nature computational science

Article

https://doi.org/10.1038/s43588-025-00939-x

A robust computational framework for methylation age and disease-risk prediction based on pairwise learning

Received: 20 February 2025

Accepted: 19 November 2025

Published online: 13 January 2026

Check for updates

Yu Zhang [1,2,7], Yichen Yao [1,7], Yuanhao Tang [1], Yuan Cheng [2,3], Yinghui Xu [1,3] ✉, Ying He [1,2] ✉, Yuan Qi [2,3,4] ✉ & Li Jin [5,6] ✉

Conventional epigenetic clocks encounter challenges in generalizability, especially when there are pronounced batch effects between the training and test datasets, restricting their clinical applicability for aging assessment. Here we present MAPLE, a robust computational framework for methylation age and disease-risk prediction through pairwise learning. MAPLE utilizes pairwise learning to discern the relative relationships between two DNA methylation profiles regarding age or disease risk. It effectively identifies aging- or disease-related biological signals while mitigating technical biases in the data. MAPLE outperforms five competing methods, achieving a median absolute error of 1.6 years across 31 benchmark tests from diverse studies, sequencing platforms, data preprocessing methods and tissue types. Furthermore, MAPLE performs well when assessing aging-related disease risk, with mean areas under the curve of 0.97 for disease identification and 0.85 for pre-disease status detection. Overall, we show that MAPLE has great potential for assessing epigenetic age and aging-related disease risk clinically.

Aging is characterized by increasing morbidity and a concomitant decline in quality of life, both of which impose substantial social and economic burdens[1]. Breakthrough studies in recent decades have indicated that it is feasible to extend both lifespan and healthspan through interventions such as caloric restriction and partial epigenetic reprogramming[2,3]. However, the clinical application of aging interventions requires precise measurements of biological age and aging rates.

DNA methylation (DNAm) is the covalent attachment of a methyl group to the fifth carbon of a cytosine residue and typically occurs at cytosine–guanine dinucleotides (CpGs) in mammals. DNAm alteration is a hallmark of aging[4]. For example, DNAm tends to increase with age at certain CpG islands, particularly at Polycomb target genes and promoters of tumor suppressor genes[5]. These cumulative changes in DNAm probably establish an age-related transcriptional program[6]. Consequently, genome-wide DNAm can serve as an estimator of biological age[6]. However, the performance of current epigenetics clocks might be influenced by multiple factors, including sequencing platform, data preprocessing method, tissue type and population, which pose challenges to their clinical application[7,8]. These clocks, such as the Horvath clock[9] and HannumAge[6], predominantly rely on linear models such as LASSO, which restrict their predictive capabilities for epigenetic age. Another line of research, such as AltumAge[10], uses deep learning models but does not adequately address batch effects between different datasets during modeling. Consequently, these clocks might have incorporated technical biases unrelated to

[1]INF Technology (Shanghai) Co. Ltd, Shanghai, China. [2]Shanghai Academy of Artificial Intelligence for Science, Shanghai, China. [3]Artificial Intelligence Innovation and Incubation Institute, Fudan University, Shanghai, China. [4]Zhongshan Hospital, Fudan University, Shanghai, China. [5]State Key Laboratory of Genetic Engineering, and Collaborative Innovation Center for Genetics and Development, School of Life Sciences, Fudan University, Shanghai, China. [6]Human Phenome Institute, Fudan University, Shanghai, China. [7]These authors contributed equally: Yu Zhang, Yichen Yao. ✉e-mail: xuyinghui@fudan.edu.cn; heying340@gmail.com; qiyuan@fudan.edu.cn; lijin@fudan.edu.cn

aging into their models, hindering their ability to generalize effectively to unseen data in complex clinical settings. In addition, in DNAm datasets used for training, the number of features often greatly exceeds the number of samples, which increases the risk of overfitting and impairs the model's generalizability. This phenomenon is commonly referred to as the high-dimensional, low-sample-size problem in machine learning. Therefore, there is a critical need for a generalizable DNAm clock that can eliminate technical biases while preserving aging-related biological signals, enabling accurate measurement of human aging processes.

Although epigenetic clocks offer a valuable measure of epigenetic age, reliance solely on epigenetic age proves insufficient for recommending targeted interventions to combat aging. Type 2 diabetes (T2D) and cardiovascular diseases (CVD) are prevalent globally, and aging is identified as the predominant risk factor for both diseases[11,12]. Early diagnosis and interventions can delay the onset of T2D and CVD, enhance the quality of life for at-risk populations and potentially reduce associated mortality[13,14]. Therefore, utilizing DNAm profiles for risk assessment of aging-related diseases, such as T2D and CVD, is crucial for personalized aging interventions and health management. Similar to epigenetic clocks, existing risk assessment models for T2D and CVD, often based on linear models such as the Cox proportional hazards regression model (Cox model)[15,16], encounter pronounced generalizability challenges when applied across different studies and sequencing platforms.

Deep pairwise learning, commonly referred to as Siamese networks, represents a class of neural network architectures that comprise two subnetworks sharing identical parameters and weights. This approach has been widely applied in various domains, including face recognition, voice recognition and visual tracking, among others[17]. Pairwise learning demonstrates promising potential in addressing the aforementioned challenges such as the high-dimensional, low-sample-size problem and the risk of overfitting to training datasets[18]. In this Article, we present MAPLE—a robust computational framework for methylation age and disease-risk prediction based on pairwise learning. MAPLE predicts the differences in epigenetic age or disease risk between any two samples by comparing their DNAm profiles, thereby encoding DNAm data from diverse sources into a unified latent space. This approach enables the effective distinction of samples with different aging or disease statuses. Consequently, MAPLE can consistently outperform traditional epigenetic clocks across varied studies, sequencing platforms, data preprocessing methods and tissues. Furthermore, MAPLE is capable of providing precise risk assessments for aging-related diseases and accurately identifying pre-disease states. Its robust and accurate performance in epigenetic age and disease risk evaluation paves the way for the clinical application of DNAm sequencing in aging assessment and intervention.

Result

Overview of MAPLE

A critical challenge in the clinical application of epigenetic clocks is the robust and accurate prediction of epigenetic age for unseen DNAm datasets, regardless of variations in sequencing platforms, data preprocessing methods, populations or tissue types (Fig. 1a). MAPLE addresses this challenge by mitigating inherent batch effects among datasets through pairwise learning. Although the distribution of DNAm profiles may differ across datasets owing to varying data sources or data preprocessing methods, the ages of the samples from different datasets remain comparable. By predicting the age difference between any two DNAm profiles, MAPLE encodes the DNAm data into a unified latent space that preserves aging-related information while eliminating batch effects resulting from data heterogeneity. In addition, the pairwise construction of samples leads to a quadratic increase in the number of training samples, enabling MAPLE to be thoroughly trained and reducing the risk of overfitting.

MAPLE is capable of predicting epigenetic age through a two-step training procedure conducted entirely on the training data. First, MAPLE uses pairwise learning to train a Siamese network as the encoder, which integrates DNAm profiles from diverse sources into a unified latent space. Specifically, two samples are randomly selected from the training data, and their DNAm profiles are input into the same encoder, generating embeddings within the latent space (Fig. 1b). These embeddings are subsequently used to predict the difference in chronological age between the two samples (Fig. 1d). Once the encoder has been adequately trained, an age predictor is constructed to estimate the age of a sample on the basis of its corresponding embedding (Fig. 1b). During inference, the DNAm profiles of the test samples, which are not used during training, are passed through the trained encoder and age predictor to generate its predicted epigenetic age (Fig. 1c).

In the training of disease risk prediction models, traditional linear models are utilized to generate disease risk scores, which serve as the learning objective for pairwise learning. Because these scores are only comparable within the same dataset (Supplementary Fig. 1), we propose to learn the risk score difference between two samples within the same dataset using pairwise learning (Fig. 1e). Analogous to the epigenetic age prediction, an encoder is trained on the training data to extract disease-related epigenetic signals. Subsequently, a disease predictor is constructed to classify the disease type on the basis of the generated embedding, thereby integrating supervised disease label information into the model (Fig. 1b). During inference, test samples not seen during training are processed through the trained encoder to generate their embeddings. Disease risk scores are then calculated based on the distances in the embedding space (Fig. 1c; for details, see the Methods).

Systematic benchmarking to study MAPLE's performance in epigenetic age prediction

A clinically valuable epigenetic clock must accurately and reliably predict epigenetic age for unseen DNAm data across diverse platforms, preprocessing pipelines and tissues. To evaluate MAPLE's generalizability, we design two comparison scenarios encompassing 31 benchmark tests under various biological and technical conditions.

In the first comparison scenario, we assess performance on blood DNAm datasets containing technical biases. Six independent test datasets derived from Illumina Infinium Methylation450 (450K) and Illumina Infinium MethylationEPIC (EPIC) arrays are preprocessed using three common methods (Subset-quantile Within Array Normalization (SWAN)[19], Beta MIxture Quantile dilation (BMIQ)[20] and Gaussian Mixture Quantile Normalization (GMQN)[21]), yielding 24 benchmark tests. In the second comparison scenario, we focus on the epigenetic age estimation in nonblood samples, which is crucial for aging biology research of different tissues. Performance was quantified by the median absolute error (MAE) and Pearson correlation coefficient (PCC) between predicted and chronological age.

To investigate the contribution of pairwise learning, we compare MAPLE with a multilayer perceptron (MLP) baseline sharing the same architecture but trained to predict absolute rather than relative age. In the first scenario, MAPLE achieves a mean MAE of 1.45 and PCC of 0.97, while the MLP baseline yields a mean MAE of 4.14 and PCC of 0.88 (Fig. 2a and Supplementary Fig. 2a). MAPLE also outperforms the MLP baseline in nonblood datasets (Fig. 2b and Supplementary Fig. 2b). These results underscore the importance of pairwise learning in improving epigenetic age prediction.

We further compare MAPLE with five state-of-the-art epigenetic clocks, including HorvathAge[9], HannumAge[6], PhenoAge[22], AltumAge[10] and cAge[23]. In the first comparison scenario, across the six raw datasets, MAPLE achieves the best performance with an average MAE of 1.40 and PCC of 0.97, whereas HorvathAge (MAE 4.90, PCC 0.87) and PhenoAge (MAE 5.63, PCC 0.81) perform worst because they cannot adapt to EPIC array data (Fig. 2a and Supplementary Fig. 2a). MAPLE remains

https://doi.org/10.1038/s43588-025-00939-x

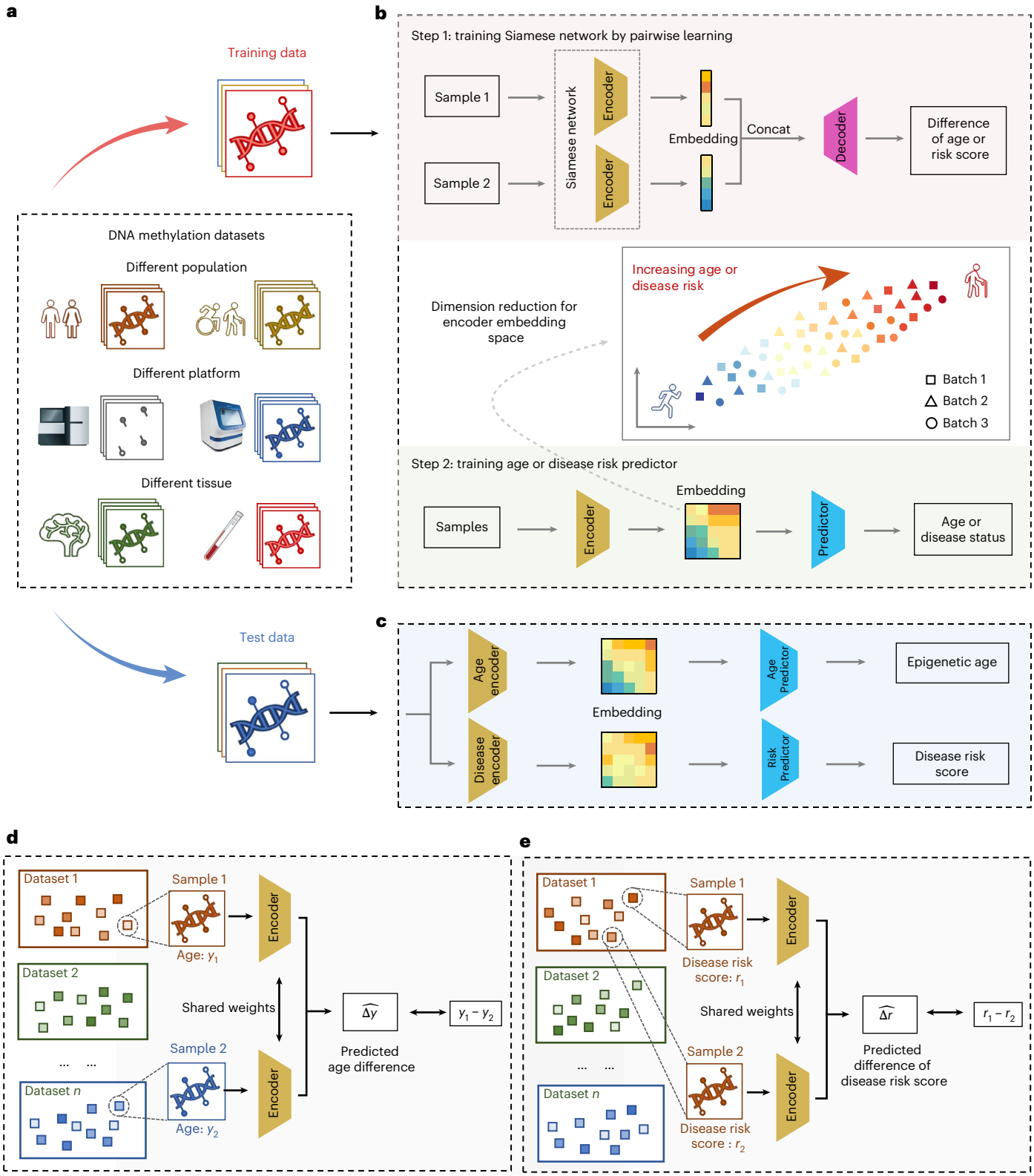

Fig. 1 | Overview of MAPLE. a, Heterogeneity in DNAm data. In clinical settings, methylation profiles are collected from diverse populations, sequencing platforms and tissue types. **b**, Training pipeline of MAPLE. An encoder is first trained using pairwise learning to map training methylation profiles from different sources into a unified embedding space. A predictor is subsequently trained on the resulting embeddings to estimate either epigenetic age or disease status. The embedding space preserves age or disease-related epigenetic signals while minimizing confounding factors. **c**, Inference with MAPLE. The trained model takes methylation data from the test set as input to predict the corresponding epigenetic ages and disease risk scores. **d**,**e**, Pairwise learning strategies used for training: pairwise learning is applied to predict chronological age differences between two samples from different datasets for epigenetic age prediction (**d**); pairwise learning is applied to predict disease risk differences between two samples within the same dataset for disease risk prediction (**e**).

https://doi.org/10.1038/s43588-025-00939-x

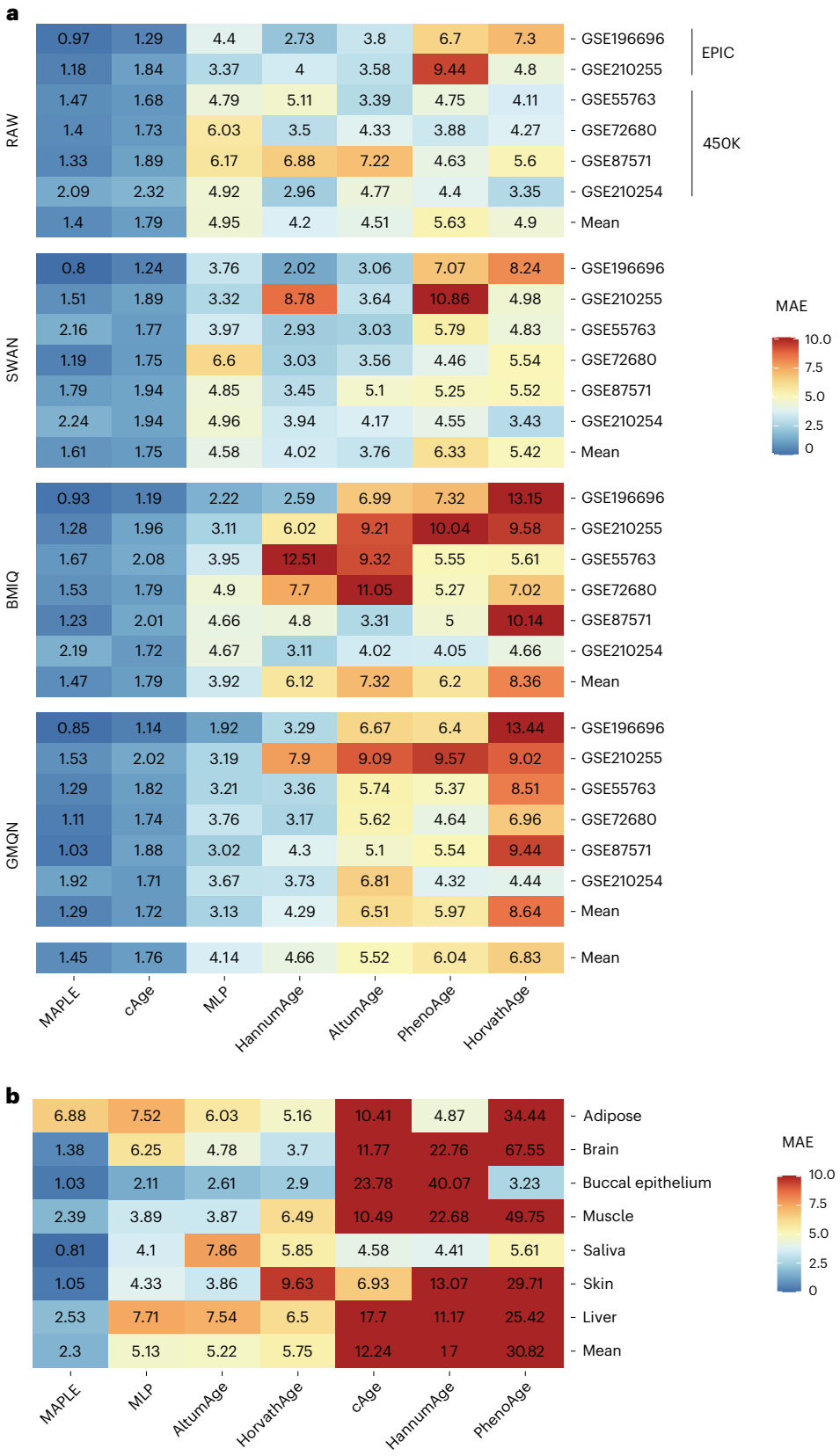

Fig. 2 | Performance of MAPLE and six competing methods across 31 benchmark tests. a, Heatmap showing the MAEs of epigenetic age prediction across blood-derived test datasets, indexed by their GEO accession numbers. Datasets GSE196696 and GSE210255 are generated using the EPIC array; all others are profiled using the 450K array. Methylation data preprocessing methods are indicated on the left. The seven methods are ordered from left to right by increasing mean MAE. **b**, Heatmap showing MAEs of age prediction across seven nonblood tissue types. Test datasets are represented by tissue types, and methods are arranged from left to right in order of increasing mean MAE.

highly stable across preprocessing pipelines, with MAEs of 1.61, 1.47 and 1.29 for the SWAN, BMIQ and GMQN methods, respectively (Fig. 2a). By contrast, HannumAge, AltumAge and HorvathAge show large performance drops when preprocessing methods change. For example, the MAEs of HannumAge, AltumAge and HorvathAge increase to 6.12, 7.32 and 8.36, respectively, in BMIQ-processed data (Fig. 2a). In the second comparison scenario, MAPLE again achieves superior performance on nonblood datasets, with an averaged MAE and an averaged PCC of 2.30 and 0.95, respectively (Fig. 2b and Supplementary Fig. 2b). In comparison, AltumAge and HorvathAge record averaged MAEs of 5.22 and 5.75, while cAge, HannumAge and PhenoAge exhibit averaged MAEs exceeding 10.0 (Fig. 2b). Across all 31 benchmarks, MAPLE achieves the lowest mean MAE (1.61), outperforming the second-ranked (cAge, MAE 3.86) and third-ranked methods (AltumAge, MAE 5.46), confirming its superior accuracy and robustness across datasets, preprocessing methods and tissue types.

To evaluate batch-effect mitigation, we perform principal component analysis (PCA) and principal variance component analysis (PVCA). PCA of the MAPLE-derived embeddings reveals a clear age-associated gradient, with samples from different tissues and studies well intermixed (Supplementary Fig. 3a–d). By contrast, PCA of the raw DNAm data shows clustering primarily driven by batch effects (Supplementary Fig. 3e,f). PVCA further confirms that, after MAPLE integration, the proportion of variance attributable to batch effects decreases markedly from 0.88 to 0.53, while the proportion of variance explained by age increases substantially from <0.01 to 0.42 (Supplementary Fig. 3). Together, these results demonstrate that MAPLE effectively mitigates batch effects while enhancing biologically meaningful age-associated signals, ensuring robust performance across heterogeneous datasets.

Aging-associated CpGs identified by MAPLE

The aim of an epigenome-wide association study (EWAS) is to examine genome-wide epigenetic variants (predominantly DNAm level at CpGs), to detect differences that are statistical associated with phenotypes of interest. EWAS is helpful for deciphering disease etiology and to identify disease-specific biomarkers. Traditionally, age-related CpGs can be identified by EWAS. We select two independent aging-related studies[24,25] and obtain their corresponding age-related CpG sets via the EWAS Atlas database[26]. Although the two CpG sets show a significant overlap (odds ratio (OR) 58.2, P value $<10^{-16}$), the genes colocalized with the overlapping CpGs are enriched only in pathways related to cell adhesion (Supplementary Fig. 4a,b).

By contrast, applying MAPLE with the integrated gradients (IG) method to the same two datasets yields a substantially greater overlap between the top 2,000 most informative CpG sites from each dataset (OR 4823.1, P value $<10^{-16}$) (Supplementary Fig. 4a; see the Methods for details on the selection strategy of attribution method and CpG site). The genes colocalized with the overlapping CpGs are significantly enriched for Gene Ontology (GO) terms related to cell adhesion, organ development and morphogenesis, cognition, cytokine production and so on (Fig. 3a). These pathway enrichment results are consistent with the literatures of aging biology[27–31]. Collectively, these findings indicate that MAPLE achieves accurate epigenetic age prediction by prioritizing biologically meaningful CpGs that are mechanistically involved in the aging process.

Sex-specific aging trajectories and condition-associated accelerated aging identified by MAPLE

Previous studies have reported sex-specific differences in aging trajectories. In particular, menopause has been associated with accelerated epigenetic aging in blood tissue, and women around the age of 50 years are known to undergo substantial physiological changes related to aging[32,33]. This phenomenon probably reflects a sharper transition in epigenetic aging during the menopausal period, in contrast to the relatively steady progression of chronological age. As a result, female individuals around menopausal age are expected to exhibit increased discordance between epigenetic and chronological age.

To test this hypothesis, we use MAPLE to predict epigenetic age on held-out blood tissue datasets and calculate a rolling mean of the absolute error between predicted epigenetic age and chronological age to quantify the level of discordance. As shown in Fig. 3b, female individuals around age 52 exhibit a pronounced increase in absolute error, whereas no comparable pattern is observed in the male subgroup. Note that most women undergo menopause between the ages of 45 and 55, with the average age being approximately 52. This sex-specific trend suggests that MAPLE effectively captures key biological processes associated with menopause-related aging dynamics.

In addition, it has been widely hypothesized that the estimated age by epigenetic clocks can be used to calculate age acceleration, defined as the difference between estimated and actual chronological age, which reflects variations in past rates of aging[6,9,10,22]. A higher value of age acceleration indicates a faster aging rate and an increased risk of mortality and age-related diseases.

To assess whether the epigenetic age predicted by MAPLE accurately reflects the rate of aging, we use several biologically well-recognized examples of accelerated aging. Populations with Down syndrome, human immunodeficiency virus (HIV) infection, smoking habits, obesity and Alzheimer's disease (AD) are all believed to experience accelerated aging and higher mortality rate[30,34–37]. In accordance with the findings of aging biology, MAPLE predicts that these populations with Down syndrome, HIV infection, smoking habits, obesity and AD exhibit significantly higher age acceleration compared with control groups (Fig. 3c–g and Supplementary Fig. 5a–c). Notably, MAPLE detects significant age acceleration in the brain tissues of patients with AD, but not in their blood samples (Fig. 3g and Supplementary Fig. 5d). This finding demonstrates that nonblood tissues can reveal unique and biologically relevant aging signals that may be missed by blood-based analyses alone, further underscoring the importance of cross-tissue epigenetic age prediction for advancing aging biology research.

CVD risk assessment with MAPLE

Complex chronic diseases, such as CVD and T2D, are generally influenced by genetic and environmental factors. DNAm is an important epigenetic modification through which genetic variants and environmental exposures impact disease risk, providing dynamic predictive information for an individual's disease risk.

In the progression of CVD, age is a strong nonmodifiable risk factor[38]. Atherosclerosis (AS) is a pathological condition characterized by the accumulation of plaques within the arterial walls, leading to reduced blood flow and an elevated risk of heart attack and stroke[39]. Coronary artery ectasia (CAE) is a condition characterized by abnormally dilated coronary arteries, predominantly caused by AS, and is associated with a worse prognosis[40]. Stroke is a severe disease that is defined as a neurological deficit attributed to an acute focal injury of the central nervous system by a vascular cause, mostly due to embolism from atherosclerotic plaque[41].

MAPLE effectively integrates DNAm data from multiple studies and can predict CVD risk. In the PCA plots of its embedding space, younger control samples cluster in the lower left, whereas older and unhealthy samples with stroke, CAE or arteriosclerosis appear in the upper right (Fig. 4a,b). Samples from different studies are well mixed (batch PVCA reduced from 0.78 to 0.53; Supplementary Fig. 6), indicating that MAPLE maps heterogeneous batches into a unified embedding capturing CVD-related epigenetic signals. Using these embeddings, MAPLE computes a CVD risk score for each sample. Among seven test groups, stroke samples show the highest scores, while those aged 18–40 have the lowest (Fig. 4c). By contrast, the Cox model produces inconsistent risk scores across datasets, failing to generalize to independent cohorts (Fig. 4c).

https://doi.org/10.1038/s43588-025-00939-x

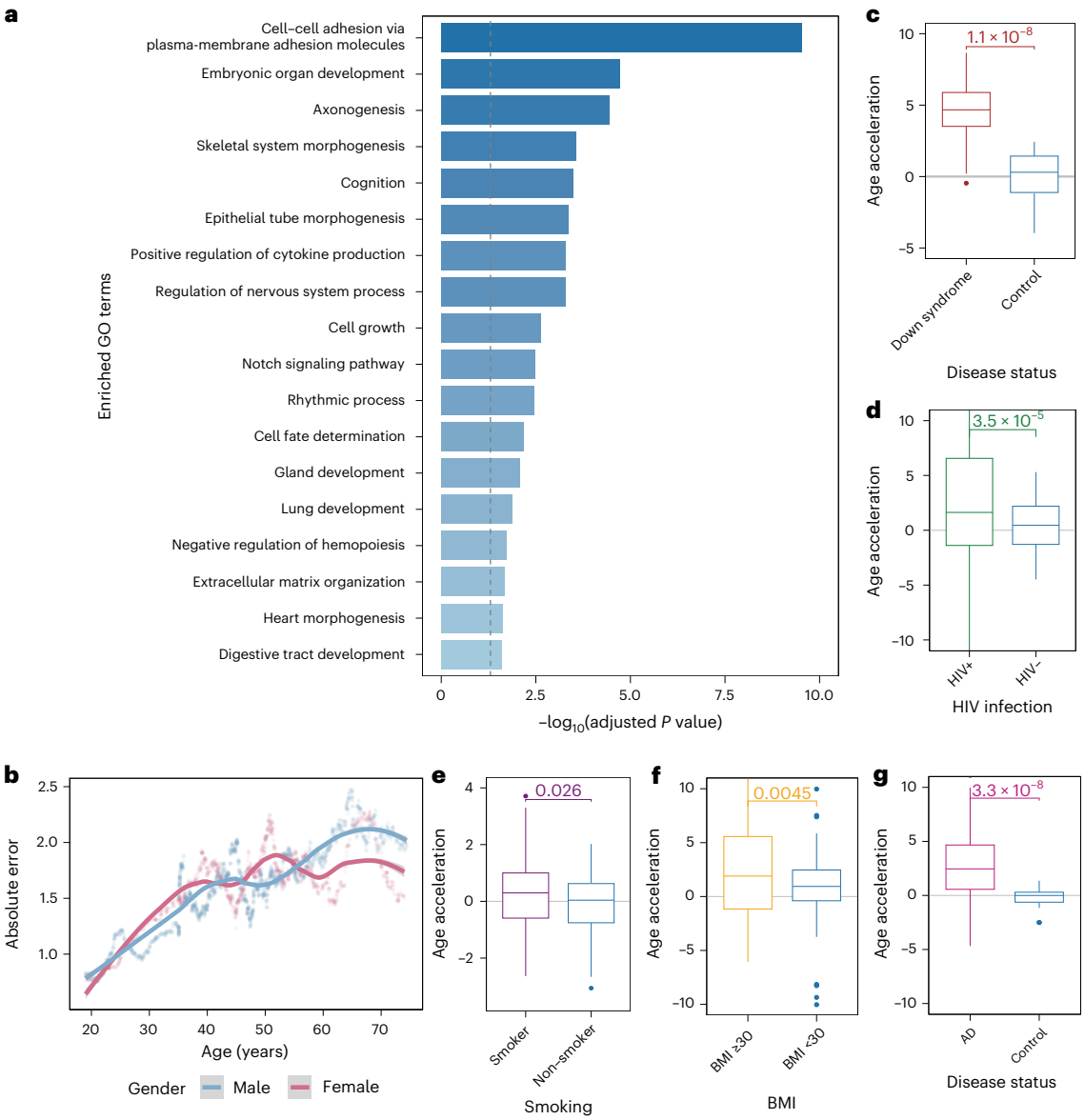

Fig. 3 | Capturing aging-related biological processes and detecting accelerated aging with MAPLE. a, Bar plots showing enriched pathways of the genes colocalized with age-related CpGs identified by MAPLE. Pathway enrichment is evaluated using a one-sided hypergeometric test, with *P* values adjusted using the Benjamini–Hochberg method. The dotted gray line denotes the adjusted significance threshold of 0.05. **b**, Dot plots showing the trajectory of the 100-sample rolling mean of the absolute error between epigenetic and chronological age across the lifespan. Each dot represents the mean absolute error within a rolling window of 100 samples, with colors indicating sex. Solid lines represent locally estimated scatterplot smoothing (LOESS) fits to the data, and the shaded gray area denotes the 95% confidence interval. **c**–**g**, Box plots showing age acceleration levels across populations with Down syndrome (**c**), HIV infection (**d**), smoking (**e**), obesity (**f**) and AD (**g**), compared with their respective control populations. Each dot represents the age-acceleration estimate derived from one independent biological sample, defined as the DNAm profile from a single individual. For **c**–**d**, the sample sizes are as follows: Down syndrome, $n = 29$ cases and $n = 19$ controls (blood); HIV, $n = 229$ cases and $n = 45$ controls (blood); smoking, $n = 72$ smokers and $n = 80$ nonsmokers (blood); obesity, $n = 84$ cases and $n = 82$ controls (muscle); AD, $n = 64$ cases and $n = 63$ controls (brain). Box plots represent the median (center line) and the 25th and 75th percentiles (box bounds), and the whiskers extend to 1.5× the interquartile range (IQR). Statistical significance is evaluated using one-sided *t*-tests.

Furthermore, we compare MAPLE with the Cox model by evaluating their risk scores in distinguishing disease and pre-disease populations from controls. For stroke identification, MAPLE achieves an area under the receiver operating characteristic curve (AUROC) of 0.98 and an area under the precision-recall curve (AUPRC) of 0.90, substantially outperforming the Cox model (AUROC 0.69, AUPRC 0.13; Fig. 4d,e). For the identification of pre-disease status including CAE and AS, MAPLE yields higher AUPRCs (0.41 for CAE and 0.42 for AS) than the Cox model (0.20 and 0.35, respectively; Fig. 4d,e). Notably, MAPLE maintains superior performance even when the age distribution between positive and negative samples is balanced (Supplementary Fig. 7a,b). We further assess whether the risk scores from both models could distinguish three nonhealthy conditions from healthy controls. For stroke classification, MAPLE achieves an AUPRC exceeding 0.99, whereas the Cox model reaches only 0.80 (Supplementary Fig. 7c,d), indicating that MAPLE more effectively reduces false-positive classifications. Overall, MAPLE consistently outperforms the Cox model in detecting both disease and pre-disease states associated with CVD.

An enrichment analysis of genes colocalized with CVD-related CpG sites identified by MAPLE highlights its ability to capture CpG

https://doi.org/10.1038/s43588-025-00939-x

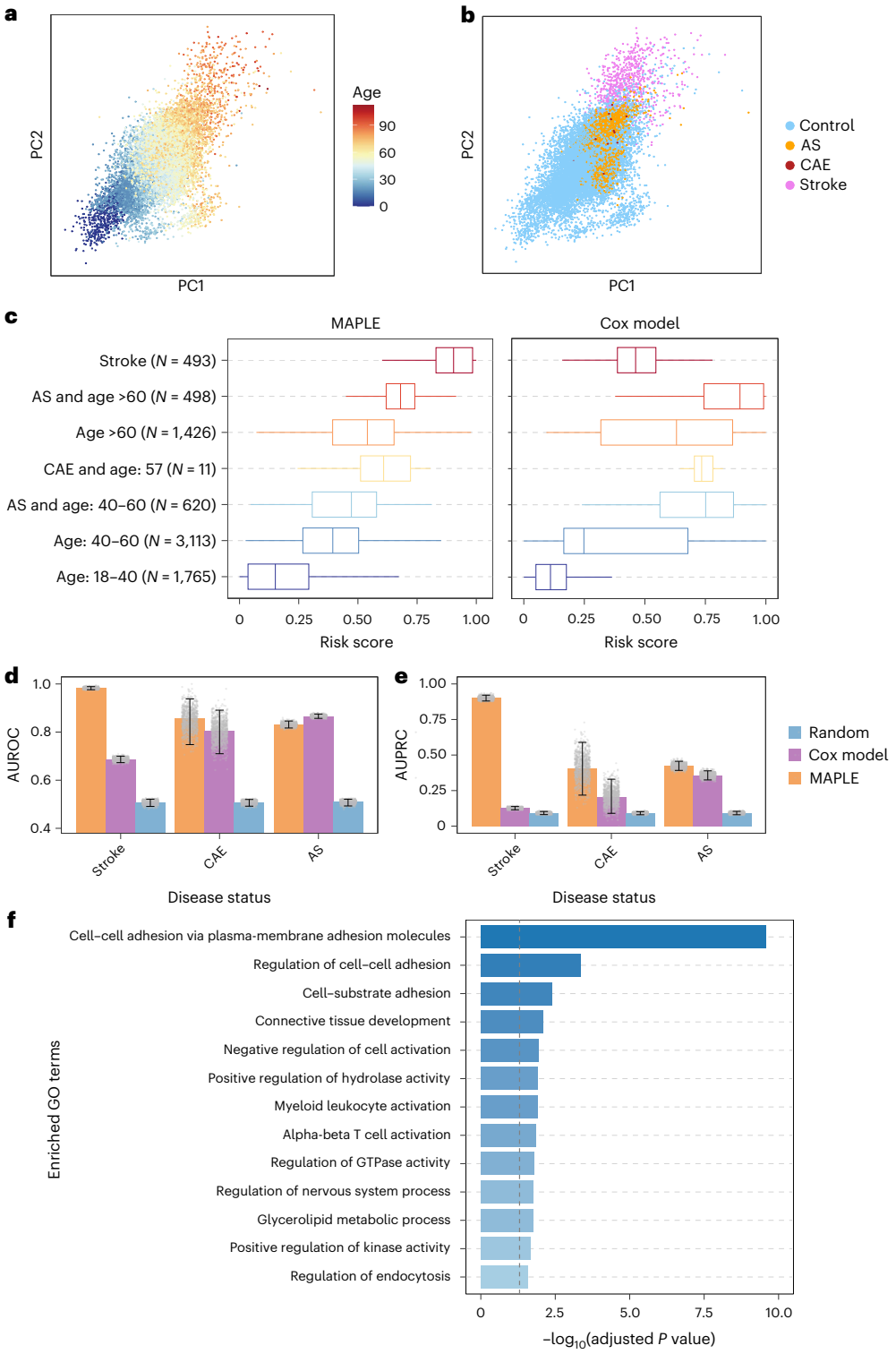

Fig. 4 | CVD risk estimation from DNAm data with MAPLE. a,**b**, PCA visualizations of methylation embeddings for training and test samples. In **a**, point colors represent sample age, whereas in **b**, point colors indicate disease status related to CVD ($N_{control}$ = 13,321, N_{AS} = 1,118, N_{CAE} = 11, N_{stroke} = 710). **c**, Box plots comparing CVD risk scores predicted by MAPLE (left) and the Cox model (right) for different groups of test samples. Each dot represents the predicted risk score for one independent biological sample, defined as the DNAm profile from a single individual. The sample sizes for each group are shown in the figure. Box plots depict the median (center line), the 25th and 75th percentiles (box bounds) and the whiskers, which extend to 1.5× the IQR. **d**,**e**, Bar plots showing the AUROC (**d**) and AUPRC (**e**) values for distinguishing stroke, CAE and AS samples from controls. Data are presented as mean values, and error bars denote the 95% confidence intervals (2.5th–97.5th percentiles), estimated from 1,000 bootstrap resamples. The gray dots represent the individual AUROC or AUPRC values obtained in each bootstrap iteration, illustrating the full distribution of bootstrap estimates. **f**, Bar plots showing enriched pathways of the genes colocalized with CVD-related CpGs identified by MAPLE. Pathway enrichment is evaluated using a one-sided hypergeometric test, with *P* values adjusted using the Benjamini–Hochberg method. The dotted gray line denotes the adjusted significance threshold of 0.05.

https://doi.org/10.1038/s43588-025-00939-x

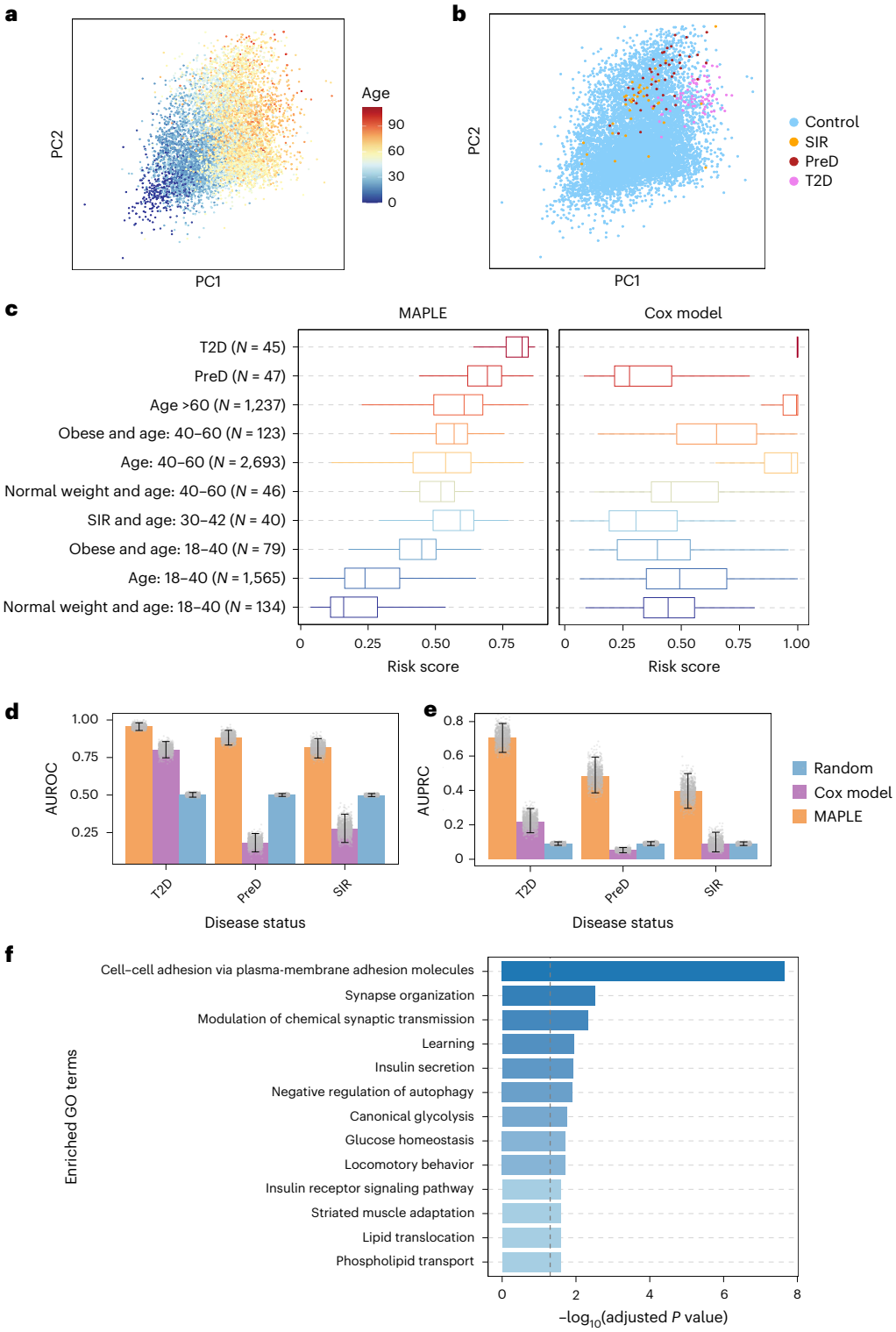

Fig. 5 | T2D risk estimation from DNAm data with MAPLE. a,**b**, PCA visualizations of methylation embeddings for training and test samples. In **a**, point colors represent sample age, whereas in **b**, point colors indicate disease status related to T2D ($N_{control}$ = 12,576, N_{SIR} = 40, N_{PreD} = 47, N_{T2D} = 74). **c**, Box plots comparing T2D risk scores predicted by MAPLE (left) and Cox model (right) for different groups of test samples. Each dot represents the predicted risk score for one independent biological sample, defined as the DNAm profile from a single individual. The sample sizes for each group are shown in the figure. Box plots depict the median (center line), the 25th and 75th percentiles (box bounds) and the whiskers, which extend to 1.5× the IQR. **d**,**e**, Bar plots show the AUROC (**d**) and AUPRC (**e**) values for distinguishing T2D, PreD and SIR samples from controls. Data are presented as mean values, and error bars denote the 95% confidence intervals (2.5th–97.5th percentiles), estimated from 1,000 bootstrap resamples. The gray dots represent the individual AUROC or AUPRC values obtained in each bootstrap iteration, illustrating the full distribution of bootstrap estimates. **f**, Bar plots showing enriched pathways of the genes colocalized with T2D-related CpGs identified by MAPLE. Pathway enrichment is evaluated using a one-sided hypergeometric test, with *P* values adjusted using the Benjamini–Hochberg method. The dotted gray line denotes the adjusted significance threshold of 0.05.

sites implicated in the pathogenesis of CVD. The genes are enriched in cell-adhesion-related terms, 'positive regulation of hydrolase activity', 'myeloid leukocyte activation', 'alpha-beta T cell activation' and so on (Fig. 4f). These enrichment results are consistent with previous reports on the molecular mechanisms underlying CVD pathophysiology[42–45].

T2D risk assessment with MAPLE

The risk assessment for T2D can similarly be conducted using the same approach as the CVD model. Both age and obesity are powerful risk factors in the progression of T2D. Systemic insulin resistance (SIR) is a pathological condition in which the body does not respond to insulin and is unable to convert the glucose into energy, and prediabetes (PreD) is a condition characterized by higher-than-normal blood sugar levels that are not yet high enough to be considered T2D. Individuals with SIR and PreD are more likely to progress to T2D. T2D is a disease characterized by dysregulation of carbohydrate, lipid and protein metabolism[11].

For T2D risk prediction, MAPLE encodes DNAm profiles into a unified embedding space that captures T2D-related variation while minimizing technical bias. In the PCA plots, younger control samples cluster in the lower left, whereas older and unhealthy samples (including SIR, PreD and T2D) occupy the upper right (Fig. 5a,b). Samples from different studies are well mixed (batch PVCA reduced from 0.84 to 0.60; Supplementary Fig. 8), indicating robust integration across datasets. MAPLE computes a T2D risk score for each sample, with the highest values in T2D, followed by PreD. In age-stratified controls, scores increase with age, and obese individuals show higher scores than those with normal weight. SIR samples also display elevated risk scores relative to age-matched controls (Fig. 5c). By contrast, the Cox model yields inconsistent and noncomparable scores across datasets.

Using these risk scores for classification, MAPLE achieves an AUROC of 0.96 and an AUPRC of 0.71 for T2D, outperforming the Cox model (AUROC 0.80, AUPRC 0.22) (Fig. 5d,e). For PreD and SIR, MAPLE attains AUROCs of 0.89 and 0.82 and AUPRCs of 0.48 and 0.39, respectively, whereas the Cox model performs below random (Fig. 5d,e). MAPLE maintains this advantage even after balancing age distributions between positive and negative samples (Supplementary Fig. 9a,b). Both MAPLE and the Cox model effectively classify T2D and healthy samples; however, for PreD and SIR, MAPLE demonstrates superior performance compared with the Cox model (Supplementary Fig. 9c,d). These results highlight MAPLE's superior ability to generalize across datasets and minimize false-positive classifications among healthy individuals.

An enrichment analysis further demonstrates that MAPLE effectively identifies T2D-caused CpG sites. These CpGs sites are localized on the genes enriched in cell-adhesion-associated terms, 'learning', 'insulin secretion', 'canonical glycoysis', 'glucose homeostasis', 'locomotory behavior', 'insulin repcetor signaling pathway' and so on (Fig. 5f). These enrichment results are consistent with previous reports on T2D pathogenesis[11,46].

Discussion

Developing clinically applicable epigenetic clocks remains challenging due to two key factors: the high dimensionality of DNAm data, where the number of CpG sites far exceeds the number of samples, and the heterogeneity introduced by variations in sequencing platforms, preprocessing pipelines and tissue sources. MAPLE addresses both challenges through pairwise learning. Using pairwise inputs greatly expands the number of effective training samples, mitigating overfitting. Instead of predicting absolute age, MAPLE predicts relative age differences between two samples, allowing it to focus on genuine aging-related signals while ignoring confounding factors such as tissue type or platform. This design also enables MAPLE to outperform conventional Cox models in disease risk prediction.

Large cohort studies often collect extensive DNAm and longitudinal clinical data, offering opportunities to build predictors for chronic disease risk. However, due to concerns regarding participant privacy, these studies typically release only risk assessment models rather than the underlying DNAm and clinical datasets, limiting the development of more advanced prediction frameworks. In this study, MAPLE overcomes this challenge by integrating publicly available DNAm data with the released risk models to predict the risk of two chronic diseases (CVD and T2D), demonstrating the practicality of privacy-preserving methylation-based risk prediction.

Furthermore, the progression of various diseases, including neurological and autoimmune disorders, is closely associated with alterations in DNAm[47]. As a versatile computational framework, MAPLE can be extended to develop predictors for a broader range of health outcomes. Therefore, MAPLE advances the clinical utility of methylation-based assessment while safeguarding participant privacy.

Recent advances in computational epigenomics have introduced some valuable approaches for DNAm analysis, including sequence-based prediction algorithms such as iDNA-ABF[48] and Methyl-GP[49], as well as foundation models such as MethylGPT[50]. These methods offer complementary directions to MAPLE. Sequence-based methods infer DNAm patterns directly from DNA sequence, focusing on intrinsic genomic determinants of methylation. By contrast, MAPLE takes DNAm profiles as input, integrating both genetic and environmental influences to assess aging and disease risk. In future work, paired DNAm and sequence data could be jointly utilized within the MAPLE framework. For instance, DNA sequence features could be processed by sequence-based models to generate predicted DNAm profiles, which could then be integrated with empirically measured DNAm data to improve downstream predictions. Such integration would facilitate the disentanglement of genetically determined versus environmentally induced DNAm changes, offering deeper insights into the mechanisms underlying individual health trajectories.

Foundation models such as MethylGPT, which learn context-aware DNAm representations from large-scale data, could also be combined with MAPLE. In this hybrid architecture, MethylGPT would provide biologically informed representations of individual samples, while MAPLE would model intersample variation to yield accurate predictions of epigenetic age or disease risk. This synergy could substantially improve the precision and robustness of methylation-based predictive modeling.

Methods

Architecture of MAPLE

We construct MAPLE and apply it to two tasks: epigenetic age prediction and disease risk prediction. This approach addresses the challenge of the batch effects among DNAm datasets and generalizes the prediction ability to unseen data. Below, we describe the workflows of the two tasks to clarify how pairwise learning is used to tackle the inherent incompatibility among datasets.

Age prediction task. The age prediction task involves training a model on DNAm datasets P to predict the epigenetic age of samples in the held-out test datasets Q. Let $\{\mathbf{x}_i^{(p)}\}_{i=1}^{N_p}$ be the DNAm profiles of samples after preprocessing from a DNAm training dataset indexed by $p \in P$, and $\{y_i^{(p)}\}_{i=1}^{N_p}$ be the corresponding chronological ages. Here, each $\mathbf{x}_i^{(p)}$ is a C-dimensional vector, where C is the number of CpG sites; $y_i^{(p)}$ is a continuous scalar value; and N_p is the number of samples in dataset p. Similarly, let $\{\mathbf{x}_i^{(q)}\}_{i=1}^{N_q}$ be the DNAm profiles of samples after preprocessing from the qth DNAm test dataset with N_q samples ($q \in Q$), whose ages are unknown.

In the age prediction task, the training of MAPLE follows a two-step procedure, during which only the training datasets P are used.

The held-out test datasets Q remain entirely unseen throughout the training process to ensure unbiased evaluation. The first step is the generation of unified DNAm embeddings using pairwise learning. MAPLE aims to utilize pairwise learning to encode the DNAm profiles $\{\mathbf{x}_i^{(p)}\}_{i=1}^{N_p}$ into a uniform embedding space where the aging-related signal is preserved but the batch effect is removed. It adopts the Siamese network architecture, where the input is a pair of DNAm profile, and the predictive objective is the age difference between them. Given the DNAm profile $\mathbf{x}_i^{(p)}$ of the ith sample and the DNAm profile $\mathbf{x}_j^{(p)}$ $(i \neq j)$ of the jth sample (sample i and sample j can come from same or different training datatsets), MAPLE uses a MLP f with the shared parameters $\theta_{\text{Enc,age}}$ as encoder to encode the two DNAm profiles to the corresponding latent representation vector $\mathbf{l}_i^{(p)}$ and $\mathbf{l}_j^{(p)}$, and the calculation process is as follows:

$$\mathbf{l}_i^{(p)} = f\left(\mathbf{x}_i^{(p)}, \theta_{\text{Enc,age}}\right); \mathbf{l}_j^{(p)} = f\left(\mathbf{x}_j^{(p)}, \theta_{\text{Enc,age}}\right).$$

Then, another MLP h is regarded as decoder to take the combination of $\mathbf{l}_i^{(p)}$ and $\mathbf{l}_j^{(p)}$ to predict the age difference between the ith sample and jth sample $\Delta\hat{y}_{ij}^{(p)}$, and the calculation process is as follows:

$$\Delta\hat{y}_{ij}^{(p)} = h\left(\left[\mathbf{l}_i^{(p)} \parallel \mathbf{l}_j^{(p)} \parallel \mathbf{l}_i^{(p)} - \mathbf{l}_j^{(p)}\right], \theta_{\text{Dec,age}}\right),$$

where $\theta_{\text{Dec,age}}$ is the parameters of decoder h.

The loss function of the first step is Huber loss between predictive and true age difference, and its formula is as follows:

$$\text{Loss}\left(y_i^{(p)} - y_j^{(p)}, \Delta\hat{y}_{ij}^{(p)}\right) = \begin{cases} \frac{1}{2}\left(y_i^{(p)} - y_j^{(p)} - \Delta\hat{y}_{ij}^{(p)}\right)^2, if\,|y_i^{(p)} - y_j^{(p)} - \Delta\hat{y}_{ij}^{(p)}| \leq \delta \\ \delta|y_i^{(p)} - y_j^{(p)} - \Delta\hat{y}_{ij}^{(p)}| - \frac{1}{2}\delta^2, if\,|y_i^{(p)} - y_j^{(p)} - \Delta\hat{y}_{ij}^{(p)}| > \delta \end{cases},$$

where δ is the hyperparameter of Huber loss. Huber loss effectively combines the advantages of both mean absolute error (MAE) and Mean Squared Error (MSE) loss functions. For small errors, it resembles MSE loss, facilitating faster convergence during training compared to MAE. However, for larger errors, it behaves more like MAE, exhibiting reduced sensitivity to outliers relative to MSE loss[51].

The second step is the training of age predictor. MAPLE aims to directly predict age $\hat{y}_i^{(p)}$ based on the well-trained encoder in the first step. Given the DNAm profile $\mathbf{x}_i^{(p)}$ of the ith sample, the age predictor g takes the output of encoder f as input to predict $\hat{y}_i^{(p)}$, and the calculation process is as follows:

$$\hat{y}_i^{(p)} = g\left(f\left(\mathbf{x}_i^{(p)}, \theta_{\text{Enc,age}}\right), \theta_{\text{Pre,age}}\right),$$

where $\theta_{\text{Pre,age}}$ is the parameters of predictor g. The loss function of the second step is Huber loss between predictive and true age.

In the inference stage, the epigenetic age $\hat{y}_i^{(q)}$ of the ith sample of the qth test dataset is predicted by the calculation process as follows:

$$\hat{y}_i^{(q)} = g\left(f\left(\mathbf{x}_i^{(q)}, \theta_{\text{Enc,age}}\right), \theta_{\text{Pre,age}}\right).$$

Disease risk prediction task. In the disease risk prediction task, a model is trained on the DNAm datasets P and evaluated on the held-out test datasets Q. Let $\{\mathbf{x}_i^{(p)}\}_{i=1}^{N_p}$ be the DNAm profiles of samples after preprocessing from a DNAm training dataset indexed by $p \in P$, and $\{r_i^{(p)}\}_{i=1}^{N_p}$ be the corresponding risk scores provided by the linear model from a large cohort study and comparable only within the dataset p. Here, each $\mathbf{x}_i^{(p)}$ is a C-dimensional vector, where C is the number of CpG sites; $r_i^{(p)}$ is a continuous scalar value; $z_i^{(p)} \in \{0, 1\}$, where 0 and 1 corresponds control and disease status, respectively; and N_p is the number of samples in dataset p. Similarly, let $\{\mathbf{x}_i^{(q)}\}_{i=1}^{N_q}$ be the DNAm profiles of samples after preprocessing from the qth DNAm test dataset with N_q samples $(q \in Q)$.

The training process of the disease risk prediction model consists of two sequential steps, both of which are performed exclusively on the training datasets P, with the held-out test datasets Q strictly excluded throughout the entire procedure. The first step is the generation of unified DNAm embeddings using pairwise learning. MAPLE aims to utilize pairwise learning to encode the DNAm profiles $\{\mathbf{x}_i^{(p)}\}_{i=1}^{N_p}$ into a uniform embedding space where the disease-risk-related signal is preserved but the batch effect is removed. Given the DNAm profile $\mathbf{x}_i^{(p)}$ of the ith sample and the DNAm profile $\mathbf{x}_j^{(p)}$ $(i \neq j)$ of the jth sample (sample i and sample j can come from the same training datatset), MAPLE uses an encoder f with the shared parameters $\theta_{\text{Enc,disease}}$ to encode the two DNAm profiles to the corresponding latent representation vector $\mathbf{l}_i^{(p)}$ and $\mathbf{l}_j^{(p)}$, and then uses a decoder to transform the combination of $\mathbf{l}_i^{(p)}$ and $\mathbf{l}_j^{(p)}$ to the risk score difference between the ith sample and jth sample $\Delta\hat{r}_{ij}^{(p)}$. In contrast to the age prediction, the ε-insensitive loss function presented by support vector regression[52] is used owing to its robustness against noise and its ability to mitigate overfitting, and its formula is as follows:

$$\text{Loss}\left(r_i^{(p)} - r_j^{(p)}, \Delta\hat{r}_{ij}^{(p)}\right) = \begin{cases} 0, if\,|r_i^{(p)} - r_j^{(p)} - \Delta\hat{r}_{ij}^{(p)}| \leq \varepsilon \\ |r_i^{(p)} - r_j^{(p)} - \Delta\hat{r}_{ij}^{(p)}| - \varepsilon, if\,|r_i^{(p)} - r_j^{(p)} - \Delta\hat{r}_{ij}^{(p)}| > \varepsilon \end{cases},$$

where ε is the slack variable of the ε-insensitive loss function.

The second step is the training of disease classifier. Given the DNAm profile $\mathbf{x}_i^{(p)}$ of the ith sample, the classifier g takes the output of encoder f as input to predict the possibility distribution $\hat{z}_i^{(p)}$ of disease status, and the calculation process is as follows:

$$\hat{z}_i^{(p)} = \text{softmax}\left(g\left(f\left(\mathbf{x}_i^{(p)}, \theta_{\text{Enc,disease}}\right), \theta_{\text{Pre,disease}}\right)\right),$$

where $\theta_{\text{Pre,age}}$ is the parameters of classifier g. The loss function of the second step is cross-entropy between predictive and true disease status, and its formula is as follows:

$$\text{CE}\left(\hat{z}_i^{(p)}, z_i^{(p)}\right) = -\frac{1}{N}\sum_{i=1}^{N}\left(z_i^{(p)}\log\left(\hat{z}_i^{(p)}\right) + \left(1 - z_i^{(p)}\right)\log\left(1 - \hat{z}_i^{(p)}\right)\right),$$

where N is the number of training samples.

In the inference stage, we calculate disease risk score in the embedding space because the classifier g can only provide disease states rather than a continuous disease risk indicator. The DNAm profiles $\{\mathbf{x}_i^{(p)}\}_{i=1}^{N_p}$ and $\{\mathbf{x}_i^{(q)}\}_{i=1}^{N_q}$ of the training and test datasets are mapped into embeddings $\{\mathbf{l}_i^{(p)}\}_{i=1}^{N_p}$ and $\{\mathbf{l}_i^{(q)}\}_{i=1}^{N_q}$. From the training datasets, we select the individuals with age <20 as the healthy group, with embeddings $\{\mathbf{l}_i^{(p)}\}_{i=1}^{N_{\text{health}}}$. We also select the individuals diagnosed with stroke or T2D as disease group, with embeddings $\{\mathbf{l}_i^{(p)}\}_{i=1}^{N_{\text{disease}}}$.

For each test sample, we compute two average Euclidean distances, distance to the healthy group $D_{\text{health},i}$ and distance to the disease group $D_{\text{disease},i}$, as follows:

$$D_{\text{health},i} = \frac{1}{N_{\text{health}}}\sum_{j=1}^{N_{\text{health}}}||\mathbf{l}_i^{(q)} - \mathbf{l}_j^{(p)}||_2, q \in Q, p \in P$$

$$D_{\text{disease},i} = \frac{1}{N_{\text{disease}}}\sum_{j=1}^{N_{\text{disease}}}||\mathbf{l}_i^{(q)} - \mathbf{l}_j^{(p)}||_2, q \in Q, p \in P,$$

where $||\cdot||_2$ denotes the Euclidean norm.

Next, we normalize these distances based on quartile statistics from the training data. For D_{health}, we defined the lower quartile of D_{health} among healthy training samples as $D_{\text{health,min}}$, and the upper quartile of D_{health} among disease training samples as $D_{\text{health,max}}$, respectively. The normalized distance the healthy group $D_{\text{health,norm},i}$ of each test sample is calculated as follows:

$$D_{\text{health,norm},i} = \min\left(\max\left(\frac{D_{\text{health},i} - D_{\text{health,min}}}{D_{\text{health,max}} - D_{\text{health,min}}}, 0\right), 1\right).$$

Similarly, for D_{disease}, we defined the upper quartile of D_{disease} among healthy training samples as $D_{\text{disease,max}}$, and the lower quartile of D_{health} among disease training samples as $D_{\text{disease,min}}$, respectively. The normalized distance the disease group $D_{\text{health,norm},i}$ of each test sample is calculated as follows:

$$D_{\text{disease,norm},i} = \min\left(\max\left(\frac{D_{\text{disease},i} - D_{\text{disease,min}}}{D_{\text{disease,max}} - D_{\text{disease,min}}}, 0\right), 1\right).$$

Finally, the disease risk score for each test sample is computed as

$$\text{Riskscore}_i = \frac{D_{\text{health},norm,i} + 1 - D_{\text{disease,norm},i}}{2}.$$

Implementation details

For all MLPs in this Article, the dropout probability is set to 0.2, and the activation function is ReLU. In the age prediction task, the encoder consists of a 9-layer MLP with an input size of 285,898, hidden layer sizes of 1,024, 1,024, 512, 512, 256, 256, 64 and 64, and an output size of 32. The decoder is a 5-layer MLP with an input size of 96 (formed by concatenating three outputs from the encoder), hidden layer sizes of 32, 32, 16 and 16, and an output size of 1. The age predictor is also a 5-layer MLP, with an input size of 32, hidden layer sizes of 32, 32, 16 and 16, and an output size of 1. Model parameters are updated using the Adam optimizer with initial learning rates of 1×10^{-4} in the first step and 2×10^{-5} in the second step. The parameter δ of Huber loss function is set to 5. In the disease risk prediction task, the encoder is a 7-layer MLP with an input size of 307,457 (for CVD) or 303,212 (for T2D), hidden layer sizes of 1,024, 1,024, 256, 256, 64 and 64, and an output size of 32. The decoder architecture remains identical to that used in the age prediction task. The disease classifier is a 5-layer MLP, with an input size of 32, hidden layer sizes of 32, 32, 16 and 16, and an output size of 2. Optimizer settings are consistent with those in the age prediction task. The number of principal components used in the disease risk predictor is set to 10. The ε of the ε-insensitive loss function is set to 0.05.

Hyperparameter selection

To evaluate the impact of hyperparameters δ (in the Huber loss) on epigenetic age prediction, we compute the average MAE of MAPLE across 31 benchmark tests under different δ values. The results indicate that δ values in the range of 3–7 offer an optimal trade-off between convergence speed and prediction accuracy (Supplementary Fig. 10a), which aligns well with theoretical expectations. Theoretically, the Huber loss is known to effectively balance the benefits of MAE and MSE. When $\delta$is small, the loss function behaves similarly to MAE, which is more robust to outliers but converges slowly due to nondifferentiability at zero. Conversely, a larger δ makes the Huber loss resemble MSE, which accelerates convergence but is more sensitive to outliers, potentially degrading generalization.

To assess the sensitivity of MAPLE to the slack variable ε in the disease risk prediction task, we vary ε in the range of 0–0.3 and evaluate model performance on stroke and pre-disease status classification (including CAE and AS) in the test data of CVD task. Our experiments show that setting ε in the range of 0.05–0.15 yields the most favorable classification results (Supplementary Fig. 10b,c). These empirical results are well supported by theoretical analysis. In ε-insensitive loss functions, the hyperparameter ε defines a 'no-penalty zone' within which prediction errors are not penalized. This mechanism introduces a margin of tolerance that can improve generalization by ignoring small deviations. However, ε that is too small leads to overly strict constraints, causing the model to overfit noise and become overly sensitive to minor variations. Conversely, a too-large ε results in overly relaxed constraints, making the model insensitive to meaningful signals and prone to underfitting.

To justify the choice of the MLP architecture in MAPLE, we conduct an ablation study using the epigenetic age prediction task as an example. We evaluate the impact of different architectural configurations based on the average MAE of MAPLE across 31 benchmark datasets. Three key hyperparameters are systematically varied: the number of encoder layers, the number of decoder layers and the network width. Considering accuracy, training speed and memory consumption, the optimal range is found to be 5 to 9 for encoder layers, 4 to 7 for decoder layers and 0.75× to 2× the current setting for network width (Supplementary Fig. 11). Therefore, the current architectural choices in MAPLE fall within the empirically optimal range.

Data preprocessing

DNAm array generally contains two kinds of probes, namely, Infinium I and II probes, and the β value distributions they produce are very different, influencing the analysis of DNAm data. In this Article, we use three common methods, namely SWAN[19], BMIQ[20] and GMQN[21], to normalize DNAm array data.

We utilize GMQN to correct batch effects and probe bias in the training data. Low-quality CpG sites, with >10% of missing values across train samples were removed. For the age, CVD risk and T2D risk prediction tasks, 285,898, 307,457 and 303,212 CpG sites are preserved. For the CpG sites used in the training stage, missing values of each CpG site were imputed with the average beta value of that CpG across other samples, using SimpleImputer from the Python package scikit-learn.

For training the disease classifier, we make data augmentation for the disease samples to balance the numbers of positive and negative samples. The numbers of stroke and T2D samples are amplified from 217 and 29 to 2152 and 2059. The method of data augmentation is that the 20% of β values are masked and the new positive samples are generated by imputing the masked values with the average beta value of each CpG across other samples.

To assess the impact of data augmentation on disease risk prediction, we compare the following variants of our model on the CVD risk prediction task: MAPLE (our original model using the masking/imputation-based augmentation method), MAPLE_w/o augmentation (the same model trained without any augmentation on disease samples), MAPLE_SMOTE (using SMOTE for disease sample augmentation), MAPLE_SVMSMOTE (using SVMSMOTE), MAPLE_ADASYN (using ADASYN). The results demonstrate that the masking/imputation method, ADASYN, and SVMSMOTE consistently yield the strongest performance across tasks. While SMOTE provides moderate improvement, all augmentation methods outperform the baseline model trained without any data augmentation (Supplementary Fig. 12).

Competing methods

In the age prediction task, MAPLE are compared with five epigenetic clocks including HorvathAge[9], HannumAge[6], PhenoAge[22], AltumAge[10] and cAge[23].

HorvathAge is a canonical multitissue epigenetic clock built using elastic net regression on DNAm profiles from 7,844 noncancerous samples spanning 51 tissues and cell types. The model regresses a nonlinear transformation of chronological age on 21,369 CpGs shared between

the Illumina 27K and 450K arrays. It selects 353 CpGs, yielding a linear combination weighted by elastic net coefficients ($\alpha = 0.5$, $\lambda = 0.0226$, optimized by cross-validation using the R package 'glmnet').

HannumAge is developed by applying an elastic net regression to whole-blood DNAm profiles from 482 individuals aged 19–101 years. A total of 71 CpG sites are selected as the most informative predictors of chronological age in blood, and their weighted linear combination constitutes the HannumAge. Optimal regularization parameters are estimated via tenfold cross-validation in the R package 'glmnet'.

PhenoAge models a composite 'Phenotypic Age' metric, which is derived from clinical biomarkers predictive of mortality, as the regression target for an elastic net model trained on whole-blood DNAm data. The model is developed using data from 9,926 adults with complete biomarker profiles from the third National Health and Nutrition Examination Survey (NHANES-III). A total of 513 CpG sites are selected during training, and their weighted linear combination constitutes the PhenoAge estimator. The penalization parameter ($\lambda = 0.0192$) is selected using the R package 'glmnet'.

AltumAge is a deep learning-based pan-tissue epigenetic clock trained on 8,050 samples from 142 DNAm datasets covering diverse tissue types. The model is a multilayer neural network with 5 hidden layers (32 neurons each), SeLU activation, batch normalization, Gaussian dropout (rate of 0.1) and activity and kernel regularization coefficients of 0.0034. The model uses 20,318 CpGs and is optimized using MSE.

cAge is a blood-based predictor trained on 24,674 samples across multiple cohorts. An EWAS is first performed to identify age-associated CpG terms, including nonlinear effects. Elastic net regression ($\alpha = 0.5$) was then applied to the selected features using a leave-one-cohort-out cross-validation framework to enhance generalizability. The final model is trained using the R package 'biglasso', with the regularization parameter set to $\lambda = 0.0308$.

We also benchmark two Cox hazards models for disease risk prediction. Cheng et al.[15] develop a model for T2D risk prediction using Cox proportional hazards-based elastic net regression model with LASSO (Cox-LASSO), trained on the Generation Scotland cohort ($n = 18{,}414$). After removing low-variance features, 200,000 variable CpGs are retained, and 145 CpGs are selected in the final model. λ is optimized within the R package 'glmnet' via internal cross-validation.

Westerman et al.[16] construct CVD risk prediction model using Cox proportional hazards regression model trained separately on the Women's Health Initiative ($n = 2{,}023$), Framingham Heart Study ($n = 2{,}587$) and Lothian Birth Cohorts ($n = 818$) cohorts. In the elastic net regression model, α is initially set to 0.05, and the model's performance is stable across $\alpha = 0.05$–0.95. λ is optimized via fivefold cross-validation. A total of 100,000 CpGs are preselected according to median absolute deviation, and 1,275 features are retained in the final ensemble model.

In this study, HorvathAge, HannumAge and PhenoAge are performed by the package ENmix. AltumAge is downloaded via GitHub at https://github.com/rsinghlab/AltumAge, cAge is downloaded via GitHub at https://github.com/elenabernabeu/cage_bage and the CVD and T2D risk prediction tools based on Cox model are downloaded via GitHub at https://github.com/kwesterman/meth_cvd and https://github.com/marioni-group/episcores-diabetes-prediction, respectively.

Evaluation metrics

We use MAE and PCC to assess the performance of an epigenetic clock in the age prediction task. MAE is defined as the median of the absolute errors between predictive and true ages and is calculated as

$$\mathrm{MAE} = \mathrm{median}(\{|\hat{y}_i - y_i|\}_{i=1}^{N}),$$

where N is the number of test samples, and $\hat{y}_i$ and y_i are the predicted and the true age of the ith sample, respectively. PCC is a correlation coefficient that measures linear correlation between predictive and true ages and is calculated as

$$\mathrm{PCC} = \frac{\sum_{i=1}^{N}\left(\hat{y}_i - \bar{\hat{y}}\right)(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}\left(\hat{y}_i - \bar{\hat{y}}\right)^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}},$$

where N is the number of test samples, $\hat{y}_i$ and y_i are the predicted and the true age of the ith sample, respectively, and $\bar{\hat{y}}$ and $\bar{y}$ are the averaged predicted and the true age of the ith sample, respectively.

We utilize the AUROC and the AUPRC to measure how well CVD or T2D risk scores distinguish between positive and negative samples. Given that, in clinical contexts, the number of negative samples often far exceeds that of positive samples, we set the number of negative samples to be ten times that of positive samples for the calculation of both metrics. Three evaluation scenarios are designed: negative samples are drawn from the entire control population, the age-matched control population and the healthy population. These scenarios reflect the model's ability to identify disease or pre-disease states among the control group, age-matched individuals, and healthy participants, respectively. In this context, the healthy population consists of control samples from individuals aged under 30 with a body mass index of less than 25 kg m^{-2}. AUROC is calculated by plotting the true positive rate against the false positive rate at different thresholds, while AUPRC is computed by plotting the precision against the recall at different thresholds. A perfect classifier has an AUROC of 1 and an AUPRC of 1, whereas a random classifier yields an AUROC of 0.5 and an AUPRC of 0.09.

Due to the limited sample sizes in some disease cohorts, we adopt a bootstrapping-based approach to improve the reliability of model performance evaluation. This method is used to compute both AUROC and AUPRC. Specifically, for each classification task, we perform 1,000 rounds of bootstrapping with replacement from the combined set of positive and negative test samples. In each round, we calculate the AUC, and the final reported value is the mean of the 1,000 AUCs. The 2.5th and 97.5th percentiles of the resulting distribution are taken as the bounds of the 95% confidence interval. This approach allows us to quantify the uncertainty arising from small sample sizes and provides a more robust estimate of model performance.

To quantitatively evaluate the effectiveness of MAPLE in mitigating batch effects, we adopted PVCA. Specifically, we first applied PCA to both the raw DNAm data and the MAPLE-derived latent representations, retaining the top 20 principal components for downstream analysis.

For each principal component PC_i, we performed analysis of variance by fitting a linear model of the form

$$\mathrm{PC}_i = \beta_0 + \beta_1 \mathrm{Age} + \beta_2 \mathrm{Batch} + \tau,$$

where PC_i denotes the scores of the ith principal component, and Age and Batch are covariates representing biological and technical variation, respectively, β_0, β_1, and β_2 are the regression, coefficients, and τ is the random error term. The sum of squares for each factor, denoted as $\mathrm{SS}_{\mathrm{Batch},i}$ and $\mathrm{SS}_{\mathrm{Age},i}$, quantifies the proportion of variance in PC_i attributable to Age and Batch, respectively. $\mathrm{SS}_{\mathrm{Total},i}$ is the total sum of squares from the analysis of variance model for PC_i.

To obtain the final PVCA scores for Age and Batch, we aggregated the variance contributions across all 20 principal components using the proportion of variance each PC explains in the total PCA decomposition as weights. Formally, the PVCA score for Batch and Age are computed as

$$\mathrm{PVCA}_{\mathrm{Batch}} = \sum_{i=1}^{20} \omega_i \frac{\mathrm{SS}_{\mathrm{Batch},i}}{\mathrm{SS}_{\mathrm{Total},i}}$$

$$\mathrm{PVCA}_{\mathrm{Age}} = \sum_{i=1}^{20} \omega_i \frac{SS_{\mathrm{Age},i}}{SS_{\mathrm{Total},i}},$$

where ω_i is the proportion of variance explained by the ith principal component.

GO analysis for CpG sites

We first identify CpG sites most strongly associated with aging or disease risk. Gene annotations for these selected CpG sites were obtained from the annotation files of the 450K array and the EPIC array. Finally, we use ClusterProfiler[53] to conduct GO enrichment analysis for gene sets colocalized with the selected CpG sites and perform multiple testing correction using the Benjamini–Hochberg method with the significance level set at 0.05.

To identify informative CpG sites, we use an attribution method called IG[54] to quantify the contribution of each site to model predictions in both epigenetic age and disease risk tasks. For the epigenetic age prediction task, the IG score for the ith CpG site is computed as

$$\mathrm{IG}_i = \left(x_i - x_i'\right) \times \int_{\alpha=0}^{1} \frac{\partial F\left(x_i' + \alpha \times \left(x_i - x_i'\right)\right)}{\partial x_i} \mathrm{d}\alpha,$$

where $F(\bullet)$ denotes the epigenetic age prediction model (comprising the encoder and predictor), x_i is the DNAm level of the ith CpG site, $\frac{\partial F\left(x_i' + \alpha \times \left(x_i - x_i'\right)\right)}{\partial x_i}$ is the gradient of $F(\bullet)$ with respect to x_i, and x_i' is the baseline value, set to 0 unless otherwise specified. IG scores for disease risk prediction tasks are computed analogously, substituting $F(\bullet)$ with the disease prediction model.

We rank CpG sites by the absolute value of their IG scores and selected the top 2,000 features most strongly associated with age or disease risk.

To support the rationale for selecting the top 2,000 IG-ranked CpG sites for model interpretation, we conducted three complementary analyses examining both the attribution method and the number of CpG sites used.

First, we compute IG scores across four representative test datasets: two aging-related cohorts (GSE61496 and GSE87571), and the test sets for the CVD and T2D risk prediction task. The resulting IG score distributions demonstrate that the vast majority of CpG sites have attribution scores near zero, indicating that only a small fraction of CpGs are strongly associated with aging or disease risk (Supplementary Fig. 13). This suggests a natural sparsity in feature importance, thereby justifying the use of a thresholding strategy to select the most informative CpG sites.

Second, we evaluate how the number of selected CpG sites affects the biological interpretability of the model. Using the epigenetic age prediction task as an example, we selecte varying numbers of CpG sites (for instance, top 250, 500, 1,000, 2,000, 3,000 and 5,000) based on IG scores for two aging-related cohorts (GSE61496 and GSE87571). We then perform GO enrichment analysis on the genes corresponding to the overlapping CpGs between the two datasets. We found that selections of 1,000 or more CpG sites yield stable and biologically meaningful enrichment results, which remained consistent up to 5,000 sites (Supplementary Fig. 14). These results justify our use of 2,000 CpG sites as a reasonable choice that captures sufficient biological signal.

Third, to assess whether the choice of attribution method affects model interpretability, we compare IG with three alternative techniques: SHAP, DeepLIFT and Feature Ablation. For each method, we identify the top 2,000 CpG sites in the epigenetic age prediction task and perform GO enrichment analysis on the genes associated with the overlapping CpG sites. All four attribution methods consistently highlight aging-related biological processes, including cell adhesion, organ development and morphogenesis, cognition and cytokine production. (Supplementary Fig. 15). These results suggest that the outcomes of the model interpretability analysis are robust to the choice of attribution method.

Run time analysis

MAPLE is trained on an NVIDIA A100 graphics processing unit, requiring 12.92 h for the epigenetic age prediction task, 26.85 h for CVD risk prediction and 22.86 h for T2D risk prediction. While training MAPLE demands substantial computational resources, inference is highly efficient. For instance, MAPLE can process 1,000 DNAm profiles and generate predictions in approximately 3 min (Supplementary Fig. 16), making it well suited for real-time clinical decision-making scenarios.

Ethical implications

While MAPLE enables robust prediction of epigenetic age and disease risk based on DNAm data, its application also raises important ethical considerations, particularly related to privacy and potential misuse.

First, DNAm profiles may encode sensitive health, behavioral and environmental information, which introduces risks of re-identification and unintended disclosure. Such risks could lead to social stigma or discrimination if not properly managed. To mitigate these concerns, all datasets used in this study were fully de-identified and obtained under appropriate institutional approvals, with strict adherence to data governance policies. Furthermore, MAPLE is designed for efficient offline deployment and does not require the transmission of raw DNAm data to centralized servers, thereby minimizing exposure risk. Our use of DNAm data and the system design of MAPLE conform to major data protection frameworks, including the General Data Protection Regulation and the Health Insurance Portability and Accountability Act (HIPAA).

Second, we acknowledge the potential misuse of epigenetic risk models in contexts such as insurance, employment or unauthorized profiling. We strongly recommend that any deployment of MAPLE or similar tools adhere to relevant legal and ethical safeguards, such as the Genetic Information Nondiscrimination Act and HIPAA, which explicitly prohibit discrimination based on genetic or health-related information.

Finally, we believe that the development and application of methylation-based predictive models must be guided by core ethical principles of beneficence, autonomy and justice. As MAPLE advances toward translational and clinical applications, we advocate for transparent communication with participants, rigorous oversight by institutional review boards or ethics committees, and continuous evaluation of the broader societal impact. These measures are essential to ensure that predictive tools like MAPLE are used responsibly and equitably.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All datasets used in this study are downloaded from public databases. The EWAS Data Hub (https://ngdc.cncb.ac.cn/ewas/datahub/index) curates a large number of well-normalized DNAm array data and the corresponding metadata[55], and the control samples from 450 K platform in our study are mostly derived from this database. The LOLIPOP project provides 2711 DNAm array data (450K) of blood samples[56], and its accession number in GEO database is GSE55763. The Grady Trauma Project provides a dataset where 422 and 795 DNAm array data of blood samples are generated from the 450K and EPIC platforms[57], respectively, and their accession numbers in the GEO database are GSE72680 and GSE132203, respectively. A dataset including 732 DNAm array data (450K) of blood samples is from the study by Johansson et al.[58], and its accession number in the GEO database is GSE87571.

The Genetic Epidemiology Network of Arteriopathy (GENOA) project provides a dataset where 418 and 1394 DNAm array data of peripheral blood leukocytes are generated from the 450K and EPIC platforms[59], respectively, and their accession numbers in the GEO database are GSE210254 and GSE210255, respectively. The Center for International Blood and Marrow Transplant Research (CIBMTR) project has generated a dataset comprising 570 DNAm array samples (EPIC) from blood samples[60], with the associated accession number in the GEO database being GSE196696. The AIRWAVE project provides 1129 DNAm array data (EPIC) of peripheral blood mononuclear cells, and its accession number in the GEO database is GSE147740. The European network of national schizophrenia networks studying gene–environment interactions (EU-GEI) study has generated a dataset comprising 558 DNAm array samples (EPIC) from blood samples[60], with the associated accession number in the GEO database being GSE152026. The adipose datasets (N_{450K} = 157,) include two studies by Horvath et al.[61] and Bonder et al.[62] and can be downloaded from the GEO database with the accession numbers GSE61257, GSE61450 and GSE61453. The brain datasets (N_{450K} = 876) include two studies by Horvath et al.[63] and Jaffe et al.[64] and can be downloaded from the GEO database with the accession numbers GSE64509 and GSE74193. The buccal epithelium dataset (N_{450K} = 27) is from the study by Lussier et al.[65], and its accession number in the GEO database is GSE109042. The muscle datasets (N_{450K} = 132) include four studies by Zykovich et al.[66], Horvath et al.[61], Bonder et al.[62] and Tobi et al.[67] and can be downloaded from the GEO database with the accession numbers GSE50498, GSE61259, GSE61452 and GSE78743. The saliva dataset (N_{450K} = 131) is from the study by Chuang et al.[68], and its accession number in the GEO database is GSE111223. The skin dataset (N_{450K} = 15) is from the study by Tobi et al.[68], and its accession number in the GEO database is GSE78743. The liver datasets (N_{450K} = 177) include three studies by Horvath et al.[61], Tobi et al.[68] and Ahrens et al.[69] and can be downloaded from the GEO database with the accession numbers GSE61258, GSE78743 and GSE48325. The AS datasets are provided by two studies including the Multi-Ethnic Study of Atherosclerosis (MESA)[70] ($N_{450K,AS}$ = 863, $N_{450K,Control}$ = 339) and Progression of Early Subclinical Atherosclerosis (PESA)[71] ($N_{EPIC,AS}$ = 255, $N_{EPIC,Control}$ = 136) and can be downloaded from the GEO database with the accession numbers GSE56046 and GSE220622. The CAE dataset (N_{450K} = 11) is from the study by Lu et al.[72], and its accession number in the GEO database is GSE87016. The stroke datasets (N_{450K} = 678, N_{EPIC} = 32) include two studies by Cullell et al.[73] and Soriano-Tarraga et al.[74] and can be downloaded from the GEO database with the accession numbers GSE203399 and GSE69138. The SIR dataset (N_{450K} = 40) is from the study by Arner et al.[75], and its accession number in the GEO database is GSE76285. The PreD dataset ($N_{EPIC,PreD}$ = 47, $N_{EPIC,Control}$ = 46) is from the study by Yumi et al.[76], and its accession number in the GEO database is GSE199700. The T2D datasets (N_{450K} = 74) include two studies by Lunnon et al.[77] and Dye et al.[78] and can be downloaded from the GEO database with the accession numbers GSE62003 and GSE197881. The above datasets are selected to be either the training or the test datasets. For details about the training and the test datasets in the epigenetic age, CVD and T2D prediction tasks, refer to Supplementary Tables 1–3. Source data are provided with this paper. In addition, source data for Supplementary Figs. 1–9 and 12–13 are available in Supplementary Data 1.

Code availability

The MAPLE source code and analysis scripts are available via GitHub at https://github.com/Drizzle-Zhang/MAPLE, together with a usage documentation[79].

References

1. Partridge, L., Deelen, J. & Slagboom, P. E. Facing up to the global challenges of ageing. *Nature* **561**, 45–56 (2018).
2. Roth, G. S. et al. Biomarkers of caloric restriction may predict longevity in humans. *Science* **297**, 811 (2002).
3. Lu, Y. et al. Reprogramming to recover youthful epigenetic information and restore vision. *Nature* **588**, 124–129 (2020).
4. Lopez-Otin, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The hallmarks of aging. *Cell* **153**, 1194–1217 (2013).
5. Michalak, E. M., Burr, M. L., Bannister, A. J. & Dawson, M. A. The roles of DNA, RNA and histone methylation in ageing and cancer. *Nat. Rev. Mol. Cell Biol.* **20**, 573–589 (2019).
6. Hannum, G. et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol. Cell* **49**, 359–367 (2013).
7. McEwen, L. M. et al. Systematic evaluation of DNA methylation age estimation with common preprocessing methods and the Infinium MethylationEPIC BeadChip array. *Clin. Epigenet.* **10**, 123 (2018).
8. El Khoury, L. Y. et al. Systematic underestimation of the epigenetic clock and age acceleration in older subjects. *Genome Biol.* **20**, 283 (2019).
9. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol.* **14**, R115 (2013).
10. de Lima Camillo, L. P., Lapierre, L. R. & Singh, R. A pan-tissue DNA-methylation epigenetic clock based on deep learning. *npj Aging* **8**, 4 (2022).
11. Ahmad, E., Lim, S., Lamptey, R., Webb, D. R. & Davies, M. J. Type 2 diabetes. *Lancet* **400**, 1803–1820 (2022).
12. Lozano, R. et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* **380**, 2095–2128 (2012).
13. Lee, C. G. et al. Effect of metformin and lifestyle interventions on mortality in the Diabetes Prevention Program and Diabetes Prevention Program Outcomes Study. *Diabetes Care* **44**, 2775–2782 (2021).
14. Navar, A. M. et al. Earlier treatment in adults with high lifetime risk of cardiovascular diseases: what prevention trials are feasible and could change clinical practice? Report of a National Heart, Lung, and Blood Institute (NHLBI) Workshop. *Am. J. Prev. Cardiol.* **12**, 100430 (2022).
15. Cheng, Y. et al. Development and validation of DNA methylation scores in two European cohorts augment 10-year risk prediction of type 2 diabetes. *Nat. Aging* **3**, 450–458 (2023).
16. Westerman, K. et al. Epigenomic assessment of cardiovascular disease risk and interactions with traditional risk metrics. *J. Am. Heart Assoc.* **9**, e015299 (2020).
17. Li, Y., Chen, C. L. P. & Zhang, T. A survey on Siamese network: methodologies, applications, and opportunities. *IEEE Trans. Artif. Intell.* **3**, 994–1014 (2022).
18. Feng, B. et al. A bioactivity foundation model using pairwise meta-learning. *Nat. Mach. Intell.* **6**, 962–974 (2024).
19. Maksimovic, J., Gordon, L. & Oshlack, A. SWAN: subset-quantile within array normalization for Illumina Infinium HumanMethylation450 BeadChips. *Genome Biol.* **13**, R44 (2012).
20. Teschendorff, A. E. et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* **29**, 189–196 (2013).
21. Xiong, Z., Li, M., Ma, Y., Li, R. & Bao, Y. GMQN: a reference-based method for correcting batch effects and probe bias in HumanMethylation BeadChip. *Front. Genet.* **12**, 810985 (2021).
22. Levine, M. E. et al. An epigenetic biomarker of aging for lifespan and healthspan. *Aging* **10**, 573–591 (2018).
23. Bernabeu, E. et al. Refining epigenetic prediction of chronological and biological age. *Genome Med.* **15**, 12 (2023).
24. Zaghlool, S. B. et al. Association of DNA methylation with age, gender, and smoking in an Arab population. *Clin. Epigenet.* **7**, 6 (2015).

25. Li, S. et al. Identification, replication and characterization of epigenetic remodelling in the aging genome: a cross population analysis. *Sci. Rep.* **7**, 8183 (2017).
26. Li, M. et al. EWAS Atlas: a curated knowledgebase of epigenome-wide association studies. *Nucleic Acids Res.* **47**, D983–D988 (2019).
27. Shin, E. Y. et al. Integrin-mediated adhesions in regulation of cellular senescence. *Sci. Adv.* **6**, eaay3909 (2020).
28. Morange, M. Development and aging. *Biol. Theory* **6**, 59–64 (2011).
29. Lu, A. T. et al. Universal DNA methylation age across mammalian tissues. *Nat. Aging* **3**, 1144–1166 (2023).
30. Wyss-Coray, T. A. geing neurodegeneration and brain rejuvenation. *Nature* **539**, 180–186 (2016).
31. Li, X. et al. Inflammation and aging: signaling pathways and intervention therapies. *Signal Transduct. Target Ther.* **8**, 239 (2023).
32. Levine, M. E. et al. Menopause accelerates biological aging. *Proc. Natl Acad. Sci. USA* **113**, 9327–9332 (2016).
33. Li, J. et al. Determining a multimodal aging clock in a cohort of Chinese women. *Med* **4**, 825–848.e813 (2023).
34. Kozlov, G., Franceschi, C. & Vedunova, M. Intricacies of aging and Down syndrome. *Neurosci. Biobehav Rev.* **164**, 105794 (2024).
35. Rodes, B., Cadinanos, J., Esteban-Cantos, A., Rodriguez-Centeno, J. & Arribas, J. R. Ageing with HIV: challenges and biomarkers. *EBioMedicine* **77**, 103896 (2022).
36. Cao, X. et al. Contribution of life course circumstances to the acceleration of phenotypic and functional aging: a retrospective study. *EClinicalMedicine* **51**, 101548 (2022).
37. Santos, A. L. & Sinha, S. Obesity and aging: molecular mechanisms and therapeutic approaches. *Ageing Res Rev.* **67**, 101268 (2021).
38. Mozaffarian, D. et al. Heart disease and stroke statistics—2016 update. *Circulation* **133**, e38–e360 (2016).
39. Libby, P. et al. Atherosclerosis. *Nat. Rev. Dis. Prim.* **5**, 56 (2019).
40. Richards, G. H. C., Hong, K. L., Henein, M. Y., Hanratty, C. & Boles, U. Coronary artery ectasia: review of the non-atherosclerotic molecular and pathophysiologic concepts. *Int. J. Mol. Sci.* **23**, 5195 (2022).
41. Campbell, B. C. V. & Khatri, P. Stroke. *Lancet* **396**, 129–142 (2020).
42. Hillis, G. S. & Flapan, A. D. Cell adhesion molecules in cardiovascular disease: a clinical perspective. *Heart* **79**, 429–431 (1998).
43. Imig, J. D. & Hammock, B. D. Soluble epoxide hydrolase as a therapeutic target for cardiovascular diseases. *Nat. Rev. Drug Discov.* **8**, 794–805 (2009).
44. Nahrendorf, M. Myeloid cell contributions to cardiovascular health and disease. *Nat. Med.* **24**, 711–720 (2018).
45. Simons, K. H. et al. T cell co-stimulation and co-inhibition in cardiovascular disease: a double-edged sword. *Nat. Rev. Cardiol.* **16**, 325–343 (2019).
46. Qiu, S. et al. Association between circulating cell adhesion molecules and risk of type 2 diabetes: a meta-analysis. *Atherosclerosis* **287**, 147–154 (2019).
47. Yousefi, P. D. et al. DNA methylation-based predictors of health: applications and statistical considerations. *Nat. Rev. Genet.* **23**, 369–383 (2022).
48. Jin, J. et al. iDNA-ABF: multi-scale deep biological language learning model for the interpretable prediction of DNA methylations. *Genome Biol.* **23**, 219 (2022).
49. Xie, H., Wang, L., Qian, Y., Ding, Y. & Guo, F. Methyl-GP: accurate generic DNA methylation prediction based on a language model and representation learning. *Nucleic Acids Res.* **53**, gkaf223 (2025).
50. Ying, K. et al. MethylGPT: a foundation model for the DNA methylome. Preprint at *bioRxiv* https://doi.org/10.1101/2024.10.30.621013 (2024).
51. Peter, J. H. Robust estimation of a location parameter. *Ann. Math. Stat.* **35**, 73–101 (1964).
52. Schölkopf, B., Smola, A. J., Williamson, R. C. & Bartlett, P. L. New support vector algorithms. *Neural Comput.* **12**, 1207–1245 (2000).
53. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).
54. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. In *Proc. 34th International Conference on Machine Learning* (eds Precup, D. & Teh, Y. W.) 3319–3328 (PMLR, 2017).
55. Xiong, Z. et al. EWAS Open Platform: integrated data, knowledge and toolkit for epigenome-wide association study. *Nucleic Acids Res.* **50**, D1004–D1009 (2022).
56. Wahl, S. et al. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature* **541**, 81–86 (2017).
57. Zannas, A. S. et al. Epigenetic upregulation of FKBP5 by aging and stress contributes to NF-kappaB-driven inflammation and cardiovascular risk. *Proc. Natl Acad. Sci. USA* **116**, 11370–11379 (2019).
58. Johansson, A., Enroth, S. & Gyllensten, U. Continuous aging of the human DNA methylome throughout the human lifespan. *PLoS ONE* **8**, e67378 (2013).
59. Shang, L. et al. meQTL mapping in the GENOA study reveals genetic determinants of DNA methylation in African Americans. *Nat. Commun.* **14**, 2711 (2023).
60. Webster, A. P. et al. Donor whole blood DNA methylation is not a strong predictor of acute graft versus host disease in unrelated donor allogeneic haematopoietic cell transplantation. *Front. Genet.* **15**, 1242636 (2024).
61. Horvath, S. et al. Obesity accelerates epigenetic aging of human liver. *Proc. Natl Acad. Sci. USA* **111**, 15538–15543 (2014).
62. Bonder, M. J. et al. Genetic and epigenetic regulation of gene expression in fetal and adult human livers. *BMC Genomics* **15**, 860 (2014).
63. Horvath, S. et al. The cerebellum ages slowly according to the epigenetic clock. *Aging* **7**, 294–306 (2015).
64. Jaffe, A. E. et al. Mapping DNA methylation across development, genotype and schizophrenia in the human frontal cortex. *Nat. Neurosci.* **19**, 40–47 (2016).
65. Lussier, A. A. et al. DNA methylation as a predictor of fetal alcohol spectrum disorder. *Clin. Epigenet.* **10**, 5 (2018).
66. Zykovich, A. et al. Genome-wide DNA methylation changes with age in disease-free human skeletal muscle. *Aging Cell* **13**, 360–366 (2014).
67. Tobi, E. W. et al. DNA methylation as a mediator of the association between prenatal adversity and risk factors for metabolic disease in adulthood. *Sci. Adv.* **4**, eaao4364 (2018).
68. Chuang, Y. H. et al. Parkinson's disease is associated with DNA methylation levels in human blood and saliva. *Genome Med.* **9**, 76 (2017).
69. Ahrens, M. et al. DNA methylation analysis in nonalcoholic fatty liver disease suggests distinct disease-specific and remodeling signatures after bariatric surgery. *Cell Metab.* **18**, 296–302 (2013).
70. Reynolds, L. M. et al. Age-related variations in the methylome associated with gene expression in human monocytes and T cells. *Nat. Commun.* **5**, 5366 (2014).
71. Sanchez-Cabo, F. et al. Subclinical atherosclerosis and accelerated epigenetic age mediated by inflammation: a multi-omics study. *Eur. Heart J.* **44**, 2698–2709 (2023).
72. Lu, T. P. et al. Genome-wide methylation profiles in coronary artery ectasia. *Clin. Sci.* **131**, 583–594 (2017).
73. Cullell, N. et al. DNA methylation and ischemic stroke risk: an epigenome-wide association study. *Thromb. Haemost.* **122**, 1767–1778 (2022).

74. Soriano-Tarraga, C. et al. Epigenome-wide association study identifies TXNIP gene associated with type 2 diabetes mellitus and sustained hyperglycemia. *Hum. Mol. Genet* **25**, 609–619 (2016).
75. Arner, P. et al. The epigenetic signature of systemic insulin resistance in obese women. *Diabetologia* **59**, 2393–2405 (2016).
76. Yumi Noronha, N. et al. 14-weeks combined exercise epigenetically modulated 118 genes of menopausal women with prediabetes. *Front. Endocrinol.* **13**, 895489 (2022).
77. Lunnon, K. et al. Blood methylomic signatures of presymptomatic dementia in elderly subjects with type 2 diabetes mellitus. *Neurobiol. Aging* **36**, 1600 e1601–1600 e1604 (2015).
78. Dye, C. K. et al. Shifts in the immunoepigenomic landscape of monocytes in response to a diabetes-specific social support intervention: a pilot study among Native Hawaiian adults with diabetes. *Clin. Epigenet.***14**, 91 (2022).
79. Zhang, Y. MAPLE code. *Zenodo* https://doi.org/10.5281/zenodo.17637977 (2025).

Acknowledgements

We gratefully acknowledge the following funding sources: National Natural Science Foundation of China (grant nos. 82394432 and 92249302 to Y.Q.), Shanghai Municipal Science and Technology Major Project (grant no. 2023SHZDZX02 to L.J.). The computations in this research were performed using the Computing for the Future at Fudan (CFFF) platform of Fudan University.

Author contributions

Conceptualization: L.J., Y.Q., Y.H., Y.X., Y.C. and Y.Z.; design of methodology: Y.H., Y.X., Y.Z. and Y.Y.; implementation of methodology: Y.Z. and Y.Y; benchmark analysis: Y.Z. and Y.T.; writing: Y.Z., Y.H., Y.C. and Y.Y.; supervision: Y.H., Y.X., Y.Q. and L.J.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43588-025-00939-x.

Correspondence and requests for materials should be addressed to Yinghui Xu, Ying He, Yuan Qi or Li Jin.

Peer review information *Nature Computational Science* thanks Qi Liu and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: Ananya Rastogi and Michelle Badri, in collaboration with the *Nature Computational Science* team. Peer reviewer reports are available.

nature portfolio

Corresponding author(s): Yuan Qi

Last updated by author(s): Nov 18, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software or code was involved in data collection.
Data analysis	MAPLE is available on GitHub (https://github.com/Drizzle-Zhang/MAPLE). Software: Python (3.10.6), R (4.3.2) Python packages: Pytorch (2.2.1), Captum (0.7.0) R packages: clusterProfiler (4.7.1.003), GMQN (0.1.0), minfi (1.48.0), Cox-LASSO model for T2D risk (https://github.com/marioni-group/episcores-diabetes-prediction), Cox-LASSO model for CVD risk (https://github.com/kwest□erman/□meth_cvd)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All datasets used in this study are downloaded from public databases and are as follows:
• The EWAS Data Hub (https://ngdc.cncb.ac.cn/ewas/datahub/index) curates a large number of well-normalized DNAm array data and the corresponding metadata55, and the control samples from 450K platform in our study are mostly derived from this database.
• The LOLIPOP project provides 2711 DNAm array data (450K) of blood samples56, and its accession number in GEO database is GSE55763.
• The Grady Trauma Project provides a dataset where 422 and 795 DNAm array data of blood samples are generated from 450K and EPIC platform57, respectively, and their accession number in GEO database are GSE72680 and GSE132203, respectively.
• A dataset including 732 DNAm array data (450K) of blood samples is from the study by Johansson et al.58, and its accession number in GEO database is GSE87571.
• The Genetic Epidemiology Network of Arteriopathy (GENOA) project provides a dataset where 418 and 1394 DNAm array data of peripheral blood leukocytes are generated from 450K and EPIC plarform59, respectively, and their accession number in GEO database are GSE210254 and GSE210255, respectively.
• The Center for International Blood and Marrow Transplant Research (CIBMTR) project has generated a dataset comprising 570 DNAm array samples (EPIC) from blood samples60, with the associated accession number in the GEO database being GSE196696.
• The AIRWAVE project provides 1129 DNAm array data (EPIC) of peripheral blood mononuclear cells, and its accession number in GEO database is GSE147740.
• The European network of national schizophrenia networks studying gene-environment interactions (EU-GEI) study has generated a dataset comprising 558 DNAm array samples (EPIC) from blood samples60, with the associated accession number in the GEO database being GSE152026.
• The adipose datasets (N450K =157,) include two studies by Horvath et al.61 and Bonder et al.62, and can be downloaded from GEO database with the accession number GSE61257, GSE61450, GSE61453.
• The brain datasets (N450K =876) include two studies by Horvath et al.63 and Jaffe et al.64, and can be downloaded from GEO database with the accession number GSE64509 and GSE74193.
• The buccal epithelium dataset (N450K =27) is from the study by Lussier et al.65, and its accession number in GEO database is GSE109042.
• The muscle datasets (N450K =132) include four studies by Zykovich et al.66, Horvath et al.61, Bonder et al.62 and Tobi et al.67, and can be downloaded from GEO database with the accession number GSE50498, GSE61259, GSE61452, GSE78743.
• The saliva dataset (N450K =131) is from the study by Chuang et al.68, and its accession number in GEO database is GSE111223.
• The skin dataset (N450K =15) is from the study by Tobi et al.68, and its accession number in GEO database is GSE78743.
• The liver datasets (N450K =177) include three studies by Horvath et al.61, Tobi et al.68 and Ahrens et al.69, and can be downloaded from GEO database with the accession number GSE61258, GSE78743, GSE48325.
• The atherosclerosis datasets are provided by two studies including the Multi-Ethnic Study of Atherosclerosis (MESA)70 (N450K,AS = 863, N450K,Control= 339) and Progression of Early Subclinical Atherosclerosis (PESA)71 (NEPIC, AS = 255, NEPIC, Control= 136), and can be downloaded from GEO database with the accession number GSE56046 and GSE220622.
• The coronary artery ectasia (CAE) dataset (N450K =11) is from the study by Lu et al.72, and its accession number in GEO database is GSE87016.
• The stroke datasets (N450K =678, NEPIC =32) include two studies by Cullell et al.73 and Soriano-Tarraga et al.74, and can be downloaded from GEO database with the accession number GSE203399 and GSE69138.
• The systemic insulin resistance (SIR) dataset (N450K =40) is from the study by Arner et al.75, and its accession number in GEO database is GSE76285.
• The prediabetes dataset (NEPIC, PreD = 47, NEPIC, Control= 46) is from the study by Yumi et al.76, and its accession number in GEO database is GSE199700.
• The type 2 diabetes (T2D) datasets (N450K = 74) include two studies by Lunnon et al.77 and Dye et al.78, and can be downloaded from GEO database with the accession number GSE62003 and GSE197881.
The above datasets are selected to be either the training or the test datasets. For details about the training and the test datasets in the epigenetic age, CVD and T2D prediction tasks, refer to Supplementary Table 1-3.
Source data for Figures 2-5 is available with this manuscript. In addition, source data for Supplementary Figures is available in Supplementary Data 1.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	We took sex into condiderations in our study and our model could be applied to both male and female. The gender-related metadata associated with the publicly accessible datasets employed in this research is fully traceable via their respective source repositories.
Reporting on race, ethnicity, or other socially relevant groupings	All socially relevant information used in this study could be obtained from their respective source repositories.
Population characteristics	See above
Recruitment	All datasets used in this study were obtained from publicly accessible repositories. Participant recruitment procedures were performed by the original studies and are fully described in their respective publications and associated metadata.
Ethics oversight	All open-access datasets referenced in this study have obtained documented institutional review board certifications through established ethical compliance protocols at their source institutions.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to calculate the sample size. In the epigenetic age prediction task, the training set contains 11313 samples and the test set contains 7749 samples. In the CVD risk prediction task, the training set contains 7234 samples and the test set contains 7926 samples. In the T2D risk prediction task, the training set contains 6728 samples and the test set contains 6009 samples.
Data exclusions	No data were excluded from analyses.
Replication	This study involves computational analyses only. All analyses were performed using publicly available datasets and fully documented code, ensuring full reproducibility.
Randomization	Samples were randomly arranged by computing program during the model training stage.
Blinding	Each sample was labeled with an numeric ID whose annotation was kept blinded during data analysis.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*