



Research
Glycomedicine—Article

Deep Reinforcement Learning-Driven Multi-Omics Integration for Constructing gtAge: A Novel Aging Clock from the IgG N-Glycome and Blood Transcriptome



Yao Xia^{a, #, *}, Syed Mohammed Shamsul Islam^{a, i, j, #}, Xingang Li^{b, #}, Abdul Baten^{c, d}, Xuerui Tan^e, Wei Wang^{b, e, f, g, h, *}

^a School of Science, Edith Cowan University, Joondalup, WA 6027, Australia

^b Nutrition and Health Innovation Research Institute & School of Medical and Health Sciences, Edith Cowan University, Joondalup, WA 6027, Australia

^c Institute of Precision Medicine and Bioinformatics, Royal Prince Alfred Hospital, Sydney, NSW 2050, Australia

^d Department of Biomedical Informatics and Digital Health, School of Medical Sciences, University of Sydney, Sydney, NSW 2050, Australia

^e Clinical Research Centre, The First Affiliated Hospital & Institute for Glycome Study, Shantou University Medical College, Shantou 515041, China

^f Chemistry and Chemical Engineering Guangdong Laboratory, Shantou 515041, China

^g Beijing Key Laboratory of Clinical Epidemiology, School of Public Health, Capital Medical University, Beijing 110069, China

^h School of Public Health, Shandong First Medical University & Shandong Academy of Medical Sciences, Tai'an 271016, China

ⁱ Department of Computing and Information System, Daffodil International University, Dhaka 1216, Bangladesh

^j Department of Computer Science and Software Engineering, The University of Western Australia, Perth, WA 6009, Australia

ARTICLE INFO

Article history:

Received 20 January 2025

Revised 15 April 2025

Accepted 4 August 2025

Available online 19 August 2025

Keywords:

Aging clock

IgG N-glycome

Transcriptome

Multimics integration

Deep reinforcement learning

Biological age

ABSTRACT

Previous studies have demonstrated that the immunoglobulin G (IgG) N-glycome and transcriptome are potential biochemical signatures of chronological and biological ages, and several aging clocks have been developed. By integrating the IgG N-glycome and transcriptome, we propose a novel aging clock, gtAge. We developed a deep reinforcement learning-based multimics integration method called AlphaSnake. The results showed that AlphaSnake achieved a predicted coefficient of determination (R^2) value of 0.853, outperforming the concatenation-based integration method ($R^2 = 0.820$). The gtAge estimated by AlphaSnake explained up to 85.3% of the variance in chronological age, which was higher than that in age predicted from IgG N-glycome solely (gAge; $R^2 = 0.290$) and age predicted from transcriptome solely (tAge; $R^2 = 0.812$). We also found that the delta age—the difference between the predicted age and chronological age—was associated with several age-related phenotypes. Both delta gtAge and tAge were negatively associated with high-density lipoprotein ($p = 0.02$ and $p = 0.022$, respectively), whereas delta gAge was positively correlated with cholesterol ($p = 0.006$), triglyceride ($p = 0.002$), fasting plasma glucose ($p = 0.014$), low-density lipoprotein ($p = 0.006$), and glycated hemoglobin ($p = 0.039$). These findings suggest that gtAge, tAge, and gAge are potential biomarkers for biological age.

© 2025 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Aging is a complex physiological process characterized by increased inflammation and a progressive decline in health, ultimately resulting in disease and death [1]. Although chronological age (CA), the time elapsed since birth, is the most direct and com-

monly used metric, it does not entirely capture individual variability in the aging process. In reality, some individuals remain healthy until they are 80 to 90 years of age, whereas others may start experiencing age-related decline significantly earlier in their lives. This discrepancy can be attributed to differences in biological age (BA), which integrates genetic, lifestyle, nutritional, disease-related, and general health factors to accurately reflect the true biological aging process. BA is an abstract concept that encompasses the internal, external, and functional aspects of whole-body aging, which cannot be reflected by a single or multiple biomarkers [2]. In our study, the estimated BA did not represent the full extent of

* Corresponding authors.

E-mail addresses: yao.xia@uwa.edu.au (Y. Xia), wei.wang@ecu.edu.au (W. Wang).

These authors contributed equally to this work.

biological aging, rather provided insights into the specific dimensions of the biological aging process. We used the term “biological age” for convenience, acknowledging that aging clocks and related biomarkers act as proxies or indicators of certain biological features associated with aging, rather than as comprehensive measures of aging itself.

Since the last century, the identification of aging biomarkers has been an attractive research topic. These biomarkers are classically defined as “biological parameters of an organism that either alone or in some multivariate composite will, in the absence of disease, better predict functional capability at some late age than will chronological age” [3,4]. The American Federation for Aging Research (AFAR) proposed that aging biomarkers should be able to predict the rate of aging, be free from the effect resulted from diseases, be easily measurable, and be validated in laboratory animals [4–6]. However, the AFAR criteria are too stringent, and no biomarkers have ultimately and strictly been able to meet those till date. According to another definition that is not so stringent [4], the task of identifying biomarkers of BA can be divided into two steps—① identify biomarkers that can predict the CA. Typical metrics, for example, predicted coefficient of determination (R^2) and mean absolute error, are used to evaluate the performance of predictive model. ② Evaluate whether the biomarkers can bring additive information in aging-related assessments. A major purpose of the aging clocks is to identify biomarkers that can act as surrogate endpoints in clinical trials. As lifespan studies in humans are practically infeasible owing to the time required, reliable BA estimators may offer an alternative way to measure the impact of interventions on the aging process. Thus, aging clocks are highly valuable tools in translational aging research.

Blood transcriptome and IgG *N*-glycans have been established as promising aging biomarkers in a wide range of studies [7–18]. Peters et al. [12] used large-scale datasets to build an age predicted from transcriptome solely (tAge) model and found that, depending on the cohorts analyzed, the model's R^2 was between 0.121 and 0.599. In a recent study, Shokhirev and Johnson [15] built a tAge model with R^2 value of 0.985 from cross validation in the discovery cohort, whereas in several external validation (replication) cohorts, the model's performance dropped significantly. However, IgG *N*-glycans have not been investigated extensively. In addition, studies based on glyocytes found that the IgG *N*-glycome explains 23.3%–58.0% of the variance in CA [19]. Krištić et al. [9] built an IgG *N*-glycome-based aging clock (referred to as gAge), with the model's R^2 value of 0.580 in the discovery cohort, whereas the R^2 dropped below 0.5 in the external validation cohorts. Yu et al. [20] built a gAge clock using a Chinese population, with the model's R^2 value of 0.294.

Furthermore, using a combination of markers from different sources, attempts have been made to identify aging biomarkers. Using the UK Biobank data, Mak et al. [21] conducted a large-scale study ($n = 308\ 156$) to investigate the association between the aging clock and cancer incidence. They calculated three biomarker-based BA measures—Klemera–Doubal method, PhenoAge, and Homeostatic Dysregulation—from 18 routine clinical biomarkers (forced expiratory volume in one second (FEV1), systolic blood pressure, serum glucose, C-reactive protein, and so on). All BA measures were associated with increased risks of lung cancer and colorectal cancer, and PhenoAge was linked to an increased risk of breast cancer. An inverse association between BA measures and prostate cancer was noted, but that weakened after adjusting for glucose-related biomarkers [21]. This study highlights the way composite biomarker-based BA estimates can modestly predict the risk of cancer and reflect the diverse physiological systems involved in the aging process. However, this task becomes challenging when the integration of multiomics data is considered. Zierer et al. [22] adopted a graphical random forest

to identify age-related features from multiomics data and combined them to predict age-related diseases. However, they did not attempt building an aging clock using integrated multiomics data. To the best of our knowledge, and based on recent reviews in this field [4,23–27], no study has attempted to build an integrated multiomics aging clock. Using feature sets consisting of different omics datasets to predict BA can improve the prediction performance and interpretability. The combined feature set can contain biological information from different omics datasets. Integrating the IgG *N*-glycome and transcriptome in blood, this study aimed to build a potential aging clock, gtAge, for BA prediction. To integrate multiomics data more effectively, we formulated the multiomics integration challenge as a feature selection process and developed a deep reinforcement learning (DRL)-based method called AlphaSnake (see details in Section 2.6). We chose the DRL-based method, as it was well suited for sequential decision-making problems, such as iterative feature selection, where the agent learnt the optimal strategy for balancing different omics sources and dynamically selecting the most informative features. Compared with traditional methods, the DRL-based method offers a more flexible and adaptive framework capable of handling high-dimensional heterogeneous omics data. Additionally, as there is limited research exploring DRL-based method for multiomics integration, our study aims to explore and evaluate the potential of DRL in this context.

2. Materials and methods

2.1. Overview of the dataset

In this study, plasma samples from a subcohort, BHAS-302, consisting of 302 individuals were collected from the Busselton Healthy Ageing Study (BHAS). This is an ongoing cohort of community-dwelling “Baby Boomers” from the Shire of Busselton, Western Australia [28]. This cohort was ethnically homogeneous and of European origin, living in a nature-friendly environment located on the coast of the Indian Ocean. The subset consisted of 134 males and 168 females, with an average age of (56.98 ± 5.23) years.

2.2. IgG *N*-glycome profiling

The IgG *N*-glycome profiling was performed on samples collected from the BHAS-302 cohort according to the protocol followed by Yu et al. [20], Pučić et al. [29], and Menni et al. [30].

2.2.1. Isolation of IgG from plasma

IgG was isolated using protein-G monolithic plates [29]. First, equilibrated protein-G monolithic plates were washed. Then, 90 μL of plasma was diluted 10 \times with the binding buffer (1 \times phosphate-buffered saline, pH 7.4), applied to the protein-G plate, and washed instantly. Finally, IgGs were eluted with 1 mL of 0.1 mol·L⁻¹ formic acid and immediately neutralized with 1 mol·L⁻¹ ammonium bicarbonate.

2.2.2. *N*-glycan release and labeling

The IgG *N*-glycans release and labeling methods were performed following Menni et al. [30]. Briefly, the isolated IgG samples were dried and denatured with the addition of 20 μL 2% sodium dodecyl sulfonate (w/v). Then, 10 μL of 4% IGEPAL CA-630 (Sigma-Aldrich, USA) and 0.5 mIU of peptide-*N*-glycosidase F in 10 μL 5 \times phosphate-buffered saline were incubated at 60 °C for 10 min and added to the samples. After adding the buffer, the samples were incubated overnight at 37 °C for *N*-glycan release. The released *N*-glycans were collected and labeled with

2-aminobenzamide, a fluorescent dye used for making glycans visible by ultra-performance liquid chromatography (UPLC) via multistage mixing with 2-aminobenzamide, dimethyl sulfoxide, glacial acetic acid, and 2-picoline borane. Labeled IgG N-glycans were cleaned and eluted by hydrophilic interaction liquid chromatography-solid phase extraction, and the combined elutes were scanned by UPLC.

2.2.3. Analysis of IgG N-glycan traits

The analysis of IgG N-glycan traits was performed following Yu et al. [20]. IgG N-glycans were split into 24 glycan chromatographic peaks (GP 1–GP 24) and quantified according to their relative contributions to individual peak (each GP was assigned a relative abundance value) to the total IgG N-glycome using hydrophilic interaction liquid chromatography on a Waters ACQUITY UPLC instrument (Waters Corporation, USA). The traits of IgG N-glycans in each peak were determined using mass spectrometry and represented using alpha numeric characters (Tables S1 and S2 in Appendix A). The minor GP, GP3, was excluded from all calculations, as, in some samples, it co-eluted with a contaminant that significantly affected its value.

In addition to basic IgG N-glycan traits (GP 1–GP 24), 54 derived GPs were calculated based on the abundance of these basic GPs. The derived GPs were referred to as “IGPs.” The R package glycanr and its iudt function with default settings were used to calculate IGPs. The structural characteristics of basic GPs and the equations for calculating the IGPs are shown in Tables S1 and S2 [31].

After profiling, an individual's IgG N-glycome was represented by a p -dimension vector $\mathcal{G}_i = [g_1, g_2, \dots, g_p]$, $i = 1, 2, \dots, N$, where g_p represented the abundance of the p th IgG N-glycan trait (or derived trait) in the i th individual out of N individuals.

2.3. Transcriptome profiling

2.3.1. Messenger RNA (mRNA) library preparation and RNA sequencing

Library preparation and RNA sequencing were performed at the Busselton Population Medical Research Institute (BPMRI). The Illumina NextSeq 500 platform (USA) was used to perform paired-end 150 bp RNA sequencing. The amount of input RNA varied from 0.6 to 2 μ g. RNA quality was checked using a LabChip GX Bioanalyzer (Caliper, USA), and samples with RNA integrity number (RIN) > 7 were included. Library preparation was performed according to the low-sample protocol of the TruSeq Stranded mRNA SamplePrep Guide 15031047_E (Illumina). Index sequences were added during library preparation to allow multiplexing. The resulting complementary DNA (cDNA) libraries were quality-checked for size and purity using a LabChip GX Bioanalyzer and quantified using a Qubit dsDNA BR Assay kit (Thermo Fisher Scientific, USA). Libraries were normalized, and pools comprising ten libraries were created. The pools were denatured and diluted according to the Denature and Dilute Libraries Guide 15048776_v02 protocol (Illumina) and then loaded onto a high-output flow cell. Automated sequencing was performed using an Illumina NextSeq 500 according to the NextSeq 500 System Guide 15046563_v02 protocol, yielding approximately 30 million paired-end reads per sample. FASTQ data were stored in BaseSpace (Illumina).

2.3.2. Bioinformatics analysis

Salmon [32] was used to quantify the gene expression levels in clean reads. The reference was Gencode.v38 transcriptome. The parameter “-l A” and flags “--validateMappings,” “--seqBias,” and “--gcBias” were passed to Salmon, whereas the other parameters and flags were default. The transcripts per kilobase million output values were summed to the gene level, and only mRNAs were

extracted for downstream analyses. After the bioinformatics analysis, an individual's IgG N-glycome was represented by a q -dimension vector, $\mathcal{T}_i = [t_1, t_2, \dots, t_q]$, $i = 1, 2, \dots, N$, where t_q represented the expression level of the q th gene in the i th individual out of N individuals. Lowly expressed genes were excluded.

2.4. Association analysis between age and IgG N-glycome and transcriptome

The feature sets were two omics data, IgG N-glycome g , and transcriptome t , and the concatenated/integrated features set was $\mathbf{c} = [g_1, \dots, g_p, t_1, \dots, t_q]$. Univariate linear regression controlling for age and sex was conducted on all features in \mathbf{c} , where the response was log-transformed feature abundance, and independent variables were age and sex.

$$\log_{10}(\text{abundance} + 1) = \beta_0 + \beta_1 \text{age} + \beta_2 \text{sex} \quad (1)$$

For a specific feature, the coefficient β_1 represents the effect of CA, β_0 is the intercept, and β_2 represents the effect of sex. p values were adjusted using the Benjamin–Hochberg method.

2.5. Least angle regression (LARS) for feature selection and age prediction

2.5.1. LARS algorithm

LARS is an efficient algorithm to optimize least absolute shrinkage and selection operator (LASSO) regression [33]. The LASSO–LARS model was used for feature selection and age prediction in this study. The most significant advantage of the LARS algorithm is its high computational efficiency. It suits the situation of “ $p \gg n$,” that is, when the number of features p is significantly greater than the sample size n [34]. However, the LARS algorithm is sensitive to noise [33], requiring noisy features to be excluded via feature selection.

2.5.2. Bootstrap-based feature selection

We used a bootstrap-based feature selection method to extract the most relevant features and construct a predictive model. This method is called Lasso Bootstrap feature selection (LB-FS). The LB-FS consists of three main steps—① feature ranking; ② forward feature selection (FFS); and ③ feature refinement. Steps ① and ② are referred to as LB-FFS, and step ③ is referred to as LB-FR.

① Feature ranking. A bootstrap sampling strategy was used to rank features. It works as follows: we assumed that the dataset has n samples, denoted as $\mathcal{D} = (x_i, y_i)$, $i = 1, 2, \dots, n$. The full features set contains p features of the i th sample, which is $\mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{ip}]$. Randomly sampling the entire dataset with replacement T times generated T sub datasets. The t th ($t \leq T$) sub dataset is $\mathcal{D}_t = (x_j, y_j)$, $j = 1, 2, \dots, m$, where $m = \lceil \alpha n \rceil$, α is the sampling rate $\alpha \in [0, 1]$. In this study, $\alpha = 0.9$ was used. Each \mathcal{D}_t value was fed into the model (LASSO–LARS) for feature selection. The model outputs coefficients set was $\beta_t = [\beta_{t1}, \beta_{t2}, \dots, \beta_{tp}]$ for \mathcal{D}_t . Owing to the shrinkage coefficient of LASSO–LARS, the coefficients of the less relevant features are zero. A vector $\mathbf{S}_t = [s_{t1}, s_{t2}, \dots, s_{tp}]$ represents the feature selection status for \mathcal{D}_t , where $s_{tp} \in \{0, 1\}$. Here, 1 implies that the p th feature is selected, and 0 implies that it is not selected. The total selected times \mathbf{S} can be calculated as follows:

$$\mathbf{S} = [s_1, s_2, \dots, s_p] = \left[\sum_{t=1}^T s_{t1}, \sum_{t=1}^T s_{t2}, \dots, \sum_{t=1}^T s_{tp} \right] \quad (2)$$

where s_p is the total selected time (TST) across t sub datasets of the p th feature. The features can be ranked according to \mathbf{S} , that is, the feature with higher TST ranks higher. This step is comparable with the classical feature-ranking methods. For example, in correlation-

based ranking, $\mathbf{C} = [c_1, c_2, \dots, c_p]$ represents the features' correlations to the response, and features having higher correlations rank higher. The “top- n features” refer to the n features having the highest metrics (e.g., correlation and TST).

② Forward feature selection. After obtaining the TST-based feature ranking, FFS was used to select the optimal feature set. Initially, the selected feature set was $\mathcal{X} = \emptyset$. The FFS iterates K times. At each iteration, the top feature is selected without replacement and is added to \mathcal{X} . In the k th iteration, $\mathcal{X}_k = \{x_{\text{top-1}}, x_{\text{top-2}}, \dots, x_{\text{top-k}}\}$, and the corresponding dataset $\mathcal{D}_k = (\mathcal{X}_{ki}, \mathcal{Y}_i)$, $i = 1, 2, \dots, n$. After each selection, a LASSO–LARS model (an independent model different from that used for feature ranking) was evaluated on \mathcal{D}_k with f -fold cross-validation (CV). At the f th fold, the \mathcal{D}_k is split into a training set, $\mathcal{D}_{\text{train}}^{f-k} = (\mathcal{X}_{kl}, \mathcal{Y}_l)$, and a testing set, $\mathcal{D}_{\text{test}}^{f-k} = (\mathcal{X}_{kh}, \mathcal{Y}_h)$, where l and h are the sample numbers in training and testing sets. After splitting, the LASSO–LARS model is trained on $\mathcal{D}_{\text{train}}^{f-k}$ and predicts on $\mathcal{D}_{\text{test}}^{f-k}$, resulting in the estimations $\widehat{\mathbf{y}}_{kf} = [\widehat{y}_{f1}, \widehat{y}_{f2}, \dots, \widehat{y}_{fn}]$ of the actual values of the responses $\mathbf{y}_{kf} = [y_{f1}, y_{f2}, \dots, y_{fn}]$. The estimations from f folds are concatenated and compared with the true values to calculate the evaluation metrics \mathbf{r} (e.g., R^2). In this study, 10-fold CV was used, that is $f = 10$. After iterating the FFS for K times, we get a set of FFS CV metrics $\mathbf{r} = [r_1, r_2, \dots, r_K]$. The selected features at the k th iteration, \mathcal{X}_k that achieved the highest metrics, were regarded as the optimal feature set \mathcal{X}^* .

$$\mathcal{X}^* = \underset{\mathcal{X}_k}{\operatorname{argmax}} E(\mathbf{y}_k, \widehat{\mathbf{y}}_k) \quad (3)$$

The performance of the model was recorded at each step, and a curve representing the performance change along with the added top features was drawn from the FFS process, referred to as the FFS curve.

③ Feature refinement. Nested CV can be used to refine the optimal feature set in the feature refinement step. We start with the dataset with the optimal feature set $\mathcal{D}^* = \{\mathcal{X}^*, \mathcal{Y}\}$, where \mathcal{X}^* represents the optimal feature set and \mathcal{Y} represents the response set. A bootstrap sampling method was adopted where the sample size was equal to U . At the u th ($u \leq U$) iteration, the sampled sub dataset was $\mathcal{D}_u^* = (x_j^*, y_j)$, $j = 1, 2, \dots, m$, where $m = \lceil \lambda n \rceil$, n is the sample number of the entire dataset, and λ is the sampling rate $\lambda \in [0, 1]$. G -fold CV was conducted on \mathcal{D}_u^* . At the g th fold, \mathcal{D}_u^* was split into a training set, $\mathcal{D}_{\text{train-g}}^*$, and a testing set, $\mathcal{D}_{\text{test-g}}^*$. The LASSO–LARS model was trained using the training set $\mathcal{D}_{\text{train-g}}^*$ and evaluated using the testing set $\mathcal{D}_{\text{test-g}}^*$. This training–testing process was repeated for fine-tuning (a grid search strategy), and when the evaluation result achieved the highest score, the hyperparameters were regarded as optimal, and the corresponding trained parameters were $\beta_u = [\beta_{u1}, \beta_{u2}, \dots, \beta_{uq}]$, $q < p$. $\beta_{uq} = 0$ implies that the corresponding feature x_{uq} is not selected. Hence a set representing the feature selection status can be $\mathbf{R}_u = [r_{u1}, r_{u2}, \dots, r_{uq}]$, $r_{uq} \in \{0, 1\}$, where 1 implies selection and 0 implies non-selection. After iterating the bootstrapping U times, we obtained the overall feature-selected set \mathbf{R} .

$$\mathbf{R} = [r_1, r_2, \dots, r_q] = \left[\sum_{u=1}^U r_{u1}, \sum_{u=1}^U r_{u2}, \dots, \sum_{u=1}^U r_{uq} \right] \quad (4)$$

$B(x_q) = r_q$ is the TST of this bootstrapping process of the q th feature x_q .

The q th feature is included in the final refined features set \mathcal{X}^{**} only when $r_q > \delta U$, $\delta \in [0, 1]$, where δ is an important hyperparam-

eter affecting the refined features set. In this study, $\delta = 0.9$ was used.

$$\forall \mathbf{x} \in \mathcal{X}^{**} (B(\mathbf{x}) > \delta U), \delta \in [0, 1] \quad (5)$$

2.5.3. Predict age from transcriptome and IgG N-glycome separately

Two predictive models were built from IgG N-glycome and transcriptome separately, and the predictive feature sets were denoted as GLYC and TRANS. The LB-FS method was used for feature selection, and the refined feature set was used as the final feature set to build the predictive models. The performance of the models was evaluated using a 10-fold CV. R^2 and Pearson's correlations were used to evaluate performance of the models. The predicted and actual ages from 10-fold CV were combined into two vectors to compute R^2 and Pearson's correlations.

For comparison, a correlation-based method was also used in this study. Briefly, the correlation and p value from univariate regression controlling for sex between each feature (gene and glycan traits) and response (age) were used for feature ranking. Features with larger absolute correlation coefficients and smaller p values were ranked higher. The other steps of the LB-FS were identical to those of the original LB-FS. The results from the FFS step of LB-FS were denoted as LB-FFS, Pearson-FFS, and LRp-FFS, representing the results of original TST ranking, Pearson correlation-based ranking, and p value of the linear regression-based ranking, respectively.

2.6. Multiomics integration with AlphaSnake

First, a simple concatenation-based method was used to integrate the transcriptome and IgG N-glycans. The concatenated/integrated features set (denoted as CONC) was $\mathbf{c} = [g_1, \dots, g_p, t_1, \dots, t_q]$, where p and q were the numbers of IgG N-glycan traits and genes. The LB-FS method was applied to the concatenated vector \mathbf{c} .

To improve age prediction using integrated multiomics data, we developed a DRL-based multiomics integration method, AlphaSnake, which has been discussed in the following section.

2.6.1. AlphaSnake algorithm

The algorithm hypothesized that “the integration of the most predictive features of each omics could be the optimal predictive feature set.” First, we need to determine the most predictive features of each omics data, that is, feature ranking, resulting in two ranked feature sets—ranked IgG N-glycome profile, $\mathbf{g}_{\text{ranked}} = [g_{\text{top-1}}, g_{\text{top-2}}, \dots, g_{\text{top-p}}]$, and ranked transcriptome profile, $\mathbf{t}_{\text{ranked}} = [t_{\text{top-1}}, t_{\text{top-2}}, \dots, t_{\text{top-q}}]$. In AlphaSnake, a FFS strategy is adopted, where the candidate integrated feature set (IS) \mathbf{c} moves forward for k steps. At each step, either the top feature from $\mathbf{g}_{\text{ranked}}$ or $\mathbf{t}_{\text{ranked}}$ feature is selected (without replacement) and put into \mathbf{c} . The forward process iterates until the stopping condition is satisfied at the k th step. The integrated set $\mathbf{c} = [g_{\text{top-1}}, g_{\text{top-2}}, t_{\text{top-1}}, g_{\text{top-3}}, t_{\text{top-2}}, \dots]$ contains k top features from both ranked feature sets. If the stop condition is when the performance of the model built from \mathbf{c} stops increasing, then the integrated set can be regarded as the optimal integrated set \mathbf{c}^* . Fig. 1 shows an example of the workflow of the AlphaSnake algorithm. At Step 0, the IgG N-glycome and transcriptome are ranked, and the IS is an empty set. At Step 1, the IS selects the top-1 feature from the IgG N-glycome set. In Steps 2 and 3, the IS then selects the top-1 and top-2 features from the transcriptome set. This workflow shows only the first four steps of a possible sequence of actions. This FFS process can be regarded as a problem when searching

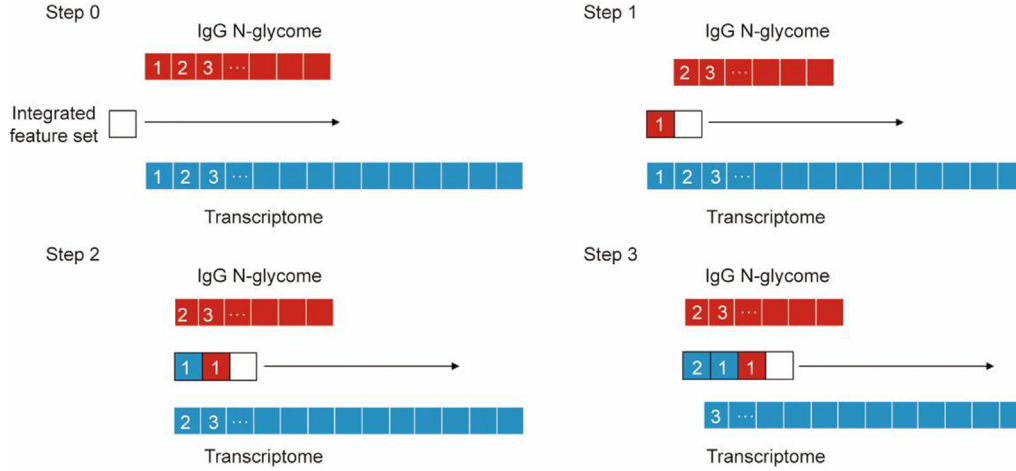


Fig. 1. The general FFS workflow of the AlphaSnake algorithm.

for optimal sequences of actions. Reinforcement learning (RL) is an ideal method for solving this problem, wherein an intelligent agent can be trained to take the optimal action sequence, allowing the predictive model (built from it) to achieve optimal performance.

A value-based RL algorithm, deep Q learning (DQN), was adopted. A Markov decision process represents the DQN, $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \text{Tr}, R, \gamma, \mathcal{O}\}$, where \mathcal{S} represents the state space, \mathcal{A} represents the action space, Tr represents the transition function, R represents the reward function, $\gamma \in [0,1]$ is the discount factor for calculating the accumulative rewards G , and \mathcal{O} is a buffer set named “replay memory” that stores the observed transitions, and helps improve the training processes of DQN. The parameters are as follows:

State space (\mathcal{S}). The state is defined as a set $\mathcal{S} = \{\mathbf{s}_k\}$, representing the selection status (selected/unselected) of combined candidates features from IgG glycome and transcriptome, where $\mathbf{s}_k = [f_1, f_2, \dots, f_h]$, $h = p + q$ represents the state at the k th step, and $f_h \in \{0, 1\}$ implies the h th candidate feature’s selection status, selected(1)/unselected(0). At the beginning of the FFS, all the elements in \mathcal{S} were zero.

Action space (\mathcal{A}). Action refers to the k th step, the top feature, from which omics are selected into \mathbf{c} . It is defined as a one-dimension vector $\mathcal{A} = [a_1, a_2, \dots, a_k]$, $a \in \{0, 1\}$, where $a_k = 0$ implies that, at the k th step, the top feature from $\mathbf{g}_{\text{ranked}}$ is selected into \mathbf{c} , whereas $a_k = 1$ implies that the top feature is from $\mathbf{t}_{\text{ranked}}$. As only two omics were used in this study, a_k could be a binary scale.

Transition function (Tr). At the k th FFS iteration, an action is performed, the state is transited, and the reward is $(\mathbf{s}_{k+1}, r_k) = \text{Tr}(\mathbf{s}_k, a_k)$. By acting, the specific element $f_h \in \mathbf{s}$ representing the corresponding selected feature $c_h \in (\mathbf{g} \cup \mathbf{t})$ changes from 0 to 1, implying that this feature’s selection status has changed from “unselected” to “selected.”

Reward function (R). The reward function is a part of the transition. It takes the state as input and outputs the rewards r , $r_k = R(\mathbf{s}_k, a_k)$, and the accumulative reward (return) as $G_k = \sum_{u=k}^K \gamma^{u-k} r_u$. The function R is from a deterministic environment, where the selected features are represented by \mathbf{s}_k , is used to build a predictive model, and 10-fold CV is used to evaluate the model and output the reward r_k —a scale reflecting the performance of the model. The metrics R^2 are used to evaluate the performance.

Replay memory (\mathcal{O}). The replay memory \mathcal{O} is a fixed-length buffer set that stores the observed transitions $(\mathbf{s}_{k+1}, r_k, \mathbf{s}_k, a_k)$. The agent randomly selects transitions from \mathcal{O} to update its policies. Initially, the agent focuses on exploring. Thus, the transitions

stored in \mathcal{O} are generated from random actions. At every training epoch, the resulting transitions are pushed into \mathcal{O} to gradually replace the random transitions. To continue exploration, the agent has a probability P_{rand} to take random actions, where P_{rand} decreases with increasing update times. As training epochs increase, the agent cares more about exploitation, which implies that the P_{rand} decreases, making the portion of observed transitions in \mathcal{O} larger. Thus, the training process becomes more stable. This strategy is known as the epsilon-greedy policy.

The environment is deterministic, that is, one action results in a deterministic transition. The aim is to train the agent to achieve the highest accumulative reward G_k from k consecutive actions (i.e., the FFS process). The principal idea of Q-learning is to find the value function, that is, the Q function, which outputs the value (expected return) of each action/state pair, so that the action achieving the highest value for the corresponding state can be chosen. π represents the agent’s policy, and the optimal policy π^* is:

$$\pi^*(\mathbf{s}) = \underset{a}{\operatorname{argmax}} Q^*(\mathbf{s}, a) \quad (6)$$

The Q function can be approximated with a deep neural network D .

$$D(\mathbf{s}) = \mathbf{w}^{\text{Tr}} \mathbf{s} + b \quad (7)$$

where \mathbf{w} is the weight vector.

The Q function considers the state as an input and outputs the value for each action. Hence, the Q function can be written as follows:

$$Q(\mathbf{s}_k, a) = D(\mathbf{s})[a] \quad (8)$$

where $[a]$ indicates $D(\mathbf{s})$ ’s output dimension of the corresponding action a , that is, the value of performing action a under state \mathbf{s} . The Q -function update process can be represented by the Bellman equation:

$$Q^\pi(\mathbf{s}_k, a_k) = r_k + \gamma Q^\pi(\mathbf{s}_{k+1}, \pi(\mathbf{s}_{k+1})) \quad (9)$$

where r_k is the immediate reward received after an action, and γ is the discount factor that weighs the importance of future rewards relative to immediate rewards. The difference between two consecutive steps is the temporal difference error δ_t .

$$\delta_t = Q^\pi(\mathbf{s}_k, a_k) - \left[r_k + \gamma \max_a Q^\pi(\mathbf{s}_{k+1}, a) \right] \quad (10)$$

The Huber loss, L , is used for minimizing the δ_t .

$$L(\delta_t) = \begin{cases} \frac{1}{2} \delta_t^2, & |\delta_t| \leq 1 \\ |\delta_t| - \frac{1}{2}, & \text{otherwise} \end{cases} \quad (11)$$

Before starting the training, we define an epsilon-greedy policy. A dynamic threshold ϵ is defined as:

$$\epsilon = \epsilon_{\text{end}} + (\epsilon_{\text{start}} - \epsilon_{\text{end}}) \exp\left(-\frac{t}{\epsilon_{\text{decay}}}\right) \quad (12)$$

where ϵ_{start} , ϵ_{end} , and ϵ_{decay} are three hyperparameters, and t is the step where the Q function has been updated. At each time when the Q function is updated, a random number $z \in [0, 1]$ is generated and used to guide whether action is being taken randomly:

$$a_k = \begin{cases} \arg \max_a Q(\mathbf{s}_k, a), z > \epsilon \\ \text{rand}(\{0, 1\}), \text{otherwise} \end{cases} \quad (13)$$

where $\text{rand}(\{0, 1\})$ represents a randomly selecting element from the set $\{0, 1\}$.

During training, at each update step, $Q(\mathbf{s}_t, a_t)$, r_k , and $\max_a Q(\mathbf{s}_{k+1}, a)$ are computed to calculate the Huber loss, L , and update the Q -function parameters. A target network is added to compute $\max_a Q(\mathbf{s}_{k+1}, a)$ for added stability, where its parameters are kept frozen most of the time, but they are updated with the policy network's parameters at specific epoch μ —a hyperparameter.

At the beginning of model training, we pre-define the maximum epoch number E and reset the environment, and initialize the state, where the elements of the initial state are zero. Then, at each epoch $e \in E$, we start the training loop. At each step within an epoch, we sample a batch of actions, execute them, observe the corresponding subsequent states and rewards, push the transitions into the replay memory, and update the model. When the loop meets the stop criterion, the epoch ends, and the best reward, r_e^* , is recorded. It should be noted that r_e^* has the best performance from an epoch and not the accumulative reward. The idea of an accumulative reward is used to train the Q function with the temporal difference algorithm.

In AlphaSnake, a DQN is used to train an intelligent agent, which maximizes the return of the action sequence from each training epoch of the DQN. The action sequence that achieves the best prediction is selected, and the state that makes the model achieve the best prediction is the optimal state, that is, the optimal selection status vector. According to the optimal state, all features, labeled as “selected,” are the final selected/IS.

2.6.2. Age prediction with the integrated multiomics data

First, feature ranking of LB-FS was performed on the simply concatenated \mathbf{c} to generate a ranked IS $\mathbf{c}_{\text{ranked}} = [c_1, c_2, \dots, c_{p+q}] = [g_{\text{top-1}}, g_{\text{top-2}}, t_{\text{top-1}}, \dots, g_{\text{top-p}}, t_{\text{top-q}}]$, where c_{p+q} represented a feature either from IgG N-glycome profile or transcriptome profile, and $g_{\text{top-p}}$ and $t_{\text{top-q}}$ represented their relative ranks in corresponding profiles. The ranked feature sets for IgG N-glycome and transcriptome profiles were extracted from $\mathbf{c}_{\text{ranked}}$ according to their relative ranks, $\mathbf{g}_{\text{ranked}} = [g_{\text{top-1}}, g_{\text{top-2}}, \dots, g_{\text{top-p}}]$ and $\mathbf{t}_{\text{ranked}} = [t_{\text{top-1}}, t_{\text{top-2}}, \dots, t_{\text{top-q}}]$. The FFS in the LB-FS strategy was performed on $\mathbf{g}_{\text{ranked}}$, $\mathbf{t}_{\text{ranked}}$, and $\mathbf{c}_{\text{ranked}}$, resulting in selected GLYC, TRNAS, and CONC feature sets separately, and their corresponding optimal FFS performance can be denoted as R^{2*} .

Then, the AlphaSnake algorithm was conducted using $\mathbf{g}_{\text{ranked}}$ and $\mathbf{t}_{\text{ranked}}$ to identify the optimal IS $\mathbf{c}_{\text{rl}}^* = [g_{\text{top-1}}, g_{\text{top-2}}, t_{\text{top-1}}, \dots]$. The feature set selected by AlphaSnake was denoted as the feature set from the trajectory that achieved the highest optimal FFS R^{2*} (CONC-RL).

After performing AlphaSnake, the feature refinement in LB-FS was adopted on the FFS-selected GLYC, TRANS, CONC, and CONC-RL, resulting in the final refined feature sets, GLYC, TRANS, CONC,

and CONC-RL. The ages predicted from the final refined feature sets were used as aging clocks. The performance of the models from the refined feature set is denoted as R^{2**} .

The performance of the models was evaluated using a 10-fold CV. R^2 , including R^{2*} and R^{2**} , and the Pearson's correlation (only for models from the refined feature sets) calculated between the predicted and actual ages, were used to evaluate the performance. The predicted and actual ages from 10-fold experiments were combined into two vectors to compute R^2 and Pearson's correlation. The performances of AlphaSnake and concatenation were compared using a paired t -test on the results of the 10-fold CV.

For comparison, Pearson's correlation-based and univariate linear regression-based FFS were performed. Briefly, the Pearson's correlation and p value from univariate regression were used for feature ranking. Features with larger absolute correlation coefficients and smaller p values were ranked higher. The FFS was the same as the FFS step in LB-FS. The results from the FFS based on different rankings are denoted as LB-FFS, Pearson-FFS, and LRp-FFS.

2.7. Feature importance and enrichment analysis

To identify the most important predictive features, the Python package, *shap*, was used to calculate the Shap values of features. Enrichment analyses were performed using the GSEAPy software. The latest Kyoto Encyclopedia of Genes and Genomes (KEGG) and gene ontology (GO) gene sets were used as references. Statistical significance for enriched pathways was set at false discovery rate (FDR) < 0.05 .

2.8. Evaluation of the additive information of aging clocks

The age predicted from a specific feature set can be regarded as a metric, that is, the aging clock, parallel to CA. The additive values of the three aging clocks were evaluated as follows: ① gAge—age predicted from the optimal GLYC feature set, ② tAge—age predicted from the optimal TRANS feature set, and ③ gtAge—age predicted from the optimal CONC-RL feature set. The corresponding delta ages for gAge, tAge, and gtAge are denoted as delta_gAge, delta_tAge, and delta_gtAge. Associations between delta age and biological parameters (e.g., blood pressure, cholesterol level, and glucose level) were explored using linear regression. The dependent variable was the phenotype of interest, the independent variable was delta age, and associations were adjusted for CA.

3. Results

3.1. Forty-three IgG N-glycan traits and six genes were associated with CA

Univariate association analyses showed that 43 IgG N-glycan traits and 6 genes were significantly associated with CA (FDR < 0.25). Most IgG glycan traits were more significant than genes according to raw p values and could explain more variability according to the adjusted R^2 . The statistics of the 49 age-associated features are presented in Table 1, and the statistics of all features are presented in Table S3 in Appendix A. The associations between the CA and top four age-associated features are shown in Fig. 2 (all 49 features are shown in Fig. S1 in Appendix A).

3.2. Integration of selected genes and IgG N-glycan traits can accurately predict age

We used LB-FS to select the most predictive features from the GLYC, TRANS, and CONC. FFS results of the original LB-FS strategy are shown in Fig. 3(a).

Table 1
Features significantly associated with age.

Feature	Description	Coefficient	CI [2.5%]	CI [97.5%]	F	Adjusted R ²	p	FDR
IGP53_glycan	FA2G2/total neutral glycans	-0.010	-0.012	-0.007	42.345	0.216	3.40 × 10 ⁻¹⁷	2.12 × 10 ⁻¹³
GP14_glycan	FA2G2	-0.009	-0.010	-0.007	42.113	0.215	3.77 × 10 ⁻¹⁷	2.12 × 10 ⁻¹³
GP6_glycan	FA2B	0.009	0.007	0.011	38.038	0.197	1.94 × 10 ⁻¹⁶	7.27 × 10 ⁻¹³
IGP45_glycan	FA2B/total neutral glycans	0.008	0.006	0.010	37.110	0.194	4.14 × 10 ⁻¹⁶	9.79 × 10 ⁻¹³
IGP57_glycan	Digalactosyl glycans/total neutral glycans	-0.009	-0.011	-0.007	39.950	0.206	4.36 × 10 ⁻¹⁶	9.79 × 10 ⁻¹³
IGP55_glycan	Agalactosylated glycans/total neutral glycans	0.007	0.006	0.009	39.627	0.204	1.10 × 10 ⁻¹⁵	2.06 × 10 ⁻¹²
GP18_glycan	FA2G2S1	-0.007	-0.009	-0.005	30.703	0.165	1.78 × 10 ⁻¹³	2.85 × 10 ⁻¹⁰
GP4_glycan	FA2	0.008	0.006	0.010	31.284	0.168	3.13 × 10 ⁻¹²	4.40 × 10 ⁻⁹
IGP43_glycan	FA2/total neutral glycans	0.007	0.005	0.009	30.220	0.163	9.81 × 10 ⁻¹²	1.22 × 10 ⁻⁸
IGP76_glycan	FG2 ⁿ /(BG2 ⁿ + FBG2 ⁿ)	-0.005	-0.006	-0.003	22.818	0.127	9.34 × 10 ⁻¹¹	1.05 × 10 ⁻⁷
IGP26_glycan	FGS/(F + FG + FGS)	-0.005	-0.006	-0.003	21.168	0.118	6.32 × 10 ⁻¹⁰	6.36 × 10 ⁻⁷
IGP77_glycan	BG2 ⁿ /(FG2 ⁿ + FBG2 ⁿ)	0.006	0.004	0.008	20.340	0.114	6.80 × 10 ⁻¹⁰	6.36 × 10 ⁻⁷
IGP69_glycan	FBG2 ⁿ /G2 ⁿ	0.004	0.003	0.005	19.069	0.107	2.40 × 10 ⁻⁹	2.07 × 10 ⁻⁶
IGP75_glycan	FBG2 ⁿ /(FG2 ⁿ + FBG2 ⁿ)	0.004	0.003	0.006	18.979	0.107	3.15 × 10 ⁻⁹	2.53 × 10 ⁻⁶
IGP74_glycan	FBG2 ⁿ /FG2 ⁿ	0.001	0	0.001	16.918	0.096	2.00 × 10 ⁻⁸	1.50 × 10 ⁻⁵
GP1_glycan	FA1	0.001	0	0.001	19.736	0.111	2.76 × 10 ⁻⁸	1.94 × 10 ⁻⁵
IGP56_glycan	G1 ⁿ	-0.002	-0.002	-0.001	15.504	0.088	1.17 × 10 ⁻⁷	7.74 × 10 ⁻⁵
IGP41_glycan	FA1/total neutral glycans	0.001	0.000	0.001	17.460	0.099	2.14 × 10 ⁻⁷	1.34 × 10 ⁻⁴
IGP36_glycan	FBS ^{total} /F ^S ^{total}	0.001	0.001	0.002	14.778	0.084	2.68 × 10 ⁻⁷	1.58 × 10 ⁻⁴
FKBP1B	FKBP prolyl isomerase 1B	0.004	0.002	0.006	13.474	0.077	1.42 × 10 ⁻⁶	7.96 × 10 ⁻⁴
IGP38_glycan	FBS1/(FS1 + FBS1)	0.005	0.003	0.007	11.848	0.067	2.36 × 10 ⁻⁶	1.26 × 10 ⁻³
IGP37_glycan	FBS1/FS1	0.001	0.001	0.001	11.891	0.067	2.69 × 10 ⁻⁶	1.37 × 10 ⁻³
IGP66_glycan	FG2 ⁿ /G2 ⁿ	0.004	0.002	0.005	12.954	0.074	3.11 × 10 ⁻⁶	1.52 × 10 ⁻³
IGP71_glycan	FB ⁿ /F ⁿ ^{total}	0.004	0.002	0.005	12.549	0.071	5.98 × 10 ⁻⁶	2.80 × 10 ⁻³
IGP54_glycan	FA2BG2/total neutral glycans	-0.003	-0.005	-0.002	14.098	0.080	7.06 × 10 ⁻⁶	3.17 × 10 ⁻³
IGP72_glycan	F ⁿ /(B ⁿ + FB ⁿ)	-0.004	-0.005	-0.002	12.305	0.070	8.20 × 10 ⁻⁶	3.54 × 10 ⁻³
IGP70_glycan	FB ⁿ /F ⁿ	0.004	0.002	0.006	12.066	0.068	8.78 × 10 ⁻⁶	3.65 × 10 ⁻³
IGP47_glycan	FA2[G]G1/total neutral glycans	-0.002	-0.003	-0.001	9.916	0.056	2.16 × 10 ⁻⁵	8.67 × 10 ⁻³
IGP48_glycan	FA2G1/total neutral glycans	-0.003	-0.004	-0.001	10.148	0.057	2.24 × 10 ⁻⁵	8.67 × 10 ⁻³
TSPAN15	Tetraspanin 15	-0.005	-0.007	-0.003	9.550	0.054	2.72 × 10 ⁻⁵	1.02 × 10 ⁻²
LTK	Leukocyte receptor tyrosine kinase	-0.008	-0.012	-0.004	9.052	0.051	6.96 × 10 ⁻⁵	2.52 × 10 ⁻²
IGP40_glycan	FBS2/(FS2 + FBS2)	0.002	0.001	0.003	13.825	0.079	7.61 × 10 ⁻⁵	2.67 × 10 ⁻²
GP11_glycan	FA2BG1	0.001	0.001	0.002	8.532	0.048	8.81 × 10 ⁻⁵	3.00 × 10 ⁻²
GP15_glycan	FA2BG2	-0.003	-0.004	-0.001	11.133	0.063	1.04 × 10 ⁻⁴	3.43 × 10 ⁻²
IGP27_glycan	FBGS/(FB + FBG + FBGS)	-0.004	-0.006	-0.002	7.694	0.043	1.16 × 10 ⁻⁴	3.73 × 10 ⁻²
IGP65_glycan	FG2 ⁿ /G2 ⁿ	-0.001	-0.001	0	9.233	0.052	1.21 × 10 ⁻⁴	3.76 × 10 ⁻²
IGP33_glycan	F ^{total} S1/F ^{total} S2	-0.002	-0.004	-0.001	7.877	0.044	1.84 × 10 ⁻⁴	5.58 × 10 ⁻²
IGP51_glycan	A2G2/total neutral glycans	-0.004	-0.007	-0.002	11.061	0.063	2.39 × 10 ⁻⁴	7.05 × 10 ⁻²
IGP39_glycan	FBS2/FS2	0.002	0.001	0.003	11.831	0.067	2.52 × 10 ⁻⁴	7.24 × 10 ⁻²
NOG	Noggin	-0.008	-0.013	-0.004	17.324	0.098	3.33 × 10 ⁻⁴	9.36 × 10 ⁻²
RORC	RAR related orphan receptor C	-0.006	-0.010	-0.003	6.884	0.038	3.67 × 10 ⁻⁴	1.01 × 10 ⁻¹
GP23_glycan	FA2G2S2	-0.003	-0.004	-0.001	7.553	0.042	4.81 × 10 ⁻⁴	1.29 × 10 ⁻¹
IGP68_glycan	FBG1 ⁿ /G1 ⁿ	0.003	0.001	0.005	7.704	0.043	6.81 × 10 ⁻⁴	1.78 × 10 ⁻¹
GP12_glycan	A2G2	-0.003	-0.006	-0.001	9.652	0.054	8.03 × 10 ⁻⁴	2.02 × 10 ⁻¹
IGP62_glycan	F ⁿ	-0.001	-0.001	0	8.099	0.045	8.19 × 10 ⁻⁴	2.02 × 10 ⁻¹
IGP30_glycan	FG2S2/(FG2 + FG2S1 + FG2S2)	0.003	0.001	0.005	9.683	0.055	8.27 × 10 ⁻⁴	2.02 × 10 ⁻¹
GP2_glycan	A2	0.004	0.002	0.006	6.136	0.033	9.71 × 10 ⁻⁴	2.28 × 10 ⁻¹
IGP24_glycan	FGS/(FG + FGS)	-0.002	-0.003	-0.001	5.543	0.029	9.79 × 10 ⁻⁴	2.28 × 10 ⁻¹
GNG11	G protein subunit gamma 11	0.005	0.002	0.008	12.494	0.071	9.93 × 10 ⁻⁴	2.28 × 10 ⁻¹

F: core fucose; A1: 1 β1–2 linked GlcNAc antenna on core; A2: 2 β1–2 linked GlcNAc antenna on core; B: bisecting GlcNAc; G: galactose; S: sialic acid; n: “neutral” glycan; CI: confidence interval; F: F-statistic.

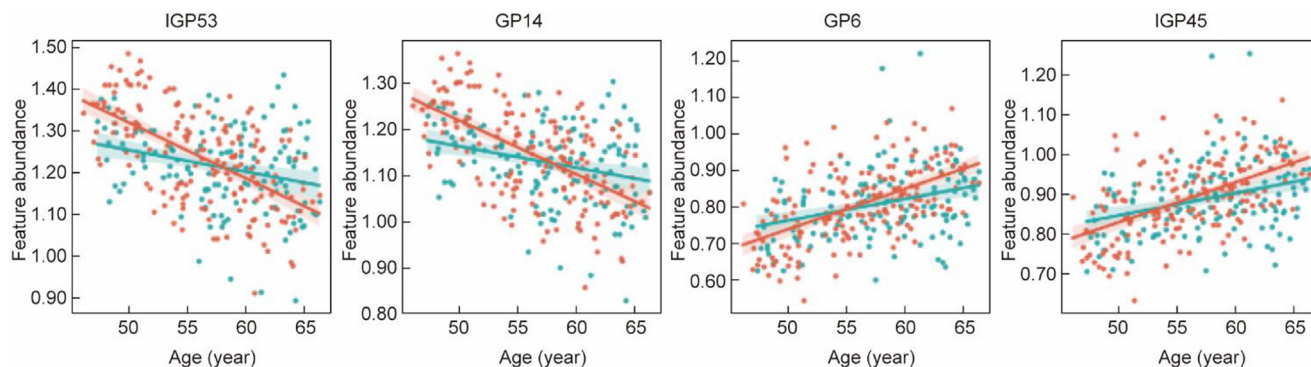


Fig. 2. Top-4 age-associated features versus age. The scatter plots and regression lines are sex-specific, with red representing females and blue representing males. The y-axis represents feature abundance, and the x-axis represents age. The study cohort consisted of 134 males and 168 females, and their average age was (56.98 ± 5.23) years.

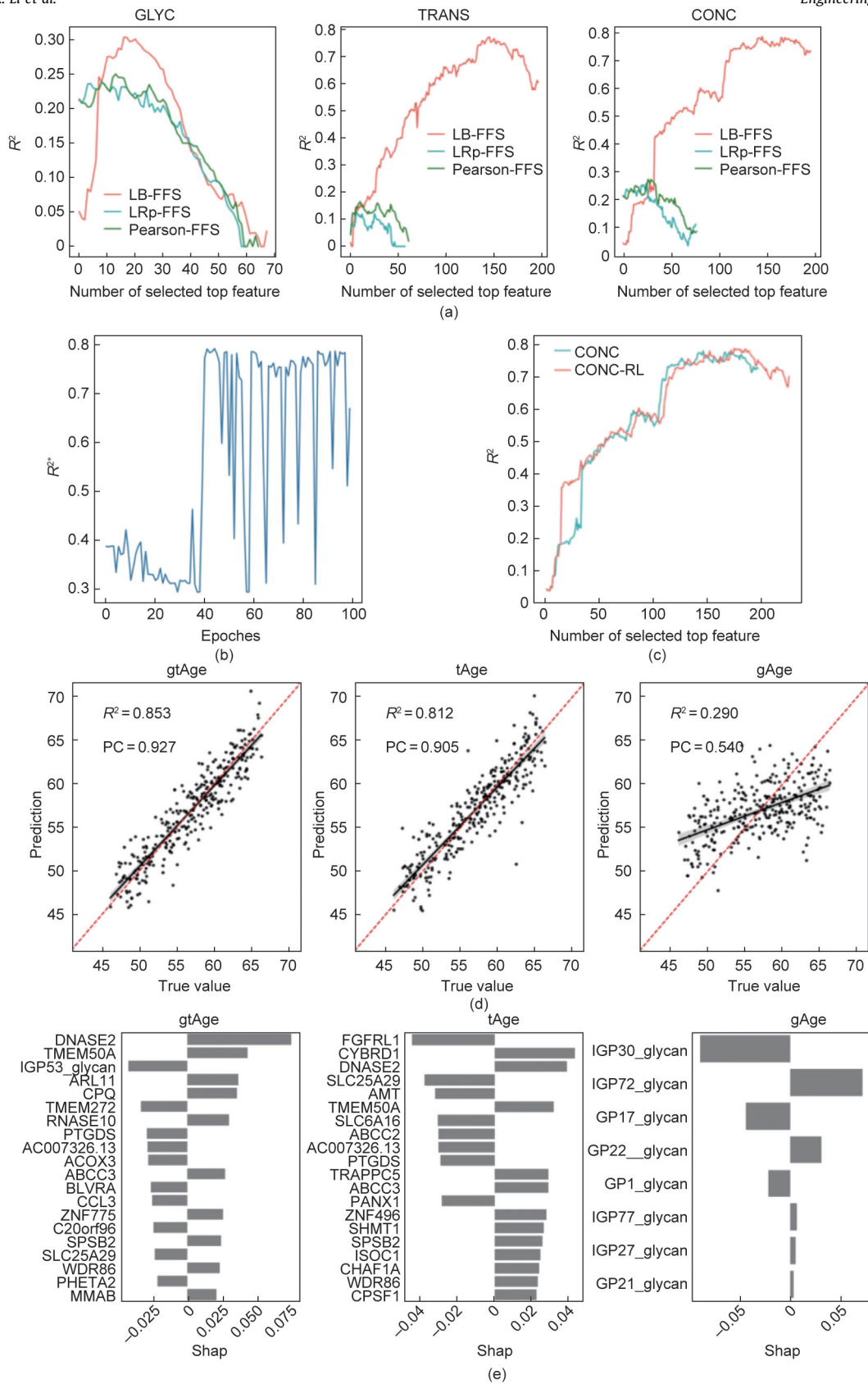


Fig. 3. (a) FFS curves for GLYC, TRANS, and CONC. The x-axes represent the number of top features selected into the models during the FFS processes; the y-axes represent the R^2 from the models built with the selected top features at each FFS step. The FFS processes stop when the performance stops increasing for consecutive 50 steps. The red curves represent the FFS curves from the LB-FS strategy. The blue curves represent the FFS curves from a feature ranking based on the p values from the association analyses. The green curves represent the FFS curves from a feature ranking based on Pearson's correlation. (b) The reward along with the training process of the AlphaSnake algorithm. The curves show the changes in optimal performances of trajectories from the AlphaSnake algorithm. (c) FFS curves from CONC-RL and CONC. The blue curve represents the FFS curves from CONC feature set and the red curve represents the FFS curve from CONC-RL feature set. (d) Predictions versus true CA for gtAge, tAge, and gAge. The x-axes represent the true CA; the y-axes represent the predicted age from different refined feature sets (i.e., aging clocks). The red dashed lines indicate the perfect prediction. (e) The top-20 features for gtAge, tAge, and gAge. The x-axes represent the mean of Shap values of features; the y-axes display the top-20 features for each aging clock (top-8 features for gAge).

The results showed that FFS in the LB-FS strategy achieved the highest optimal performance for GLYC, TRANS, and CONC. For GLYC, the optimal performance R^{2*} was approximately 0.3, indicating that age could not be accurately predicted from the IgG N-glycome alone in our cohort. The optimal performance for TRANS was $R^{2*} = 0.771$, and that for CONC was $R^{2*} = 0.784$, suggesting that the transcriptome could accurately predict age and that the IgG N-glycome might improve the prediction.

3.3. DRL-based multiomics data integration method improves age prediction from integrated multiomics data

We used the proposed AlphaSnake to select the optimal feature set, referred to as CONC-RL. The changes in optimal performances (i.e., reward) of trajectories from the AlphaSnake algorithm, along with the training epochs, are shown in Fig. 3(b).

The curve exhibited high fluctuation, a typical characteristic of model-free RL. The optimal performance of AlphaSnake-selected feature set, CONC-RL, was $R^{2*} = 0.793$, which was marginally higher than that of the model built from CONC. The results showed that AlphaSnake improved the prediction of CONC. Fig. 3(c) compares the FFS curves of CONC and CONC-RL.

Although both curves showed a comparable increasing trend, CONC-RL achieved a higher plateau and included more features than CONC, suggesting the potential of AlphaSnake in terms of delivering better performance and novel findings.

Feature refinement of LB-FS was conducted using the FFS-selected feature sets to generate the final refined feature sets to predict gtAge, tAge, and gAge, where gtAge was predicted from two refined feature sets, CONC and CONC-RL.

The results of the predictive models built using the refined feature sets are presented in Table 2.

The performance of the model for gtAge (CONC) was $R^{2**} = 0.820$, which was higher than that of the models for tAge ($R^{2**} = 0.812$) and gAge ($R^{2**} = 0.290$). The model for gtAge (CONC-RL) increased the R^{2**} by 0.33 to $R^{2**} = 0.853$, and the difference between CONC and CONC-RL was significant under a paired *t*-test of the 10-fold CV results ($p < 0.01$), demonstrating the advantage of AlphaSnake in exploring the optimally IS for age prediction. Therefore, gtAge (CONC-RL) was used as the final gtAge in the following analyses. Although refined CONC-RL and CONC contained seven glycans, those were different. IGP53 and IGP51 were unique to the refined CONC-RL, whereas IGP57 and GP1 were found only in the refined CONC. After excluding glycan traits from the refined feature sets for CONC and CONC-RL, R^{2**} for gtAge decreased significantly, indicating that glycan traits played an essential role in age prediction for the ISS. Plots of aging clocks versus CA are shown in Fig. 3(d). The results showed that tAge and gtAge were highly correlated with CA, whereas gAge was not.

3.4. Importance and enrichment analysis of predictive features

The final predictive features for gtAge, tAge, and gAge were analyzed for feature importance. The Shap values for the top-

Table 2
The prediction results from the refined feature sets.

Aging clock	R^{2**}	Num	R^{2**} w/o glycan	R^{2**} glycan	Glycans
tAge	0.812	130	0.812	0.000	N/A
gAge	0.290	8	0.000	0.290	IGP72; GP22; GP21; IGP27; IGP30; GP1; IGP77; GP17
gtAge (CONC)	0.820	135	0.598	0.274	IGP72; IGP57; IGP27; GP22; GP21; GP5; GP1
gtAge (CONC-RL)	0.853	144	0.611	0.275	IGP72; IGP27; GP22; GP21; GP5; IGP53; IGP51

R^{2**} : R^2 from the optimal feature set; Num: Number of features included in the optimal feature set; R^{2**} w/o glycan: R^2 from the optimal feature set excluding features from the IgG N-glycome set; R^{2**} glycan: R^2 from the IgG N-glycan traits in the optimal feature set; glycans: IgG N-glycan traits selected into the optimal feature set; N/A: not available.

20 features (only eight features for gAge) are shown in Fig. 3(e).

The identified genes in the final feature sets for gtAge and tAge were different, whereas the models built on both feature sets achieved $R^{2**} > 0.8$, indicating that more than one feature set could accurately predict CA. In the predictive feature set for gtAge, only one glycan trait, IGP53, was present in the top-20 feature list, suggesting its importance in gtAge. However, IGP53 was not selected as the final feature set for gAge. This could be because IGP53 alone did not contribute sufficiently to age prediction when only glycan data are used; however, its combined effect with transcriptomic features enhanced its predictive value in an integrated model.

Enrichment analyses showed no significantly enriched pathways for the genes in the final feature set for gtAge (FDR < 0.05). In contrast, 15 significant pathways were found for tAge. The results are summarized in Table 3.

3.5. Predicted age brings additive information to phenotypes

Delta age, indicating the difference between the predicted age and CA, was calculated for gtAge, tAge, and gAge separately (see Section 2). A positive delta age can be interpreted in term of reflecting more rapid biological aging. We found that delta gtAge and delta tAge were negatively associated with high-density lipoprotein (HDL) and that delta gAge was positively associated with total cholesterol (TC), triglycerides (TG), fasting plasma glucose (FPG), low-density lipoprotein (LDL), and glycated hemoglobin (Hba1c) (Table 4). Additionally, the *p* values for the associations of delta gtAge with FPG and LDL were not only lower than 0.25 but also were lower than those for delta tAge. Significant associations between delta age and HDL are shown in Fig. 4 (all age-related phenotypes are shown in Fig. S2).

4. Discussion

This study provides evidence on the involvement of IgG N-glycan traits and genes in aging. Using traditional linear association analysis, this study found that 43 IgG N-glycan traits and 6 genes were significantly associated (FDR < 0.25) with CA. Using the abundance of N-glycan traits and gene expression to build aging clocks, more genes and glycan traits contributing to CA prediction were identified. To construct a transcriptomics-based aging clock, tAge, 130 predictive genes were selected. Of them, 91 were included in a large-scale analysis by Peters et al. [12], and 12 were found to be age-associated. These results partially overlapped age-associated genes with those of previous large-scale analyses, suggesting that building a machine learning-based model is another way to identify, screen, and validate novel and existing associations. In practice, if an outcome of interest can be predicted with a machine learning model built from a set of features, these features can be associated with the outcome [35].

Through pathway enrichment analysis of 130 predictive genes for tAge, 15 enriched pathways were identified. One of the most significantly enriched pathways was associated with chemokine functions, including chemokine activity and chemokine receptor

binding (GOs: 0008009, 0042379, 0031726, and 0031730). Certain chemokines are considered pro-inflammatory markers. They act primarily as mediators in the immune response to recruit immune cells, such as monocytes, neutrophils, and other effector cells in the blood, to sites of infection or tissue injury [36]. They are also known as inflammatory chemokines. In this study, *CXCL8*, *CCL3*, *CCL4*, and *PF4V1* were identified as important components of these inflammatory responses. *CXCL8* gene, which encodes a protein and is referred to as interleukin (IL)-8, is universally secreted by mononuclear cells (neutrophils, T lymphocytes, eosinophils, macrophages, epithelial cells, and fibroblasts) and acts as a guide molecule directing neutrophils to inflammatory sites. *CCL3* and *CCL4* are C–C motif chemokine ligands that play a role in inflamma-

tory responses by binding to their receptors [37]. *PF4V1* is also a chemokine gene that is highly comparable with platelet factor 4 and displays a key antiangiogenic function. Other identified pathways were mainly related to immunity, including the regulation of natural killer cell-mediated immunity (GO: 0002716), leukocyte-mediated cytotoxicity (GO: 0001911), and natural killer cell-mediated cytotoxicity (GO: 0045953).

For the integrated omics-based aging clock, *gtAge*, 144 features, including 137 genes and 7 glycan traits, were selected. Although no significant pathways were identified from the selected genes, the genes encoded core proteins of the immune processes—*CLEC12B*, *KLRC1*, and *LGALS9*—met the selection criteria for *gtAge*. *CLEC12B* is a member of the C-type lectin domain family, which enables protein phosphatase binding activity and signaling receptor inhibitor activity, and is involved in the natural killer cell inhibitory signaling pathway [38]. *KLRC1* belongs to the killer cell lectin-like receptor family, also known as the natural killer group 2 (NKG2) family, and acts as a key inhibitory receptor on natural killer cells, regulating their activation and effector functions. *LGALS9* is a galactoside-binding lectin, which is implicated in several classic immune pathways for modulating cell–cell and cell–matrix interactions. For example, the activation of extracellular signal-related kinase 1/2 (ERK1/2) phosphorylation in mast cells and dendritic cells induces the production of cytokines (IL-6, IL-8, and IL-12) and chemokines (CCL2). The inflammatory pathways identified in the current study support the new inflammaging hypothesis [39], a developing inflammation status in most older individuals, and offer a novel tool for the estimation of BA. The predictive feature set for *gtAge* contains potential anti-aging therapeutic targets. For example, by searching for the genes identified in the current study in DrugBank [40], ten medications (reserpine, gadoxetic acid, taurocholic acid, lamivudine, probenecid, verapamil, indinavir, vincristine, sulfapyrazone, and tetrahydrofolic acid) have been developed and marketed targeting three proteins (encoded by *ABCC2*, *ABCC3*, and *SLC22A1* genes) to treat high blood pressure, chronic gout, and hepatitis B infection, and prevent human immunodeficiency virus (HIV) infection. These medications are used to treat inflammation-mediated diseases, which further implies that the gene set identified for *gtAge* represents a broader target set for the development of anti-inflammatory drugs. From the feature importance, we observed that *IGP53* contributed significantly to the prediction of *gtAge* and was shown to be negatively associated with age, implying that a decrease in *IGP53* could be a marker of biological aging. *IGP 53* was derived as the percentage of digalactosylated with core fucose glycan (FA2G2, GP14) in the total neutral IgG glycans (GP1–GP15). Taken together, the integrated features of *gtAge* showed a coherent aging mechanism driven by immune modulation and chronic inflammation. The convergence of inhibitory immune genes (*CLEC12B*, *KLRC1*, and *LGALS9*) and pro-inflammatory glycan alterations (such as the age-associated

Table 3
Significantly enriched terms for tAge.

Enriched functional pathway	Genes in the enriched term
Chemokine activity (GO: 0008009)	<i>CXCL8</i> ; <i>CCL4</i> ; <i>CCL3</i> ; <i>PF4V1</i>
Chemokine receptor binding (GO: 0042379)	<i>CXCL8</i> ; <i>CCL4</i> ; <i>CCL3</i> ; <i>PF4V1</i>
CCR1 chemokine receptor binding (GO: 0031726)	<i>CCL4</i> ; <i>CCL3</i>
CCR5 chemokine receptor binding (GO: 0031730)	<i>CCL4</i> ; <i>CCL3</i>
ABC-type glutathione S-conjugate transporter activity (GO: 0015431)	<i>ABCC3</i> ; <i>ABCC2</i>
Renal filtration (GO: 0097205)	<i>MYO1E</i> ; <i>ITGA3</i> ; <i>TMEM63C</i>
Negative regulation of natural killer cell mediated immunity (GO: 0002716)	<i>CLEC12B</i> ; <i>KLRC1</i> ; <i>LGALS9</i>
Glycine metabolic process (GO: 0006544)	<i>SHMT1</i> ; <i>AMT</i> ; <i>GCAT</i>
Negative regulation of leukocyte mediated cytotoxicity (GO: 0001911)	<i>CLEC12B</i> ; <i>KLRC1</i> ; <i>LGALS9</i>
Negative regulation of natural killer cell mediated cytotoxicity (GO: 0045953)	<i>CLEC12B</i> ; <i>KLRC1</i> ; <i>LGALS9</i>
Response to interleukin-1 (GO: 0070555)	<i>CXCL8</i> ; <i>CCL4</i> ; <i>CCL3</i> ; <i>LGALS9</i> ; <i>NKX3-1</i>
Receptor internalization (GO: 0031623)	<i>CXCL8</i> ; <i>MX1</i> ; <i>CD9</i> ; <i>RAMP1</i>
Positive regulation of CD4-positive, CD25-positive, alpha–beta regulatory T cell differentiation (GO: 0032831)	<i>LGALS9</i> ; <i>FOXP3</i>
Positive regulation of natural killer cell chemotaxis (GO: 2000503)	<i>CCL4</i> ; <i>CCL3</i>
Chemokine-mediated signaling pathway (GO: 0070098)	<i>CXCL8</i> ; <i>CCL4</i> ; <i>CCL3</i> ; <i>PF4V1</i>

CXCL8: C–X–C motif chemokine ligand 8; *CCL4*: C–C motif chemokine ligand 4; *CCL3*: C–C motif chemokine ligand 3; *PF4V1*: platelet factor 4 variant 1; *ABCC3*: ATP binding cassette subfamily C member 3; *ABCC2*: ATP binding cassette subfamily C member 2; *MYO1E*: myosin 1E; *ITGA3*: integrin subunit alpha 3; *TMEM63C*: transmembrane protein 63C; *CLEC12B*: C-type lectin domain family 12 member B; *KLRC1*: killer cell lectin like receptor C1; *LGALS9*: galectin 9; *SHMT1*: serine hydroxymethyltransferase 1; *AMT*: aminomethyltransferase; *GCAT*: glycine C-acetyltransferase; *NKX3-1*: NK3 homeobox 1; *MX1*: MX dynamin like GTPase 1; *CD9*: CD9 molecule; *RAMP1*: receptor activity modifying protein 1; *FOXP3*: forkhead box P3.

Table 4
Univariate analyses for disease status.

Phenotype	CA		Delta <i>gtAge</i>		Delta tAge		Delta <i>gAge</i>	
	Coefficient (95% CI)	<i>p</i>	Coefficient (95% CI)	<i>p</i>	Coefficient (95% CI)	<i>p</i>	Coefficient (95% CI)	<i>p</i>
SBP	0.006 (0.003, 0.008)	0	0.002 (–0.004, 0.009)	0.454	–0.002 (–0.008, 0.003)	0.446	0.003 (–0.002, 0.008)	0.213
DBP	0.004 (0.001, 0.006)	0.009	0.003 (–0.004, 0.01)	0.445	–0.003 (–0.010, 0.003)	0.293	0.003 (–0.002, 0.009)	0.217
TC	–0.001 (–0.005, 0.003)	0.594	0.004 (–0.007, 0.015)	0.480	–0.004 (–0.013, 0.006)	0.421	0.011 (0.003, 0.019)	0.006
TG	0.008 (–0.002, 0.017)	0.111	0.013 (–0.013, 0.039)	0.315	0.006 (–0.017, 0.029)	0.606	0.030 (0.011, 0.049)	0.002
FPG	0.005 (0.002, 0.008)	0.001	0.005 (–0.003, 0.013)	0.229	0 (–0.008, 0.007)	0.943	0.008 (0.002, 0.014)	0.014
HDL	–0.003 (–0.008, 0.002)	0.255	–0.016 (–0.029, 0.003)	0.020	–0.014 (–0.026, –0.002)	0.022	–0.007 (–0.017, 0.003)	0.172
LDL	–0.002 (–0.008, 0.004)	0.512	0.010 (–0.006, 0.026)	0.234	–0.002 (–0.017, 0.012)	0.744	0.017 (0.005, 0.030)	0.006
HbA1c	0.003 (0.001, 0.004)	0.006	0.001 (–0.004, 0.006)	0.709	0.001 (–0.003, 0.006)	0.552	0.004 (0, 0.008)	0.039
CRP	0.016 (–0.005, 0.037)	0.144	0.011 (–0.046, 0.067)	0.707	–0.005 (–0.055, 0.045)	0.846	0.025 (–0.019, 0.068)	0.268

SBP: systolic blood pressure; DBP: diastolic blood pressure; CRP: C-reactive protein.

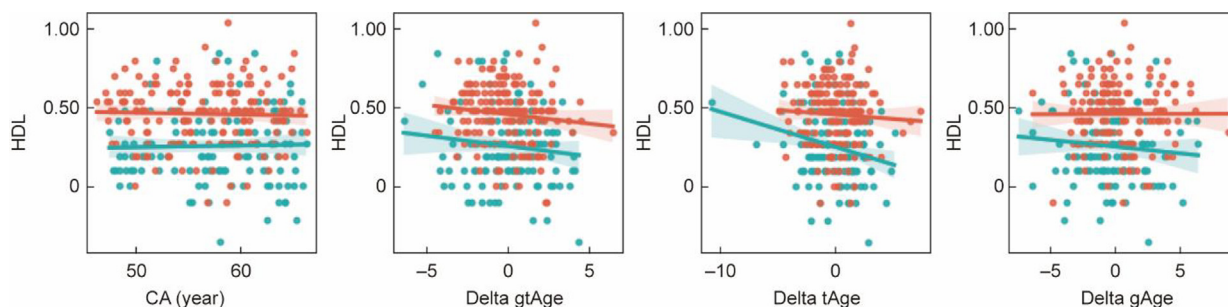


Fig. 4. The additive information about HDL from gAge, tAge, and gtAge. The plots show the relationships between six aging-related phenotypes and the delta ages for gAge, tAge, and gtAge. Red dots and lines represent females, and blue dots and lines represent males.

decline in IGP53/FA2G2) indicates a unified process of immunosenescence and inflammation. This highlights the value of combining transcriptomic and glycomic data in capturing multifaceted biological aging landscapes. Therefore, the gtAge model not only estimates BA with high resolution but also uncovers actionable targets within inflammation-related pathways, reinforcing its potential both as a biomarker and tool for identifying anti-inflammatory therapeutic strategies.

In contrast to the transcriptome analysis, we detected more age-associated IgG N-glycan traits using the linear method. For the original glycan traits, we found that FA2G2 (GP14), agalactosylated bisecting GlcNAc with core fucose glycan (FA2B, GP6), digalactosylated and sialylated bisecting GlcNAc with core fucose glycan (FA2G2S1, GP18), and agalactosylated bisecting GlcNAc with core fucose glycan (FA2, GP4) have the strongest associations with CA, which is consistent with the findings from all replicated cohorts in the study by Krištić et al. [9]. The other three age-associated glycan traits found in our study—agalactosylated glycan (FA1, GP1), digalactosylated with core fucose and bisecting GlcNAc glycan (FA2BG1, GP11), and digalactosylated bisecting GlcNAc with core fucose glycan (FA2BG2, GP15)—also show strong associations with age in at least two validation cohorts in the study by Krištić et al. [9]. Krištić et al. [9] detected three strongly age-associated glycan traits, namely, agalactosylated bisecting GlcNAc glycan (A2, GP2), digalactosylated and disialylated with core fucose glycan (FA2G2S2, GP23), and digalactosylated glycan (A2G2, GP12). In the current study, the p values for these three glycan traits were less than 0.001, and their FDR values were less than 0.25, implying their contributions to aging. In another study by Yu et al. [20], the exact raw p values from the association analyses were not reported. Therefore, we could not compare strengths of our associations with those of Yu et al. [20]. However, if we use the same significant cut-off score as those in the study by Yu et al. (i.e., $p < 0.05$) [20], we can consider monogalactosylated with core fucose glycan (FA2G1, GP9), GP20 (structure not determined), monogalactosylated with core fucose and bisecting GlcNAc glycan (FA2[6]BG1, GP10), monogalactosylated with core fucose glycan (FA2[6]G1, GP8), and high mannose (M5, GP5) as age-associated glycans in addition to the traits mentioned previously. In total, 15 glycan traits were significantly age-associated under this cutoff, and all these traits were age-associated in the study by Yu et al. [20]. The significant association between adjusted galactosylation and sialylation levels of IgG and CA can be explained by inflammation, which contributes to the pathogenesis of aging and related diseases through chronic, sterile, low-grade inflammation [41]. Both Krištić et al. [9] and Yu et al. [20] did not investigate the associations between CA and derivative glycan traits (IGP1–IGP54). One potential strength of our study is that we explored the age-associated derivative glycan traits. As to the predictive model built from IgG N-glycome, the best $R^2 = 0.290$, which is comparable with

the model in Yu et al.'s study [20] while lower than Krištić et al.'s study [9].

The current study found very few age-associated genes and several age-associated glycan traits, implying that aging has a larger effect on the IgG N-glycome than on the transcriptome, indicating that the interaction between genetic and epigenetic factors plays an essential role. Aging can change the IgG N-glycome via transcription and translation processes while also changing the IgG N-glycome by influencing the epigenetic factors. In fact, both these effects exist. Further *in vitro* and *in vivo* validations of age-associated glycans should control for non-genetic factors if genetic factors are the main research targets or vice versa. Furthermore, changes in inflammatory states have been proved to be involved in the aging process, and IgG glycosylation plays an essential role in the inflammatory cascade [19,20,42,43]. Thus, the IgG glycan-involved aging clock may show the inflammatory aging states. An intriguing aspect of aging highlighted in our study is the potential role of the IgG glycome in influencing age-related processes. Recent findings suggest that the glycosylation profile of IgG can act as a critical modulator of inflammation and senescence.

Based on the associations identified, we built a novel aging clock, gtAge, by integrating IgG N-glycome and transcriptome data for BA prediction. We used two methods to integrate omics and build models for gtAge. The essential difference between the two models is the selected feature set. The feature set for the first model was selected by performing LB-FS on a concatenated IgG N-glycome and transcriptome (CONC), and the corresponding model achieved $R^2 = 0.820$. The second model adopted a novel method, AlphaSnake, to integrate the IgG N-glycome and transcriptome, and adopted LB-FS to select the final feature set (denoted as CONC-RL). The corresponding model achieved $R^2 = 0.853$. Both models showed superior performance compared with the model built from the transcriptome ($R^2 = 0.812$) or IgG N-glycome ($R^2 = 0.290$). The model from CONC-RL performed better than that from CONC, indicating AlphaSnake could be a promising strategy for multiomics integration. The advantage of AlphaSnake is that it leverages a DRL method to determine the optimal trajectory for FFS through trial and error. Compared with the simple concatenation method, exploration by AlphaSnake brings more possible combinations of rankings from integrated omics. Therefore, AlphaSnake increases the probability of achieving a higher optimal FFS performance and includes more potential predictive features for FFS to achieve the optimal performance. This, finally, results in a better model for the refined feature set than LB-FS. In addition, AlphaSnake relies on feature ranking and selects the most important features from different omics. Thus, the search space is limited to a practical range. For example, if we do not use feature-ranking information, that is, assuming all features are equally important, we need to search for 2^n possible feature combinations (n is the number of features), which is not practical for high-dimensional

data. In our experiment, AlphaSnake identified the optimal FFS trajectory at approximately 40 epochs, which is not computationally expensive. AlphaSnake can be adopted for other omics data and research questions, and requires to be studied further.

The additive information for gAge, tAge, and gtAge was preliminarily evaluated. We observed that delta gtAge and delta tAge were negatively associated with HDL, indicating that biologically younger individuals have higher HDL levels. However, CA was not significantly associated with HDL. Therefore, gtAge and tAge provide additional information for investigating HDL levels. Compared with tAge and gtAge, gAge had a lower correlation with CA. However, delta gAge was positively associated with more phenotypes, including TC, TG, FPG, LDL, and HbA1c, whereas it was negatively associated with HDL ($p = 0.172$). In addition, we observed that the p values of the associations between gtAge and FPG and LDL were lower than 0.25, whereas those of the same associations for tAge were obviously higher, suggesting that gtAge could learn information from the glycan part of the IS. Hence, we propose that integrating different omics to build an aging clock can improve predictions and benefit from different biological information contained in both omics. A limitation of the current study is that the cohort only covered a middle-aged population (aged between 46 and 66 years), but a vital period of life, for example menopause for women and andropause for men [16,19]. However, further studies are required to validate the proposed gtAge.

We observed that gAge provided additional information on more phenotypes than gtAge or tAge. Our gAge is positively associated with CA and can explain 29% of the variance in CA, indicating that, to a certain extent, it can measure the rate of aging. Simultaneously, it learns information related to the phenotypes of the study interest, enlarging the difference between the predicted and CAs. Therefore, a poorer predictive model for CA may not always be disadvantageous, as it may provide additional information for measuring BA. The IgG N-glycome has been proven to be associated with several non-communicable diseases, such as type 2 diabetes (T2D), in this study population [10]. Further studies should explore whether gAge is a better predictive feature of IgG N-glycome-related diseases than CA. In addition, we found that gAge was highly sex-specific for several phenotypes, which might result from the fact that sex and hormone levels were associated with notable changes in IgG N-glycome [16,19].

In summary, we have identified age-associated genes and IgG N-glycan traits. In addition, we created a novel aging clock, gtAge, by building a model to predict CA from integrated IgG N-glycome and transcriptome data, leveraging a LASSO-LARS bootstrap feature selection method and a DRL-based multiomics integration method. The absolute difference between gtAge and CA was positively associated with T2D in the study population, suggesting that gtAge may be a potential indicator of BA. To the best of our knowledge, this is the first study that integrated IgG N-glycans and transcriptomes to predict age. We found that integrating multiomics data can improve age prediction and that, depending on different biomarkers, the predicted aging clock may reflect different information. As the pilot cohort in this study is limited by the narrow population age range and small sample size, the proposed gtAge might not have sufficient generalizability and transferability. Poor generalizability and transferability of age prediction models were also observed in other studies. For example, Peters et al. [12] found that the R^2 of age prediction models varied from 0.121 to 0.599 depending on the different cohorts analyzed. Shokhirev and Johnson [15] found that the best model could achieve $R^2 > 0.95$ in the discovery cohort, whereas after applying the model to external validation datasets, the performance decreased to $R^2 = 0.48$. Therefore, more studies involving larger cohorts and more diverse populations are needed to further validate gtAge. In addition, although glycomics and transcriptomics offer valuable insights, their use in

real-world settings is limited by high costs, complex workflows, and scalability-related challenges. In particular, glycan profiling lacks standardization and automation, making large-scale clinical applications difficult. As technology advances and costs decrease, broader implementation may become more feasible.

In this study, we propose AlphaSnake, a novel RL-based feature selection algorithm specifically designed for multiomics data integration. Unlike traditional concatenation-based integration, AlphaSnake employs DRL to intelligently select and prioritize features from diverse omics datasets, thereby enhancing predictive performance by focusing on the most informative variables. Our results show that AlphaSnake significantly outperforms conventional concatenation-based approaches. A major advantage of AlphaSnake is its ability to preserve the original format of each feature, thereby enhancing its interpretability. In addition, AlphaSnake efficiently reduces the computational overhead by leveraging pre-calculated feature rankings, which allows it to focus dynamically on selecting the most impactful features. Its flexibility allows the integration of any ranking methodology, making it adaptable to various multiomics integration scenarios beyond aging research. However, AlphaSnake also possesses certain limitations. Its dependence on the initial feature-ranking information indicates that inaccurate or biased rankings can affect the downstream model performance. Moreover, compared with transformation-based methods (e.g., dimensionality reduction techniques, such as principal component analysis (PCA) or autoencoders), AlphaSnake may require a more detailed preliminary analysis to generate reliable feature rankings. Although model-based approaches may inherently capture complex interactions without explicit rankings, AlphaSnake explicitly depends on accurate ranking mechanisms that could constrain its effectiveness in certain contexts. Future research should validate AlphaSnake more extensively across diverse populations, such as African and Asian cohorts, and across broader age ranges. Such validation is crucial to confirm its utility as a robust and generalizable tool for biological aging assessments and other complex biological questions involving multiomics data integration.

CRedit authorship contribution statement

Yao Xia: Writing – review & editing, Writing – original draft, Visualization, Validation, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Syed Mohammed Shamsul Islam:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Conceptualization. **Xingang Li:** Writing – review & editing, Supervision, Investigation, Formal analysis, Conceptualization. **Abdul Baten:** Writing – review & editing, Supervision, Methodology. **Xuerui Tan:** Writing – review & editing, Conceptualization. **Wei Wang:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was funded by an Australia–China International Collaborative Grant (NHMRC APP1112767-NSFC 81561128020), the European Union's Horizon 2020 Research and Innovation Program under grant agreement (779238), the Edith Cowan University Higher Degree

by Research Scholarship (ECU-HDR 10492768), the Western Australian Future Health Research and Innovation Funds (WANMA/EL2023-24/2 and WANMA/Ideas2024-25/5), and the Edith Cowan University Early-Mid Career Researcher Grant Scheme (G1006465).

The authors appreciate the Busselton Population Medical Research Institute (BPMRI) in Western Australia for sharing the summary statistics publicly available for the benefit of this study.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.eng.2025.08.016>.

References

- [1] López-Otín C, Blasco MA, Partridge L, Serrano M, Kroemer G. The hallmarks of aging. *Cell* 2013;153(6):1194–217.
- [2] Johnson AA, Shokhirev MN. Contextualizing aging clocks and properly describing biological age. *Aging Cell* 2024;23(12):e14377.
- [3] Baker GT, Sprott RL. Biomarkers of aging. *Exp Gerontol* 1988;23(4–5):223–39.
- [4] Jylhävä J, Pedersen NL, Hägg S. Biological age predictors. *EBioMedicine* 2017;21:29–36.
- [5] Butler RN, Sprott R, Warner H, Bland J, Feuers R, Forster M, et al. Biomarkers of aging: from primitive organisms to humans. *J Gerontol A Biol Sci Med Sci* 2004;59(6):B560–7.
- [6] Johnson TE. Recent results: biomarkers of aging. *Exp Gerontol* 2006;41(12):1243–6.
- [7] Fleischer JG, Schulte R, Tsai HH, Tyagi S, Ibarra A, Shokhirev MN, et al. Predicting age from the transcriptome of human dermal fibroblasts. *Genome Biol* 2018;19(1):221.
- [8] Gudelj I, Lauc G, Pezer M. Immunoglobulin G glycosylation in aging and diseases. *Cell Immunol* 2018;333:65–79.
- [9] Krištić J, Vučković F, Menni C, Klarić L, Keser T, Beceheli I, et al. Glycans are a novel biomarker of chronological and biological ages. *J Gerontol A Biol Sci Med Sci* 2014;69(7):779–89.
- [10] Li X, Wang H, Russell A, Cao W, Wang X, Ge S, et al. Type 2 diabetes mellitus is associated with the immunoglobulin G N-glycome through putative proinflammatory mechanisms in an Australian population. *OMICS* 2019;23(12):631–9.
- [11] Meyer DH, Schumacher B. BiT age: a transcriptome-based aging clock near the theoretical limit of accuracy. *Aging Cell* 2021;20(3):e13320.
- [12] Peters MJ, Joehanes R, Pilling LC, Schurmann C, Conneely KN, Powell J, et al. The transcriptional landscape of age in human peripheral blood. *Nat Commun* 2015;6(1):8570.
- [13] Russell AC, Šimurina M, Garcia MT, Novokmet M, Wang Y, Rudan I, et al. The N-glycosylation of immunoglobulin G as a novel biomarker of Parkinson's disease. *Glycobiology* 2017;27(5):501–10.
- [14] Schmidt M, Hopp L, Arakelyan A, Kirsten H, Engel C, Wirkner K, et al. The human blood transcriptome in a large population cohort and its relation to aging and health. *Front Big Data* 2020;3:548873.
- [15] Shokhirev MN, Johnson AA. Modeling the human aging transcriptome across tissues, health status, and sex. *Aging Cell* 2021;20(1):e13280.
- [16] Štambuk J, Nakić N, Vučković F, Pučić-Baković M, Razdorov G, Trbojević-Akmačić I, et al. Global variability of the human IgG glycome. *Aging* 2020;12(15):15222–59.
- [17] Wang F, Yang J, Lin H, Li Q, Ye Z, Lu Q, et al. Improved human age prediction by using gene expression profiles from multiple tissues. *Front Genet* 2020;11:1025.
- [18] Zaytseva OO, Sharapov SZ, Perola M, Esko T, Landini A, Hayward C, et al. Investigation of the causal relationships between human IgG N-glycosylation and 12 common diseases associated with changes in the IgG N-glycome. *Hum Mol Genet* 2022;31(10):1545–59.
- [19] Wang W. Glycomedicine: the current state of the art. *Engineering* 2023;26(7):12–5.
- [20] Yu X, Wang Y, Kristic J, Dong J, Chu X, Ge S, et al. Profiling IgG N-glycans as potential biomarker of chronological and biological ages: a community-based study in a Han Chinese population. *Medicine* 2016;95(28):e4112.
- [21] Mak JKL, McMurrin CE, Kuja-Halkola R, Hall P, Czene K, Jylhävä J, et al. Clinical biomarker-based biological aging and risk of cancer in the UK Biobank. *Br J Cancer* 2023;129(1):94–103.
- [22] Zierer J, Pallister T, Tsai PC, Krumsiek J, Bell JT, Lauc G, et al. Exploring the molecular basis of age-related disease comorbidities using a multi-omics graphical model. *Sci Rep* 2016;6(1):37646.
- [23] Hartmann A, Hartmann C, Secci R, Hermann A, Fuellen G, Walter M. Ranking biomarkers of aging by citation profiling and effort scoring. *Front Genet* 2021;12:686320.
- [24] Kudryashova KS, Burka K, Kulaga AY, Vorobyeva NS, Kennedy BK. Aging biomarkers: from functional tests to multi-omics approaches. *Proteomics* 2020;20(5–6):e1900408.
- [25] Rivero-Segura NA, Bello-Chavolla OY, Barrera-Vázquez OS, Gutierrez-Robledo LM, Gomez-Verjan JC. Promising biomarkers of human aging: in search of a multi-omics panel to understand the aging process from a multidimensional perspective. *Ageing Res Rev* 2020;64:101164.
- [26] Solovev I, Shaposhnikov M, Moskalev A. Multi-omics approaches to human biological age estimation. *Mech Ageing Dev* 2020;185:111192.
- [27] Wu L, Xie X, Liang T, Ma J, Yang L, Yang J, et al. Integrated multi-omics for novel aging biomarkers and antiaging targets. *Biomolecules* 2021;12(1):39.
- [28] James A, Hunter M, Straker L, Beilby J, Bucks R, Davis T, et al. Rationale, design and methods for a community-based study of clustering and cumulative effects of chronic disease processes and their effects on ageing: the Busselton healthy ageing study. *BMC Public Health* 2013;13(1):936.
- [29] Pucić M, Knezević A, Vidic J, Adamczyk B, Novokmet M, Polasek O, et al. High throughput isolation and glycosylation analysis of IgG—variability and heritability of the IgG glycome in three isolated human populations. *Mol Cell Proteomics* 2011;10(10):M111-010090.
- [30] Menni C, Keser T, Mangino M, Bell JT, Erte I, Akmačić I, et al. Glycosylation of immunoglobulin G: role of genetic and epigenetic influences. *PLoS One* 2013;8(12):e82558.
- [31] Lauc G, Huffman JE, Pučić M, Zgaga L, Adamczyk B, Mužinić A, et al. Loci associated with N-glycosylation of human immunoglobulin G show pleiotropy with autoimmune diseases and haematological cancers. *PLoS Genet* 2013;9(1):e1003225.
- [32] Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* 2017;14(4):417–9.
- [33] Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *Ann Stat* 2004;32(2):407–99.
- [34] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. 2nd ed. New York City: Springer; 2009.
- [35] Sun Y, Clarke B, Clarke J, Li X. Predicting antibiotic resistance gene abundance in activated sludge using shotgun metagenomics and machine learning. *Water Res* 2021;202:117384.
- [36] Mantovani A, Garlanda C. Humoral innate immunity and acute-phase proteins. *N Engl J Med* 2023;388(5):439–52.
- [37] Ren M, Guo Q, Guo L, Lenz M, Qian F, Koenen RR, et al. Polymerization of MIP-1 chemokine (CCL3 and CCL4) and clearance of MIP-1 by insulin-degrading enzyme. *EMBO J* 2010;29(23):3952–66.
- [38] Hoffmann SC, Schellack C, Textor S, Konold S, Schmitz D, Cerwenka A, et al. Identification of CLEC12B, an inhibitory receptor on myeloid cells. *J Biol Chem* 2007;282(31):22370–5.
- [39] Ferrucci L, Fabbri E. Inflammageing: chronic inflammation in ageing, cardiovascular disease, and frailty. *Nat Rev Cardiol* 2018;15(9):505–22.
- [40] Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res* 2008;36(Database issue):D901–6.
- [41] Franceschi C, Garagnani P, Parini P, Giuliani C, Santoro A. Inflammaging: a new immune-metabolic viewpoint for age-related diseases. *Nat Rev Endocrinol* 2018;14(10):576–90.
- [42] Wang W. Can DNA be glycosylated? *Engineering*. In press.
- [43] Wu Z, Guo Z, Zheng Y, Wang Y, Zhang H, Pan H, et al. IgG N-glycosylation cardiovascular age tracks cardiovascular risk beyond calendar age. *Engineering* 2023;26:99–107.