Research Paper

Precious1GPT: multimodal transformer-based transfer learning for aging clock development and feature importance analysis for aging and age-related disease target discovery

Anatoly Urban[1], Denis Sidorenko[1], Diana Zagirova[1], Ekaterina Kozlova[1], Aleksandr Kalashnikov[2], Stefan Pushkov[1], Vladimir Naumov[1], Viktoria Sarkisova[1], Geoffrey Ho Duen Leung[1], Hoi Wing Leung[1], Frank W. Pun[1], Ivan V. Ozerov[1], Alex Aliper[1,2], Feng Ren[3], Alex Zhavoronkov[1,2]

[1]Insilico Medicine, Pak Shek Kok, New Territories, Hong Kong
[2]Insilico Medicine, Masdar City, United Arab Emirates
[3]Insilico Medicine, Shanghai, China

Correspondence to: Alex Zhavoronkov; **email:** alex@insilico.com
Keywords: transformers, deep learning, therapeutic target discovery, aging biomarkers, human aging
Received: April 21, 2023 **Accepted:** May 24, 2023 **Published:** June 13, 2023

ABSTRACT

Aging is a complex and multifactorial process that increases the risk of various age-related diseases and there are many aging clocks that can accurately predict chronological age, mortality, and health status. These clocks are disconnected and are rarely fit for therapeutic target discovery. In this study, we propose a novel approach to multimodal aging clock we call Precious1GPT utilizing methylation and transcriptomic data for interpretable age prediction and target discovery developed using a transformer-based model and transfer learning for case-control classification. While the accuracy of the multimodal transformer is lower within each individual data type compared to the state of art specialized aging clocks based on methylation or transcriptomic data separately it may have higher practical utility for target discovery. This method provides the ability to discover novel therapeutic targets that hypothetically may be able to reverse or accelerate biological age providing a pathway for therapeutic drug discovery and validation using the aging clock. In addition, we provide a list of promising targets annotated using the PandaOmics industrial target discovery platform.

INTRODUCTION

Aging is a complex, multifactorial process that results from a multitude of interacting biological mechanisms occurring at different levels within an organism [1]. The development of accurate, physiologically meaningful biomarkers of aging is crucial for assessing the efficacy of potential anti-aging therapies and advancing the field of aging research [2, 3]. Deep neural networks (DNNs) have demonstrated remarkable success in various applications, including biomedical research [4, 5]. Population-specific aging clocks have been developed using large datasets from diverse ethnic groups, enabling more accurate predictions of chronological age and biological age, as well as assessment of all-cause mortality [6]. Moreover, artificial intelligence (AI)-driven platforms, such as PandaOmics, have facilitated the identification and prioritization of novel aging-associated targets for drug discovery and repurposing [7]. Recent studies have also demonstrated the value of AI in advancing longevity research by harnessing the power of next-generation sequencing data and omics technologies [8].

Insilico Medicine has been at the forefront of using generative AI in biology since 2016 [9–11]. Their research has led to the development of generative biology approaches that utilize generative systems to generate synthetic biological data, including their first successful demonstration taking place at the National Institute of Aging [12]. In addition to target discovery, Insilico has also developed capabilities in generative chemistry [4, 13–15]. These approaches have been successfully applied to various diseases and aging and have shown potential in identifying novel compounds and accelerating drug development [16–18]. As the aging population continues to grow, there is an urgent need for new therapeutic targets to delay and treat age-related diseases. Therefore, the application of generative biology approaches in exploring the complex interplay between aging and diseases holds great promise for identifying potential novel targets and accelerating drug development efforts.

Deep aging clocks have been developed for various applications in pharmaceutical research and development. For example, DeepMAge, a methylation aging clock developed using deep learning, shows remarkable accuracy and biological relevance in predicting human age and identifying health-related conditions [19]. Moreover, deep aging clocks could potentially be used for target identification, drug discovery, data quality control, and synthetic patient data generation [3]. Additionally, the use of AI to comprehend the intricate interplay between the microenvironment within the human body and the external environment has shown promise in revealing the role of external factors in aging [8]. The integration of deep learning techniques with genomics and other omics data has enabled comprehensive comparisons of DNA repair transcriptomes in species with extreme lifespan differences, shedding light on the potential role of DNA repair as a longevity assurance system [20].

Despite the progress made in developing deep aging clocks and AI-based biomarkers, there are still several challenges and opportunities for improvement in the field of biohorology [16]. The development of deep learning, DNNs, and generative approaches is expected to significantly advance the field, leading to more accurate and robust aging biomarkers [16]. Furthermore, *in silico* methods for screening and ranking potential geroprotective candidates, based on their ability to regulate age-related changes in signaling pathway clouds, hold promise for accelerating the discovery of effective interventions and reducing the time and cost of pre-clinical work and clinical trials [21]. For instance, GeroScope could predict novel geroprotectors from existing human gene expression data by mapping expression differences between young and old subjects to age-related signaling pathways and ranking known substances (potential geroprotector candidates) based on their likelihood to target differential pathways and mimic the young signalome [22]. Similarly, the human gut microbiome has been shown to have a strong association with host age, and deep learning-based models have been developed to predict host age based on gut microflora taxonomic profiles, further providing insights into potential aging biomarkers [17].

To identify aging biomarkers associated with age-related diseases, in the present work, we combined the ability of aging clocks to predict biological age and thus grasp molecular changes accompanied by senescence and our target ID approach to establish genes that are related to the development of diseases. This provides us with a novel perspective on uncovering the molecular mechanisms of diseases in the context of aging, allowing us to identify promising strategies to delay and treat age-related diseases.

RESULTS

Performance of the transformer-based multimodal aging clock

In the current study, we have developed a comprehensive pipeline, as illustrated in Figure 1. The pipeline consists of several key steps, including training a multimodal transformer-based regressor on normal sample data for age prediction and subsequently using the learned weights to fine-tune a transformer-based classifier for distinguishing between case and control samples. Next, we perform gene prioritization by employing the feature importance values obtained from the regressor to rank genes based on their relevance to aging and utilizing the importance values from the classifier to rank genes in terms of their relevance to both aging and disease. Finally, we analyze the resulting gene lists using the PandaOmics TargetID Platform to gain insights into potential targets for age-related diseases. In this study, we implemented a transformer-based architecture to accommodate both numerical and categorical data as input for our predictive model. This strategy, which we call Precious1GPT, enables the construction of multimodal classifiers and regressors that can effectively process diverse data types, such as RNA-seq expression data and epigenomics methylation data taking into account data type and tissue type. Consequently, our model demonstrates proficiency in age prediction and case-control classification, showcasing its versatility in handling multifaceted inputs.

We employed Optuna [23], a hyperparameter optimizer, to optimize the parameters for each model. We optimized L1 and L2 regularization, the activation function, dropout value, batch size, the number of neurons in hidden layers, and the gradient update used. The final regression metrics for the optimized models are shown in Table 1, calculated for the epigenetic and expression sub-datasets and for the whole dataset, respectively. The metrics were calculated by splitting the data to train (80%) and test (20%) set with stratification by sample tissue. For five-fold cross-validation stratified by tissue and data modality, results are shown in Supplementary Table 1. Metrics for individual tissues are shown in Supplementary Table 2. The methylation data subset, expression data subset, and all the test data combined were also calculated for each metric. Learning curves depicting MAE during training are shown in Supplementary Figure 1.

Important genes for age prediction

SHapley Additive exPlanations (SHAP) values [24] are a technique for explaining the output of machine learning models. From the regressor model, we obtained a list of features ranked by their SHAP values representing their importance for age prediction (Supplementary Table 3). Pathway enrichment analysis was subsequently performed for the top-100 genes ranked based on the SHAP values, which showed that these genes are implicated in multiple pathways associated with aging and age-related diseases (Table 2).

Identification of potential targets for age-related diseases through feature importance analysis

Utilizing the feature importance analysis based on SHAP values, we generated a list of genes associated with aging (Supplementary Table 3). We then compared these genes with known drug targets in our in-house database to identify potential therapeutic interventions for 4 selected age-related diseases, namely idiopathic pulmonary fibrosis (Supplementary Table 4), chronic obstructive pulmonary disease (COPD) (Supplementary Table 5), Parkinson's disease (PD) (Supplementary Table 6) and heart failure (Supplementary Table 7). Evaluation metrics for case-control classifiers are shown in Supplementary Table 8.

We adopted a transfer learning approach to identify genes involved in disease development in the context of aging. We first trained a DL-model as a regressor to predict age using an age dataset. Subsequently, we fine-tuned the model by re-training it as a case-control classifier while keeping the previously learned weights frozen, with the exception of the last layer. The SHAP values generated from this analysis were then used to determine the relative importance of molecular features in driving disease development in the context of aging. This allowed us to identify specific genes that are involved in disease development in the context of aging and determine their relative importance. To establish a baseline, we trained the same classifiers on the complete feature set. For a number of diseases, we

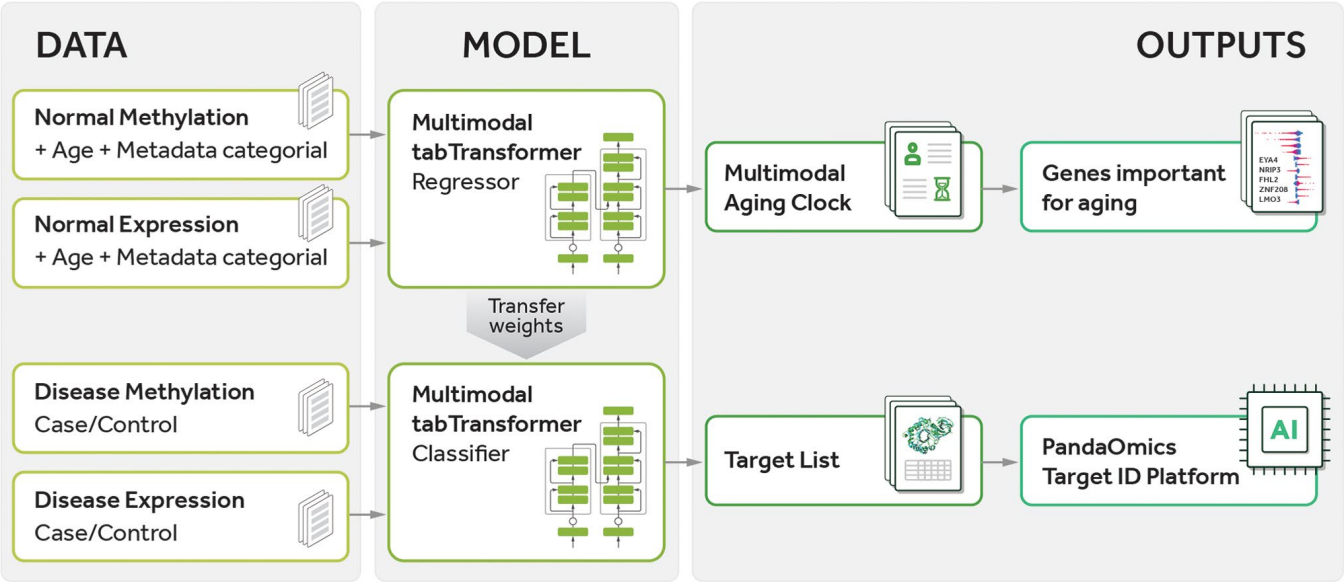


Figure 1. Pipeline of the current study. The pipeline involves training a multimodal transformer-based regressor on normal sample data to predict age, followed by transferring the learned weights to a transformer-based classifier for distinguishing between case and control samples. Gene prioritization is then performed using feature importance values obtained from the regressor to rank genes according to their relevance to aging and using importance values from the classifier to rank genes according to their relevance to both aging and disease. Finally, the gene lists are analyzed using the PandaOmics TargetID Platform.

Table 1. Multimodal transformer-based regressor metrics were evaluated on the hold-out test dataset (20% of all data).

Metric	Methylation data	Expression data	Combined
MAE	4.227	6.287	5.622
RMSE	6.129	8.155	7.560
R^2	0.934	0.584	0.807
MdAE	2.880	5.098	4.336
Number of samples in test set	4,019	2,730	6,749

Table 2. Reactome pathway analysis results for the top-100 genes selected based on the SHAP values.

Pathway	*P*-value	Odds ratio	Combined score
tRNA Processing in Mitochondrion R-HSA-6785470	0.036	33.99	112.78
Amino Acid Transport Across Plasma Membrane R-HSA-352230	0.002	13.77	86.96
Suppression Of Apoptosis R-HSA-9635465	0.043	27.19	85.36
Vasopressin-like Receptors R-HSA-388479	0.043	27.19	85.36
Highly Sodium Permeable Postsynaptic Acetylcholine Nicotinic Receptors R-HSA-629587	0.050	22.66	67.72
Cytosolic Sulfonation of Small Molecules R-HSA-156584	0.011	13.68	61.37

observed a slight but significant increase in classification metrics (Supplementary Table 8). These results indicated that the last layer of the neural network, which was trained to predict biological age, contains sufficient information to differentiate between case and control.

To validate the performance of our model and to establish its ability to accurately estimate age based on methylation data, we acquired two methylation datasets from the Gene Expression Omnibus (GEO) repository - one on cells that were reprogrammed to induced pluripotent stem cells (iPSC) (dataset GSE54848) and the other on fibroblasts of the developing fetuses (dataset GSE76641). The selection of the independent dataset on the developing cells, along with reverse aging cells data, allows us to provide another level of validation evidence confirming the potency of the trained model. These datasets were processed using the same methods as our primary dataset on age-related methylation, and we used the model to predict age. In order to increase the robustness of our validation, we used the same processing methods for both the iPSC and fetus datasets as we had for our primary dataset. This ensured consistency and minimized the possibility of any discrepancies or variations in our results due to differences in processing methods. The results were consistent with expectations, as iPSCs become younger during induction (Figure 2, Left) and fetal tissue becomes older during development (Figure 2, Right).

Manual analysis of the resulting targets

Utilizing a transfer learning approach, we have built aging-aware case-control classifiers and extracted feature importance values from them. The lists of top-200 genes ranked by expression classifiers were retrieved (Supplementary Tables 4–7) and considered as a starting point for further target identification and prioritization techniques offered by the AI-powered PandaOmics platform to propose a list of promising novel targets for age-related diseases. According to PandaOmics TargetID platform, APLNR was ranked top-20 for all 4 diseases, while IL23R was ranked top-20 for COPD, PD, and heart failure (Figure 3, Supplementary Figures 2–4). APLNR and IL23R were therefore selected as the most promising targets for treating multiple age-related diseases. In general, APLNR, a receptor for Apelin and Elabela peptide ligands, is involved in regulating several important physiological processes, including cardiovascular function, fluid balance, and metabolism. IL23R is a receptor for the pro-inflammatory cytokine IL-23 and is associated with chronic inflammation, which is considered to be one of the hallmarks of aging [25]. Accumulating evidence demonstrated that the effectiveness of our approach in addressing various age-related diseases, as documented in existing literature, provides further validation for our approach. Therefore, our unique approach with the amplification of PandaOmics allows us to identify various potential targets associated with essential aging-driven tissue

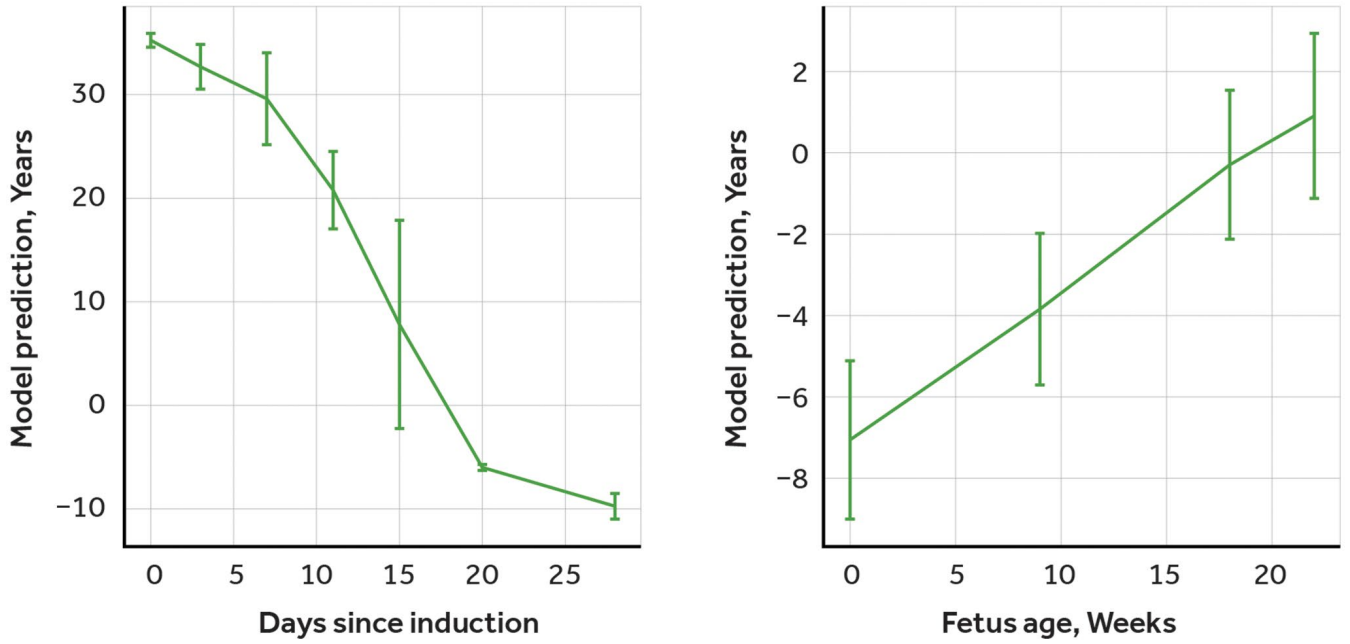


Figure 2. Validation of age-predictor model using induced pluripotent stem cells (iPSC) and fetal tissues methylation data. Left: Predictions of the multimodal transformer for iPSC induction dataset, days after transfection with reprogramming factors. Right: Predictions of the multimodal transformer for embryonic tissue dataset, weeks after last menstruation, averaged across tissues.

Figure 3. Example of Target ID output for chronic obstructive pulmonary disease. Top-200 genes from expression classifiers were applied as a gene list in PandaOmics corresponding project for COPD, and a filter for small molecules was applied to identify druggable targets. Twenty genes highly ranked by PandaOmics are shown.

dysfunction, which may be useful for the delay and treatment of multiple age-related diseases.

DISCUSSION

The development of "aging clocks," based on machine learning models that predict age based on biological data, has become a major milestone in aging research. Nevertheless, such an approach has severe limitations, such as a lack of ability to explain biological processes accompanied by aging and, thus, the ability to propose therapeutic interventions to compensate for age-related deterioration [26]. Additionally, aging clocks can be used to monitor the effectiveness of interventions and therapies designed to target age-related diseases [27]. For example, if an intervention is able to slow down the aging process, as measured by the aging clock, it may be more likely to be effective in delaying or treating age-related diseases. This can be done by comparing the aging clock values of an individual before and after the intervention and measuring the change in the aging clock value, which may indicate the effectiveness of the intervention. The development of Precious1GPT, a multimodal aging clock using a transformer-based model and transfer learning for case-control classification, as well as the identification of potential therapeutic targets for age-related diseases through feature importance analysis, has demonstrated the potential of our approach in deciphering the molecular mechanisms of aging. The transformer-based model allowed for the integration of multi-omics data and improved the accuracy of the aging clock, while the transfer learning approach facilitated the identification of disease-related genes in the context of aging. However, our study has several limitations, including the reliance on publicly available datasets, which may contain noisy and low-quality data. Future research should focus on validating the identified targets using experimental methods and exploring the potential of new drug targets.

Several aging clocks have been proposed in the literature, each with its own strengths and limitations. Some of the most prominent aging clocks include the Epigenetic Clock, DNAm PhenoAge, and the transcriptomic-based Aging.AI clock [28–31]. These clocks utilize various molecular markers, such as DNA methylation or gene expression patterns, to predict an individual's chronological age. Our proposed multimodal aging clock uses a transformer-based model and transfers learning to integrate diverse data sources, including epigenetic and transcriptomic data, and to predict age with high accuracy. When compared to existing aging clocks, our approach demonstrates several advantages. First, the multimodal nature of our approach enables the integration of different omics data types, leading to a more comprehensive and accurate assessment of an individual's biological age. By incorporating multiple data types, our aging clock can capture a wider range of molecular changes associated with aging, leading to a more reliable and informative model. Second, the use of transformer-based deep learning models allows our approach to capturing complex relationships between features, which can lead to improved age prediction accuracy. In contrast, traditional aging clocks like the Epigenetic Clock and DNAm PhenoAge rely on linear regression models, which may not be able to fully capture the complexity of age-related molecular changes. Third, our approach employs transfer learning for case-control classification, enabling the identification of potential targets for age-related diseases. This aspect of our method offers a significant advantage of Precious1GPT over existing aging clocks, as it not only allows for accurate age prediction but also contributes to the discovery of novel therapeutic targets for age-related diseases.

Unexpectedly, our model could not identify the genes which play key roles in known age-related pathways (SIRT, mTOR, and AMPK) as important genes (i.e., top-200) for age prediction. However, such a phenomenon was also observed in published DL age prediction models [32]. To further test if the most important genes (i.e., top-100) share any biological features, we performed pathway enrichment analysis which revealed that the list of identified genes is significantly enriched in the following pathways that are associated with the aging process:

1. tRNA Processing in Mitochondrion (R-HSA-6785470): This pathway is involved in the processing of transfer RNAs (tRNAs) within the mitochondria, which is essential for proper mitochondrial protein synthesis and overall mitochondrial function. Mitochondrial dysfunction has been implicated in the aging process and age-related diseases, such as neurodegenerative disorders and metabolic syndromes [33].

2. Amino Acid Transport Across Plasma Membrane (R-HSA-352230): This pathway describes the transport of amino acids across the plasma membrane, a crucial process for maintaining cellular homeostasis and protein synthesis. Dysregulation of amino acid transport may lead to imbalances in protein synthesis and degradation, which could contribute to cellular senescence, a hallmark of aging [25].

3. Suppression of Apoptosis (R-HSA-9635465): This pathway is involved in the regulation of apoptosis, a crucial cellular process that controls cell death and tissue homeostasis. Dysregulation of apoptosis has been linked to aging and age-related diseases such as cancer,

neurodegenerative disorders, and cardiovascular diseases [34].

4. Vasopressin-like Receptors (R-HSA-388479): This pathway focuses on the signaling of vasopressin-like receptors, which play a role in water homeostasis, blood pressure regulation, and stress response. Alterations in these processes have been associated with age-related physiological changes, such as decreased stress resilience and increased risk of hypertension [35].

5. Highly Sodium Permeable Postsynaptic Acetylcholine Nicotinic Receptors (R-HSA-629587): This pathway deals with the function of acetylcholine nicotinic receptors, which are involved in neurotransmission and neuromuscular function. Impairment of neurotransmission and synaptic function has been implicated in aging and age-related neurodegenerative disorders, such as Alzheimer's and Parkinson's diseases [36].

6. Cytosolic Sulfonation of Small Molecules (R-HSA-156584): This pathway describes the process of cytosolic sulfonation, a phase II detoxification reaction that helps to maintain cellular redox homeostasis and protects cells from oxidative stress. Oxidative stress has been widely recognized as a major contributor to the aging process and the development of age-related diseases.

The transfer learning approach utilized in this study enabled the construction of aging-centered case-control classifiers. These models were used to obtain lists of genes ranked by both their association with aging and diseases. Fibrotic disease, inflammatory disease, neurological disease, and cardiovascular disease are common disease classes in humans. To represent these categories, we have selected different age-related diseases, including idiopathic pulmonary fibrosis, COPD, PD, and heart failure, for each disease class. With the application of PandaOmics, APLNR, and IL23R are identified as the most potential aging targets for delaying and treating multiple age-related diseases. APLNR was in top-20 predictions for all four selected diseases. A declining Apelin/APLNR signaling promotes aging, whereas its restoration extended healthspan [37], and endogenous Apelin is protective against age-related loss of retinal ganglion cells in mice [38], further revealing its critical role in regulating aging. While the expression of both Apelin and APLNR decreases with increasing age [37], agonism of apelin receptors produces beneficial effects in fibrotic, cardiovascular, and cognitive disorders [39–41]. Taken together, targeting Apelin-APLNR signaling represents a very promising approach for the treatment of multiple age-related complications. Another potential multi-disease target that was in top-20 predictions for COPD, PD, and heart failure is IL23R, a receptor for IL-23 pro-inflammatory cytokine. Upregulation of the p19 subunit expression and IL-23 protein production in dendritic cells was observed in aged mice and may represent a potential mechanism for inadequate inflammatory responses in aging [42]. In the COPD murine model, IL-23$^{-/-}$ mice developed significantly lower static compliance values and decreased emphysematous changes in the lung tissue compared to WT mice [43]. Though the role of IL-23 is understudied in PD, neuroinflammation is a typical pathological feature of many neurodegenerative diseases, while emerging evidence indicates that sustained activation of microglia and astrocytes is central to dopaminergic degeneration in PD [44]. IL-23 can also enhance age-associated inflammation in Alzheimer's disease [45], likely to cause the accumulation of cellular damage and compromise the body's ability to repair itself. Local production of IL-23 in the Central Nervous System has been demonstrated for astrocytes and infiltrating macrophages under inflammatory conditions [46]. Altogether, agonizing Apelin/APLNR signaling and antagonizing IL23/IL23R axis may serve as potential therapeutic strategies for delaying and treating multiple age-related complications. These findings provide insights into potential targets for the delay and treatment of age-related diseases and demonstrate the utility of the transfer learning approach in identifying important genes associated with age-related dysfunction.

As there is a huge bet on AI and transformer applications in biomedicine, we expect future studies to focus on developing further the approach proposed in this paper, including possible integration of larger proprietary disease-specific datasets and validation of the identified targets in the wet lab setting. Moreover, exploring the potential of new drug targets and optimizing our model's performance will be crucial for advancing our understanding of the molecular mechanisms of aging and developing reliable interventions for age-related diseases. Ultimately, there is a great hope that Precious1GPT and its continued development and refinement will contribute significantly to the improvement of human health and longevity.

MATERIALS AND METHODS

Data sources for multimodal aging clock development

To train our models, we used several datasets, including publicly available and in-house-built ones. For training age prediction models based on the epigenetic status of tissue, we employed 450 k Illumina Methylation array

data from EWAS Data Hub [47] (number of samples = 8,374).

For building age prediction models based on the transcriptomics status of tissue, we employed RNA-Sequencing data from the GTEx project [48] (number of samples = 12,453). We trained our models in a tissue-agnostic fashion. For methylation, data distribution of samples across ages is shown in Supplementary Figure 5A and across tissues in Supplementary Figure 5B. For expression data, distribution of samples across ages is shown in Supplementary Figure 6A, across tissues in Supplementary Figure 6B.

For assessing prediction results and predicting disease targets using age-pretrained models, we used custom datasets from the PandaOmics software [49] for 4 selected age-related diseases: idiopathic pulmonary fibrosis, COPD, PD, and heart failure from where we have obtained samples annotated as carrying disease (case samples) and healthy ones (control samples).

To evaluate possible interventions to prevent senescence development, we have constructed datasets containing only features corresponding to genes for which approved drugs exist. For this, we have used an in-house constructed database of approved drugs and their targets based on the information from [50].

As input features for our age prediction models were either beta values averaged across CpG probes annotated as TS200 region from Illumina 450k Methylation Array for epigenomic data or TPM values for protein coding (genes) for expression datasets accordingly.

For the DNA methylation data, we obtained the ß-values from the CNCB data hub, where the raw data were obtained using the GMQN package developed by the CNCB [47].

In our study, we have opted for the TSS200 region as the most interpretable for age prediction. Following the aggregation of corresponding beta-values, we are left with approximately 14,000 features representing average methylation of proximal promoter regions.

This choice of region is based on its potential to offer a more accurate and comprehensive assessment of age-related changes in methylation patterns. The TSS200 region, situated within 200 base pairs upstream of the transcription start site, is known to play a crucial role in gene regulation [51]. As such, it is expected to exhibit significant age-related changes in methylation patterns, providing a robust basis for predicting biological age.

To further enhance the interpretability of our age prediction model, we utilized machine learning techniques to identify and select the most informative features from the 14,000-feature dataset. This allowed us to refine our model, ensuring that it captures the most relevant age-related methylation changes in the TSS200 region. In addition, the selected features can potentially shed light on the molecular mechanisms underlying aging, as well as the development of age-related diseases.

By focusing on the TSS200 region, we aim to not only improve the accuracy of our aging clock but also gain a deeper understanding of the complex relationship between DNA methylation, aging, and age-dependent diseases. This knowledge can then be used to develop targeted therapeutic interventions aimed at mitigating the impact of age-related diseases and improving overall health and quality of life in aging populations.

For gene expression models, we have used TPM values which were back-corrected using ComBat [52] and quantile-normalized using qnorm [53]. For performance evaluations, we employed a shuffle split stratified by sample tissue, leaving 20% of all data for the test set.

Transformer-based model for multimodal aging clock

Given the abundance of omics data available for various experimental conditions and the distinct challenges of predicting one omic data type from another, we propose the development of a transformer-based model to estimate sample age across different sample types. This multi-tissue, multi-omics transformer-based age prediction model aims to harness the power of deep learning to effectively integrate diverse data sources and improve age prediction accuracy.

Transformers have demonstrated remarkable success in various applications, particularly in natural language processing tasks. Their capacity to model long-range dependencies and capture complex relationships among features makes them well-suited for multi-omics data integration. By leveraging the transformer architecture, our proposed model is able to identify and exploit relevant information from different omics data types, such as genomics, transcriptomics, proteomics, and metabolomics, as well as different tissue types.

By developing a multi-tissue, multi-omics transformer-based age prediction model, we hope to enhance the accuracy and generalizability of aging clocks, ultimately contributing to a better understanding of the aging process and facilitating the development of targeted therapies for age-related diseases.

More formally about the transformer model. Let X be the input matrix of size (N, D) containing epigenetic (methylation) and expression data, and Z be the input matrix of size (N, M) containing categorical data. Let Y be the output matrix of size (N, 1) containing the predicted age values. Since we used TabTransformer (LINK), we represented the model as a function F that takes X and Z as input and returns Y as output: $Y = F(X, Z)$. The model consists of multiple linear and attention layers, each of which applies a set of learnable parameters to the input and produces an output. We represent each layer as a function G that takes an input matrix A and a set of learnable parameters W and b, and produces an output matrix:

$$B : B = G(A,W,b) =$$
$$ReLU(AW + b)B = G(A,W,b) =$$
$$ReLU(AW + b)$$

The TabTransformer [54] model consists of multiple layers, including self-attention layers and feedforward layers. Let A be the output of the previous layer, and let W_q, W_k, and W_v be learnable weight matrices of size (D, d_k). The self-attention layer computes the attention matrix A' as follows:

$$A' = soft\,max\left(\frac{AW_q(W_k)^T}{\sqrt{d_k}}\right)W_vO = A'W_v$$

where W_q, W_k, and W_v are weight matrices of size (D, d_k). The feedforward layer computes the output matrix H as follows:

$$H = G(O,W_1,b_1) = ReLU(OW_1 + b_1)$$

where W_1 is a weight matrix of size (h, d_h), and b_1 is a bias vector of size (1, d_h). We repeat the self-attention and feedforward layers multiple times to create a deep TabTransformer model. Next, we apply a feedforward layer with weight matrix W_1 and bias b_1 to the output of the self-attention layer: $H = ReLU(OW_1 + b_1)$. Finally, we apply a linear layer with weight matrix W_2 and bias b_2 to the output of the last feedforward layer to produce the final output matrix Y:

$$Y = HW_2 + b_2$$

To train the TabTransformer model, we used mean squared error (MSE) loss that measures the difference between the predicted age values and the true age values. Let $\hat{Y}$ be the predicted age values and Y_{true} be the true age values:

$$L = \frac{1}{N}\sum_{i=1}^{N}(\hat{Y}_l - Y_{true,i})^2$$

Overall, the experiment involves training a TabTransformer model on a dataset containing epigenetic (methylation) and expression data, as well as categorical data (tissue and dataset type), to predict age values. The model consists of multiple layers, including self-attention and feedforward layers, and is trained using a loss function such as MSE.

In the present study, we utilized PyTorch Tabular [55], which is built on top of PyTorch, for all the work with the transformer. PyTorch Tabular provides a highly optimized and efficient way of handling tabular data with PyTorch. We used PyTorch Tabular's various functions and modules to preprocess the input data, construct the transformer architecture, and train it on the data.

In the model, all the hyperparameters, such as learning rate, dropout rate, number of hidden layers, and activation functions, were determined through the use of Optuna [23], a hyperparameter optimization framework. The values of these parameters were chosen based on their performance during multiple rounds of training and validation. Based on our experiments, the optimal hyperparameters for TabTransformer are a model architecture with hidden layers of size 128, 2048, and 128, dropout probability of 0, "ELU" activation function, a learning rate of 0.00023, "AdamW" optimizer, weight decay of 0, and a batch size of 96.

Transfer learning for case-control classification

To build models which take into account both senescent and clinical status, we employed a transfer learning strategy. Initially, we trained a model as a regressor to predict age in an age dataset. Subsequently, we froze the model weights, excluding the final layer, and re-trained the resulting model as a case-control classifier. The derived SHAP values indicate the significance of each molecular feature in disease development concerning aging. This approach enabled us to determine the relative importance of various molecular features in driving disease development within the aging context. The pipeline is depicted in Figure 1.

Feature importance analysis

SHAP values, a mathematical method of feature importance analysis that constitutes a robust and interpretable technique for elucidating the contributions of individual features in complex predictive models, is based on the Shapley value from game theory and involves computing the contribution of each feature to

explain the final predictions of machine learning models [24]. SHAP values explain individual predictions and identify the most important features in the model. Additionally, the employment of SHAP values for feature analysis offers a considerable advantage in multimodal settings, where discerning the interplay between various factors is indispensable for the development of accurate and efficacious aging clocks.

Pathway enrichment analysis

Pathway enrichment analysis was performed on the list of top-100 genes ranked by their SHAP values obtained from the regressor model with the pathways available on the Reactome database. R package Enrichr was used to calculate the enrichment levels and *p*-values. Pathways with p-value < 0.05 were considered as significantly enriched.

Manual analysis of resulting targets

Combination of aging clocks and target ID represents an interesting approach to identifying targets for aging-associated diseases. To illustrate the applicability of this approach for target identification, we have investigated 4 diseases associated with aging: idiopathic pulmonary fibrosis, chronic obstructive pulmonary disease (COPD), Parkinson's disease (PD), and heart failure in PandaOmics. Gene lists of top-200 genes ranked by expression classifiers were used in Target ID projects for the mentioned diseases, along with a filter for small molecules to identify potential druggable proteins across these lists.

Abbreviations

AI: Artificial intelligence; COPD: Chronic Obstructive Pulmonary Disease; DNN: Deep Neural Network; DL: Deep Learning; iPSC: induced Pluripotent Stem Cells; ML: Machine Learning; PD: Parkinson's Disease; SHAP: SHapley Additive exPlanations.

AUTHOR CONTRIBUTIONS

AU – Software, methodology, project administration, writing (reviewing and editing), model optimization, DS – Writing, software, methodology, data processing, DZ – Data curation, manuscript reviewing, EK – Data curation, reviewing, AK – Data collection, resources, SP – Software, methodology, VN – Writing (reviewing and editing), supervision, VS – Writing, methodology, formal analysis, GL and HL – Writing (reviewing and editing), FP – Data curation, Writing (reviewing and editing), IO – Writing (reviewing and editing), supervision, AA – Writing (reviewing and editing), supervision, FR – Conceptualization, resources, supervision, AZ – Conceptualization, resources, supervision.

ACKNOWLEDGMENTS

We greatly thank Elizaveta Ekimova for her kind help with the preparation of the figures.

CONFLICTS OF INTEREST

The authors are affiliated with Insilico Medicine, a commercial company developing and using generative artificial intelligence and other next-generation AI technologies and robotics for drug discovery, drug development, and aging research. Utilizing its generative AI platform and a range of deep aging clocks, Insilico Medicine has developed a portfolio of multiple therapeutic programs targeting fibrotic diseases, cancer, immunological diseases, and a range of age-related diseases.

FUNDING

This study received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

1. Moskalev AA, Aliper AM, Smit-McBride Z, Buzdin A, Zhavoronkov A. Genetics and epigenetics of aging and longevity. Cell Cycle. 2014; 13:1063–77. https://doi.org/10.4161/cc.28433 PMID:24603410

2. Putin E, Mamoshina P, Aliper A, Korzinkin M, Moskalev A, Kolosov A, Ostrovskiy A, Cantor C, Vijg J, Zhavoronkov A. Deep biomarkers of human aging: Application of deep neural networks to biomarker development. Aging (Albany NY). 2016; 8:1021–33. https://doi.org/10.18632/aging.100968 PMID:27191382

3. Zhavoronkov A, Mamoshina P. Deep Aging Clocks: The Emergence of AI-Based Biomarkers of Aging and Longevity. Trends Pharmacol Sci. 2019; 40:546–9. https://doi.org/10.1016/j.tips.2019.05.004 PMID:31279569

4. Mamoshina P, Vieira A, Putin E, Zhavoronkov A. Applications of Deep Learning in Biomedicine. Mol Pharm. 2016; 13:1445–54. https://doi.org/10.1021/acs.molpharmaceut.5b00982 PMID:27007977

5. Zhavoronkov A, Li R, Ma C, Mamoshina P. Deep biomarkers of aging and longevity: from research to applications. Aging (Albany NY). 2019; 11:10771–80.

https://doi.org/10.18632/aging.102475
PMID:31767810

6. Mamoshina P, Kochetov K, Putin E, Cortese F, Aliper A, Lee WS, Ahn SM, Uhn L, Skjodt N, Kovalchuk O, Scheibye-Knudsen M, Zhavoronkov A. Population Specific Biomarkers of Human Aging: A Big Data Study Using South Korean, Canadian, and Eastern European Patient Populations. J Gerontol A Biol Sci Med Sci. 2018; 73:1482–90. https://doi.org/10.1093/gerona/gly005 PMID:29340580

7. Pun FW, Leung GHD, Leung HW, Liu BHM, Long X, Ozerov IV, Wang J, Ren F, Aliper A, Izumchenko E, Moskalev A, de Magalhães JP, Zhavoronkov A. Hallmarks of aging-based dual-purpose disease and age-associated targets predicted using PandaOmics AI-powered discovery engine. Aging (Albany NY). 2022; 14:2475–506. https://doi.org/10.18632/aging.203960 PMID:35347083

8. Marino N, Putignano G, Cappilli S, Chersoni E, Santuccione A, Calabrese G, Bischof E, Vanhaelen Q, Zhavoronkov A, Scarano B, Mazzotta AD, Santus E. Towards AI-driven longevity research: An overview. Front Aging. 2023; 4:1057204. https://doi.org/10.3389/fragi.2023.1057204 PMID:36936271

9. Aliper A, Plis S, Artemov A, Ulloa A, Mamoshina P, Zhavoronkov A. Deep Learning Applications for Predicting Pharmacological Properties of Drugs and Drug Repurposing Using Transcriptomic Data. Mol Pharm. 2016; 13:2524–30. https://doi.org/10.1021/acs.molpharmaceut.6b00248 PMID:27200455

10. Shayakhmetov R, Kuznetsov M, Zhebrak A, Kadurin A, Nikolenko S, Aliper A, Polykovskiy D. Erratum: Addendum: Molecular Generation for Desired Transcriptome Changes With Adversarial Autoencoders. Front Pharmacol. 2020; 11:1236. https://doi.org/10.3389/fphar.2020.01236 PMID:32973498

11. Aliper A, Zavoronkovs A, Zhebrak A, Kadurin A, Polykovskiy D, Shayakhmetov R. Mutual information adversarial autoencoder. US11403521B2. 2022. https://patents.google.com/patent/US11403521B2/en/.

12. Moore JH, Raghavachari N, and Workshop Speakers. Artificial Intelligence Based Approaches to Identify Molecular Determinants of Exceptional Health and Life Span-An Interdisciplinary Workshop at the National Institute on Aging. Front Artif Intell. 2019; 2:12. https://doi.org/10.3389/frai.2019.00012 PMID:33733101

13. Kadurin A, Nikolenko S, Khrabrov K, Aliper A, Zhavoronkov A. druGAN: An Advanced Generative Adversarial Autoencoder Model for de Novo Generation of New Molecules with Desired Molecular Properties in Silico. Mol Pharm. 2017; 14:3098–104. https://doi.org/10.1021/acs.molpharmaceut.7b00346 PMID:28703000

14. Putin E, Asadulaev A, Ivanenkov Y, Aladinskiy V, Sanchez-Lengeling B, Aspuru-Guzik A, Zhavoronkov A. Reinforced Adversarial Neural Computer for de Novo Molecular Design. J Chem Inf Model. 2018; 58:1194–204. https://doi.org/10.1021/acs.jcim.7b00690 PMID:29762023

15. Polykovskiy D, Zhebrak A, Vetrov D, Ivanenkov Y, Aladinskiy V, Mamoshina P, Bozdaganyan M, Aliper A, Zhavoronkov A, Kadurin A. Entangled Conditional Adversarial Autoencoder for de Novo Drug Discovery. Mol Pharm. 2018; 15:4398–405. https://doi.org/10.1021/acs.molpharmaceut.8b00839 PMID:30180591

16. Galkin F, Mamoshina P, Aliper A, de Magalhães JP, Gladyshev VN, Zhavoronkov A. Biohorology and biomarkers of aging: Current state-of-the-art, challenges and opportunities. Ageing Res Rev. 2020; 60:101050. https://doi.org/10.1016/j.arr.2020.101050 PMID:32272169

17. Galkin F, Mamoshina P, Aliper A, Putin E, Moskalev V, Gladyshev VN, Zhavoronkov A. Human Gut Microbiome Aging Clock Based on Taxonomic Profiling and Deep Learning. iScience. 2020; 23:101199. https://doi.org/10.1016/j.isci.2020.101199 PMID:32534441

18. Zhavoronkov A, Mamoshina P, Vanhaelen Q, Scheibye-Knudsen M, Moskalev A, Aliper A. Artificial intelligence for aging and longevity research: Recent advances and perspectives. Ageing Res Rev. 2019; 49:49–66. https://doi.org/10.1016/j.arr.2018.11.003 PMID:30472217

19. Galkin F, Mamoshina P, Kochetov K, Sidorenko D, Zhavoronkov A. DeepMAge: A Methylation Aging Clock Developed with Deep Learning. Aging Dis. 2021; 12:1252–62. https://doi.org/10.14336/AD.2020.1202 PMID:34341706

20. MacRae SL, Croken MM, Calder RB, Aliper A, Milholland B, White RR, Zhavoronkov A, Gladyshev VN, Seluanov A, Gorbunova V, Zhang ZD, Vijg J. DNA repair in species with extreme lifespan differences.

Aging (Albany NY). 2015; 7:1171–84. https://doi.org/10.18632/aging.100866 PMID:26729707

21. Zhavoronkov A, Buzdin AA, Garazha AV, Borisov NM, Moskalev AA. Signaling pathway cloud regulation for in silico screening and ranking of the potential geroprotective drugs. Front Genet. 2014; 5:49. https://doi.org/10.3389/fgene.2014.00049 PMID:24624136

22. Aliper A, Belikov AV, Garazha A, Jellen L, Artemov A, Suntsova M, Ivanova A, Venkova L, Borisov N, Buzdin A, Mamoshina P, Putin E, Swick AG, et al. In search for geroprotectors: in silico screening and in vitro validation of signalome-level mimetics of young healthy state. Aging (Albany NY). 2016; 8:2127–52. https://doi.org/10.18632/aging.101047 PMID:27677171

23. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A Next-generation Hyperparameter Optimization Framework. 2019. https://doi.org/10.48550/arXiv.1907.10902

24. Lundberg S, Lee SI. A Unified Approach to Interpreting Model Predictions. 2017. https://doi.org/10.48550/arXiv.1705.07874

25. Canfield CA, Bradshaw PC. Amino acids in the regulation of aging and aging-related diseases. Transl Med Aging. 2019; 3:70–89. https://doi.org/10.1016/j.tma.2019.09.001

26. Holzscheck N, Falckenhayn C, Söhle J, Kristof B, Siegner R, Werner A, Schössow J, Jürgens C, Völzke H, Wenck H, Winnefeld M, Grönniger E, Kaderali L. Modeling transcriptomic age using knowledge-primed artificial neural networks. NPJ Aging Mech Dis. 2021; 7:15. https://doi.org/10.1038/s41514-021-00068-5 PMID:34075044

27. Buckley MT, Sun ED, George BM, Liu L, Schaum N, Xu L, Reyes JM, Goodell MA, Weissman IL, Wyss-Coray T, Rando TA, Brunet A. Cell-type-specific aging clocks to quantify aging and rejuvenation in neurogenic regions of the brain. Nat Aging. 2023; 3:121–37. https://doi.org/10.1038/s43587-022-00335-4 PMID:37118510

28. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 2013; 14:R115. https://doi.org/10.1186/gb-2013-14-10-r115 PMID:24138928

29. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, Klotzle B, Bibikova M, Fan JB, Gao Y, Deconde R, Chen M, Rajapakse I, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol Cell. 2013; 49:359–67. https://doi.org/10.1016/j.molcel.2012.10.016 PMID:23177740

30. Levine ME, Lu AT, Quach A, Chen BH, Assimes TL, Bandinelli S, Hou L, Baccarelli AA, Stewart JD, Li Y, Whitsel EA, Wilson JG, Reiner AP, et al. An epigenetic biomarker of aging for lifespan and healthspan. Aging (Albany NY). 2018; 10:573–91. https://doi.org/10.18632/aging.101414 PMID:29676998

31. Liu Z, Leung D, Thrush K, Zhao W, Ratliff S, Tanaka T, Schmitz LL, Smith JA, Ferrucci L, Levine ME. Underlying features of epigenetic aging clocks in vivo and in vitro. Aging Cell. 2020; 19:e13229. https://doi.org/10.1111/acel.13229 PMID:32930491

32. de Lima Camillo LP, Lapierre LR, Singh R. A pan-tissue DNA-methylation epigenetic clock based on deep learning. npj Aging. 2022; 8:1–15. https://doi.org/10.1038/s41514-022-00085-y

33. Amorim JA, Coppotelli G, Rolo AP, Palmeira CM, Ross JM, Sinclair DA. Mitochondrial and metabolic dysfunction in ageing and age-related diseases. Nat Rev Endocrinol. 2022; 18:243–58. https://doi.org/10.1038/s41574-021-00626-7 PMID:35145250

34. Tower J. Programmed cell death in aging. Ageing Res Rev. 2015; 23:90–100. https://doi.org/10.1016/j.arr.2015.04.002 PMID:25862945

35. Birder LA, Wolf-Johnston AS, Jackson EK, Wein AJ, Dmochowski R. Aging increases the expression of vasopressin receptors in both the kidney and urinary bladder. Neurourol Urodyn. 2019; 38:393–7. https://doi.org/10.1002/nau.23830 PMID:30311671

36. Takata K, Kimura H, Yanagisawa D, Harada K, Nishimura K, Kitamura Y, Shimohama S, Tooyama I. Nicotinic Acetylcholine Receptors and Microglia as Therapeutic and Imaging Targets in Alzheimer's Disease. Molecules. 2022; 27:2780. https://doi.org/10.3390/molecules27092780 PMID:35566132

37. Rai R, Ghosh AK, Eren M, Mackie AR, Levine DC, Kim SY, Cedernaes J, Ramirez V, Procissi D, Smith LH, Woodruff TK, Bass J, Vaughan DE. Downregulation of the Apelinergic Axis Accelerates Aging, whereas Its Systemic Restoration Improves the Mammalian Healthspan. Cell Rep. 2017; 21:1471–80. https://doi.org/10.1016/j.celrep.2017.10.057 PMID:29117554

38. Ishimaru Y, Sumino A, Shibagaki F, Yamamuro A, Yoshioka Y, Maeda S. Endogenous Apelin Is Protective Against Age-Associated Loss of Retinal Ganglion Cells

in Mice. Front Aging Neurosci. 2020; 12:58. https://doi.org/10.3389/fnagi.2020.00058 PMID:32296325

39. Wang H, Cong L, Yin X, Zhang N, Zhu M, Sun T, Fan J, Xue F, Fan X, Gong Y. The Apelin-APJ axis alleviates LPS-induced pulmonary fibrosis and endothelial mesenchymal transformation in mice by promoting Angiotensin-Converting Enzyme 2. Cell Signal. 2022; 98:110418. https://doi.org/10.1016/j.cellsig.2022.110418 PMID:35882286

40. Zhou Q, Chen L, Tang M, Guo Y, Li L. Apelin/APJ system: A novel promising target for anti-aging intervention. Clin Chim Acta. 2018; 487:233–40. https://doi.org/10.1016/j.cca.2018.10.011 PMID:30296443

41. Luo H, Han L, Xu J. Apelin/APJ system: A novel promising target for neurodegenerative diseases. J Cell Physiol. 2020; 235:638–57. https://doi.org/10.1002/jcp.29001 PMID:31254280

42. El Mezayen R, El Gazzar M, Myer R, High KP. Aging-dependent upregulation of IL-23p19 gene expression in dendritic cells is associated with differential transcription factor binding and histone modifications. Aging Cell. 2009; 8:553–65. https://doi.org/10.1111/j.1474-9726.2009.00502.x PMID:19624579

43. Fujii U, Miyahara N, Taniguchi A, Waseda K, Morichika D, Kurimoto E, Koga H, Kataoka M, Gelfand EW, Cua DJ, Yoshimura A, Tanimoto M, Kanehiro A. IL-23 Is Essential for the Development of Elastase-Induced Pulmonary Inflammation and Emphysema. Am J Respir Cell Mol Biol. 2016; 55:697–707. https://doi.org/10.1165/rcmb.2016-0015OC PMID:27351934

44. MacMahon Copas AN, McComish SF, Fletcher JM, Caldwell MA. The Pathogenesis of Parkinson's Disease: A Complex Interplay Between Astrocytes, Microglia, and T Lymphocytes? Front Neurol. 2021; 12:666737. https://doi.org/10.3389/fneur.2021.666737 PMID:34122308

45. Mohammadi Shahrokhi V, Ravari A, Mirzaei T, Zare-Bidaki M, Asadikaram G, Arababadi MK. IL-17A and IL-23: plausible risk factors to induce age-associated inflammation in Alzheimer's disease. Immunol Invest. 2018; 47:812–22. https://doi.org/10.1080/08820139.2018.1504300 PMID:30081688

46. Nitsch L, Schneider L, Zimmermann J, Müller M. Microglia-Derived Interleukin 23: A Crucial Cytokine in Alzheimer's Disease? Front Neurol. 2021; 12:639353. https://doi.org/10.3389/fneur.2021.639353 PMID:33897596

47. Xiong Z, Li M, Yang F, Ma Y, Sang J, Li R, Li Z, Zhang Z, Bao Y. EWAS Data Hub: a resource of DNA methylation array data and metadata. Nucleic Acids Res. 2020; 48:D890–5. https://doi.org/10.1093/nar/gkz840 PMID:31584095

48. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N, Foster B, Moser M, Karasik E, et al, and GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013; 45:580–5. https://doi.org/10.1038/ng.2653 PMID:23715323

49. PandaOmics Insilico Medicine. https://insilico.com/pandaomics.

50. Home - ClinicalTrials.gov. https://clinicaltrials.gov/.

51. Tang J, Zou J, Zhang X, Fan M, Tian Q, Fu S, Gao S, Fan S. PretiMeth: precise prediction models for DNA methylation based on single methylation mark. BMC Genomics. 2020; 21:364. https://doi.org/10.1186/s12864-020-6768-9 PMID:32414326

52. Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genom Bioinform. 2020; 2:lqaa078. https://doi.org/10.1093/nargab/lqaa078 PMID:33015620

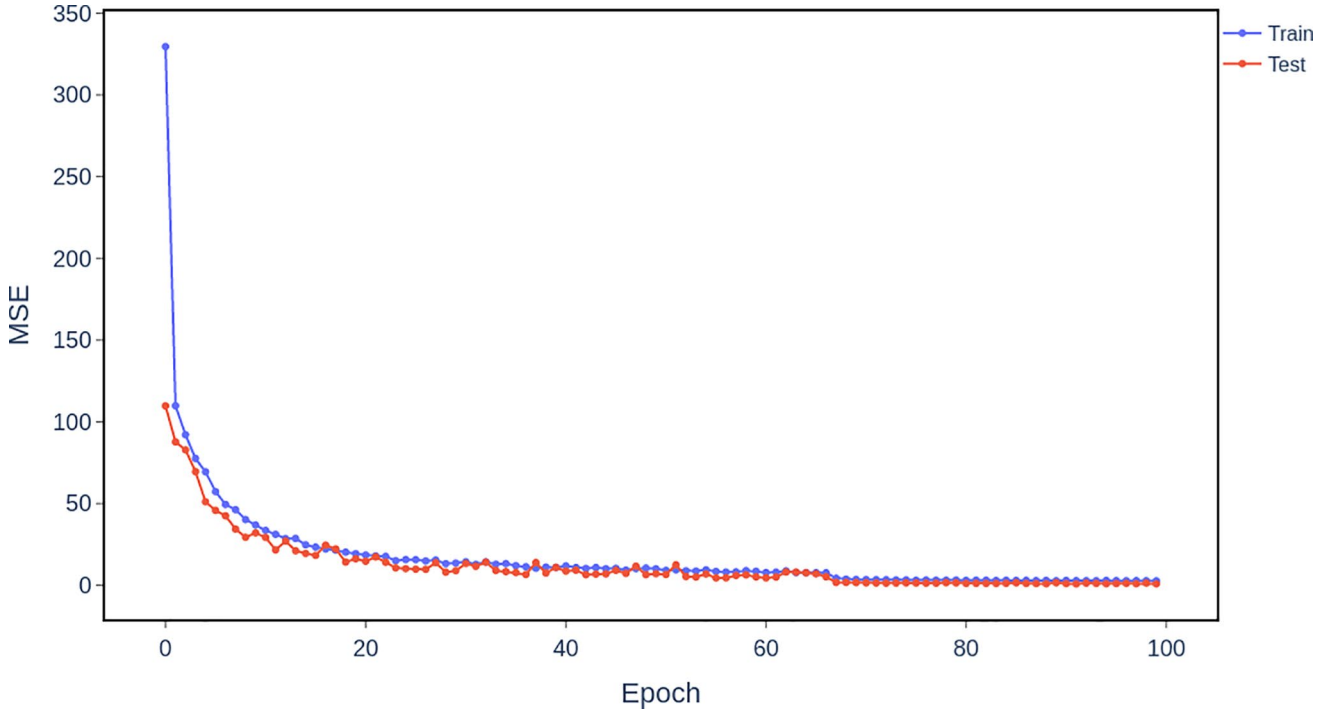
53. van der Sande M, van Heeringen S. qnorm. 2021. https://github.com/Maarten-vd-Sande/qnorm.

54. Huang X, Khetan A, Cvitkovic M, Karnin Z. TabTransformer: Tabular Data Modeling Using Contextual Embeddings. 2020. https://doi.org/10.48550/arXiv.2012.06678

55. Joseph M. PyTorch Tabular: A Framework for Deep Learning with Tabular Data. 2021. https://doi.org/10.48550/arXiv.2104.13638

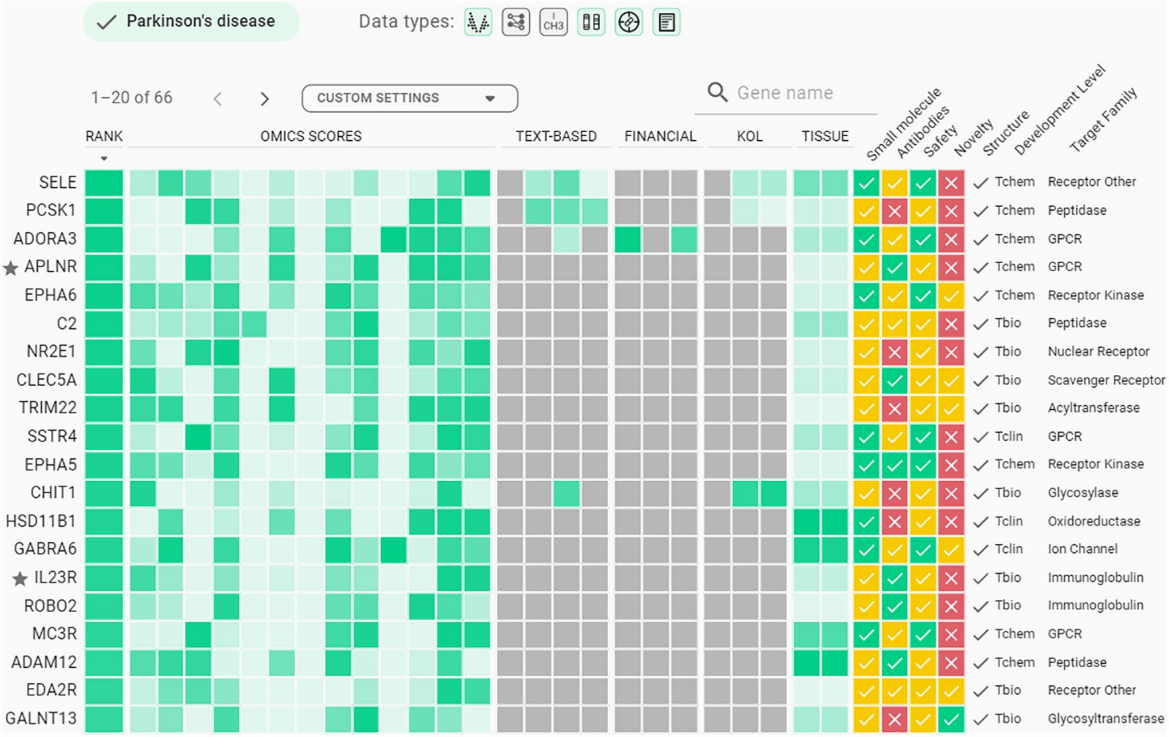
SUPPLEMENTARY MATERIALS

Supplementary Figures

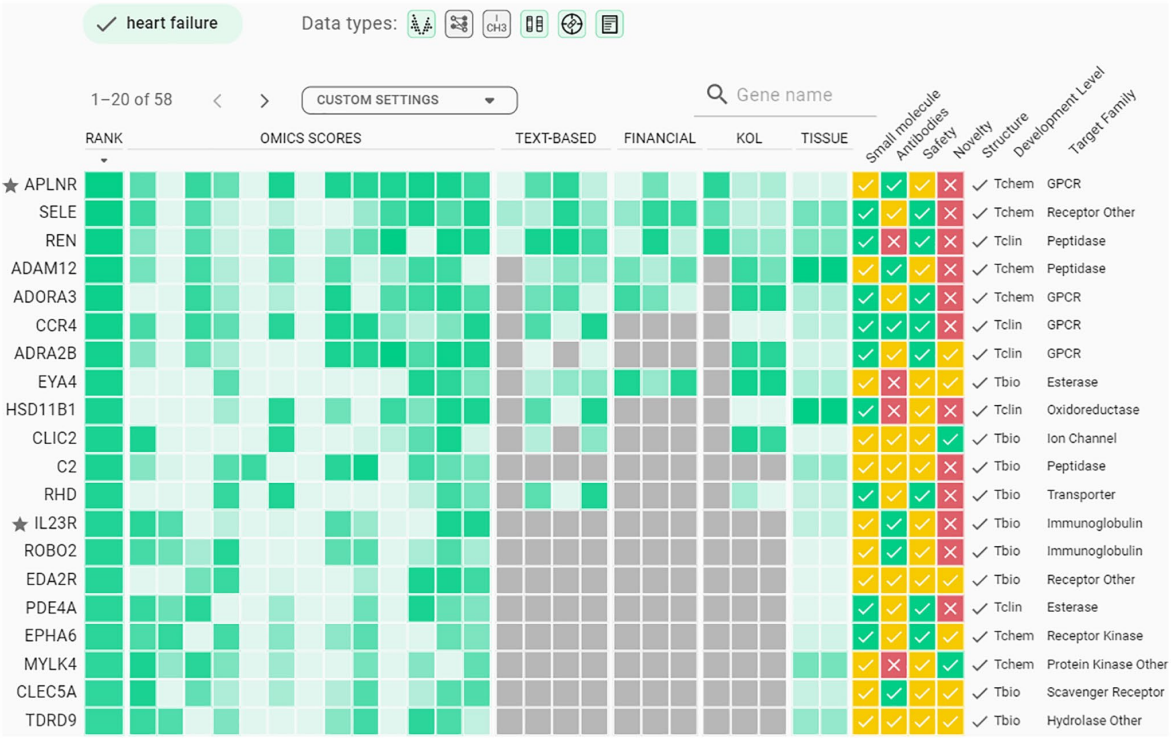


Supplementary Figure 1. Multimodal transformer learning curves on the train and 20% hold-out test datasets.

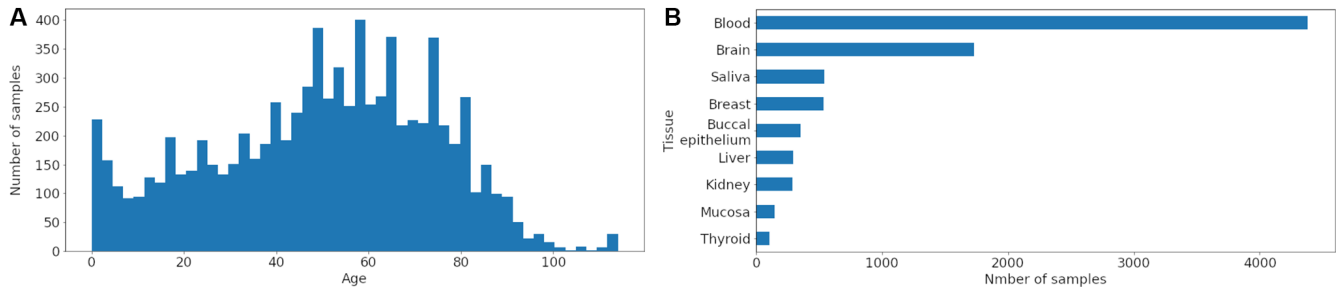
Supplementary Figure 2. Example of target ID output for idiopathic pulmonary fibrosis. Top-200 genes from expression classifiers were applied as a gene list in PandaOmics corresponding project for idiopathic pulmonary fibrosis, and a filter for small molecules was applied to identify druggable targets. Twenty genes highly ranked by PandaOmics are shown.



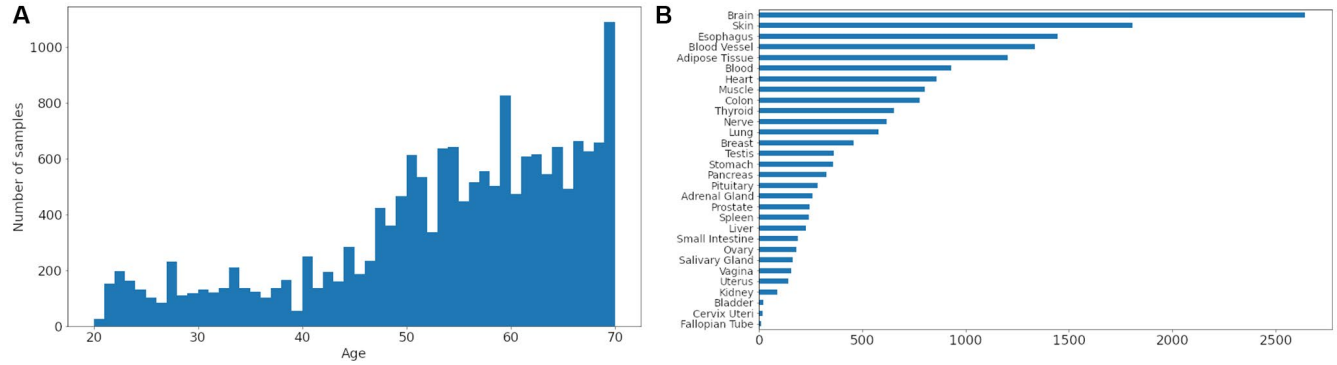
Supplementary Figure 3. Example of target ID output for Parkinson's disease. Top-200 genes from expression classifiers were applied as a gene list in PandaOmics corresponding project for PD, and a filter for small molecules was applied to identify druggable targets. Twenty genes highly ranked by PandaOmics are shown.



Supplementary Figure 4. Example of target ID output for heart failure. Top-200 genes from expression classifiers were applied as a gene list in PandaOmics corresponding project for heart failure, and a filter for small molecules was applied to identify druggable targets. Twenty genes highly ranked by PandaOmics are shown.



Supplementary Figure 5. Distribution by age (**A**) and tissues (**B**) for DNAm samples. Data was obtained from CNCB EWAS data hub. Ages distributed from 0 to 110 years. Most of the samples are blood samples.



Supplementary Figure 6. Distribution by age (**A**) and tissues (**B**) for RNA-seq samples. Data are obtained from the GTEx project. Ages are distributed between 20 and 70 years. Brain and Skin samples comprise a bigger part of the dataset.

Supplementary Tables

Please browse Full Text version to see the data of Supplementary Tables 3–8.

Supplementary Table 1. 5-fold cross-validation for multimodal transformer age prediction.

Metric	Combined	Methylation	Expression
MAE	5.800+/−0.437	4.815+/−0.458	6.469+/−0.427
RMSE	7.665+/−0.436	6.680+/−0.520	8.266+/−0.403
R2	0.823+/−0.021	0.923+/−0.013	0.572+/−0.041
MdAE	4.546+/−0.517	3.569+/−0.486	5.335+/−0.588

Supplementary Table 2. Performance of multimodal model on different combinations of tissues and data modalities. Estimates on 20% tissue-stratified hold-out test dataset.

TISSUE	MODALITY	MAE	R2	RMSE	MSE	MdAE	MAD	TEST_SAMPLES
Thyroid	METHYLATION	2.456	0.900	4.287	18.381	0.567	0.609	21
Buccal epithelium	METHYLATION	2.656	0.972	3.618	13.092	1.707	1.669	69
Saliva	METHYLATION	3.170	0.959	4.237	17.948	2.249	2.069	107
Mucosa	METHYLATION	3.358	0.896	3.988	15.903	2.879	3.010	29
Brain	METHYLATION	3.749	0.949	6.381	40.716	1.851	1.922	345
Blood	METHYLATION	4.291	0.928	5.990	35.880	3.168	3.177	863
Brain	EXPRESSION	4.483	0.640	5.954	35.446	3.539	3.372	528
Blood Vessel	EXPRESSION	5.229	0.715	6.680	44.621	4.159	4.515	267
Thyroid	EXPRESSION	5.367	0.698	6.967	48.539	4.406	4.195	129
Nerve	EXPRESSION	5.461	0.656	7.075	50.061	4.479	4.448	124
Testis	EXPRESSION	5.714	0.673	7.298	53.268	5.423	5.411	72
Breast	METHYLATION	5.797	0.738	7.738	59.876	4.477	4.443	106
Liver	METHYLATION	6.036	0.744	7.749	60.047	5.300	4.857	59
Ovary	EXPRESSION	6.070	0.672	8.488	72.052	4.460	4.425	36
Adrenal Gland	EXPRESSION	6.085	0.683	7.431	55.216	5.729	5.729	52
Pituitary	EXPRESSION	6.154	0.199	7.633	58.267	5.666	5.091	56
Small Intestine	EXPRESSION	6.204	0.708	7.533	56.739	6.014	6.007	38
Adipose Tissue	EXPRESSION	6.210	0.529	7.976	63.624	5.219	5.279	241
Salivary Gland	EXPRESSION	6.296	0.532	8.413	70.781	4.595	4.454	33
Kidney	METHYLATION	6.315	0.908	7.951	63.222	5.968	5.826	58
Skin	EXPRESSION	6.602	0.538	8.664	75.069	5.031	4.817	362
Esophagus	EXPRESSION	6.647	0.626	8.317	69.169	5.484	5.507	289
Prostate	EXPRESSION	6.670	0.669	8.349	69.711	5.096	5.345	49
Uterus	EXPRESSION	6.724	0.619	8.063	65.016	5.714	5.905	29
Lung	EXPRESSION	7.018	0.399	9.223	85.067	5.069	5.393	115
Breast	EXPRESSION	7.037	0.521	8.919	79.541	5.798	5.230	92
Muscle	EXPRESSION	7.040	0.558	8.604	74.027	5.981	5.584	160
Heart	EXPRESSION	7.059	0.403	8.799	77.421	6.302	6.405	172
Colon	EXPRESSION	7.204	0.590	9.143	83.598	6.241	6.118	156
Stomach	EXPRESSION	7.242	0.469	8.924	79.640	6.177	6.177	72

Pancreas	EXPRESSION	7.291	0.429	9.043	81.774	6.549	6.176	65
Vagina	EXPRESSION	7.436	0.311	9.372	87.826	5.883	5.364	31
Liver	EXPRESSION	7.592	0.243	9.372	87.829	7.066	5.343	45
Blood	EXPRESSION	8.158	0.397	10.617	112.724	6.288	6.303	186
Spleen	EXPRESSION	9.050	0.346	11.204	125.526	6.770	7.227	48
Kidney	EXPRESSION	9.715	−0.118	12.731	162.079	8.318	7.508	18

Supplementary Table 3. Feature importance analysis for Aging clock.

Supplementary Table 4. Feature importance analysis for Pulmonary fibrosis case-control classifier.

Supplementary Table 5. Feature importance analysis for Chronic Obstructive Pulmonary Disease case-control classifier.

Supplementary Table 6. Feature importance analysis for Parkinson's disease case-control classifier.

Supplementary Table 7. Feature importance analysis for Heart failure case-control classifier.

Supplementary Table 8. Metrics for multimodal transformer-based case-control classifiers with and without pretraining on aging data.