Marios Giannakis, Genevieve M Boland, Andrew J Aguirre, Ana C Anderson, Orit Rozenblatt-Rosen, Aviv Regev, and Nir Hacohen. "Multicellular immune hubs and their organization in MMRd and MMRp colorectal cancer." In: *The Journal of Immunology* 206.1_Supplement (2021), pp. 68–13.

[52] Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system." In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794.

[53] Yiqun Chen and James Zou. "GenePT: a simple but effective foundation model for genes and cells built from ChatGPT." In: *bioRxiv* (2024), pp. 2023–10.

[54] Jo-Fan Chien, Hanqing Liu, Bang-An Wang, Chongyuan Luo, Anna Bartlett, Rosa Castanon, Nicholas D Johnson, Joseph R Nery, Julia Osteen, Junhao Li, et al. "Cell-type-specific effects of age and sex on human cortical neurons." In: *Neuron* 112.15 (2024), pp. 2524–2539.

[55] Kaare Christensen, Gabriele Doblhammer, Roland Rau, and James W Vaupel. "Ageing populations: the challenges ahead." In: *The lancet* 374.9696 (2009), pp. 1196–1208.

[56] Matthew J Christopher, Allegra A Petti, Michael P Rettig, Christopher A Miller, Ezhilarasi Chendamarai, Eric J Duncavage, Jeffery M Klco, Nicole M Helton, Michelle O'Laughlin, Catrina C Fronick, et al. "Immune escape of relapsed AML cells after allogeneic transplantation." In: *New England Journal of Medicine* 379.24 (2018), pp. 2330–2341.

[57] JW Clark, L Snell, RPC Shiu, FW Orr, N Maitre, CPH Vary, DJ Cole, and PH Watson. "The potential role for prolactin-inducible protein (PIP) as a marker of human breast cancer micrometastasis." In: *British Journal of Cancer* 81.6 (1999), pp. 1002–1008.

[58] James H Cole, Stuart J Ritchie, Mark E Bastin, Valdés Hernández, S Muñoz Maniega, Natalie Royle, Janie Corley, Alison Pattie, Sarah E Harris, Qian Zhang, et al. "Brain age predicts mortality." In: *Molecular psychiatry* 23.5 (2018), pp. 1385–1392.

[59] Pierre Comon. "Independent component analysis, a new concept?" In: *Signal processing* 36.3 (1994), pp. 287–314.

[60] Gene Ontology Consortium. "The Gene Ontology (GO) database and informatics resource." In: *Nucleic acids research* 32.suppl_1 (2004), pp. D258–D261.

[61] UniProt Consortium. "UniProt: a worldwide hub of protein knowledge." In: *Nucleic acids research* 47.D1 (2019), pp. D506–D515.

[62] Maria-Chiara Corti, Jack M Guralnik, Marcel E Salive, and John D Sorkin. "Serum albumin level and physical disability as predictors of mortality in older persons." In: *Jama* 272.13 (1994), pp. 1036–1042. ISSN: 0098-7484.

[63] Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L Moreira, Narges Razavian, and Aristotelis Tsirigos. "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning." In: *Nature medicine* 24.10 (2018), pp. 1559–1567.

[64] David R Cox. "Regression models and life-tables." In: *Journal of the Royal Statistical Society: Series B (Methodological)* 34.2 (1972), pp. 187–202.

[65] Eileen M Crimmins. "Lifespan and healthspan: past, present, and promise." In: *The Gerontologist* 55.6 (2015), pp. 901–911.

[66] Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. "scGPT: toward building a foundation model for single-cell multi-omics using generative AI." In: *Nature methods* 21.8 (2024), pp. 1470–1480.

[67] Lester R Curtin, Leyla K Mohadjer, Sylvia M Dohrmann, Deanna Kruszon-Moran, Lisa B Mirel, Margaret D Carroll, Rosemarie Hirsch, Vicki L Burt, and Clifford L Johnson. "National Health and Nutrition Examination Survey: sample design, 2007-2010." In: *Vital and health statistics. Series 2, Data evaluation and methods research* 160 (2013), pp. 1–23.

[68] Lester R Curtin, Leyla K Mohadjer, Sylvia M Dohrmann, Jill M Montaquila, Deanna Kruszan-Moran, Lisa B Mirel, Margaret D Carroll, Rosemarie Hirsch, Susan Schober, and Clifford L Johnson. "The National Health and Nutrition Examination Survey: Sample Design, 1999-2006." In: *Vital and health statistics. Series 2, Data evaluation and methods research* 155 (2012), pp. 1–39.

[69] Teresa Davoli, Hajime Uno, Eric C Wooten, and Stephen J Elledge. "Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy." In: *Science* 355.6322 (2017), eaaf8399.

[70] Renée De Mutsert, Dinanda J De Jager, Kitty J Jager, Carmine Zoccali, and Friedo W Dekker. "Interaction on an additive scale." In: *Nephron Clinical Practice* 119.2 (2011), pp. c154–c157.

[71] Rosario Donato, Guglielmo Sorci, and Ileana Giambanco. "S100A6 protein: functional roles." In: *Cellular and Molecular Life Sciences* 74.15 (2017), pp. 2749–2760.

[72] VE Dunlock, AB Arp, E Jansen, S Charrin, SJ van Deventer, MD Wright, L Querol-Cano, E Rubinstein, and AB van Spriel. "Dynamic regulation of CD45 by tetraspanin CD53." In: *bioRxiv* (2019), p. 854323.

[73] Ron Edgar, Michael Domrachev, and Alex E Lash. "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository." In: *Nucleic acids research* 30.1 (2002), pp. 207–210.

[74] Camilla Engblom, Christina Pfirschke, and Mikael J Pittet. "The role of myeloid cells in cancer therapies." In: *Nature Reviews Cancer* 16.7 (2016), pp. 447–462.

[75] Gökcen Eraslan, Lukas M Simon, Maria Mircea, Nikola S Mueller, and Fabian J Theis. "Single-cell RNA-seq denoising using a deep count autoencoder." In: *Nature communications* 10.1 (2019), p. 390.

[76] Gabriel Erion, Joseph D Janizek, Pascal Sturmfels, Scott M Lundberg, and Su-In Lee. "Improving performance of deep learning models with axiomatic attribution priors and expected gradients." In: *Nature machine intelligence* 3.7 (2021), pp. 620–631.

[77] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. "Dermatologist-level classification of skin cancer with deep neural networks." In: *Nature* 542.7639 (2017), pp. 115–118.

[78] Antonio Fabregat, Steven Jupe, Lisa Matthews, Konstantinos Sidiropoulos, Marc Gillespie, Phani Garapati, Robin Haw, Bijay Jassal, Florian Korninger, Bruce May, et al. "The reactome pathway knowledgebase." In: *Nucleic acids research* 46.D1 (2018), pp. D649–D655.

[79] Chao Fan, Diwei Liu, Rui Huang, Zhigang Chen, and Lei Deng. "PredRSA: a gradient boosted regression trees approach for predicting protein solvent accessibility." In: *Bmc Bioinformatics*. Vol. 17. 1. BioMed Central. 2016, pp. 85–95.

[80] Daniel I Feig, Duk-Hee Kang, and Richard J Johnson. "Uric acid and cardiovascular risk." In: *New England Journal of Medicine* 359.17 (2008), pp. 1811–1821. ISSN: 0028-4793.

[81] G Michael Felker, Larry A Allen, Stuart J Pocock, Linda K Shaw, John J V McMurray, Marc A Pfeffer, Karl Swedberg, Duolao Wang, Salim Yusuf, and Eric L Michelson. "Red cell distribution width as a novel prognostic marker in heart failure: data from the CHARM Program and the Duke Databank." In: *Journal of the American College of Cardiology* 50.1 (2007), pp. 40–47. ISSN: 0735-1097.

[82] Jason G Fleischer, Roberta Schulte, Hsiao H Tsai, Swati Tyagi, Arkaitz Ibarra, Maxim N Shokhirev, Ling Huang, Martin W Hetzer, and Saket Navlakha. "Predicting age from the transcriptome of human dermal fibroblasts." In: *Genome biology* 19.1 (2018), pp. 1–8.

[83] Kevin Flurkey, Joanne M Currer, and DE Harrison. "Mouse models in aging research." In: *The mouse in biomedical research*. Elsevier, 2007, pp. 637–672.

[84] Earl S Ford, Chaoyang Li, Stephen Cook, and Hyon K Choi. "Serum concentrations of uric acid and the metabolic syndrome among US children and adolescents." In: *Circulation* 115.19 (2007), pp. 2526–2532. ISSN: 0009-7322.

[85] Claudio Franceschi, Paolo Garagnani, Paolo Parini, Cristina Giuliani, and Aurelia Santoro. "Inflammaging: a new immune–metabolic viewpoint for age-related diseases." In: *Nature Reviews Endocrinology* 14.10 (2018), pp. 576–590.

[86] Peter D Fransquet, Jo Wrigglesworth, Robyn L Woods, Michael E Ernst, and Joanne Ryan. "The epigenetic clock as a predictor of disease and mortality risk: a systematic review and meta-analysis." In: *Clinical epigenetics* 11.1 (2019), pp. 1–17.

[87] Yu Fukuda, Yao Wang, Shangli Lian, John Lynch, Shinjiro Nagai, Bruce Fanshawe, Ayten Kandilci, Laura J Janke, Geoffrey Neale, Yiping Fan, et al. "Upregulated heme biosynthesis, an exploitable vulnerability in MYCN-driven leukemogenesis." In: *JCI insight* 2.15 (2017), e92409.

[88] Joshua J Gagne, Robert J Glynn, Jerry Avorn, Raisa Levin, and Sebastian Schneeweiss. "A combined comorbidity score predicted mortality in elderly patients better than existing scores." In: *Journal of clinical epidemiology* 64.7 (2011), pp. 749–759.

[89] Fedor Galkin, Polina Mamoshina, Kirill Kochetov, Denis Sidorenko, and Alex Zhavoronkov. "DeepMAge: a methylation aging clock developed with deep learning." In: *Aging and disease* 12.5 (2021), p. 1252.

[90] Andrea Ganna and Erik Ingelsson. "5 year mortality predictors in 498 103 UK Biobank participants: A prospective population-based study." In: *The Lancet* 386.9993 (2015), pp. 533–540. ISSN: 1474547X. DOI: 10.1016/S0140-6736(15)60175-1. URL: http://dx.doi.org/10.1016/S0140-6736(15)60175-1.

[91] Andrea Ganna and Erik Ingelsson. "5 year mortality predictors in 498 103 UK Biobank participants: a prospective population-based study." In: *The Lancet* 386.9993 (2015), pp. 533–540.

[92] Adam Gayoso, Zoë Steier, Romain Lopez, Jeffrey Regier, Kristopher L Nazor, Aaron Streets, and Nir Yosef. "Joint probabilistic modeling of single-cell multi-omic data with totalVI." In: *Nature methods* 18.3 (2021), pp. 272–282.

[93] Andrew J Gentles, Aaron M Newman, Chih Long Liu, Scott V Bratman, Weiguo Feng, Dongkyoon Kim, Viswam S Nair, Yue Xu, Amanda Khuong, Chuong D Hoang, et al. "The prognostic landscape of genes and infiltrating immune cells across human cancers." In: *Nature medicine* 21.8 (2015), pp. 938–945.

[94] Abdelaziz Ghanemi, Mayumi Yoshioka, and Jonny St-Amand. "Genetic Expression between Ageing and Exercise: Secreted Protein Acidic and Rich in Cysteine as a Potential "Exercise Substitute" Antiageing Therapy." In: *Genes* 13.6 (2022), p. 950.

[95] Alessandro Gialluisi, Augusto Di Castelnuovo, Simona Costanzo, Marialaura Bonaccio, Mariarosaria Persichillo, Sara Magnacca, Amalia De Curtis, Chiara Cerletti, Maria Benedetta Donati, Giovanni de Gaetano, et al. "Exploring domains, clinical implications and environmental associations of a deep learning marker of biological ageing." In: *European Journal of Epidemiology* 37.1 (2022), pp. 35–48.

[96] Philip Goldwasser and Joseph Feldman. "Association of serum albumin and mortality risk." In: *Journal of clinical epidemiology* 50.6 (1997), pp. 693–703. ISSN: 0895-4356.

[97] Abhinav Goyal, John A Spertus, Kensey Gosch, Lakshmi Venkitachalam, Philip G Jones, Greet Van den Berghe, and Mikhail Kosiborod. "Serum potassium levels and mortality in acute myocardial infarction." In: *Jama* 307.2 (2012), pp. 157–164. ISSN: 0098-7484.

[98] Robert C Green, J Scott Roberts, L Adrienne Cupples, Norman R Relkin, Peter J Whitehouse, Tamsen Brown, Susan LaRusse Eckert, Melissa Butson, A Dessa Sadovnick, Kimberly A Quaid, et al. "Disclosure of APOE genotype for risk of Alzheimer's disease." In: *New England Journal of Medicine* 361.3 (2009), pp. 245–254.

[99] Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. "A kernel statistical test of independence." In: *Advances in neural information processing systems* 20 (2007).

[100] Ani Grigoryan, Johannes Pospiech, Stephen Krämer, Daniel Lipka, Thomas Liehr, Hartmut Geiger, Hiroshi Kimura, Medhanie A Mulaw, and Maria Carolina Florian. "Attrition of X chromosome inactivation in aged hematopoietic stem cells." In: *Stem cell reports* 16.4 (2021), pp. 708–716.

[101] Roxane Gruel, Baukje Bijnens, Johanna Van Den Daele, Sofie Thys, Roland Willems, Dirk Wuyts, Debby Van Dam, Peter Verstraelen, Rosanne Verboven, Jana Roels, et al. "S100A8-enriched microglia populate the brain of tau-seeded and accelerated aging mice." In: *Aging Cell* 23.5 (2024), e14120.

[102] Ishaan Gulrajani, Kundan Kumar, Faruk Ahmed, Adrien Ali Taiga, Francesco Visin, David Vazquez, and Aaron Courville. "Pixelvae: A latent variable model for natural images." In: *arXiv preprint arXiv:1611.05013* (2016).

[103] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs." In: *Jama* 316.22 (2016), pp. 2402–2410.

[104] Jiaming Guo, Wei Qiu, Xiang Li, Xuandong Zhao, Ning Guo, and Quanzheng Li. "Predicting Alzheimer's disease by hierarchical graph convolution from positron emission tomography imaging." In: *2019 IEEE international conference on big data (big data)*. IEEE. 2019, pp. 5359–5363.

[105] Jing Guo, Iwata Ozaki, Jinghe Xia, Takuya Kuwashiro, Motoyasu Kojima, Hirokazu Takahashi, Kenji Ashida, Keizo Anzai, and Sachiko Matsuhashi. "PDCD4 knockdown induces senescence in hepatoma cells by up-regulating the p21 expression." In: *Frontiers in oncology* 8 (2019), p. 661.

[106] Laleh Haghverdi, Aaron TL Lun, Michael D Morgan, and John C Marioni. "Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors." In: *Nature biotechnology* 36.5 (2018), pp. 421–427.

[107] Greg Hamerly and Charles Elkan. "Learning the k in k-means." In: *Advances in neural information processing systems* 16 (2003).

[108] Gregory Hannum, Justin Guinney, Ling Zhao, LI Zhang, Guy Hughes, SriniVas Sadda, Brandy Klotzle, Marina Bibikova, Jian-Bing Fan, Yuan Gao, et al. "Genome-wide methylation profiles reveal quantitative views of human aging rates." In: *Molecular cell* 49.2 (2013), pp. 359–367.

[109] BL Heitmann, H Erikson, BM Ellsinger, KL Mikkelsen, and B Larsson. "Mortality associated with body fat, fat-free mass and body mass index among 60-year-old Swedish men—a 22-year follow-up. The study of men born in 1913." In: *International journal of obesity* 24.1 (2000), pp. 33–37.

[110] Zachary R Hettinger, Amy L Confides, Peter W Vanderklish, Silvana Sidhom, Michal M Masternak, and Esther E Dupont-Versteegden. "Skeletal muscle RBM3 expression is associated with extended lifespan in Ames Dwarf and calorie restricted mice." In: *Experimental gerontology* 146 (2021), p. 111214.

[111] Irina Higgins, Loic Matthey, Xavier Glorot, Arka Pal, Benigno Uria, Charles Blundell, Shakir Mohamed, and Alexander Lerchner. "Early visual concept learning with unsupervised deep learning." In: *arXiv preprint arXiv:1606.05579* (2016).

[112] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. "beta-vae: Learning basic visual concepts with a constrained variational framework." In: *International conference on learning representations*. 2017.

[113] Geoffrey E Hinton and Ruslan R Salakhutdinov. "Reducing the dimensionality of data with neural networks." In: *science* 313.5786 (2006), pp. 504–507.

[114] Katherine A Hoadley, Christina Yau, Denise M Wolf, Andrew D Cherniack, David Tamborero, Sam Ng, Max DM Leiserson, Beifang Niu, Michael D McLellan, Vladislav Uzunangelov, et al. "Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin." In: *Cell* 158.4 (2014), pp. 929–944.

[115] Nicholas Holzscheck, Cassandra Falckenhayn, Jörn Söhle, Boris Kristof, Ralf Siegner, André Werner, Janka Schössow, Clemens Jürgens, Henry Völzke, Horst Wenck, et al. "Modeling transcriptomic age using knowledge-primed artificial neural networks." In: *npj Aging and Mechanisms of Disease* 7.1 (2021), pp. 1–13.

[116] Benjamin D Horne, Heidi T May, Joseph B Muhlestein, Brianna S Ronnow, Donald L Lappé, Dale G Renlund, Abdallah G Kfoury, John F Carlquist, Patrick W Fisher, Robert R Pearson, et al. "Exceptional mortality prediction by risk scores from common laboratory tests." In: *The American journal of medicine* 122.6 (2009), pp. 550–558.

[117] Steve Horvath. "DNA methylation age of human tissues and cell types." In: *Genome biology* 14.10 (2013), pp. 1–20.

[118] Steve Horvath and Kenneth Raj. "DNA methylation-based biomarkers and the epigenetic clock theory of ageing." In: *Nature Reviews Genetics* 19.6 (2018), pp. 371–384.

[119] Yufei Hu, Zhilin Wang, Ting Liu, and Wei Zhang. "Piezo-like gene regulates locomotion in Drosophila larvae." In: *Cell reports* 26.6 (2019), pp. 1369–1377.

[120] Weijun Huang, LaTonya J Hickson, Alfonso Eirin, James L Kirkland, and Lilach O Lerman. "Cellular senescence: the good, the bad and the unknown." In: *Nature Reviews Nephrology* 18.10 (2022), pp. 611–627.

[121] Joseph D Janizek, Ayse B Dincer, Safiye Celik, Hugh Chen, William Chen, Kamila Naxerova, and Su-In Lee. "Uncovering expression signatures of synergistic drug responses via ensembles of explainable machine-learning models." In: *Nature biomedical engineering* 7.6 (2023), pp. 811–829.

[122] Jianqin Jiao, Kanisha Kavdia, Vishwajeeth Pagala, Lance Palmer, David Finkelstein, Yiping Fan, Junmin Peng, and Fabio Demontis. "An age-downregulated ribosomal RpS28 protein variant regulates the muscle proteome." In: *G3* 11.7 (2021), jkab165.

[123] Ana Jimenez-Pascual and Florian A. Siebzehnrubl. "Fibroblast growth factor receptor functions in glioblastoma." In: *Cells* 8.7 (2019), p. 715.

[124] Chenghao Jin, Yijie Shao, Xiaotao Zhang, Jiani Xiang, Ruize Zhang, Zeyu Sun, Shuhao Mei, Jingyi Zhou, Jianmin Zhang, and Ligen Shi. "A unique type of highly-activated microglia evoking brain inflammation via Mif/Cd74 signaling axis in aged mice." In: *Aging and disease* 12.8 (2021), p. 2125.

[125] Clifford Leroy Johnson, Sylvia M Dohrmann, Vicki L Burt, and Leyla Kheradmand Mohadjer. *National health and nutrition examination survey: sample design, 2011-2014*. 2014. US Department of Health and Human Services, Centers for Disease Control and . . ., 2014.

[126] W Evan Johnson, Cheng Li, and Ariel Rabinovic. "Adjusting batch effects in microarray expression data using empirical Bayes methods." In: *Biostatistics* 8.1 (2007), pp. 118–127.

[127] Graham Jones and Antony Barker. "Reference intervals." In: *The Clinical Biochemist Reviews* 29.Suppl 1 (2008), S93.

[128] Caroline Jose, Nadège Bellance, and Rodrigue Rossignol. "Choosing between glycolysis and oxidative phosphorylation: a tumor's dilemma?" In: *Biochimica et Biophysica Acta (BBA)-Bioenergetics* 1807.6 (2011), pp. 552–561.

[129] Brian T Joyce, Tao Gao, Yinan Zheng, Jiantao Ma, Shih-Jen Hwang, Lei Liu, Drew Nannini, Steve Horvath, Ake T Lu, Norrina Bai Allen, et al. "Epigenetic age acceleration reflects long-term cardiovascular health." In: *Circulation research* 129.8 (2021), pp. 770–781.

[130] Mechthild Jung, Robert Sabat, Jörn Krätzschmar, Henrik Seidel, Kerstin Wolk, Christiane Schönbein, Sabine Schütt, Markus Friedrich, Wolf-Dietrich Döcke, Khusru Asadullah, et al. "Expression profiling of IL-10-regulated genes in human monocytes and peripheral blood mononuclear cells from psoriatic patients during IL-10 therapy." In: *European journal of immunology* 34.2 (2004), pp. 481–493.

[131] Juulia Jylhävä, Nancy L Pedersen, and Sara Hägg. "Biological age predictors." In: *EBioMedicine* 21 (2017), pp. 29–36.

[132] Minoru Kanehisa and Susumu Goto. "KEGG: kyoto encyclopedia of genes and genomes." In: *Nucleic acids research* 28.1 (2000), pp. 27–30.

[133] Edward L Kaplan and Paul Meier. "Nonparametric estimation from incomplete observations." In: *Journal of the American statistical association* 53.282 (1958), pp. 457–481.

[134] Tanya T Karagiannis, Todd W Dowrey, Carlos Villacorta-Martin, Monty Montano, Eric Reed, Anna C Belkina, Stacy L Andersen, Thomas T Perls, Stefano Monti, George J Murphy, et al. "Multi-modal profiling of peripheral blood cells across the human lifespan reveals distinct immune cell signatures of aging and longevity." In: *EBioMedicine* 90 (2023).

[135] Alex Katayev, Claudiu Balciza, and David W Seccombe. "Establishing reference intervals for clinical laboratory test results: is there a better way?" In: *American journal of clinical pathology* 133.2 (2010), pp. 180–186.

[136] Tobias Kaufmann, Dennis van der Meer, Nhat Trung Doan, Emanuel Schwarz, Martina J Lund, Ingrid Agartz, Dag Alnæs, Deanna M Barch, Ramona Baur-Streubel, Alessandro Bertolino, et al. "Common brain disorders are associated with heritable patterns of apparent aging of the brain." In: *Nature neuroscience* 22.10 (2019), pp. 1617–1623.

[137] Brian K Kennedy, Shelley L Berger, Anne Brunet, Judith Campisi, Ana Maria Cuervo, Elissa S Epel, Claudio Franceschi, Gordon J Lithgow, Richard I Morimoto, Jeffrey E Pessin, et al. "Geroscience: linking aging to chronic disease." In: *Cell* 159.4 (2014), pp. 709–713.

[138] Nahye Kim, Hui Kwon Kim, Sungtae Lee, Jung Hwa Seo, Jae Woo Choi, Jinman Park, Seonwoo Min, Sungroh Yoon, Sung-Rae Cho, and Hyongbum Henry Kim. "Prediction of the sequence-specific cleavage activity of Cas9 variants." In: *Nature Biotechnology* 38.11 (2020), pp. 1328–1336.

[139] Sunkyu Kim, Keonwoo Kim, Junseok Choe, Inggeol Lee, and Jaewoo Kang. "Improved survival analysis by learning shared genomic information from pan-cancer data." In: *Bioinformatics* 36.Supplement_1 (2020), pp. i389–i398.

[140] Diederik P Kingma. "Adam: A method for stochastic optimization." In: *arXiv preprint arXiv:1412.6980* (2014).

[141] Diederik P Kingma, Max Welling, et al. *Auto-encoding variational bayes*. 2013.

[142] Petr Klemera and Stanislav Doubal. "A new approach to the concept and computation of biological age." In: *Mechanisms of ageing and development* 127.3 (2006), pp. 240–248.

[143] Jennifer L Kuk, Peter T Katzmarzyk, Milton Z Nichaman, Timothy S Church, Steven N Blair, and Robert Ross. "Visceral fat is an independent predictor of all-cause mortality in men." In: *Obesity* 14.2 (2006), pp. 336–341.

[144] Solomon Kullback and Richard A Leibler. "On information and sufficiency." In: *The annals of mathematical statistics* 22.1 (1951), pp. 79–86.

[145] Chia-Ling Kuo, Luke C Pilling, Zuyun Liu, Janice L Atkins, and Morgan E Levine. "Genetic associations for two biological age measures point to distinct aging phenotypes." In: *Aging cell* 20.6 (2021), e13376.

[146] Brittany N Lafaver, Li Lee, Chloe E Derocher, Lawrence F Levin, Erin M Carter, Krish Sardesai, Julian A Vallejo, Ali McAllister-Day, Tara K Crawford, Isabel M Chapman, et al. "Cardiac health, type I collagen, and aging in the oim/oim mouse model of osteogenesis imperfecta and a cohort of adults with OI." In: *American Journal of Physiology-Heart and Circulatory Physiology* 328.3 (2025), H565–H580.

[147] Edward G Lakatta and Daniel Levy. "Arterial and cardiac aging: major shareholders in cardiovascular disease enterprises: Part I: aging arteries: a "set up" for vascular disease." In: *Circulation* 107.1 (2003), pp. 139–146.

[148] Siu Sylvia Lee, Raymond YN Lee, Andrew G Fraser, Ravi S Kamath, Julie Ahringer, and Gary Ruvkun. "A systematic RNAi screen identifies a critical role for mitochondria in C. elegans longevity." In: *Nature genetics* 33.1 (2003), pp. 40–48.

[149] Benoit Lehallier, David Gate, Nicholas Schaum, Tibor Nanasi, Song Eun Lee, Hanadie Yousef, Patricia Moran Losada, Daniela Berdnik, Andreas Keller, Joe Verghese, et al. "Undulating changes in human plasma proteome profiles across the lifespan." In: *Nature medicine* 25.12 (2019), pp. 1843–1850.

[150] Seppo Lehto, Leo Niskanen, Tapani Ronnemaa, and Markku Laakso. "Serum uric acid is a strong predictor of stroke in patients with non–insulin-dependent diabetes mellitus." In: *Stroke* 29.3 (1998), pp. 635–639. ISSN: 0039-2499.

[151] Kenneth Levenberg. "A method for the solution of certain non-linear problems in least squares." In: *Quarterly of applied mathematics* 2.2 (1944), pp. 164–168.

[152] Morgan E Levine. "Modeling the rate of senescence: can estimated biological age predict mortality more accurately than chronological age?" In: *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences* 68.6 (2013), pp. 667–674.

[153] Morgan E Levine, Ake T Lu, Austin Quach, Brian H Chen, Themistocles L Assimes, Stefania Bandinelli, Lifang Hou, Andrea A Baccarelli, James D Stewart, Yun Li, et al. "An epigenetic biomarker of aging for lifespan and healthspan." In: *Aging (Albany NY)* 10.4 (2018), p. 573.

[154] Ji Li, Peter S Choi, Christine L Chaffer, Katherine Labella, Justin H Hwang, Andrew O Giacomelli, Jong Wook Kim, Nina Ilic, John G Doench, Seav Huong Ly, et al. "An alternative splicing switch in FLNB promotes the mesenchymal cell state in human breast cancer." In: *Elife* 7 (2018), e37184.

[155] Mengyuan Li, Qingrong Sun, and Xiaosheng Wang. "Transcriptional landscape of human cancers." In: *Oncotarget* 8.21 (2017), p. 34534.

[156] Xia Li, Alexander Ploner, Yunzhang Wang, Patrik KE Magnusson, Chandra Reynolds, Deborah Finkel, Nancy L Pedersen, Juulia Jylhävä, and Sara Hägg. "Longitudinal trajectories, correlations and mortality associations of nine biological ages across 20-years follow-up." In: *Elife* 9 (2020), e51507.

[157] Yuanyuan Li, Kai Kang, Juno M Krahn, Nicole Croutwater, Kevin Lee, David M Umbach, and Leping Li. "A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data." In: *BMC genomics* 18 (2017), pp. 1–13.

[158] Pin-Chao Liao, Hung-Yu Lin, Chiou-Hwa Yuh, Lin-Kwei Yu, and Horng-Dar Wang. "The effect of neuronal expression of heat shock proteins 26 and 27 on lifespan, neurodegeneration, and apoptosis in Drosophila." In: *Biochemical and biophysical research communications* 376.4 (2008), pp. 637–641.

[159] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P Mesirov. "Molecular signatures database (MSigDB) 3.0." In: *Bioinformatics* 27.12 (2011), pp. 1739–1740.

[160] Emilly M Lima, Antônio H Ribeiro, Gabriela MM Paixão, Manoel Horta Ribeiro, Marcelo M Pinto-Filho, Paulo R Gomes, Derick M Oliveira, Ester C Sabino, Bruce B Duncan, Luana Giatti, et al. "Deep neural network-estimated electrocardiographic age as a mortality predictor." In: *Nature communications* 12.1 (2021), p. 5117.

[161] Yan Lin, Ying Li, Guangyu Liang, Xiao Yang, Jiankun Yang, Qi Hu, Jian Sun, Cuntai Zhang, Haoshu Fang, and Anding Liu. "Single-cell transcriptome analysis of aging mouse liver." In: *The FASEB Journal* 38.4 (2024), e23473.

[162] Zuyun Liu, Pei-Lun Kuo, Steve Horvath, Eileen Crimmins, Luigi Ferrucci, and Morgan Levine. "A new aging measure captures morbidity and mortality risk across diverse subpopulations from NHANES IV: a cohort study." In: *PLoS medicine* 15.12 (2018), e1002718.

[163] Po-Ru Loh, George Tucker, Brendan K Bulik-Sullivan, Bjarni J Vilhjalmsson, Hilary K Finucane, Rany M Salem, Daniel I Chasman, Paul M Ridker, Benjamin M Neale, Bonnie Berger, et al. "Efficient Bayesian mixed-model analysis increases association power in large cohorts." In: *Nature genetics* 47.3 (2015), pp. 284–290.

[164] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. "The genotype-tissue expression (GTEx) project." In: *Nature genetics* 45.6 (2013), pp. 580–585.

[165] Álvaro López-Janeiro, Carlos Padilla-Ansala, Carlos E de Andrea, David Hardisson, and Ignacio Melero. "Prognostic value of macrophage polarization markers in epithelial neoplasms and melanoma. A systematic review and meta-analysis." In: *Modern Pathology* 33.8 (2020), pp. 1458–1465.

[166] Carlos López-Otín, Maria A Blasco, Linda Partridge, Manuel Serrano, and Guido Kroemer. "The hallmarks of aging." In: *Cell* 153.6 (2013), pp. 1194–1217.

[167] Carlos López-Otín, Maria A Blasco, Linda Partridge, Manuel Serrano, and Guido Kroemer. "Hallmarks of aging: An expanding universe." In: *Cell* 186.2 (2023), pp. 243–278.

[168] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. "Deep generative modeling for single-cell transcriptomics." In: *Nature methods* 15.12 (2018), pp. 1053–1058.

[169] Romain Lopez, Jeffrey Regier, Michael I Jordan, and Nir Yosef. "Information constraints on auto-encoding variational bayes." In: *Advances in neural information processing systems* 31 (2018).

[170] Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. "scGen predicts single-cell perturbation responses." In: *Nature methods* 16.8 (2019), pp. 715–721.

[171] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. "Causal effect inference with deep latent-variable models." In: *Advances in neural information processing systems* 30 (2017).

[172] Ake T Lu, Austin Quach, James G Wilson, Alex P Reiner, Abraham Aviv, Kenneth Raj, Lifang Hou, Andrea A Baccarelli, Yun Li, James D Stewart, et al. "DNA methylation GrimAge strongly predicts lifespan and healthspan." In: *Aging (Albany NY)* 11.2 (2019), p. 303.

[173] Tzu-Chiao Lu, Maria Brbić, Ye-Jin Park, Tyler Jackson, Jiaye Chen, Sai Saroja Kolluru, Yanyan Qi, Nadja Sandra Katheder, Xiaoyu Tracy Cai, Seungjae Lee, et al. "Aging Fly Cell Atlas identifies exhaustive aging features at cellular resolution." In: *Science* 380.6650 (2023), eadg0934.

[174] Amanda L Lumsden, Anwar Mulugeta, Ang Zhou, and Elina Hyppönen. "Apolipoprotein E (APOE) genotype-associated disease risks: a phenome-wide, registry-based, case-control study utilising the UK Biobank." In: *EBioMedicine* 59 (2020), p. 102954.

[175] Arian Lundberg, Linda S Lindström, Joel S Parker, Elinor Löverli, Charles M Perou, Jonas Bergh, and Nicholas P Tobin. "A pan-cancer analysis of the frequency of DNA alterations across cell cycle activity levels." In: *Oncogene* 39.32 (2020), pp. 5430–5440.

[176] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. "Explainable AI for Trees: From Local Explanations to Global Understanding." In: (2019), pp. 1–72. arXiv: 1905.04610. URL: http://arxiv.org/abs/1905.04610.

[177] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. "From local explanations to global understanding with explainable AI for trees." In: *Nature machine intelligence* 2.1 (2020), pp. 2522–5839.

[178] Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions." In: *Advances in neural information processing systems*. 2017, pp. 4765–4774.

[179] Mark Lustberg and Ellen Silbergeld. "Blood lead levels and mortality." In: *Archives of internal medicine* 162.21 (2002), pp. 2443–2449. ISSN: 0003-9926.

[180] Monika Majchrzak-Gorecka, Pawel Majewski, Beata Grygier, Krzysztof Murzyn, and Joanna Cichy. "Secretory leukocyte protease inhibitor (SLPI), a multifunctional protein in the host defense response." In: *Cytokine & growth factor reviews* 28 (2016), pp. 79–93.

[181] Marilyne Malbouyres, Alexandre Guiraud, Christel Lefrancc ois, Mélanie Salamito, Pauline Nauroy, Laure Bernard, Frédéric Sohm, Bruno Allard, and Florence Ruggiero. "Lack of the myotendinous junction marker col22a1 results in posture and locomotion disabilities in zebrafish." In: *Matrix Biology* 109 (2022), pp. 1–18.

[182] Polina Mamoshina, Kirill Kochetov, Evgeny Putin, Franco Cortese, Alexander Aliper, Won-Suk Lee, Sung-Min Ahn, Lee Uhn, Neil Skjodt, Olga Kovalchuk, et al. "Population specific biomarkers of human aging: a big data study using South Korean, Canadian, and Eastern European patient populations." In: *The Journals of Gerontology: Series A* 73.11 (2018), pp. 1482–1490.

[183] Shulin Mao, Jiayu Su, Longteng Wang, Xiaochen Bo, Cheng Li, and Hebing Chen. "A transcriptome-based single-cell biological age model and resource for tissue-specific aging measures." In: *Genome Research* 33.8 (2023), pp. 1381–1394.

[184] Nathaniel S Marshall, Keith KH Wong, Peter Y Liu, Stewart RJ Cullen, Matthew W Knuiman, and Ronald R Grunstein. "Sleep apnea as an independent risk factor for all-cause mortality: the Busselton Health Study." In: *Sleep* 31.8 (2008), pp. 1079–1085.

[185] Fernando O Martinez, Siamon Gordon, Massimo Locati, and Alberto Mantovani. "Transcriptional profiling of the human monocyte-to-macrophage differentiation and polarization: new molecules and patterns of gene expression." In: *The Journal of immunology* 177.10 (2006), pp. 7303–7311.

[186] Alexandra Maslova, Ricardo N Ramirez, Ke Ma, Hugo Schmutz, Chendi Wang, Curtis Fox, Bernard Ng, Christophe Benoist, Sara Mostafavi, and Immunological Genome Project. "Deep learning of immune cell differentiation." In: *Proceedings of the National Academy of Sciences* 117.41 (2020), pp. 25655–25666.

[187] Mitsuyoshi Matsumoto, Kentaro Uchida, Ryo Tazawa, Tomonori Kenmoku, Kosuke Inoue, Gen Inoue, and Masashi Takaso. "Impact of Aging and Estrogen Deficiency on Extracellular Matrix-Related Gene Expression in Rotator Cuff Tendons: in vitro and in vivo rat model." In: *JSES International* (2025).

[188] Susanne May and Carol Bigelow. "Modeling nonlinear dose-response relationships in epidemiologic studies: statistical approaches and practical challenges." In: *Dose-Response* 3.4 (2005), dose–response. ISSN: 1559-3258.

[189] Daniel L McCartney, Josine L Min, Rebecca C Richmond, Ake T Lu, Maria K Sobczyk, Gail Davies, Linda Broer, Xiuqing Guo, Ayoung Jeong, Jeesun Jung, et al. "Genome-wide association studies identify 137 genetic loci for DNA methylation biomarkers of aging." In: *Genome biology* 22.1 (2021), pp. 1–25.

[190] Cathal McCrory, Giovanni Fiorito, Belinda Hernandez, Silvia Polidoro, Aisling M O'Halloran, Ann Hever, Cliona Ni Cheallaigh, Ake T Lu, Steve Horvath, Paolo Vineis, et al. "GrimAge outperforms other epigenetic clocks in the prediction of age-related clinical phenotypes and all-cause mortality." In: *The Journals of Gerontology: Series A* 76.5 (2021), pp. 741–749.

[191] Domhnall McHugh and Jesús Gil. "Senescence and aging: Causes, consequences, and therapeutic avenues." In: *Journal of Cell Biology* 217.1 (2018), pp. 65–77.

[192] Andy Menke, Paul Muntner, Vecihi Batuman, Ellen K Silbergeld, and Eliseo Guallar. "Blood lead below 0.48 mmol/L (10 mg/dL) and mortality among US adults." In: *Circulation* 114.13 (2006), pp. 1388–1394.

[193] Daniele Merico, Ruth Isserlin, Oliver Stueker, Andrew Emili, and Gary D Bader. "Enrichment map: a network-based method for gene-set enrichment visualization and interpretation." In: *PloS one* 5.11 (2010), e13984.

[194] David H Meyer and Björn Schumacher. "BiT age: A transcriptome-based aging clock near the theoretical limit of accuracy." In: *Aging cell* 20.3 (2021), e13320.

[195] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2019. ISBN: 0244768528.

[196] Clara Morral, Jelena Stanisavljevic, Xavier Hernando-Momblona, Elisabetta Mereu, Adrián Álvarez-Varela, Carme Cortina, Diana Stork, Felipe Slebe, Gemma Turon, Gavin Whissell, et al. "Zonation of ribosomal DNA transcription defines a stem cell hierarchy in colorectal cancer." In: *Cell stem cell* 26.6 (2020), pp. 845–861.

[197] Yuki Nakamura, Hisataka Iwata, Takehito Kuwayama, and Koumei Shirasuna. "S100A8, which increases with age, induces cellular senescence-like changes in bovine oviduct epithelial cells." In: *American Journal of Reproductive Immunology* 82.3 (2019), e13163.

[198] Georges N Nakhoul, Haiquan Huang, Susana Arrigain, Stacey E Jolly, Jesse D Schold, Joseph V Nally Jr, and Sankar D Navaneethan. "Serum potassium, end-stage renal disease and mortality in chronic kidney disease." In: *American journal of nephrology* 41.6 (2015), pp. 456–463. ISSN: 0250-8095.

[199] Paul G Nelson, Daniel EL Promislow, and Joanna Masel. "Biomarkers for aging identified in cross-sectional studies tend to be non-causative." In: *The Journals of Gerontology: Series A* 75.3 (2020), pp. 466–472.

[200] Teresa Niccoli and Linda Partridge. "Ageing as a risk factor for disease." In: *Current biology* 22.17 (2012), R741–R752.

[201] Emma Nichols, Cassandra EI Szoeke, Stein Emil Vollset, Nooshin Abbasi, Foad Abd-Allah, Jemal Abdela, Miloud Taki Eddine Aichour, Rufus O Akinyemi, Fares Alahdab, Solomon W Asgedom, et al. "Global, regional, and national burden of Alzheimer's disease and other dementias, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016." In: *The Lancet Neurology* 18.1 (2019), pp. 88–106.

[202] Mengting Niu, Quan Zou, and Chen Lin. "CRBPDL: Identification of circRNA-RBP interaction sites using an ensemble neural network approach." In: *PLoS computational biology* 18.1 (2022), e1009798.

[203] Rezvan Noroozi, Soudeh Ghafouri-Fard, Aleksandra Pisarek, Joanna Rudnicka, Magdalena Spólnicka, Wojciech Branicki, Mohammad Taheri, and Ewelina Pośpiech. "DNA methylation-based age clocks: from age prediction to age reversion." In: *Ageing Research Reviews* (2021), p. 101314.

[204] Gherman Novakovsky, Nick Dexter, Maxwell W Libbrecht, Wyeth W Wasserman, and Sara Mostafavi. "Obtaining genetics insights from deep learning via explainable artificial intelligence." In: *Nature Reviews Genetics* 24.2 (2023), pp. 125–137.

[205] Valéria Sutti Nunes, Guilherme da Silva Ferreira, and Eder Carlos Rocha Quintão. "Cholesterol metabolism in aging simultaneously altered in liver and nervous system." In: *Aging (Albany NY)* 14.3 (2022), p. 1549.

[206] Martin Oft. "IL-10: master switch from tumor-promoting inflammation to antitumor immunity." In: *Cancer immunology research* 2.3 (2014), pp. 194–199.

[207] Hamilton Se-Hwee Oh, Yann Le Guen, Nimrod Rappoport, Deniz Yagmur Urey, Amelia Farinas, Jarod Rutledge, Divya Channappa, Anthony D Wagner, Elizabeth Mormino, Anne Brunet, et al. "Plasma proteomics links brain and immune system aging with healthspan and longevity." In: *Nature Medicine* (2025), pp. 1–9.

[208] S Jay Olshansky. "From lifespan to healthspan." In: *Jama* 320.13 (2018), pp. 1323–1324.

[209] Kanyin Liane Ong, Lauryn K Stafford, Susan A McLaughlin, Edward J Boyko, Stein Emil Vollset, Amanda E Smith, Bronte E Dalton, Joe Duprey, Jessica A Cruz, Hailey Hagins, et al. "Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021." In: *The Lancet* 402.10397 (2023), pp. 203–234.

[210] Yesim Ozarda, Victoria Higgins, and Khosrow Adeli. "Verification of reference intervals in routine clinical laboratories: practical challenges and recommendations." In: *Clinical Chemistry and Laboratory Medicine (CCLM)* 57.1 (2018), pp. 30–37.

[211] Joel S Parker, Michael Mullins, Maggie CU Cheang, Samuel Leung, David Voduc, Tammi Vickery, Sherri Davies, Christiane Fauron, Xiaping He, Zhiyuan Hu, et al. "Supervised risk predictor of breast cancer based on intrinsic subtypes." In: *Journal of clinical oncology* 27.8 (2009), pp. 1160–1167.

[212] Linda Partridge, Matias Fuentealba, and Brian K Kennedy. "The quest to slow ageing through drug discovery." In: *Nature Reviews Drug Discovery* 19.8 (2020), pp. 513–532.

[213] Kushang V Patel, Luigi Ferrucci, William B Ershler, Dan L Longo, and Jack M Guralnik. "Red blood cell distribution width and the risk of death in middle-aged and older adults." In: *Archives of internal medicine* 169.5 (2009), pp. 515–523. ISSN: 0003-9926.

[214] Kushang V Patel, Richard D Semba, Luigi Ferrucci, Anne B Newman, Linda P Fried, Robert B Wallace, Stefania Bandinelli, Caroline S Phillips, Binbing Yu, and Stephanie Connelly. "Red cell distribution width and mortality in older adults: a meta-analysis." In: *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences* 65.3 (2010), pp. 258–265. ISSN: 1758-535X.

[215] Nancye M Peel, Roderick J McClure, and Helen P Bartlett. "Behavioral determinants of healthy aging." In: *American journal of preventive medicine* 28.3 (2005), pp. 298–304.

[216] Itsik Pe'er, Roman Yelensky, David Altshuler, and Mark J Daly. "Estimation of the multiple testing burden for genomewide association studies of nearly all common variants." In: *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* 32.4 (2008), pp. 381–385.

[217] Todd S Perlstein, Jennifer Weuve, Marc A Pfeffer, and Joshua A Beckman. "Red blood cell distribution width and mortality risk in a community-based prospective cohort." In: *Archives of internal medicine* 169.6 (2009), pp. 588–594. ISSN: 0003-9926.

[218] Marjolein J Peters, Roby Joehanes, Luke C Pilling, Claudia Schurmann, Karen N Conneely, Joseph Powell, Eva Reinmaa, George L Sutphin, Alexandra Zhernakova, Katharina Schramm, et al. "The transcriptional landscape of age in human peripheral blood." In: *Nature communications* 6.1 (2015), pp. 1–14.

[219] Martin Petkovich and Glenville Jones. "CYP24A1 and kidney disease." In: *Current opinion in nephrology and hypertension* 20.4 (2011), pp. 337–344.

[220] Niklas Pfister, Peter Bühlmann, Bernhard Schölkopf, and Jonas Peters. "Kernel-based tests for joint independence." In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 80.1 (2018), pp. 5–31.

[221] Andrew Phillips, A Gerald Shaper, and PeterH Whincup. "Association between serum albumin and mortality from cardiovascular disease, cancer, and other causes." In: *The Lancet* 334.8677 (1989), pp. 1434–1436. ISSN: 0140-6736.

[222] Zoe Piran, Niv Cohen, Yedid Hoshen, and Mor Nitzan. "Disentanglement of single-cell data with biolord." In: *Nature Biotechnology* 42.11 (2024), pp. 1678–1683.

[223] Zoe Piran and Mor Nitzan. "SiFT: uncovering hidden biological processes by probabilistic filtering of single-cell data." In: *Nature Communications* 15.1 (2024), p. 760.

[224] Jason N Pitt, Nolan L Strait, Elena M Vayndorf, Benjamin W Blue, Christina H Tran, Brendon EM Davis, Karen Huang, Brock J Johnson, Keong Mu Lim, Sophie Liu, et al. "WormBot, an open-source robotics platform for survival and behavior analysis in C. elegans." In: *GeroScience* 41.6 (2019), pp. 961–973.

[225] Brian Popko, Dennis K Pearl, Diane M Walker, Theodore C Comas, Kristine D Baerwald, Peter C Burger, Bernd W Scheithauer, and Allan J Yates. "Molecular markers that identify human astrocytomas and oligodendrogliomas." In: *Journal of Neuropathology & Experimental Neurology* 61.4 (2002), pp. 329–338.

[226] Vinay Prasad, Victoria Kaestner, and Sham Mailankody. "Cancer drugs approved based on biomarkers and not tumor type—FDA approval of pembrolizumab for mismatch repair-deficient solid cancers." In: *JAMA oncology* 4.2 (2018), pp. 157–158.

[227] Evgeny Putin, Polina Mamoshina, Alexander Aliper, Mikhail Korzinkin, Alexey Moskalev, Alexey Kolosov, Alexander Ostrovskiy, Charles Cantor, Jan Vijg, and Alex Zhavoronkov. "Deep biomarkers of human aging: application of deep neural networks to biomarker development." In: *Aging (Albany NY)* 8.5 (2016), p. 1021.

[228] Bin-Zhi Qian and Jeffrey W Pollard. "Macrophage diversity enhances tumor progression and metastasis." In: *Cell* 141.1 (2010), pp. 39–51.

[229] Chengxiang Qiu, Beth K Martin, Ian C Welsh, Riza M Daza, Truc-Mai Le, Xingfan Huang, Eva K Nichols, Megan L Taylor, Olivia Fulton, Diana R O'Day, et al. "A single-cell time-lapse of mouse prenatal development from gastrula to birth." In: *Nature* 626.8001 (2024), pp. 1084–1093.

[230] Wei Qiu, Chris Arian, Ethan Weinberger, Soo R Yun, Alexander R Mendenhall, Jessica E Young, Maria Brbic*, and Su-In Lee*. "An explainable AI framework for identifying universal aging signatures in cell embeddings." In: (Under review).

[231] Wei Qiu, Hugh Chen, Ayse Berceste Dincer, Scott Lundberg, Matt Kaeberlein, and Su-In Lee. "Interpretable machine learning prediction of all-cause mortality." In: *Communications medicine* 2.1 (2022), p. 125.

[232] Wei Qiu, Hugh Chen, Matt Kaeberlein, and Su-In Lee. "ExplaiNAble BioLogical Age (ENABL Age): an artificial intelligence framework for interpretable biological age." In: *The lancet Healthy longevity* 4.12 (2023), e711–e723.

[233] Wei Qiu, Ayse B Dincer, Joseph D Janizek, Safiye Celik, Mikael J Pittet, Kamila Naxerova*, and Su-In Lee*. "Deep profiling of gene expression across 18 human cancers." In: *Nature biomedical engineering* 9.3 (2025), pp. 333–355.

[234] Wei Qiu, Jiaming Guo, Xiang Li, Mengjia Xu, Mo Zhang, Ning Guo, and Quanzheng Li. "Multi-label detection and classification of red blood cells in microscopic images." In: *2020 IEEE International Conference on Big Data (Big Data)*. IEEE. 2020, pp. 4257–4263.

[235] Daniela F Quail and Johanna A Joyce. "Microenvironmental regulation of tumor progression and metastasis." In: *Nature medicine* 19.11 (2013), pp. 1423–1437.

[236] Stephen R Quake and Tabula Sapiens Consortium. "Tabula Sapiens reveals transcription factor expression, senescence effects, and sex-specific features in cell types from 28 human organs and tissues." In: *bioRxiv* (2024), pp. 2024–12.

[237] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. "Scalable and accurate deep learning with electronic health records." In: *NPJ digital medicine* 1.1 (2018), p. 18.

[238] Nicole A Rapicavoli, Kun Qu, Jiajing Zhang, Megan Mikhail, Remi-Martin Laberge, and Howard Y Chang. "A mammalian pseudogene lncRNA at the interface of inflammation and anti-inflammatory therapeutics." In: *elife* 2 (2013), e00762.

[239] Michel Raymond and FranccoisRousset. "An exact test for population differentiation." In: *Evolution* (1995), pp. 1280–1283.

[240] Xudie Ren, Haonan Guo, Shenghong Li, Shilin Wang, and Jianhua Li. "A novel image classification method with CNN-XGBoost model." In: *International Workshop on Digital Watermarking*. Springer. 2017, pp. 378–390.

[241] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why should i trust you?" Explaining the predictions of any classifier." In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.

[242] Spencer R Rosario, Mark D Long, Hayley C Affronti, Aryn M Rowsam, Kevin H Eng, and Dominic J Smiraglia. "Pan-cancer analysis of transcriptional metabolic dysregulation using The Cancer Genome Atlas." In: *Nature communications* 9.1 (2018), p. 5330.

[243] Andrew D Rouillard, Gregory W Gundersen, Nicolas F Fernandez, Zichen Wang, Caroline D Monteiro, Michael G McDermott, and Avi Ma'ayan. "The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins." In: *Database* 2016 (2016), baw100.

[244] Antoine Emile Roux, Han Yuan, Katie Podshivalova, David Hendrickson, Rex Kerr, Cynthia Kenyon, and David Kelley. "Individual cell types in C. elegans age differently and activate distinct cell-protective responses." In: *Cell Reports* 42.8 (2023).

[245] Ling Ruan, Bharati Mendhe, Emily Parker, Andrew Kent, Carlos M Isales, William D Hill, Meghan McGee-Lawrence, Sadanand Fulzele, and Mark W Hamrick. "Long non-coding RNA MALAT1 is depleted with age in skeletal muscle in vivo and MALAT1 silencing increases expression of TGF-β1 in vitro." In: *Frontiers in Physiology* 12 (2022), p. 742004.

[246] Mariangela Russo, Giovanni Crisafulli, Alberto Sogari, Nicole M Reilly, Sabrina Arena, Simona Lamba, Alice Bartolini, Vito Amodio, Alessandro Magrì, Luca Novara, et al. "Adaptive mutability of colorectal cancers in response to targeted therapies." In: *Science* 366.6472 (2019), pp. 1473–1480.

[247] Jarod Rutledge, Hamilton Oh, and Tony Wyss-Coray. "Measuring biological age using omics data." In: *Nature Reviews Genetics* (2022), pp. 1–13.

[248] Seungjin Ryu, Sviatoslav Sidorov, Eric Ravussin, Maxim Artyomov, Akiko Iwasaki, Andrew Wang, and Vishwa Deep Dixit. "The matricellular protein SPARC induces inflammatory interferon-response in macrophages during aging." In: *Immunity* 55.9 (2022), pp. 1609–1626.

[249] Marina Salvadores and Fran Supek. "Cell cycle gene alterations associate with a redistribution of mutation risk across chromosomal domains in human cancers." In: *Nature Cancer* 5.2 (2024), pp. 330–346.

[250] Nazish Sayed, Yingxiang Huang, Khiem Nguyen, Zuzana Krejciova-Rajaniemi, Anissa P Grawe, Tianxiang Gao, Robert Tibshirani, Trevor Hastie, Ayelet Alpert, Lu Cui, et al. "An inflammatory aging clock (iAge) based on deep learning tracks multimorbidity, immunosenescence, frailty and cardiovascular aging." In: *Nature aging* 1.7 (2021), pp. 598–615.

[251] Laura A Schaap, Tara Quirke, Hanneke AH Wijnhoven, and Marjolein Visser. "Changes in body mass index and mid-upper arm circumference in relation to all-cause mortality in older adults." In: *Clinical Nutrition* 37.6 (2018), pp. 2252–2259.

[252] Karl P Schlingmann, Martin Kaufmann, Stefanie Weber, Andrew Irwin, Caroline Goos, Ulrike John, Joachim Misselwitz, Günter Klaus, Eberhard Kuwertz-Bröking, Henry Fehrenbach, et al. "Mutations in CYP24A1 and idiopathic infantile hypercalcemia." In: *New England Journal of Medicine* 365.5 (2011), pp. 410–421.

[253] Sebastian Schneeweiss, Jeremy A Rassen, Robert J Glynn, Jerry Avorn, Helen Mogun, and M Alan Brookhart. "High-dimensional propensity score adjustment in studies of treatment effects using health care claims data." In: *Epidemiology (Cambridge, Mass.)* 20.4 (2009), p. 512.

[254] Susan E Schober, Lisa B Mirel, Barry I Graubard, Debra J Brody, and Katherine M Flegal. "Blood lead levels and death from all causes, cardiovascular disease, and cancer: results from the NHANES III mortality study." In: *Environmental health perspectives* 114.10 (2006), pp. 1538–1541. ISSN: 0091-6765.

[255] Mara A Schonberg, Roger B Davis, Ellen P McCarthy, and Edward R Marcantonio. "Index to predict 5-year mortality of community-dwelling adults aged 65 and older using data from the National Health Interview Survey." In: *Journal of general internal medicine* 24.10 (2009), p. 1115.

[256] Katharina Schultebraucks, Arieh Y Shalev, Vasiliki Michopoulos, Corita R Grudzen, Soo-Min Shin, Jennifer S Stevens, Jessica L Maples-Keller, Tanja Jovanovic, George A Bonanno, Barbara O Rothbaum, et al. "A validated predictive algorithm of posttraumatic stress course following emergency department admission after a traumatic stressor." In: *Nature medicine* 26.7 (2020), pp. 1084–1088.

[257] Raghav Sehgal, Yaroslav Markov, Chenxi Qin, Margarita Meer, Courtney Hadley, Aladdin H Shadyab, Ramon Casanova, JoAnn E Manson, Parveen Bhatti, Ann Z Moore, et al. "Systems Age: A single blood methylation test to quantify aging heterogeneity across 11 physiological systems." In: *Nature Aging* (2025), pp. 1–17.

[258] Ki-Hyeon Seong, Takumi Ogashiwa, Takashi Matsuo, Yoshiaki Fuyama, and Toshiro Aigaki. "Application of the gene search system to screen for longevity genes in Drosophila." In: *Biogerontology* 2.3 (2001), pp. 209–217.

[259] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. "Cytoscape: a software environment for integrated models of biomolecular interaction networks." In: *Genome research* 13.11 (2003), pp. 2498–2504.

[260] Yan Shi, Zuishuang Guo, Fengxun Liu, Shaokang Pan, Dan Gao, Sijie Zhou, Zhenjie Liu, Feng Wang, Dongwei Liu, and Zhangsuo Liu. "Analysis of potential biomarkers for diabetic kidney disease based on single-cell RNA-sequencing integrated with a single-cell sequencing assay for transposase-accessible chromatin." In: *Aging (Albany NY)* 15.19 (2023), p. 10681.

[261] Yui Shinozaki, Nobutomo Ikarashi, Keito Tabata, Ayuka Miyazawa, Risako Kon, Hiroyasu Sakai, and Tomoo Hosoe. "Expression analysis of genes important for maintaining skin function in a senescence-accelerated mouse prone model." In: *Geriatrics & Gerontology International* 23.12 (2023), pp. 951–957.

[262] Param Priya Singh, Brittany A Demmitt, Ravi D Nath, and Anne Brunet. "The genetics of aging: a vertebrate perspective." In: *Cell* 177.1 (2019), pp. 200–220.

[263] Anna Ślusarz, LaNita A Nichols, Elizabeth A Grunz-Borgmann, Gang Chen, Adebayo D Akintola, Jeffery M Catania, Robert C Burghardt, Jerome P Trzeciakowski, and Alan R Parrish. "Overexpression of MMP-7 increases collagen 1A2 in the aging kidney." In: *Physiological reports* 1.5 (2013).

[264] Lucas K Smith, Yingbo He, Jeong-Soo Park, Gregor Bieri, Cedric E Snethlage, Karin Lin, Geraldine Gontier, Rafael Wabl, Kristopher E Plambeck, Joe Udeochu, et al. "β2-microglobulin is a systemic pro-aging factor that impairs cognitive function and neurogenesis." In: *Nature medicine* 21.8 (2015), pp. 932–937.

[265] Ugo Sorrentino, Ilaria Gabbiato, Chiara Canciani, Davide Calosci, Chiara Rigon, Daniela Zuccarello, and Matteo Cassina. "Homozygous TNNI3 mutations and severe early onset dilated cardiomyopathy: patient report and review of the literature." In: *Genes* 14.3 (2023), p. 748.

[266] Cyril Statzer, Ji Young Cecilia Park, and Collin Y Ewald. "Extracellular matrix dynamics as an emerging yet understudied hallmark of aging and longevity." In: *Aging and Disease* 14.3 (2023), p. 670.

[267] Daniel J Stekhoven and Peter Bühlmann. "MissForest—non-parametric missing value imputation for mixed-type data." In: *Bioinformatics* 28.1 (2012), pp. 112–118.

[268] Blanka Stiburkova, Jitka Stekrova, Makiko Nakamura, and Kimiyoshi Ichida. "Hereditary renal Hypouricemia type 1 and autosomal dominant polycystic kidney disease." In: *The American journal of the medical sciences* 350.4 (2015), pp. 268–271. ISSN: 0002-9629.

[269] Alexander M Strasak, Kilian Rapp, Wolfgang Hilbe, Willi Oberaigner, Elfriede Ruttmann, Hans Concin, Günter Diem, Karl P Pfeiffer, Hanno Ulmer, and VHM&PP Study Group. "Serum uric acid and risk of cancer mortality in a large prospective male cohort." In: *Cancer causes & control* 18.9 (2007), pp. 1021–1029. ISSN: 0957-5243.

[270] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." In: *Proceedings of the National Academy of Sciences* 102.43 (2005), pp. 15545–15550.

[271] Mohamed E Suliman, Richard J Johnson, Elvia García-López, A Rashid Qureshi, Hadi Molinaei, Juan Jesús Carrero, Olof Heimbürger, Peter Bárány, Jonas Axelsson, and Bengt Lindholm. "J-shaped mortality relationship for uric acid in CKD." In: *American Journal of Kidney Diseases* 48.5 (2006), pp. 761–771. ISSN: 0272-6386.

[272] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks." In: *International conference on machine learning*. PMLR. 2017, pp. 3319–3328.

[273] Hsiao H Sung, Wyatt J Spresser, Joseph P Hoffmann, Zongrui Dai, Peter M Van der Kraan, Michelle S Caird, Esmeralda Blaney Davidson, and Kenneth M Kozloff. "Collagen mutation and age contribute to differential craniofacial phenotypes in mouse models of osteogenesis imperfecta." In: *JBMR plus* 8.1 (2024), ziad004.

[274] Zoltán Szabó and Bharath K Sriperumbudur. "Characteristic and Universal Tensor Product Kernels." In: *J. Mach. Learn. Res.* 18.233 (2017), pp. 1–29.

[275] Zoltán Szabó and Bharath K Sriperumbudur. "Characteristic and universal tensor product kernels." In: *Journal of Machine Learning Research* 18.233 (2018), pp. 1–29.

[276] Damian Szklarczyk, Annika L Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T Doncheva, John H Morris, Peer Bork, et al. "STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets." In: *Nucleic acids research* 47.D1 (2019), pp. D607–D613.

[277] Rayna Takaki, Susan R Watson, and Lewis L Lanier. "DAP12: an adapter protein with dual functionality." In: *Immunological reviews* 214.1 (2006), pp. 118–129.

[278] Jie Tan, Georgia Doing, Kimberley A Lewis, Courtney E Price, Kathleen M Chen, Kyle C Cady, Barret Perchuk, Michael T Laub, Deborah A Hogan, and Casey S Greene. "Unsupervised extraction of stable expression signatures from public compendia with an ensemble of neural networks." In: *Cell systems* 5.1 (2017), pp. 63–71.

[279] Toshiko Tanaka, Angelique Biancotto, Ruin Moaddel, Ann Zenobia Moore, Marta Gonzalez-Freire, Miguel A Aon, Julián Candia, Pingbo Zhang, Foo Cheung, Giovanna Fantoni, et al. "Plasma proteomic signature of age in healthy humans." In: *Aging cell* 17.5 (2018), e12799.

[280] Andrei E Tarkhov, Thomas Lindstrom-Vautrin, Sirui Zhang, Kejun Ying, Mahdi Moqri, Bohan Zhang, Alexander Tyshkovskiy, Orr Levy, and Vadim N Gladyshev. "Nature of epigenetic aging from a single-cell perspective." In: *Nature Aging* 4.6 (2024), pp. 854–870.

[281] Catherine Tcheandjieu, Xiang Zhu, Austin T Hilliard, Shoa L Clarke, Valerio Napolioni, Shining Ma, Kyung Min Lee, Huaying Fang, Fei Chen, Yingchang Lu, et al. "Large-scale genome-wide association study of coronary artery disease in genetically diverse populations." In: *Nature medicine* 28.8 (2022), pp. 1679–1692.

[282] Rumiana Tenchov, Janet M Sasso, Xinmei Wang, and Qiongqiong Angela Zhou. "Aging hallmarks and progression and age-related diseases: a landscape view of research advancement." In: *ACS Chemical Neuroscience* 15.1 (2023), pp. 1–30.

[283] Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, X Shirley Liu, et al. "Transfer learning enables predictions in network biology." In: *Nature* 618.7965 (2023), pp. 616–624.

[284] Vésteinn Thorsson, David L Gibbs, Scott D Brown, Denise Wolf, Dante S Bortone, Tai-Hsien Ou Yang, Eduard Porta-Pardo, Galen F Gao, Christopher L Plaisier, James A Eddy, et al. "The immune landscape of cancer." In: *Immunity* 48.4 (2018), pp. 812–830.

[285] Ye Ella Tian, Vanessa Cropley, Andrea B Maier, Nicola T Lautenschlager, Michael Breakspear, and Andrew Zalesky. "Heterogeneous aging across multiple organ systems and prediction of chronic disease and mortality." In: *Nature medicine* 29.5 (2023), pp. 1221–1231.

[286] Paul RHJ Timmers, James F Wilson, Peter K Joshi, and Joris Deelen. "Multivariate genomic scan implicates novel loci and haem metabolism in human ageing." In: *Nature communications* 11.1 (2020), p. 3570.

[287] Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. "What clinicians want: contextualizing explainable machine learning for clinical end use." In: *Machine learning for healthcare conference*. PMLR. 2019, pp. 359–380.

[288] L Torlay, Marcela Perrone-Bertolotti, Elizabeth Thomas, and Monica Baciu. "Machine learning–XGBoost analysis of language networks to classify patients with epilepsy." In: *Brain informatics* 4.3 (2017), pp. 159–169.

[289] John Tower. "Heat shock proteins and Drosophila aging." In: *Experimental gerontology* 46.5 (2011), pp. 355–362.

[290] Alexandre Trapp, Csaba Kerepesi, and Vadim N Gladyshev. "Profiling epigenetic age in single cells." In: *Nature Aging* 1.12 (2021), pp. 1189–1201.

[291] Alan C Tsai and Tsui-Lan Chang. "The effectiveness of BMI, calf circumference and mid-arm circumference in predicting subsequent mortality risk in elderly Taiwanese." In: *British Journal of Nutrition* 105.2 (2011), pp. 275–281. ISSN: 1475-2662.

[292] Alan C Tsai and Tsui-Lan Chang. "The effectiveness of BMI, calf circumference and mid-arm circumference in predicting subsequent mortality risk in elderly Taiwanese." In: *British Journal of Nutrition* 105.2 (2011), pp. 275–281.

[293] Michele Tumminello, Fabrizio Lillo, and Rosario N Mantegna. "Correlation, hierarchies, and networks in financial markets." In: *Journal of economic behavior & organization* 75.1 (2010), pp. 40–58.

[294] Bauyrzhan Umbayev, Yuliya Safarova, Aislu Yermekova, Assem Nessipbekova, Aizhan Syzdykova, and Sholpan Askarova. "Role of a small GTPase Cdc42 in aging and age-related diseases." In: *Biogerontology* 24.1 (2023), pp. 27–46.

[295] Peter JM Valk, Roel GW Verhaak, M Antoinette Beijen, Claudia AJ Erpelinck, Sahar Barjesteh van Waalwijk van Doorn-Khosrovani, Judith M Boer, H Berna Beverloo, Michael J Moorhouse, Peter J Van Der Spek, Bob Löwenberg, et al. "Prognostically useful gene-expression profiles in acute myeloid leukemia." In: *New England Journal of Medicine* 350.16 (2004), pp. 1617–1628.

[296] Laurens Van Der Maaten, Eric O Postma, H Jaap Van Den Herik, et al. "Dimensionality reduction: A comparative review." In: *Journal of machine learning research* 10.66-71 (2009), p. 13.

[297] Roel GW Verhaak, Bas J Wouters, Claudia AJ Erpelinck, Saman Abbas, H Berna Beverloo, Sanne Lugthart, Bob Löwenberg, Ruud Delwel, and Peter JM Valk. "Prediction of molecular subtypes in acute myeloid leukemia based on gene expression profiling." In: *haematologica* 94.1 (2008), p. 131.

[298] Genaro R Villa, Jonathan J Hulce, Ciro Zanca, Junfeng Bi, Shiro Ikegami, Gabrielle L Cahill, Yuchao Gu, Kenneth M Lum, Kenta Masui, Huijun Yang, et al. "An LXR-cholesterol axis creates a metabolic co-dependency for brain cancers." In: *Cancer cell* 30.5 (2016), pp. 683–693.

[299] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. "Extracting and composing robust features with denoising autoencoders." In: *Proceedings of the 25th international conference on Machine learning*. 2008, pp. 1096–1103.

[300] Anna Wade, Aaron E Robinson, Jane R Engler, Claudia Petritsch, C David James, and Joanna J Phillips. "Proteoglycans and their roles in brain cancer." In: *The FEBS journal* 280.10 (2013), pp. 2399–2417.

[301] Minako Wakasugi, Junichiro James Kazama, Ichiei Narita, Tsuneo Konta, Shouichi Fujimoto, Kunitoshi Iseki, Toshiki Moriyama, Kunihiro Yamagata, Kazuhiko Tsuruya, and Koichi Asahi. "Association between hypouricemia and reduced kidney function: a cross-sectional population-based study in Japan." In: *American journal of nephrology* 41.2 (2015), pp. 138–146. ISSN: 0250-8095.

[302] Stefan Walter, Johan Mackenbach, Zoltán Vokó, Stefan Lhachimi, M Arfan Ikram, André G Uitterlinden, Anne B Newman, Joanne M Murabito, Melissa E Garcia, Vilmundur Gudnason, et al. "Genetic, physiological, and lifestyle predictors of mortality in the general population." In: *American journal of public health* 102.4 (2012), e3–e10.

[303] Thomas R Walters, Frederick H Welland, T John Gribble, and Herbert C Schwartz. "Biosynthesis of heme in leukemic leukocytes." In: *Cancer* 20.7 (1967), pp. 1117–1123.

[304] Quan Wan, Hayley Dingerdissen, Yu Fan, Naila Gulzar, Yang Pan, Tsung-Jung Wu, Cheng Yan, Haichen Zhang, and Raja Mazumder. "BioXpress: an integrated RNA-seq-derived gene expression database for pan-cancer analysis." In: *Database* 2015 (2015), bav019.

[305] Jiaxuan Wang, Jenna Wiens, and Scott Lundberg. "Shapley flow: A graph-based approach to interpreting model predictions." In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 721–729.

[306] Kang Wang, Huicong Liu, Qinchao Hu, Lingna Wang, Jiaqing Liu, Zikai Zheng, Weiqi Zhang, Jie Ren, Fangfang Zhu, and Guang-Hui Liu. "Epigenetic regulation of aging: implications for interventions of aging and diseases." In: *Signal transduction and targeted therapy* 7.1 (2022), p. 374.

[307] Gregory P Way and Casey S Greene. "Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders." In: *PACIFIC SYMPOSIUM on BIOCOMPUTING 2018: Proceedings of the Pacific Symposium*. World Scientific. 2018, pp. 80–91.

[308] Gregory P Way, Francisco Sanchez-Vega, Konnor La, Joshua Armenia, Walid K Chatila, Augustin Luna, Chris Sander, Andrew D Cherniack, Marco Mina, Giovanni Ciriello, et al. "Machine learning detects pan-cancer ras pathway activation in the cancer genome atlas." In: *Cell reports* 23.1 (2018), pp. 172–180.

[309] Gregory P Way, Michael Zietz, Vincent Rubinetti, Daniel S Himmelstein, and Casey S Greene. "Compressing gene expression data using multiple latent space dimensionalities learns complementary biological representations." In: *Genome biology* 21 (2020), pp. 1–27.

[310] Ethan Weinberger, Nicasia Beebe-Wang, and Su-In Lee. "Moment Matching Deep Contrastive Latent Variable Models." In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2022, pp. 2354–2371.

[311] Ethan Weinberger, Chris Lin, and Su-In Lee. "Isolating salient variations of interest in single-cell data with contrastiveVI." In: *Nature Methods* 20.9 (2023), pp. 1336–1345.

[312] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. "The cancer genome atlas pan-cancer analysis project." In: *Nature genetics* 45.10 (2013), pp. 1113–1120.

[313] Bo Wen, Wen-Feng Zeng, Yuxing Liao, Zhiao Shi, Sara R Savage, Wen Jiang, and Bing Zhang. "Deep learning in proteomics." In: *Proteomics* 20.21-22 (2020), p. 1900335.

[314] Stephen F Weng, Jenna Reps, Joe Kai, Jonathan M Garibaldi, and Nadeem Qureshi. "Can machine-learning improve cardiovascular risk prediction using routine clinical data?" In: *PloS one* 12.4 (2017), e0174944.

[315] Stephen F Weng, Luis Vaz, Nadeem Qureshi, and Joe Kai. "Prediction of premature all-cause mortality: A prospective general population cohort study comparing machine-learning and standard epidemiological approaches." In: *PloS one* 14.3 (2019), e0214365.

[316] Mary C White, Dawn M Holman, Jennifer E Boehm, Lucy A Peipins, Melissa Grossman, and S Jane Henley. "Age and cancer risk: a potentially modifiable relationship." In: *American journal of preventive medicine* 46.3 (2014), S7–S15.

[317] Hanneke AH Wijnhoven, Marian AE van Bokhorst-de van der Schueren, Martijn W Heymans, Henrica CW de Vet, Hinke M Kruizenga, Jos W Twisk, and Marjolein Visser. "Low mid-upper arm circumference, calf circumference, and body mass index and mortality in older persons." In: *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences* 65.10 (2010), pp. 1107–1114.

[318] Gareth H Williams and Kai Stoeber. "The cell cycle and cancer." In: *The Journal of pathology* 226.2 (2012), pp. 352–364.

[319] Svante Wold, Kim Esbensen, and Paul Geladi. "Principal component analysis." In: *Chemometrics and intelligent laboratory systems* 2.1-3 (1987), pp. 37–52.

[320] Thomas R Wood, Christopher Kelly, Megan Roberts, and Bryan Walsh. "An interpretable machine learning model of biological age." In: *F1000Research* 8.17 (2019), p. 17.

[321] Li-Wei Wu, Yuan-Yung Lin, Tung-Wei Kao, Chien-Ming Lin, Chung-Ching Wang, Gia-Chi Wang, Tao-Chun Peng, and Wei-Liang Chen. "Mid-arm circumference and all-cause, cardiovascular, and cancer mortality among obese and non-obese US adults: the national health and nutrition examination survey III." In: *Scientific reports* 7.1 (2017), pp. 1–8.

[322] Daniela Xhindoli, Sabrina Pacor, Monica Benincasa, Marco Scocchi, Renato Gennaro, and Alessandro Tossi. "The human cathelicidin LL-37—A pore-forming antibacterial peptide and host-cell modulator." In: *Biochimica et Biophysica Acta (BBA)-Biomembranes* 1858.3 (2016), pp. 546–566.

[323] Xian Xia, Yiyang Wang, Zhengqing Yu, Jiawei Chen, and Jing-Dong J Han. "Assessing the rate of aging to monitor aging itself." In: *Ageing Research Reviews* 69 (2021), p. 101350.

[324] Yanping Xiao and Gordon J Freeman. "The microsatellite instable subset of colorectal cancer is a particularly good candidate for checkpoint blockade immunotherapy." In: *Cancer discovery* 5.1 (2015), pp. 16–18.

[325] Chenling Xu, Romain Lopez, Edouard Mehlman, Jeffrey Regier, Michael I Jordan, and Nir Yosef. "Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models." In: *Molecular systems biology* 17.1 (2021), e9620.

[326] Qinghua Xu, Jinying Chen, Shujuan Ni, Cong Tan, Midie Xu, Lei Dong, Lin Yuan, Qifeng Wang, and Xiang Du. "Pan-cancer transcriptome analysis reveals a gene expression signature for the identification of tumor tissue origin." In: *Modern Pathology* 29.6 (2016), pp. 546–556.

[327] Ying Xu and Tao Wang. "LOVIT is a putative vesicular histamine transporter required in Drosophila for vision." In: *Cell reports* 27.5 (2019), pp. 1327–1333.

[328] Hui Yang, Tereza Clarence, Madeline R Scott, NM Prashant, Xinyi Wang, Milos Pjanic, Sanan Venkatesh, Aram Hong, Clara Casey, Zhiping Shao, et al. "A single-cell transcriptomic atlas of the prefrontal cortex across the human lifespan." In: *medRxiv* (2024), pp. 2024–11.

[329] Jian Yang, Teresa Ferreira, Andrew P Morris, Sarah E Medland, Pamela AF Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, Michael N Weedon, Ruth J Loos, et al. "Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits." In: *Nature genetics* 44.4 (2012), pp. 369–375.

[330] Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. "GCTA: a tool for genome-wide complex trait analysis." In: *The American Journal of Human Genetics* 88.1 (2011), pp. 76–82.

[331] Ali Yazdanyar and Anne B Newman. "The burden of cardiovascular disease in the elderly: morbidity, mortality, and costs." In: *Clinics in geriatric medicine* 25.4 (2009), p. 563.

[332] SO Yesylevskyy, VN Kharkyanen, and AP Demchenko. "Hierarchical clustering of the correlation patterns: new method of domain identification in proteins." In: *Biophysical chemistry* 119.1 (2006), pp. 84–93.

[333] Seong-Keun Yoo, Young Shin Song, Eun Kyung Lee, Jinha Hwang, Hwan Hee Kim, Gyeongseo Jung, Young A Kim, Su-jin Kim, Sun Wook Cho, Jae-Kyung Won, et al. "Integrative analysis of genomic and transcriptomic characteristics associated with progression of aggressive thyroid cancer." In: *Nature communications* 10.1 (2019), p. 2764.

[334] Baohu Zhang, Haoteng Yan, Xiaoqian Liu, Liang Sun, Shuai Ma, Si Wang, Jing Qu, Guang-Hui Liu, and Weiqi Zhang. "SenoIndex: S100A8/S100A9 as a novel aging biomarker." In: *Life Medicine* 2.4 (2023), lnad022.

[335] Dahai Zhang, Liyang Qian, Baijin Mao, Can Huang, Bin Huang, and Yulin Si. "A data-driven design for fault detection of wind turbines using random forests and XGboost." In: *IEEE Access* 6 (2018), pp. 21020–21031.

[336] Martin Jinye Zhang, Angela Oliveira Pisco, Spyros Darmanis, and James Zou. "Mouse aging cell atlas analysis reveals global and cell type-specific aging signatures." In: *Elife* 10 (2021), e62293.

[337] Qian Zhang, Costanza L Vallerga, Rosie M Walker, Tian Lin, Anjali K Henders, Grant W Montgomery, Ji He, Dongsheng Fan, Javed Fowdar, Martin Kennedy, et al. "Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing." In: *Genome medicine* 11.1 (2019), pp. 1–11.

[338] Alex Zhavoronkov, Ricky Li, Candice Ma, and Polina Mamoshina. "Deep biomarkers of aging and longevity: from research to applications." In: *Aging (Albany NY)* 11.22 (2019), p. 10771.

[339] Jian Zhou and Olga G Troyanskaya. "Predicting effects of noncoding variants with deep learning–based sequence model." In: *Nature methods* 12.10 (2015), pp. 931–934.

[340] Shankuan Zhu, Moonseong Heo, Michael Plankey, Myles S Faith, and David B Allison. "Associations of body mass index and anthropometric indicators of fat mass and fat free mass with all-cause mortality among women in the first and second National Health and Nutrition Examination Surveys follow-up studies." In: *Annals of epidemiology* 13.4 (2003), pp. 286–293.

[341] Polina Zjablovskaja, Miroslava Kardosova, Petr Danek, Pavla Angelisova, Touati Benoukraf, Alexander A Wurm, Tomas Kalina, Stephanie Sian, Martin Balastik, Ruud Delwel, et al. "EVI2B is a C/EBPα target gene required for granulocytic differentiation and functionality of hematopoietic progenitors." In: *Cell Death & Differentiation* 24.4 (2017), pp. 705–716.

[342] Peter-James H Zushin, Souhrid Mukherjee, Joseph C Wu, et al. "FDA Modernization Act 2.0: transitioning beyond animal models with human cells, organoids, and AI/ML-based approaches." In: *The Journal of clinical investigation* 133.21 (2023).