A lipidome Aging Clock shows Age Acceleration in individuals with Autism

Djakim Latumalea[1-3], Maximilian Unfried[1,3], Diogo Barardo[1,3], Jan Gruber[1,3,6], Brian K. Kennedy[1,3-5,*]

[1]Healthy Longevity Translational Research Programme, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

[2]Institute for Life Sciences & Technology, Hanze University of Applied Sciences, Groningen, The Netherlands

[3]Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

[4]Department of Physiology, Yong Loo Lin School of Medicine, National University

of Singapore, Singapore, Singapore

[5]Centre for Healthy Longevity, National University of Singapore, Singapore, Singapore

[6]Science Division, Yale-NUS College, Singapore, Singapore

[*]Corresponding author: bkennedy@nus.edu.sg

Abbreviations

SHAP (SHapley Additive exPlanations), WND (without neurological disease), DS (Down syndrome), ASD (Autism Spectrum Disorder), SZ (schizophrenia), PMI (post-mortem interval), LM IDs (LIPID MAP IDs), m/z (molecular weight), RT (retention time), FA (fatty acyls), GL (glycerolipids), GP (glycerophospholipids), SP (sphingolipids), ST (sterol lipids), PK (polyketides), PCA (principal component analysis), SVD (singular value decomposition), MAD (mean absolute deviation), ME (mean error), r (Pearson correlation coefficient), RMSE (root mean squared error), SE (standard error)

Abstract

Recent advancements in lipidomics and machine learning have been leveraged to investigate the prediction of biological age in individuals. This study delves into age acceleration patterns, entropy, and the potential role of dolichol as an aging biomarker. We introduce a novel aging clock combined with explainable AI that utilizes the lipid composition of the prefrontal cortex to predict the biological age of individuals, both those without known neurological conditions and those with autism, schizophrenia, or Down syndrome. Notably, significant age acceleration was observed in individuals with autism. Furthermore, entropy exhibits a significant increase around the age of 40, indicating potential dysregulation in the mevalonate pathway. Lastly, dolichol emerges as a potential biomarker. These findings underscore the feasibility of predicting biological age using lipidomics data, paving the way for further investigation into the intricate relationship between lipid alterations and prefrontal cortex aging, while offering valuable insights into the associated molecular mechanisms.

Keywords

Brain lipids; Cell signaling; Transport; Isoprenoids; Lipids; N-glycosylation; biomarkers; Prefrontal cortex; Neurological disease

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

1. Introduction

Neurological syndromes including Parkinson's and Alzheimer's disproportionately affect the older population (1). Moreover, these disorders (and others) are major contributors to mortality (2) and cause a significant financial burden (3). The primary risk factor for developing most neurological disorders is aging (4). While aging is characterized by a loss of physiological function and an exponential increase in mortality (5), aging is also a highly heterogeneous process (6) influenced by a range of intrinsic and extrinsic factors (7).

Lopez-Otin et al. (8) proposed a set of hallmarks of aging, which can be further classified into three distinct categories: primary, antagonistic, and integrative hallmarks. These hallmarks aim to capture the underlying causes of damage, the key responses to such damage, and the resulting effects responsible for the ultimate functional decline of aging organisms. This review has recently been updated to include further hallmarks (8).

Alterations in lipid composition play a critical role as they are involved in metabolic energy, homeostasis, and cell signaling (9, 10). In addition, aging is known to significantly alter the brain lipidome (11), which will impact the hallmarks of aging, such as proteostasis (12), and result in decreased resilience and neuronal plasticity (13). It therefore stands to reason that age-dependent alterations in the lipidome may contribute to brain aging and increase the risk of developing neurological disorders (4). Evidence already exists that lipid alterations contribute to specific age-dependent neurological disorders. For instance, dysregulated lipid homeostasis has been linked to inflammaging which plays a role in the etiology of Alzheimer's disease (14). In addition, lipid peroxidation is positively correlated with Alzheimer's disease, Down syndrome, and other neurological disorders (15). Although neurological disorders are mostly found in aged individuals, they can develop at a younger age under certain contexts. Prenatal stress and other environmental stressors can increase the risk of developing neurological disorders, such as schizophrenia, and autism spectrum disorders (16–18). People with Down syndrome, schizophrenia, and autism also tend to die at a younger age (19–21), which may indicate age acceleration.

Age acceleration can be measured using aging clocks, which may provide insight into the underlying factors that contribute to the aging process. Numerous aging clocks have been developed to predict biological age. The first prominent clock, developed by Hannum et al. (22), was based on DNA methylation data from whole blood. Subsequently, various other aging clocks using DNA methylation data have been developed and proposed (23, 24). This has been followed by the development of aging clocks utilizing other omics data, such as transcriptomics (25), proteomics (26), lipidomics (27), or a combination of various data sources (28). Most of these models are considered first-generation aging clocks as they predict chronological age. However, chronological age may not always be the most accurate marker, which has prompted researchers to directly predict mortality (29, 30). These clocks are known as second-generation clocks.

Most aging clocks use linear models, which have the advantage of being easily interpretable. However, these models are not able to capture non-linear signals, which are prominent in biological data (31), given that there is a complex interplay between molecules and regulatory mechanisms that govern biological processes. To address this limitation, researchers have developed aging clocks based on deep learning (32, 33). An alternative is the use of tree-based methods as these are able to capture complex and non-linear relationships due to their hierarchical structure. In general, there is a trade-off between predictive power and interpretability in the field of aging clocks. The field of explainable artificial intelligence (XAI) aims to make models interpretable (34), and recent breakthroughs have been made in the interpretability of machine learning and deep learning techniques (35–37). One framework that has gained significant attention is the SHapley Additive exPlanations (SHAP) framework (35), which is particularly well suited for tree-based methods. The framework can provide insight in how features contribute to the prediction at an individual and population level.

Horvath et al. (38) also developed an aging clock using brain tissue from individuals with Down syndrome showing a significant age acceleration effect. In addition, Cole et al. (39) utilized neuroimaging data to identify factors associated with age acceleration in individuals with Down Syndrome. Both studies found

significant age acceleration and showed that age acceleration is measurable through DNA methylation and neuroimaging.

Age acceleration has also been proposed in schizophrenia, although studies are conflicting (40). Higgins Chen et al. (41) explored 14 epigenetic clocks and found that 3 mortality-based clocks were able to identify significant acceleration in schizophrenia patients. Conversely, other studies utilizing epigenetic clocks based on chronological age or mortality data did not observe significantly accelerated tissue specific brain aging (42, 43). This may be caused by the relatively low amount of samples. It seems that DNA methylation clocks lack the capability to measure aging in individuals with schizophrenia, whereas mortality clocks are proficient, although they depend on clinical parameters. Developing an aging clock based on lipids provides another perspective. Also of note, there is very limited data on brain aging in individuals with Autism (20, 44).

In this study, we aimed to determine if lipidomics data from the prefrontal cortex could predict biological age. While several clocks have been devised for human brain tissue (38, 45, 46), they exclusively rely on DNA methylation data. To our knowledge, no model based on lipids has been developed thus far. Lipids are integral to understanding the relationship between aging and neurological disorders, given that they constitute approximately 40% of the dry-weight gray matter (47), with the brain exhibiting the highest diversity of lipid species (48). These findings suggest that a lipid-based clock might pinpoint aging-associated lipids, potentially shedding light on conditions like Down syndrome, schizophrenia, and autism. Additionally, we identify molecules strongly linked with age and propose them as biomarkers, a critical aspect given the post-mortem nature of the clock. Our study demonstrates that variations in brain lipids suffice for estimating biological age.

2. Materials and Methods

2.1 Samples

This study utilized the first dataset published by Yu et al. in 2018 (49). A comprehensive description of these data and the methodology employed by the authors can be found in the associated publication. Briefly, the dataset comprised 452 samples which were acquired from several brain banks, including NICHD Brain and Tissue Bank for Development Disorders, Maryland Psychiatric Research Center at the University of Maryland, Maryland Brain Collection Center, Netherlands Brain Bank, Chinese Brain Bank Center, Harvard Brain Tissue Resource Center, and the Autism Tissue Program. Samples consist of gray matter from the anterior region of the prefrontal cortex, with a mean weight of 12.55 mg (±1.65). The dataset is composed of 403 samples without neurological disorder (WND), 5 samples with Down syndrome (DS), 17 samples with autism spectrum disorder (ASD), and 27 samples with Schizophrenia (SZ). The WND samples have an age range of 0 to 99 years, with a median age of 24 years (excluding 12 prenatal samples). DS samples range from 58 to 65 years with a median age of 61 years, ages of ASD samples range from 18 to 60 years with a median age of 30 years, and SZ samples range from 23 to 64 years with a median age of 48 years. The post-mortem interval (PMI) for all samples ranges from 0 to 44 hours, with median values of 14, 6.2, 20.3, and 19.5 hours for WND, DS, ASD, and SZ samples respectively. A detailed description of sample statistics can be found in **Supplementary Table 1**. The ethnicity of subjects is not known for all samples, thus, labeled as unknown, limiting the use of this parameter for matching case and control groups.

2.2 Data Analysis

2.2.1 Data preprocessing

To align the date of birth with a time point of zero, the gestational period of 0.767 years (280/365) was subtracted from the reported age of each subject. Samples from individuals below the age of 20 tend to cluster together in feature space as visualized in **Supplementary Figure 1**, suggesting the brain is still

developing. Samples from subjects older than 80 years only consisted of individuals of Caucasian descent. To mitigate these biases, only samples between the ages of 20 and 80, inclusive, were included in the analysis. In total, 5024 lipid species were analyzed. Of these lipid species, 2222 were annotated with LIPID MAP IDs (LM IDs). It should be noted that a single lipid species may map to one or more LM IDs. The presence of degenerate LM IDs causes ambiguity. To address this issue, a conformity index was created to measure the coherence of categories, classes, or molecules (see section 2.2.7). Only lipid species unambiguously associated with a single LM ID were selected, resulting in a dataset comprising 360 lipid species.

After this initial selection, we mapped each lipid species in the dataset to the associated LM ID, molecular weight (m/z), and retention time (RT). It was discovered that several lipid species mapped to the same LM ID but had different m/z values. One explanation for these discrepancies could be that molecules have the same elements but different isotopes. To determine specific modifications, a mass search against a list of adducts, provided by LIPID MAPS, was performed. The mass search returned the LM ID and corresponding chemical formula. This search was performed for every LIPID MAPS category: fatty acyls (FA), glycerolipids (GL), glycerophospholipids (GP), sphingolipids (SP), sterol lipids (ST), prenol lipids (PR), saccharolipids (SL), and polyketides (PK). Both positive and negative ionization modes were considered. The positive mode included [M+H]+, [M+H-H2O], [M+Na]+, [M+NH4]+, [M+K]+, [M+2H]2+, [M+2Na]2+, [M+2Na-H]+. The negative ionization mode included [M-H]-, [M-Cl]-, [M+HCOO]-, [M+OAc]-, [M-CH3]-, [M-2H]2-, [M-3H]3-. After merging the lipid species according to the LM ID and corresponding chemical formula, any values that could not be matched returned a value of 0.0 and were then reassigned as missing values. Following pre-processing, the final dataset comprised a feature matrix of 242 samples and 163 features.

2.2.2 Principal component analysis and mutual information

Principal component analysis (PCA) is an eigenvector technique that can be used for dimensionality reduction. PCA results in a linear coordinate system transformation maximizing covariance along the

transformed coordinate axes (50). This can be accomplished by singular value decomposition (SVD), decomposing the covariance or correlation matrix of the data into its corresponding eigenvalues and eigenvectors, such that $S = Q\Lambda Q^T$, where Q represents the eigenvectors and Λ represents the eigenvalues. This study used the PCA class provided by the scikit-learn library (51) to perform PCA. The resulting matrix can be altered and used by various algorithms or for visualization. Additionally, eigenvectors can be rescaled by mutual information for visualization purposes and may show other relationships (52). The implementation of mutual information in this context is described in Platt et al. (53) and is given by

$$IM(X,V)_{i,j} = -sng\left(U_{i,j}\right)\log_2\sqrt{1 - U_{i,j}^2}$$

, where X is a matrix and V contains the set of singular eigenvalues.

2.2.3 Elastic net

Elastic net regression is a technique to fit a linear model predicting a dependent variable based on high-dimensional explanatory data. Elastic net avoids over-fitting while being able to select a sparse set of variables that explain the target variable (54). The algorithm combines regularization techniques from ridge regression and lasso regression. While ridge regression maintains all variables in the model, lasso regression can identify a subset of variables with predictive power.

This study used the ElasticNet class provided by the scikit-learn library (51). To optimize the *alpha* penalty term and *l1_ratio* L1 and L2 mixing parameter, we use nested stratified cross-validation in combination with grid search (see 2.2.5).

2.2.4 XGBoost

XGBoost is a highly scalable and optimized tree-boosting algorithm (55). A tree-boosting algorithm combines several weak models, which are trained on residuals, into a strong model. Tree-based algorithms, such as XGBoost, can detect both linear and nonlinear signals in data, making them useful for analyzing biological data, which often contains nonlinear signals (31).

This study uses the XGBoostRegressor class provided by the XGBoost library (55). Key model parameters, including the number of estimators, maximum depth, subsample ratio, column sample ratio, regularization lambda, and regularization alpha were optimized using nested stratified cross-validation in combination with grid search (see 2.2.5). Missing data was imputed by k-nearest neighbor imputation using the KNNImputer class provided by the scikit-learn library (51). The number of neighbors was set to 10, as it has been shown to be a robust method for biological data, in general (56). Additionally, Yeo-Johnson (57) power transformation was applied on the imputed data using the PowerTransformer class provided by the scikit-learn library to normalize the data. This transformation attempts to find a λ parameter using maximum likelihood to minimize skewness and stabilize the variance to make the data more Gaussian-like, which is a desirable property for many statistical methods.

A boosting algorithm such as XGBoost creates several models that are sequentially correcting errors made by previous models. This approach enables XGBoost to continuously refine its predictions, particularly focusing on challenging data instances.

To calculate age_{Δ} we subtracted the true chronological age from the chronological age predicted by the model:

$$age_{\Delta} = y_{pred} - y_{true}$$

age_{Δ} from samples with a neurological disease were compared with the age_{Δ} from WND samples using the Mann-Whitney U test as tree-based models are non-parametric.

2.2.5 Stratified nested cross-validation

A combination of grid search and stratified nested cross-validation was employed to identify the optimal hyperparameters for Elastic PCA and XGBoost. To ensure that the training and test sets follow the same underlying age distribution, samples are divided into age bins of 5 years, totaling 12 bins, to stratify on. The search space for each model is represented in **Table 1**. The best-performing parameters for each model are recorded and used for model development.

2.2.6 Shannon entropy

In information theory, entropy is a quantitative measure of information. One metric to calculate the entropy of a sample is Shannon entropy. The metric measures the amount of information in a variable (58). Shannon entropy requires the probability distribution of a discrete variable. Under the assumption that individuals between the ages of 20 and 40 years have a stable and healthy prefrontal cortex (see **Supplementary Figure 2**), we log-transformed the lipids using the Yeo-Johnson PowerTransformer class provided by the scikit-learn library (51). Subsequently, a 0 was assigned to values between two standard deviations from the mean and 1 otherwise. Probabilities were calculated using a 10-year moving window, where the assigned values were counted within groups and divided by the total number of values, resulting in a normalized probability between 0 and 1. The entropy was calculated for each sample using

$$H_i = -\sum_{j=1}^{n} p_j \log(p_j)$$

, where i is the ith sample and j is the jth lipid.

2.2.7 Conformity Index

The conformity index measures the coherency between LM IDs of the same type. The index was calculated by dividing the most abundant value for each sample by the total amount of values for that sample. A conformity index of 1 indicates that all values are consistent. A conformity index of 0 indicates that none of the given values are consistent.

2.2.8 Interpretation

To interpret the results of tree-based models, SHAP values were used. SHAP values (35) provide feature importance on both an individual and a global level. SHAP values were calculated using feature permutations. The TreeExplainer class was utilized using the SHAP library (35) to provide feature importance and aid in the explanation of both global and local feature importance. In addition, visualization tools were utilized to aid in the interpretation of the results.

To interpret the Elastic PCA ensemble we calculated the PCA loadings and multiplied it by the coefficients of the elastic net model. The sum was then taken to calculate the importance of each feature. We repeated this process for every model in the ensemble and calculated the average to calculate the overall importance of each feature.

2.2.9 Implementation details

The analysis pipeline was implemented in Python (v3.9.7, (59)). Numpy (v1.26.4, (60)), Pandas (v2.1.1, (61)) and Scipy (v1.11.3, (62)) were used for computational purposes and scikit-learn (v1.3.0, (51)) was used for the development of machine learning models. XGBoost (v2.0.0, (55)) was used to develop a tree-based model. For visualization, plotly (v5.17.0, (63)) and matplotlib (v3.8.0, (64)) were used. For interpretation of models the SHAP library (v0.43.0, (35)) was used.

2.2.10 Statistical tests

We used the Wilcoxon Rank Sum Test to compare entropy values and lipid concentration levels between groups as we observed a nongaussian distribution. We applied a Bonferroni correction ($P<.00004$) to account for multiple hypothesis testing. To calculate trendlines, the ordinary least squares method was utilized. T-tests were used to compare sloped of trendlines.

2.3 Experimental design

The chosen models were developed in two phases. In the first phase, an ensemble model was developed using the reference cohort to predict the chronological age of every individual. The features contain lipid species and optional metadata, and only include samples between 20 and 80 years old, inclusive. The ensemble model consisted of multiple models that demonstrated the greatest efficacy during stratified nested cross-validation. In the second phase, the model is used to predict samples with and without neurological disorders, and SHAP values were employed to explain features on both a global and local level. A general overview is presented in **Figure 1**.

3. Results

In the present study, the dataset was pre-processed, resulting in the retention of 163 lipid species and 242 samples. The rationale behind choosing these lipids lies in our objective to include only those with unique chemical formulas. This selection ensures a high granularity in our lipid dataset, allowing for a more detailed analysis. Of the 39,446 lipid concentration values, 1.98% were identified as missing, with 47% of all lipids having at least one missing value. Missing values were imputed using K-nearest neighbors. The dataset includes 195 samples of individuals without neurological disease, 27 samples of individuals diagnosed with schizophrenia, 15 samples of individuals diagnosed with autism spectrum disorder, and 5 samples diagnosed with Down syndrome. The dataset was split into a training and test set containing 175 and 20 samples respectively and was stratified by age and ethnicity. The training set was used for nested-stratified cross-validation. A comprehensive breakdown of the preprocessed dataset is presented in **Supplementary Table 2**. Despite the identification of 5 outliers through the application of the isolation forest algorithm, it was decided to retain these outliers in the dataset, to not exclude meaningful biological variance. As highlighted by Shieh and Hung (65), outliers can be caused by various factors, including biological variation. As the specific cause of the outlier samples in the dataset is uncertain, it is assumed that the outliers are a result of biological and not technical variation. For comparative analysis, a model excluding outliers was generated; however, no statistically significant differences were observed.

We trained 28 different machine learning models to find the model that would be able to detect most of the signal from the dataset. The examination indicated that gradient boosting and linear regression models demonstrated the most favorable performance in detecting the signal from the dataset, as illustrated in **Supplementary Figure 3**. This suggests that the dataset contains both linear and nonlinear signals. As a result, two models were developed for comparative analysis. The first model, referred to as Elastic PCA, utilizes elastic net in combination with PCA, which is a widely utilized approach in literature (27, 45, 66). PCA is used to reduce the dimensions of the dataset and remove noise. The second model is an XGBoost model, which is also able to pick up nonlinear signals and can be interpreted using SHAP values. Both

models were trained on the samples of individuals without Neurological disease. A comparison of the performance and the feature importance of these models was conducted. Finally, a novel model, named DoliClock, was created that utilizes only dolichol molecules to predict chronological age.

3.1 Comparison of Elastic PCA and XGBoost

To identify the optimal parameters for both models, nested stratified cross-validation with grid search was employed. The total search space consisted of 140 parameters for Elastic PCA and 768 parameters for XGBoost. Multiplying these values with the outer and inner loop results in 4,200, and 23,040 trained models, respectively. Currently, there is no standard set of metrics for measuring the model performance of aging clocks based on feature space similarity. Therefore, a variety of metrics are reported. This study reported the mean absolute deviation (MAD), mean error (ME), Pearson correlation coefficient with the corresponding p-value (r, *P*), root mean squared error (RMSE), and R2. In this study, we use MAD as the primary metric. **Table 2** presents the average metric scores for each model.

In general, Elastic PCA demonstrated superior performance compared to the XGBoost model when trained on the full dataset. Elastic PCA exhibited a high correlation, 0.76 ($P<.001$), with a MAD of 7.02 years and a standard error of 0.38. Additionally, Elastic PCA demonstrated a higher degree of generalizability in the dataset compared to the XGBoost model. The XGBoost model was found to not generalize well on the dataset, which may be attributed to overfitting on the training dataset as the cross-validation training dataset may be too small. Given that the generalizability of the model strongly depends on the subset of the samples (as seen by the minimum and maximum scores in **Table 2**), it was decided to create an ensemble model to capture most of the variance between samples, this is a similar approach to bagging, but without bootstrapping the dataset. We chose to use this approach, because bootstrapping would use a subset of samples that resulted in decreased performance.

The ensemble model was trained on the whole training dataset, using the parameters outlined in **Supplementary Table 3**. The results are presented in **Table 3**. It can be observed that the XGBoost model was able to capture most of the signal in the train dataset. This is not surprising, as tree-based algorithms

tend to overfit. This property can be leveraged to identify signals that are most predictive for a certain dataset. These signals should then be validated to be certain that the model does not fit to noise. Therefore, we compared the features identified by the XGBoost model with the most important features from the Elastic PCA ensemble model, which demonstrated stronger generalizability.

3.1.1 Elastic PCA and principal components

Principal component analysis was conducted, and the projections were correlated with the available metadata. The result of this analysis revealed that the first principal component exhibited no substantial correlations with any of the metadata. However, it was observed that PG(0-20:0/22:4) was enriched within the first principal component. Notably, this lipid belongs to the class of glycerophospholipids. This lipid species plays a role in membrane signaling, which may explain the high variance in the concentration levels. Conversely, the second and third principal components demonstrated significant correlations with Shannon entropy of the samples, with Pearson correlation coefficients of 0.36 ($P<.001$) and 0.31 ($P<.001$), respectively.

Subsequently, an investigation was conducted to ascertain which lipids contributed predominantly to the sample entropy. This analysis revealed that dolichol-19 C95H157O and C95H160NO, were the primary contributors in 25 percent of the samples, while the remaining percentage was distributed among 55 other lipids. Consequently, we recalculated entropy based only on dolichols. This resulted in a strong correlation with a Pearson correlation coefficient of 0.82 ($P<.001$) between entropy and chronological age. We hypothesized that entropy is significantly higher in older samples and samples with a neurological disorder. To investigate this hypothesis, samples were divided into bins of 10 years, yielding six distinct bins. Samples from a specific age group were compared to those from the subsequent older age group (e.g., samples aged 20-30 years compared to those aged 30-40 years) using the Wilcoxon Rank Sum test. The results indicated significantly higher entropy levels in samples from the 30-40, 40-50, 50-60, and 60-70 age groups compared to their younger counterparts ($P<.001$ for each group), implying a substantial entropy increase through time, especially around the age of 40, as illustrated in **Figure 2**. Furthermore, the

study evaluated entropy levels in samples associated with neurological disorders such as ASD, SZ, and DS, comparing them to their respective control groups. We did not find significantly higher entropy levels for ASD, SZ and DS samples in comparison with their control group.

Importantly, age exhibited a significant correlation with principal component five (r=.58, *P*<.001), and principal component six (r=-.32, *P*<.001). While Elastic PCA utilizes dolichol as seen by its coefficients, principal component five also encompassed other molecules such as PG(17:1(9Z)/0:0), and PG(15:1(9Z)/0:0, implying the presence of additional aging-related processes not solely driven by dolichol or entropy.

Ethnicity was captured in both principal component five and principal component six. Both principal components exhibited a significant correlation with Han Chinese (r=0.48, *P*<.001, and r=-0.43, *P*<.001, respectively) and Caucasians (r=-0.32, *P*<.001, and r=0.31, *P*<.001, respectively). Consequently, it can be inferred that lipid concentration values may exhibit distinct baseline levels based on ethnicity. Conversely, sex exhibited no significant correlation with the principal components, suggesting, perhaps surprisingly, that sex does not account for substantial variance among the samples.

3.1.2 Overlapping features

We sought to identify commonalities in feature importance between the Elastic PCA and XGBoost ensemble. The purpose of this comparison was to gain a comprehensive understanding of the most important features in the dataset. As Elastic PCA has been demonstrated to possess relatively strong generalization abilities, whereas XGBoost is known for its tendency to overfit, we hypothesized that a combination of the top-ranking features identified by both models would result in an optimal feature set for the final model. The result of this comparison, presented in **Table 4**, reveals that dolichol molecules were consistently identified as the most significant features out of 163 features in two different models. In addition, we calculated the correlation between all molecules and age to determine the molecules that significantly correlate with age. Our initial findings revealed that numerous molecules exhibited significant increases with age in the training cohort, with most increases occurring between group 1 (0-20

years) and group 2 (20-40 years). All the results are shown in **Supplementary Table 4**. Notably, dolichols displayed a significant increase across multiple age groups. We found that dolichol increases linearly with age with increasing variance as shown in **Figure 3**. Specifically, dolichol-19 C95H160NO and C95H157O, and dolichol-20 C100H164ONa, C100H165O, C100H168NO increased significantly between group 1 and 2 ($P<.00004$ for each), as well as between group 2 and 3 ($P<.00004$ for each). However, only dolichol-20 C100H164ONa increased significantly between group 3 and 4 ($P<.00004$), and no molecules were found to increase significantly between group 4 and 5. The lack of significance might be due to the few samples and the high variance. Based on the findings, it was hypothesized that dolichol may serve as a biomarker that can be used to predict biological age as it increases several-fold.

3.2 DoliClock

Four regression models were constructed utilizing dolichol due to its significant age-related increase. Two models exclusively featured dolichol, while the other two models incorporated both dolichol and ethnicity as variables. The inclusion of ethnicity was motivated by the substantial correlation observed between ethnicity and principal components, indicating a potential influence of ethnicity on lipid concentration levels. We did not include sex, as we found no significant differences in lipid concentrations between males and females. The results are shown in **Table 5**. The XGBoost ensemble model based on dolichol and ethnicity performed best according to MAD. Therefore, we trained DoliClock, an XGBoost ensemble model, on the whole training dataset after nested stratified cross-validation comprising only WND samples using the parameters as provided in **Supplementary Table 5**. DoliClock achieved a MAD of 2.90, ME of 0.55, r of 0.95 ($P<.001$), RMSE of 6.23, and an R2 of 0.87 on the training set. On the test set, it achieved a MAD of 6.1, ME of 1.15, r of 0.72 ($P<.001$), RSE of 12.26, and an R2 of 0.49 on the test set. DoliClock was used to predict the age of samples with a neurological disorder as shown in **Figure 4a**.

We compared the intercept and slopes of the trendline based on WND, SZ, and ASD samples using a t-test. A trendline was not created for DS samples, as they contain relatively few samples in the same age range. It can be observed that the slopes of SZ and ASD samples are greater than the slope of WND

samples as shown in **Figure 4a**. The slope of SZ samples is not significantly greater compared to the slope of WND samples (P=0.72). However, the slope of ASD samples is significantly greater than the slope of WND samples (P<0.05). Specifically, we observed that the slope of ASD samples (1.13) is around 1.8 times the slope of WND samples (0.64). This indicates that, by this measure, ASD samples have an increased rate of aging, compared to WND samples. Furthermore, we calculated age_{Δ}, which is the difference between chronological and biological age, as shown in **Figure 4b**. We compared the age_{Δ} of ASD, SZ, and DS samples with WND samples. We found that ASD and SZ samples have a significantly greater age_{Δ} than WND samples (P<.05, P<.05, respectively). DS samples did not have a significant greater age_{Δ} (P<.08). We suspect that this is due to the sample size.

Additionally, we analyzed the feature importance of DoliClock and found that the dolichol molecules dolichol-20 C1001640Na, dolichol-19 C95H160NO, and dolichol-20 C100H168NO were most influential in determining biological age, as visualized in **Figure 5a-b**. A lower concentration level of these molecules corresponds to a lower predicted age, whereas a moderate or high concentration level corresponds to a higher predicted age. An exception is dolichol-20 C100H165O, which shows that lower dolichol levels are associated with a higher predicted age, whereas higher dolichol levels are associated with a lower predicted age. The same conclusion can be drawn for dolichol-19 C95H157O. Thus, it is possible that different molecules have different functions. Ethnicity seems to have a minor effect; however, ethnicities do affect the model differently. For instance, Han Chinese will be predicted to be younger while being Caucasian does not impact the prediction significantly. We also searched for interaction effects and suspected that dolichol molecules interact with each other as there is a strong correlation between their concentration levels, as presented in **Figure 6f**. The interaction effects are presented in **Figure 7**. We observed that generally low concentrations correspond to a younger predicted age. As there is vertical dispersion, the same concentration levels affect individuals differently. Vertical dispersion results from non-linear interaction effects. Interactions do not show causality; thus, we are not able to explain the cause of vertical dispersion. We also observed nonlinear effects. For instance, the

influence of dolichol concentration seems to wear off at higher concentration levels for dolichol-19 C95H160NO and dolichol-20 C100H164ONa. In addition, it seems that very low values can increase predicted age, while generally low values can decrease predicted age in dolichol-20 C100H168NO. At last, we analyzed the individual contributions, as presented in **Figure 5c**. The result showed that there are in general two groups when grouping by similarity. A group where dolichol concentration levels contribute to a lower age, and a group where dolichol concentration levels contribute to an older age. We can observe that most of the dolichol will contribute to a higher predicted age.

Also, the variance tends to increase with age, as presented in **Figure 3** and **Figure 6a-e**. The variance tends to increase throughout the whole lifetime. Especially around the age of 40, there is a significant difference in variance ($P<.001$ using the F-statistic).

3.2.1 Concentration levels in neurological disorders

Samples with neurological disorders were compared with WND samples. The samples were matched based on age and ethnicity. The results are presented in **Table 6**. We hypothesized that dolichol concentration levels would be significantly elevated in DS samples in comparison with WND samples. To test this hypothesis, the Wilcoxon Rank Sum Test was used with a Bonferroni correction of 0.003. After correction, we did not find any concentration level that was significantly greater in DS, ASD, and SZ samples. We must note that PMI was significantly different between WND samples and DS ($P<.001$), ASD ($P<.001$), and SZ ($P=0.008$) samples after correction. We suspect that this may affect the lipid concentration levels.

4. Discussion

4.1 Age acceleration and age_{Δ}

Our findings indicate a significant increase in age_{Δ} among samples with autism and schizophrenia compared to those without neurological diseases, while samples with Down syndrome do not show a significant increase in age_{Δ} compared to samples without neurological diseases. This discrepancy in age_{Δ} among samples with Down syndrome contradicts prior research findings that relied on epigenetic markers and neuroimaging (38, 39), which suggested premature aging in individuals with Down syndrome characterized by distinct DNA methylation patterns and protein expressions. However, the lack of significant age_{Δ} increase in our Down syndrome samples may be attributed to limitations in sample size, highlighting the importance of additional datasets to validate these results. Conflicting outcomes regarding schizophrenia have been reported in previous studies utilizing epigenetic clocks based on chronological age and mortality (41-43). Our findings suggest that lipidomic profiles may effectively capture discernible age-related signatures. While further exploration is warranted in brain aging research among individuals with autism (20, 44), our results reveal significant age acceleration in this group. These findings align with those of Mason et al. (67), suggesting potential involvement of genetic factors interlinked with both aging and autistic traits, or the influence of these traits on lifestyle, thereby impacting the aging process. To comprehensively understand the underlying mechanisms of age acceleration, further investigations, potentially integrating epigenetics, proteomics, and lipidomics, are essential.

4.2 Entropy

In addition to age acceleration, our findings provide evidence of a pronounced escalation in entropy, notably around the age of 40 years. This observation suggests the potential presence of a progressively dysregulated mevalonate pathway, culminating in the accumulation of dolichol. It is noteworthy that individuals afflicted with neurological disorders typically exhibit values that fall beyond the statistical range of two standard deviations when compared with the entropy levels of their healthier counterparts

below the age of 40. This implies that individuals with neurological disorders already manifest an elevated state of entropy. However, the precise underlying mechanistic basis for this phenomenon remains obscure, and it is anticipated that forthcoming research endeavors will yield greater insights into this matter.

Further, it was hypothesized that the strong correlation observed between entropy and chronological age reflects age-dependent dysregulation of pathways involving this lipid class. Dolichols, composed of a long-chain isoprenoid alcohol comprising saturated alpha-isoprene units, belong to the class of polyprenols, and are involved in intracellular molecular transport (68), N-glycosylation (69), and serve as a key product within the mevalonate pathway (70). The mevalonate pathway is responsible for the synthesis of cholesterol, dolichol, and ubiquinone among other molecules (71). Alterations in the concentration levels of these molecules, alongside an increase in dolichol during aging (72-76), suggest dysregulation of this pathway, potentially mediated by an age-related increase in HMG-CoA reductase, the rate-limiting enzyme of this pathway (77). These findings support the hypothesis of an association between entropy and age-dependent dysregulation of the mevalonate pathway.

4.3 Dolichol as an aging biomarker

Extensive research has been conducted on dolichol and aging across a variety of organisms, including mice (78-80), rats (72, 81-86), drosophila (87), and humans (73–75, 88–91). The amount of isoprene units can range from 15 to 23 in eukaryotic cells (92), depending on species. The chemical structure is shown in **Figure 6g-h**. Dolichol has been proposed as a biomarker of aging (93, 94). The studies generally reported an accumulation of dolichol in multiple tissues, although the concentration of dolichol can vary among different types of tissues. For example, Marino et al. (82), through studying the liver of rats and measuring the effects of caloric restriction on dolichol accumulation, demonstrated that dolichol concentration only increases if HMG-CoA reductase increases. They concluded that caloric restriction retards the rate of dolichol accumulation. This finding was also confirmed by other studies (83, 84). Sakakihara and Volpe (72) showed that dolichol concentration in brain tissue increases several-fold with aging. They also showed that dolichol increases during development, suggesting that it may have a developmental purpose.

Carroll, Guthrie, and Ravi (88) pointed out that dolichol concentration increases dramatically in adults compared to neonatal tissue, also suggesting a developmental purpose. The most abundant dolichol molecule in brain tissue has isoprene unit 19 (85). Although these studies have been conducted in rats, studies in humans have yielded similar results. The most commonly found dolichol molecules have isoprene unit 17 up to 21 in the human brain (89). As there are several isoprene units, and as the abundance can vary among them, individual isoprenologues may have different functions (90, 91). Sakakihara et al. (73) linked higher dolichol concentrations to neurodegenerative diseases, suspecting it is related to lysosomal dysfunction. Elevated levels of dolichol have also been observed in individuals with Down syndrome and autism (74, 75).

Despite the established correlation between dolichol and aging, the precise role of dolichol remains uncertain. Previous research has demonstrated that dolichol is involved in the transport between various organelles (68), N-glycosylation (69), and that it serves as the end product of the mevalonate pathway (70). The involvement of dolichols as lipid carriers for glycan precursors in protein N-glycosylation presents a plausible connection between Glycan aging biomarkers and biomarkers associated with dolichols, thereby suggesting a potential mechanistic relationship between the two (95, 96).

Furthermore, research indicates that caloric restriction can reduce dolichol accumulation (82). Nevertheless, the magnitude of this effect might vary based on the duration and timing of caloric restriction (83).

Current research has identified several potential explanations for the role of dolichol in aging. One possibility is that dolichol serves a protective mechanism to safeguard older membranes against the damage of free radicals (84). If this system becomes compromised or unstable, it may contribute to an accelerated rate of biological aging (97). Another possibility is that low-density lipoprotein (LDL) may be involved in dolichol accumulation (98) As this may impair lysosomal function, dolichol accumulation may be the result of autophagic degradation (84).

4.4 Machine Learning

The dataset used to develop DoliClock consisted of molecular entities that are identifiable by their chemical formula. This means the dataset has a high level of granularity. However, this level of granularity may obscure system-level behavior. For example, unused lipid species may hold more predictive power, and could therefore enhance the interpretability of aging of the prefrontal cortex. Therefore, incorporating more molecular categories and classes into the dataset may provide additional information. The use of ratios may further enhance efficacy.

DoliClock was developed using an ensemble of XGBoost models, whose predictions were then averaged. This ensemble consists of several models that demonstrated optimal performance during the nested cross-validation procedure. The advantage of this architecture lies in its ability to combine weaker models to obtain better results in terms of generalizability. However, to assess the generalizability more robustly, it is necessary to evaluate the model on an independent dataset. In addition, it is important to acknowledge that the results may be influenced by the lipidomics approach used. Hence, it is advisable to utilize the same methodology as described by Yu et al. (49).

Additionally, DoliClock consists of homogeneous models, namely XGBoost models. Another approach to improve the efficacy is stacking, which uses a combination of heterogeneous models, such as XGBoost and Elastic Net, and a meta-estimator to produce an outcome. In this technique, a meta-model is trained on the estimators produced by the heterogenous models to predict an output. However, the strength of our approach is the interpretability of our model. The use of heterogeneous models can make the interpretation of the model challenging or computationally intractable.

4.5 Biological relevance

As entropy is highly correlated with age, we developed a model that incorporated entropy and dolichol, showing superior performance. We observed a lower MAD in predicting chronological age. However, the slopes of schizophrenia and autism tended to converge (**Supplementary Figure 4**), resembling the pattern

seen in samples without neurological disease. This convergence, however, highlights a limitation: as models become more adept at predicting chronological age, they tend to lose biological interpretability and instead fit to noise. Complex models that can overfit to low-dimensional data might be especially prone as shown by our model.

Our findings indicate that using chronological age as the sole regressand may not be appropriate when training complex models and that it is not representative of biological age. Consequently, traditional metrics like the median absolute deviation might not adequately preserve biological relevance in these models. We propose exploring alternative regressands, such as mortality scores or other metrics that are a more robust measure of biological age, thereby enhancing the interpretability and reliability of complex predictive models.

Moreover, there exists a degree of ambiguity whether the measurements conducted pertain to tissue-specific aging or capture systemic aging. Despite the focus on brain tissue, the regression on chronological age implies a representation of the broader systemic aging process. Additional research might further elucidate and refine this aspect.

4.6 Samples and background information

As data is collected post-mortem, the cause of death may be an important factor to consider. The cause may affect lipid concentration levels. Also, it may be possible that the post-mortem interval affects the lipid stability and thus, the concentration levels. In addition, medicine usage, something that is common among individuals with schizophrenia, is unknown in this cohort and may affect lipid composition (99).

4.7 Future directions

There are several avenues for future research. First, the findings of this study should be validated using an external dataset with other independent samples, through a larger dataset, and/or using biological experiments. It must be noted that the availability of post-mortem tissues is very limited.

Secondly, further research could be conducted to better understand the mechanisms underlying the strong association between dolichol and aging, specifically focusing on the differences between dolichol-19 and dolichol-20 as they are likely to behave differently. Another approach would be to investigate potential links to autophagic degradation and lipid peroxidation.

A third direction would involve a more detailed assessment of dolichol levels in different parts of the body to identify potential clinical markers for aging or find a surrogate to measure brain aging. However, it should be noted that dolichol values can vary among different types of tissue, so it may be necessary to include different features.

The last direction would be to investigate the relationship between entropy, dolichol, and aging. Entropy and dolichol-19 concentration were significantly greater around the age of 40. As variance also increases around that age, it might be interesting to combine information theory with lipids to find new leads from a different perspective.

5. Conclusion

The primary objective of this study was to examine the feasibility of utilizing lipidomics data for predicting the biological age of the prefrontal cortex and identify the most important molecule due to the post-mortem nature of the dataset. To achieve this goal, we developed a novel tool called DoliClock, specifically designed for age prediction based on lipids extracted from the prefrontal cortex. DoliClock demonstrated the capacity to discern distinct age-related impacts resulting from variations in isotopic composition within the dolichol molecule. This observation represents a novel contribution to the existing body of literature, as prior research has not elucidated such effects associated with isotopes. Furthermore, DoliClock exhibited the capability to estimate accelerated aging in individuals with neurological disorders, such as Down syndrome, schizophrenia, and autism, which are commonly associated with accelerated aging phenotypes. Our findings demonstrated significant age acceleration in samples with autism and significant age_{Δ} for samples with autism and schizophrenia. In addition, our results suggest

that entropy increases significantly around the age of 40 and that individuals with neurological disorders already suffer from heightened levels of entropy. Also, DoliClock offers a visualization of key features, facilitating a deeper understanding of age-related characteristics at both individual and population levels. These outcomes provide compelling evidence supporting the feasibility of predicting the biological age of the prefrontal cortex utilizing lipidomics data and lay the foundation for future investigations aiming to identify specific molecules using an interpretable machine learning framework.

Data availability statement

Data repository will be made available after acceptance.

References

(1) E. L. Schneider and J. M. Guralnik, “The aging of America. Impact on health care costs,” JAMA, vol. 263, no. 17, pp. 2335–2340, May 1990.

(2) V. L. Feigin et al., “The global burden of neurological disorders: Translating evidence into policy,” The Lancet Neurology, vol. 19, no. 3, pp. 255–265, 2020, issn: 1474-4422. doi: https: / / doi . org / 10 . 1016 / S1474 - 4422(19) 30411 - 9. [Online]. Available: https : / / www . sciencedirect.com/science/article/pii/S1474442219304119.

(3) C. for Disease Control and P. (CDC), “Trends in aging–United States and worldwide,” MMWR Morb Mortal Wkly Rep, vol. 52, no. 6, pp. 101–104, Feb. 2003.

(4) Y. Hou et al., “Ageing as a risk factor for neurodegenerative disease,” Nat Rev Neurol, vol. 15, no. 10, pp. 565–581, Oct. 2019.

(5) B. R. Troen, “The biology of aging,” Mt Sinai J Med, vol. 70, no. 1, pp. 3–22, Jan. 2003.

(6) Q. D. Nguyen, E. M. Moodie, M. F. Forget, P. Desmarais, M. R. Keezer, and C. Wolfson, “Health Heterogeneity in Older Adults: Exploration in the Canadian Longitudinal Study on Aging,” J Am Geriatr Soc, vol. 69, no. 3, pp. 678–687, Mar. 2021.

(7) N. R. Balcombe and A. Sinclair, “Ageing: definitions, mechanisms and the magnitude of the problem,” Best Pract Res Clin Gastroenterol, vol. 15, no. 6, pp. 835–849, Dec. 2001.

(8) C. L´opez-Ot´ın, M. A. Blasco, L. Partridge, M. Serrano, and G. Kroemer, “Hallmarks of aging: An expanding universe,” Cell, vol. 186, no. 2, pp. 243–278, 2023, issn: 0092-8674. doi: https : / / doi . org / 10 . 1016 / j . cell . 2022 . 11 . 001. [Online]. Available: https : / / www .

sciencedirect.com/science/article/pii/S0092867422013770.

(9) J. R. Sargent, D. R. Tocher, and J. G. Bell, “4 - the lipids,” in Fish Nutrition (Third Edition), J. E. Halver and R. W. Hardy, Eds., Third Edition, San Diego: Academic Press, 2003, pp. 181–257, isbn: 978-0-12-319652-1. doi: https://doi.org/10.1016/B978- 012319652-1/50005-7. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780123196521500057.

(10) H. Sunshine and M. L. Iruela-Arispe, “Membrane lipids and cell signaling,” en, Curr. Opin. Lipidol., vol. 28, no. 5, pp. 408–413, Oct. 2017.

(11) Q. Yu et al., “Lipidome alterations in human prefrontal cortex during development, aging, and cognitive disorders,” Molecular Psychiatry, vol. 25, no. 11, pp. 2952–2969, Nov. 2020, issn: 1476-5578. doi: 10.1038/s41380-018-0200-8. [Online]. Available: https://doi.org/10.1038/s41380-018-0200-8.

(12) M. Jov´e, N. Mota-Martorell, P. Torres, M. Portero-Otin, I. Ferrer, and R. Pamplona, “New insights into human prefrontal cortex aging with a lipidomics approach,” en, Expert Rev. Proteomics, vol. 18, no. 5, pp. 333–344, May 2021.

(13) B. S. McEwen and J. H. Morrison, “The brain on stress: Vulnerability and plasticity of the prefrontal cortex over the life course,” Neuron, vol. 79, no. 1, pp. 16–29, 2013, issn: 0896-6273. doi: https : / / doi . org / 10 . 1016 / j . neuron . 2013 . 06 . 028. [Online]. Available: https : //www.sciencedirect.com/science/article/pii/S0896627313005448.

(14) Y.-C. Kao, P.-C. Ho, Y.-K. Tu, I.-M. Jou, and K.-J. Tsai, “Lipids and alzheimer’s disease,”

en, Int. J. Mol. Sci., vol. 21, no. 4, p. 1505, Feb. 2020.

(15) M. Shichiri, “The role of lipid peroxidation in neurological disorders,” Journal of Clinical Biochemistry and Nutrition, vol. advpub, pp. 14–10, 2014. doi: 10.3164/jcbn.14-10.

(16) R. John, P. Bruce D., M. Andrew, and C. Jan, “The contribution of early traumatic events to schizophrenia in some patients: A traumagenic neurodevelopmental model,” Psychiatry, vol. 64, no. 4, pp. 319–345, 2001, PMID: 11822210. doi: 10.1521/psyc.64.4.319.18602. eprint: https://www.tandfonline.com/doi/pdf/10.1521/psyc.64.4.319.18602. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1521/psyc.64.4.319.18602.

(17) L. D. Selemon and N. Zecevic, “Schizophrenia: A tale of two critical periods for prefrontal cortical development,” Translational Psychiatry, vol. 5, no. 8, e623–e623, Aug. 2015, issn: 2158-3188. doi: 10.1038/tp.2015.115. [Online]. Available: https://doi.org/10.1038/tp.2015.115.

(18) C. T. Wong, J. Wais, and D. A. Crawford, “Prenatal exposure to common environmental factors affects brain lipids and increases risk of developing autism spectrum disorders,” European Journal of Neuroscience, vol. 42, no. 10, pp. 2742–2760, 2015. doi: https://doi.org/10.1111/ejn.13028. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ejn.13028. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/ejn.13028.

(19) A. H. Bittles and E. J. Glasson, “Clinical, social, and ethical implications of changing life expectancy in down syndrome,” en, Dev. Med. Child Neurol., vol. 46, no. 4, pp. 282–286, Apr.

2004.

(20) E. A. Perkins and K. A. Berkman, “Into the unknown: Aging with autism spectrum disorders,” en, Am. J. Intellect. Dev. Disabil., vol. 117, no. 6, pp. 478–496, Nov. 2012.

(21) C. Hjorthøj, A. E. St urup, J. J. McGrath, and M. Nordentoft, “Years of potential life lost and life expectancy in schizophrenia: A systematic review and meta-analysis,” en, Lancet Psychiatry, vol. 4, no. 4, pp. 295–301, Apr. 2017.

(22) G. Hannum et al., “Genome-wide methylation profiles reveal quantitative views of human aging rates,” Mol Cell, vol. 49, no. 2, pp. 359–367, Jan. 2013.

(23) S. Horvath, “DNA methylation age of human tissues and cell types,” Genome Biol, vol. 14, no. 10, R115, 2013.

(24) S. Horvath and K. Raj, “DNA methylation-based biomarkers and the epigenetic clock theory of ageing,” Nat Rev Genet, vol. 19, no. 6, pp. 371–384, Jun. 2018.

(25) D. H. Meyer and B. Schumacher, “BiT age: A transcriptome-based aging clock near the theoretical limit of accuracy,” Aging Cell, vol. 20, no. 3, e13320, Mar. 2021.

(26) A. A. Johnson, M. N. Shokhirev, and B. Lehallier, “The protein inputs of an ultra-predictive aging clock represent viable anti-aging drug targets,” Ageing Res Rev, vol. 70, p. 101 404, Sep. 2021.

(27) M. Unfried et al., “LipidClock: A Lipid-Based Predictor of Biological Age,” Front Aging, vol. 3, p. 828 239, 2022.

(28) N. Hwangbo et al., “A Metabolomic Aging Clock Using Human Cerebrospinal Fluid,” J Geron-

tol A Biol Sci Med Sci, vol. 77, no. 4, pp. 744–754, Apr. 2022.

(29) M. E. Levine et al., “An epigenetic biomarker of aging for lifespan and healthspan,” Aging (Albany NY), vol. 10, no. 4, pp. 573–591, Apr. 2018.

(30) A. T. Lu et al., “DNA methylation GrimAge strongly predicts lifespan and healthspan,” Aging (Albany NY), vol. 11, no. 2, pp. 303–327, Jan. 2019.

(31) A. E. Tarkhov, K. A. Denisov, and P. O. Fedichev, “Aging clocks, entropy, and the limits of age-reversal,” bioRxiv, 2022. doi: 10.1101/2022.02.06.479300. eprint: https://www. biorxiv . org / content / early / 2022 / 10 / 11 / 2022 . 02 . 06 . 479300 . full . pdf. [Online]. Available: https://www.biorxiv.org/content/early/2022/10/11/2022.02.06.479300.

(32) F. Galkin, P. Mamoshina, K. Kochetov, D. Sidorenko, and A. Zhavoronkov, “DeepMAge: A Methylation Aging Clock Developed with Deep Learning,” Aging Dis, vol. 12, no. 5, pp. 1252–1262, Aug. 2021.

(33) E. Bobrov et al., “PhotoAgeClock: deep learning algorithms for development of non-invasive visual biomarkers of aging,” Aging (Albany NY), vol. 10, no. 11, pp. 3249–3259, Nov. 2018.

(34) D. Doran, S. Schulz, and T. R. Besold, “What does explainable ai really mean? a new conceptualization of perspectives,” 2017. doi: 10.48550/ARXIV.1710.00794. [Online]. Available: https://arxiv.org/abs/1710.00794.

(35) S. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” 2017. doi: 10.48550/ARXIV.1705.07874. [Online]. Available: https://arxiv.org/abs/1705.07874.

(36) A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, “Grad-cam++: Gen-

eralized gradient-based visual explanations for deep convolutional networks," in 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 2018, pp. 839–847. doi: 10.1109/WACV.2018.00097.

(37) S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. M uller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," PLOS ONE, vol. 10, no. 7, pp. 1–46, Jul. 2015. doi: 10.1371/journal.pone.0130140. [Online]. Available: https://doi.org/10.1371/journal.pone.0130140.

(38) S. Horvath et al., "Accelerated epigenetic aging in Down syndrome," Aging Cell, vol. 14, no. 3, pp. 491–495, Jun. 2015.

(39) J. H. Cole et al., "Brain-predicted age in down syndrome is associated with beta amyloid deposition and cognitive decline," en, Neurobiol. Aging, vol. 56, pp. 41–49, Aug. 2017.

(40) B. Kirkpatrick and B. K. Kennedy, "Accelerated aging in schizophrenia and related disorders: Future research," Schizophrenia Research, vol. 196, pp. 4–8, 2018, Accelerated Aging in Schizophrenia, issn: 0920-9964. doi: https : / / doi . org / 10 . 1016 / j . schres . 2017 . 06 . 034. [Online]. Available: https : / / www . sciencedirect . com / science / article / pii / S0920996417303730.

(41) A. T. Higgins-Chen, M. P. Boks, C. H. Vinkers, R. S. Kahn, and M. E. Levine, "Schizophrenia and epigenetic aging biomarkers: Increased mortality, reduced cancer risk, and unique clozapine effects," en, Biol. Psychiatry, vol. 88, no. 3, pp. 224–235, Aug. 2020.

(42) J. Teeuw et al., "Accelerated aging in the brain, epigenetic aging in blood, and polygenic risk

for schizophrenia," en, Schizophr. Res., vol. 231, pp. 189–197, May 2021.

(43) S. Okazaki et al., "Epigenetic clock analysis of blood samples from japanese schizophrenia patients," en, NPJ Schizophr., vol. 5, no. 1, p. 4, Feb. 2019.

(44) F. Happ´e and R. A. Charlton, "Aging in autism spectrum disorders: A mini-review," en, Gerontology, vol. 58, no. 1, pp. 70–78, 2012.

(45) K. L. Thrush et al., "Aging the brain: multi-region methylation principal component based clock in the context of Alzheimer's disease," Aging (Albany NY), vol. 14, no. 14, pp. 5641–5668, Jul. 2022.

(46) M. E. Levine, A. T. Lu, D. A. Bennett, and S. Horvath, "Epigenetic age of the pre-frontal cortex is associated with neuritic plaques, amyloid load, and alzheimer's disease related cognitive functioning," en, Aging (Albany NY), vol. 7, no. 12, pp. 1198–1211, Dec. 2015.

(47) J. S. O'Brien and E. L. Sampson, "Lipid composition of the normal human brain: Gray matter, white matter, and myelin*," Journal of Lipid Research, vol. 6, no. 4, pp. 537–544, 1965, issn: 0022-2275. doi: https://doi.org/10.1016/S0022-2275(20)39619-X. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S002222752039619X.

(48) A. Naud´ı et al., "Chapter five - lipidomics of human brain aging and alzheimer's disease pathology," in Omic Studies of Neurodegenerative Disease: Part B, ser. International Review of Neurobiology, M. J. Hurley, Ed., vol. 122, Academic Press, 2015, pp. 133–189. doi: https://doi.org/10.1016/bs.irn.2015.05.008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0074774215000240.

(49) Q. Yu et al., “Lipidome alterations in human prefrontal cortex during development, aging, and cognitive disorders,” Mol Psychiatry, vol. 25, no. 11, pp. 2952–2969, Nov. 2020.

(50) I. T. Jolliffe, Principal Component Analysis. Berlin; New York: Springer-Verlag, 1986.

(51) F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

(52) D. P. Nassim Nicholas Taleb Pierre Zalloua, Informational rescaling of pca maps with application to genetics.

(53) D. E. Platt et al., “Autosomal genetics and y-chromosome haplogroup l1b-m317 reveal mount lebanon maronites as a persistently non-emigrating population,” European Journal of Human Genetics, vol. 29, no. 4, pp. 581–592, Apr. 2021, issn: 1476-5438. doi: 10.1038/s41431-020-00765-x. [Online]. Available: https://doi.org/10.1038/s41431-020-00765-x.

(54) H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 67, no. 2, pp. 301–320, 2003.

(55) T. Chen and C. Guestrin, “XGBoost,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, Aug. 2016. doi: 10 . 1145 / 2939672.2939785. [Online]. Available: https://doi.org/10.1145%2F2939672.2939785.

(56) K. T. Do et al., “Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies,” Metabolomics, vol. 14, no. 10, p. 128, Sep. 2018.

(57) I.-K. Yeo and R. A. Johnson, “A new family of power transformations to improve normality or symmetry,” Biometrika, vol. 87, no. 4, pp. 954–959, 2000, issn: 00063444. [Online]. Available: http://www.jstor.org/stable/2673623 (visited on 12/29/2022).

(58) C. E. Shannon, “A mathematical theory of communication,” The Bell System Technical Journal, vol. 27, pp. 379–423, 1948. [Online]. Available: http://plan9.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf (visited on 04/22/2003).

(59) Anaconda software distribution, 2020. [Online]. Available: https://docs.anaconda.com/.

(60) C. R. Harris et al., “Array programming with NumPy,” Nature, vol. 585, no. 7825, pp. 357–362, Sep. 2020. doi: 10.1038/s41586-020-2649-2. [Online]. Available: https://doi.org/10.1038/s41586-020-2649-2.

(61) W. McKinney, “Data Structures for Statistical Computing in Python,” in Proceedings of the 9th Python in Science Conference, S. van der Walt and J. Millman, Eds., 2010, pp. 56–61. doi: 10.25080/Majora-92bf1922-00a.

(62) P. Virtanen et al., “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python,” Nature Methods, vol. 17, pp. 261–272, 2020. doi: 10.1038/s41592-019-0686-2.

(63) P. T. Inc. “Collaborative data science.” (2015), [Online]. Available: https://plot.ly.

(64) J. D. Hunter, “Matplotlib: A 2d graphics environment,” Computing in Science & Engineering, vol. 9, no. 3, pp. 90–95, 2007. doi: 10.1109/MCSE.2007.55.

(65) A. D. Shieh and Y. S. Hung, “Detecting outlier samples in microarray data,” Statistical Applications in Genetics and Molecular Biology, vol. 8, no. 1, 2009. [Online]. Available: http:

//www.bepress.com/sagmb/vol8/iss1/art13/.

(66) A. T. Higgins-Chen et al., “A computational solution for bolstering reliability of epigenetic clocks: Implications for clinical trials and longitudinal tracking,” Nature Aging, vol. 2, no. 7, pp. 644–661, Jul. 2022, issn: 2662-8465. doi: 10 . 1038 / s43587 - 022 - 00248 - 2. [Online]. Available: https://doi.org/10.1038/s43587-022-00248-2.

(67) D. Mason et al., “Autistic traits are associated with faster pace of aging: Evidence from the dunedin study at age 45,” Autism Research, vol. 14, no. 8, pp. 1684–1694, 2021. doi: https: //doi.org/10.1002/aur.2534. eprint: https://onlinelibrary.wiley.com/doi/pdf/ 10.1002/aur.2534. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10. 1002/aur.2534.

(68) T. Chojnacki and G. Dallner, “The biological role of dolichol,” Biochem J, vol. 251, no. 1, pp. 1–9, Apr. 1988.

(69) M. Welti, “Regulation of dolichol-linked glycosylation,” Glycoconj J, vol. 30, no. 1, pp. 51–56, Jan. 2013.

(70) J. L. Goldstein and M. S. Brown, “Regulation of the mevalonate pathway,” Nature, vol. 343, no. 6257, pp. 425–430, Feb. 1990, issn: 1476-4687. doi: 10.1038/343425a0. [Online]. Available: https://doi.org/10.1038/343425a0.

(71) H. Karlic and F. Varga, “Mevalonate pathway,” 2019.

(72) Y. Sakakihara and J. J. Volpe, “Dolichol deposition in developing mammalian brain: content of free and fatty-acylated dolichol and proportion of specific isoprenologues,” Brain Res, vol. 316,

no. 2, pp. 255–262, Jun. 1984.

(73) Y. Sakakihara, T. Imabayashi, Y. Suzuki, and S. Kamoshita, “Elevated levels of dolichol in the brains of mucopolysaccharidosis and related disorders,” Molecular and Chemical Neuropathology, vol. 22, no. 2, pp. 97–103, Jun. 1994, issn: 1044-7393. doi: 10.1007/BF03160098. [Online]. Available: https://doi.org/10.1007/BF03160098.

(74) B. W. L. Brooksbank and M. Martinez, “Lipid abnormalities in the brain in adult down’s syndrome and alzheimer’s disease,” Molecular and Chemical Neuropathology, vol. 11, no. 3, pp. 157–185, Dec. 1989, issn: 1044-7393. doi: 10 . 1007 / BF03160049. [Online]. Available: https://doi.org/10.1007/BF03160049.

(75) R. K. Kurup and P. A. Kurup, “A hypothalamic digoxin-mediated model for autism,” Int J Neurosci, vol. 113, no. 11, pp. 1537–1559, Nov. 2003.

(76) Y. Zhang, E.-L. Appelkvist, K. Kristensson, and G. Dallner, “The lipid compositions of different regions of rat brain during development and aging,” Neurobiology of Aging, vol. 17, no. 6, pp. 869–875, 1996, issn: 0197-4580. doi: https://doi.org/10.1016/S0197-4580(96) 00076- 0. [Online]. Available: https://www.sciencedirect.com/science/article/pii/ S0197458096000760.

(77) G. P. Hooff et al., “Brain isoprenoids farnesyl pyrophosphate and geranylgeranyl pyrophosphate are increased in aged mice,” Molecular Neurobiology, vol. 46, no. 1, pp. 179–185, Jun. 2012. doi: 10.1007/s12035-012-8285-6. [Online]. Available: https://doi.org/10.1007/ s12035-012-8285-6.

(78) D. C. Crick and J. W. Rip, “Age-associated changes in dolichol and dolichyl phosphate metabolism in the kidneys and liver of mice,” Biochim Biophys Acta, vol. 1004, no. 2, pp. 180–186, Aug. 1989.

(79) R. K. Pullarkat, H. Reha, and P. S. Pullarkat, “Age-associated increase of free dolichol levels in mice,” Biochimica et Biophysica Acta (BBA) - Lipids and Lipid Metabolism, vol. 793, no. 3, pp. 494–496, 1984, issn: 0005-2760. doi: https : / / doi . org / 10 . 1016 / 0005 - 2760(84) 90269- 8. [Online]. Available: https://www.sciencedirect.com/science/article/pii/ 0005276084902698.

(80) W. G. Wood, G. Y. Sun, and F. Schroeder, “Membrane properties of dolichol in different age groups of mice,” Chem Phys Lipids, vol. 51, no. 3-4, pp. 219–226, Nov. 1989.

(81) B. Dini et al., “Effects of ageing and increased haemolysis on the levels of dolichol in rat spleen,” Exp Gerontol, vol. 37, no. 1, pp. 99–105, Dec. 2001.

(82) M. Marino, V. Pallottini, C. D’Eramo, G. Cavallini, E. Bergamini, and A. Trentalance, “Age-related changes of cholesterol and dolichol biosynthesis in rat liver,” Mech Ageing Dev, vol. 123, no. 8, pp. 1183–1189, Apr. 2002.

(83) C. Dolfi et al., “The age-related accumulation of dolichol in rat liver may correlate with expectation of life,” Biogerontology, vol. 4, no. 2, pp. 113–118, 2003.

(84) G. Cavallini et al., “Dolichol: A Component of the Cellular Antioxidant Machinery,” Lipids, vol. 51, no. 4, pp. 477–486, Apr. 2016.

(85) M. Ishinaga, “Effects of aging and diet on the accumulation of dolichol in rat tissues,” Biochem

Cell Biol, vol. 74, no. 2, pp. 265–270, 1996.

(86) R. K. Keller and S. W. Nellis, “Quantitation of dolichyl phosphate and dolichol in major organs of the rat as a function of age,” Lipids, vol. 21, no. 5, pp. 353–355, May 1986.

(87) S. M. Sternick, H. R. Massie, and S. Whitney, “Changes with ageing in total dolichol and dolichol fractions in drosophila,” Mechanisms of Ageing and Development, vol. 67, no. 1, pp. 91–99, 1993, issn: 0047-6374. doi: https : / / doi . org / 10 . 1016 / 0047 - 6374(93) 90114- 7. [Online]. Available: https://www.sciencedirect.com/science/article/pii/ 0047637493901147.

(88) K. K. Carroll, N. Guthrie, and K. Ravi, “Dolichol: function, metabolism, and accumulation in human tissues,” Biochem Cell Biol, vol. 70, no. 6, pp. 382–384, Jun. 1992.

(89) C. Edlund, M. S oderberg, K. Kristensson, and G. Dallner, “Ubiquinone, dolichol, and cholesterol metabolism in aging and alzheimer’s disease,” Biochemistry and Cell Biology, vol. 70, no. 6, pp. 422–428, 1992, PMID: 1449707. doi: 10.1139/o92-065. eprint: https://doi.org/ 10.1139/o92-065. [Online]. Available: https://doi.org/10.1139/o92-065.

(90) M. Andersson, E. L. Appelkvist, K. Kristensson, and G. Dallner, “Distribution of dolichol and dolichyl phosphate in human brain,” J Neurochem, vol. 49, no. 3, pp. 685–691, Sep. 1987.

(91) Y. Sakakihara and J. J. Volpe, “Dolichol in human brain: regional and developmental aspects,” J Neurochem, vol. 44, no. 5, pp. 1535–1540, May 1985.

(92) H. Zhang et al., “Dolichol biosynthesis and its effects on the unfolded protein response and

abiotic stress resistance in arabidopsis," en, Plant Cell, vol. 20, no. 7, pp. 1879–1898, Jul. 2008.

(93) I. Parentini, G. Cavallini, A. Donati, Z. Gori, and E. Bergamini, "Accumulation of dolichol in older tissues satisfies the proposed criteria to be qualified a biomarker of aging," J Gerontol A Biol Sci Med Sci, vol. 60, no. 1, pp. 39–43, Jan. 2005.

(94) M. Marino et al., "Age-dependent accumulation of dolichol in rat liver: is tissue dolichol a biomarker of aging?" J Gerontol A Biol Sci Med Sci, vol. 53, no. 2, pp. 87–93, Mar. 1998.

(95) J. Kriˇsti´c et al., "Glycans are a novel biomarker of chronological and biological ages," en, J. Gerontol. A Biol. Sci. Med. Sci., vol. 69, no. 7, pp. 779–789, Jul. 2014.

(96) V. Cantagrel and D. J. Lefeber, "From glycosylation disorders to dolichol biosynthesis defects: A new class of metabolic diseases," en, J. Inherit. Metab. Dis., vol. 34, no. 4, pp. 859–867, Aug. 2011.

(97) E. Bergamini et al., "Ageing and oxidative stress: a role for dolichol in the antioxidant machinery of cell membranes?" J Alzheimers Dis, vol. 6, no. 2, pp. 129–135, Apr. 2004.

(98) J. W. Rip, M. M. Blais, and L. W. Jiang, "Low-density lipoprotein as a transporter of dolichol intermediates in the mammalian circulation," Biochem J, vol. 297 (Pt 2), no. Pt 2, pp. 321–325, Jan. 1994.

(99) A. K. Mantel-Teeuwisse, J. M. Kloosterman, A. H. Maitland-van der Zee, O. H. Klungel, A. J. Porsius, and A. de Boer, "Drug-Induced lipid changes: a review of the unintended effects of some commonly used drugs on serum lipid levels," Drug Saf, vol. 24, no. 6, pp. 443–456, 2001.

Table legend

Table 1. Search space used by grid search for elastic net with PCA, and XGBoost.

Table 2. Results of nested stratified cross-validation. The best-performing model is shown in bold. Standard error of the mean is shown in brackets. Maximum and minimum scores per model are provided. The discrepancy between maximum and minimum scores is caused by high levels of variability.

Table 2 footnotes: [a]$P<.001$ for Pearson correlation coefficient; [b]$P=.005$ for Pearson correlation coefficient;

Table 3. Performance of the ensemble models trained on WND test samples. The best-performing model is shown in bold.

Table 3 footnotes: [a]$P<.001$ for Pearson correlation coefficient

Table 4. A comparison of the top 5 features identified by Elastic PCA ensemble and XGBoost ensemble with their corresponding feature ranking among all features.

Table 5. Results of nested stratified cross-validation based on only dolichol and dolichol and ethnicity. The best-performing model is shown in bold. Standard error of the mean is shown in brackets. Maximum and minimum scores per model are provided. The discrepancy between maximum and minimum scores is caused by high levels of variability.

Table 6. Dolichol concentration levels $x10^3$ among disorders. Delta shows the average change between case and control.

Supplementary Table 1. Demographics and count of samples of the original dataset.

Supplementary Table 2. Demographics and count of samples of the processed dataset.

Supplementary Table 3. Chosen hyperparameters for Elastic PCA and XGBoost. The final model consists of an ensemble of these parameters. Duplicates were removed.

Supplementary Table 4. Statistical significance of molecules increasing with age. Molecules are sorted by frequency of being statistically significant.

Supplementary Table 5. Chosen hyperparameters DoliChlock. The final model consists of an ensemble of these parameters. Duplicates were removed.

Figure legend

Figure 1. **General framework of model development. (a) The model is trained on a feature matrix. The feature matrix contains several features from different independent samples. (b) Model predictions are compared with the corresponding chronological age. (c) Explainable AI makes it possible to explain the behavior of the model and what contributes to the prediction on a global level. (d) Explainable AI makes it possible to explain feature importance per individual. For instance, one feature may contribute more towards aging than another feature for a given individual.**

Figure 2. Shannon entropy divided into age bins. Calculations are based on dolichol.

Figure 3. Summed dolichol concentration levels with a 95% confidence band using quantile regression.

Figure 4. **(a) Prediction of DoliClock for each sample. A trendline was calculated for each diagnosis using OLS. In addition, a 95% confidence interval band was calculated samples without neurological disease using a quantile gradient boosting regressor. Blue refers to samples without neurological disease, red to samples with autism, green to samples with schizophrenia, and purple to**

samples with Down syndrome. (b) Age delta of samples without neurological disease and samples with autism, schizophrenia or Down syndrome.

Figure 5. (a) Average contribution of features to predicting age. A higher concentration value generally contributes to an older age. A low concentration level generally contributes to a younger age. (b) The average contribution of features to predicting age, stratified by age groups. If a feature has a blue color, it contributes to a younger age. If it has a red color, it contributes to an older age. Dolichol-20 C100H164ONa has the most predictive power, given the range. (c) Contribution of dolichol to predicted age among individuals without neurological disease. Dolichol concentration levels are colored blue if they contribute to a younger age and red if they contribute to an older age. Samples are clustered using hierarchical clustering.

Figure 6. (a-e) Concentration levels of dolichol on the y-axis and age on the x-axis. A quadratic spline is fitted with a 95% confidence bound. Overall variance increases over the years. The correlation with age is also reported. (f) A heatmap showing the correlation between the features. (g) Chemical structure of dolichol-19. (h) Chemical structure of dolichol-20.

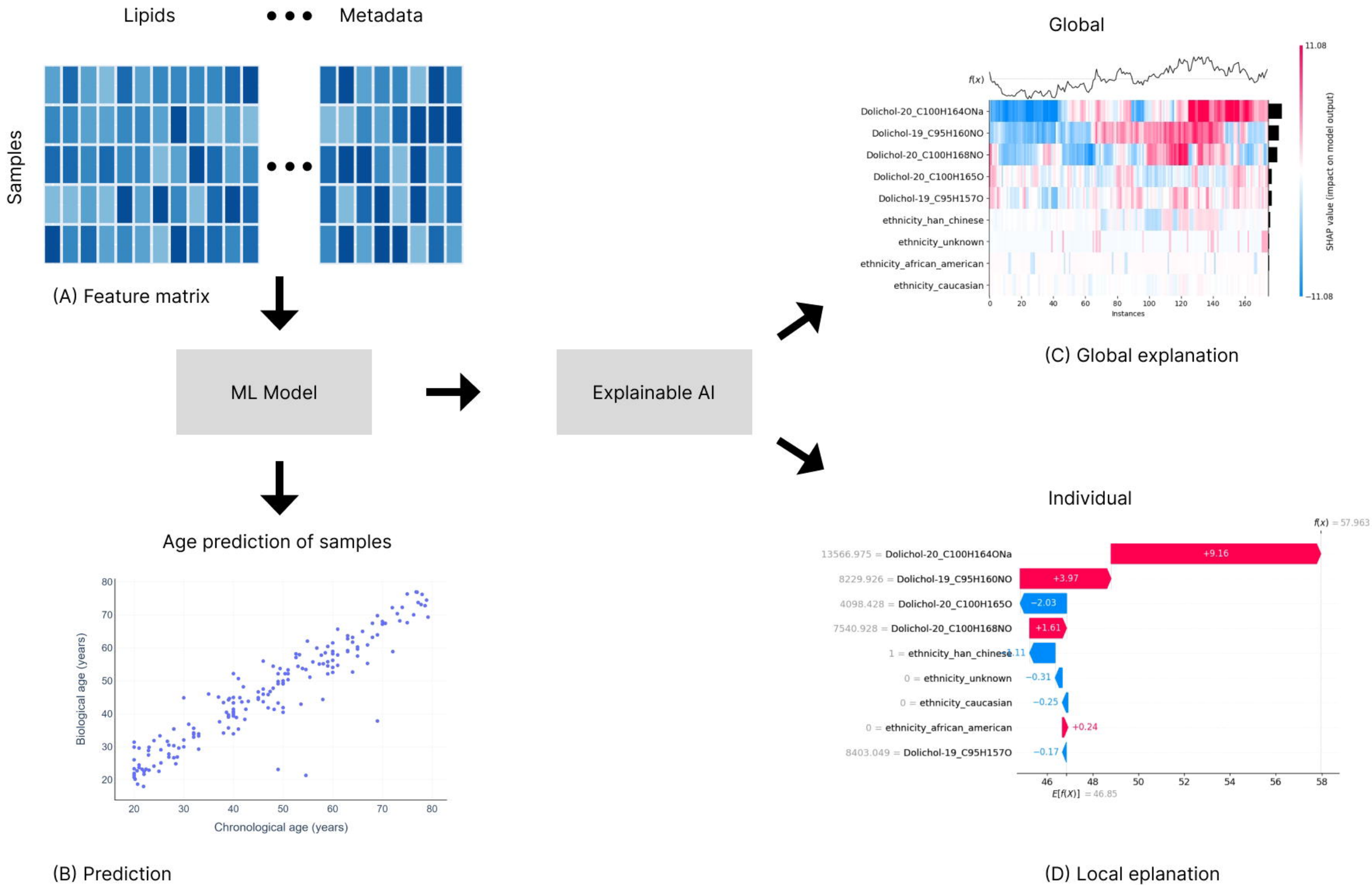
Figure 7. The interactions between different dolichol molecules based on vertical dispersion. Vertical dispersion is present when the same concentration level of a molecule can affect samples differently. This is caused by non-linear interactions between features. The feature used for coloring is chosen automatically based on the feature that is most likely to drive the interactions with another feature.

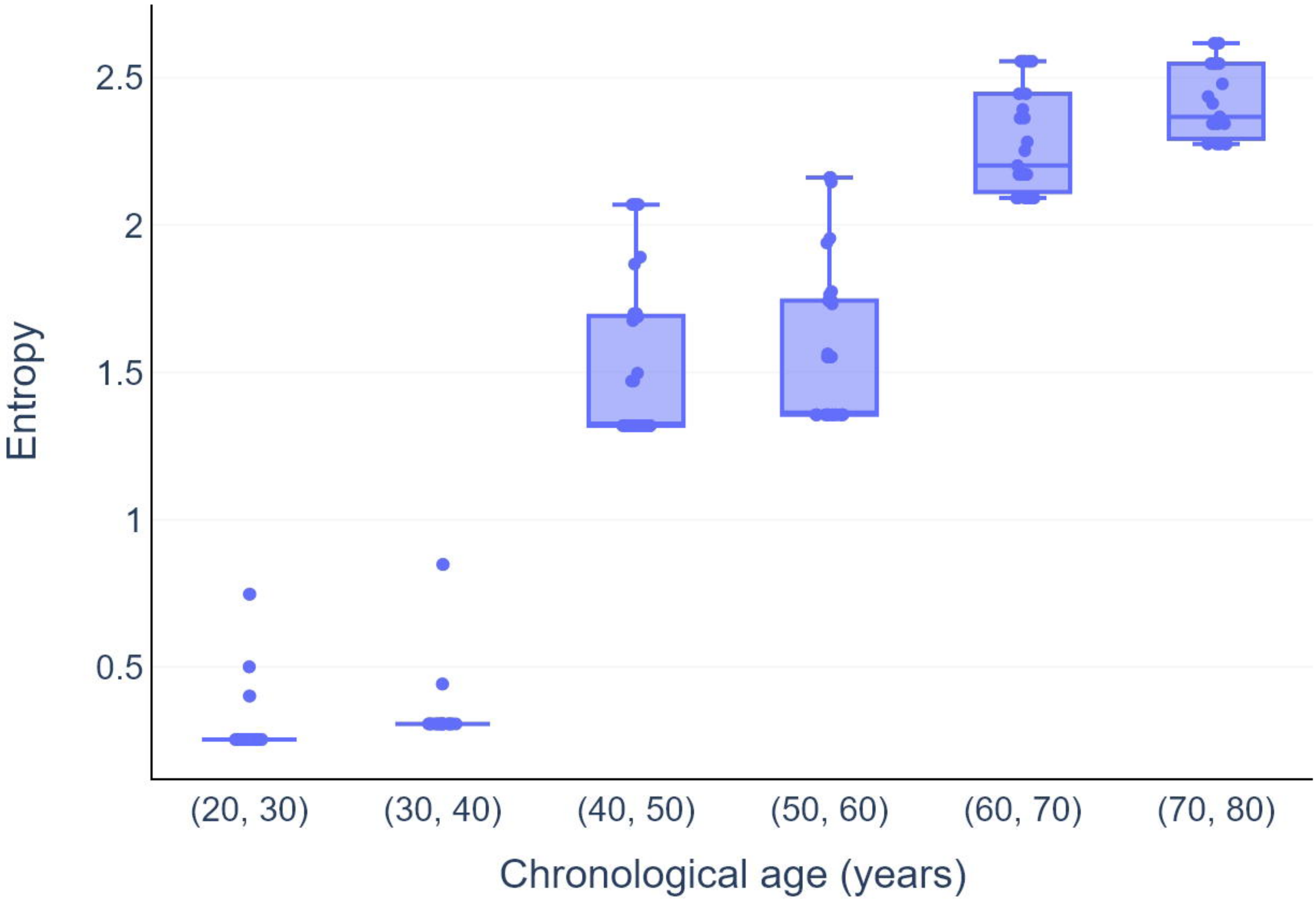
Supplementary Figure 1. Dimensionality reduction techniques to discover potential clusters. Samples with age 0 till 99 were used. (a-b) Shows the first 2 principal components. A small cluster is formed on the top left. The same pattern can be observed on the bottom left when rescaled with mutual information, which represents relative correlations. (c) TSNE shows a cluster of young samples on the top left. (d) UMAP shows a group of younger samples on the top left. In general, samples do not form clear clusters, except young samples.

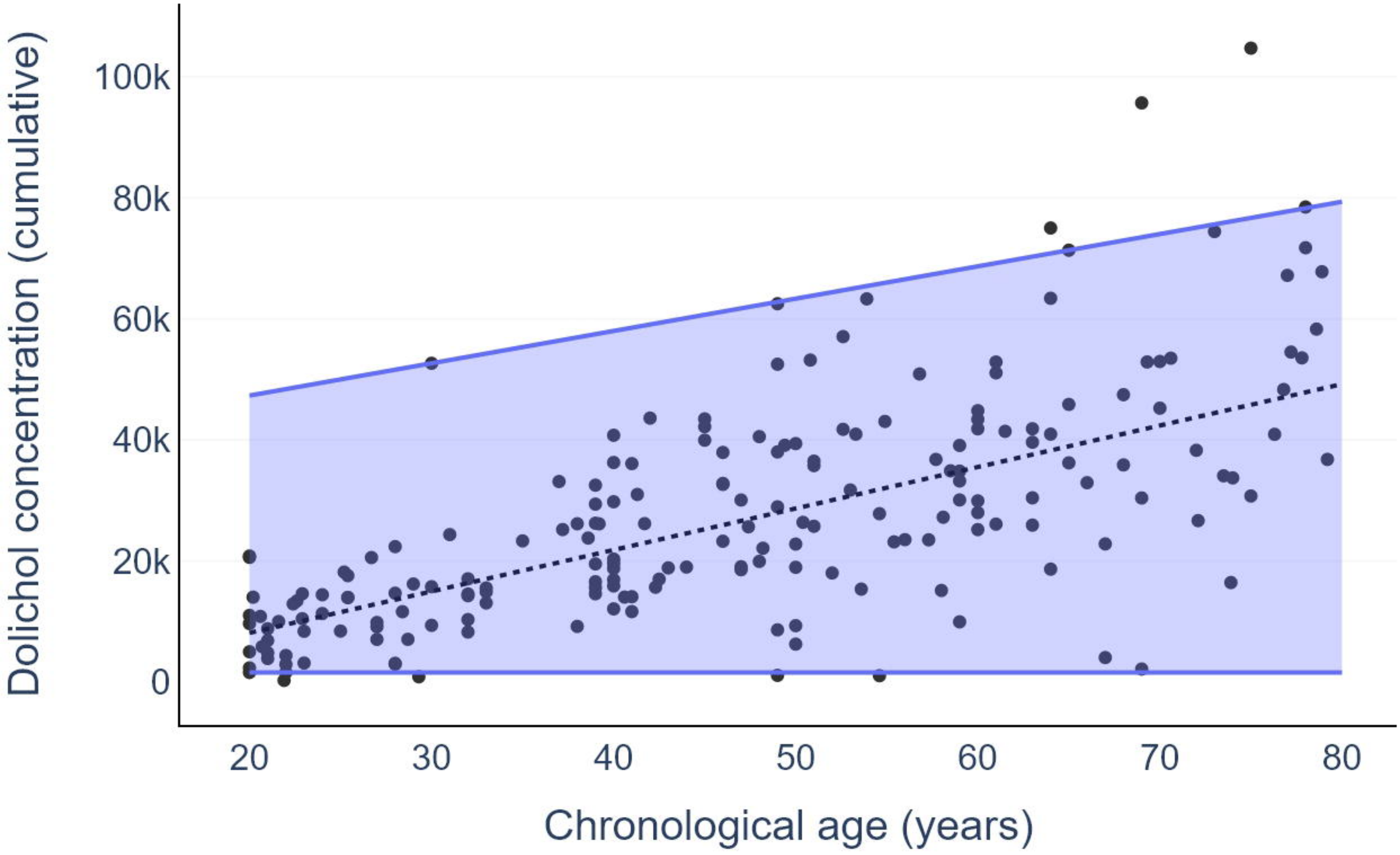
Supplementary Figure 2. Scatter plot showing the distribution of dolichol. The orange dashed line splits the data in two groups, before 40 years and after 40 years. Dolichol concentration seem to cluster around the same level before 40 years. After 40 years, dolichol concentrations seem to increase with an increase in variance. The grey and black dashed line represents two standard deviations from the mean using dolichol concentration before 40 years as a baseline.

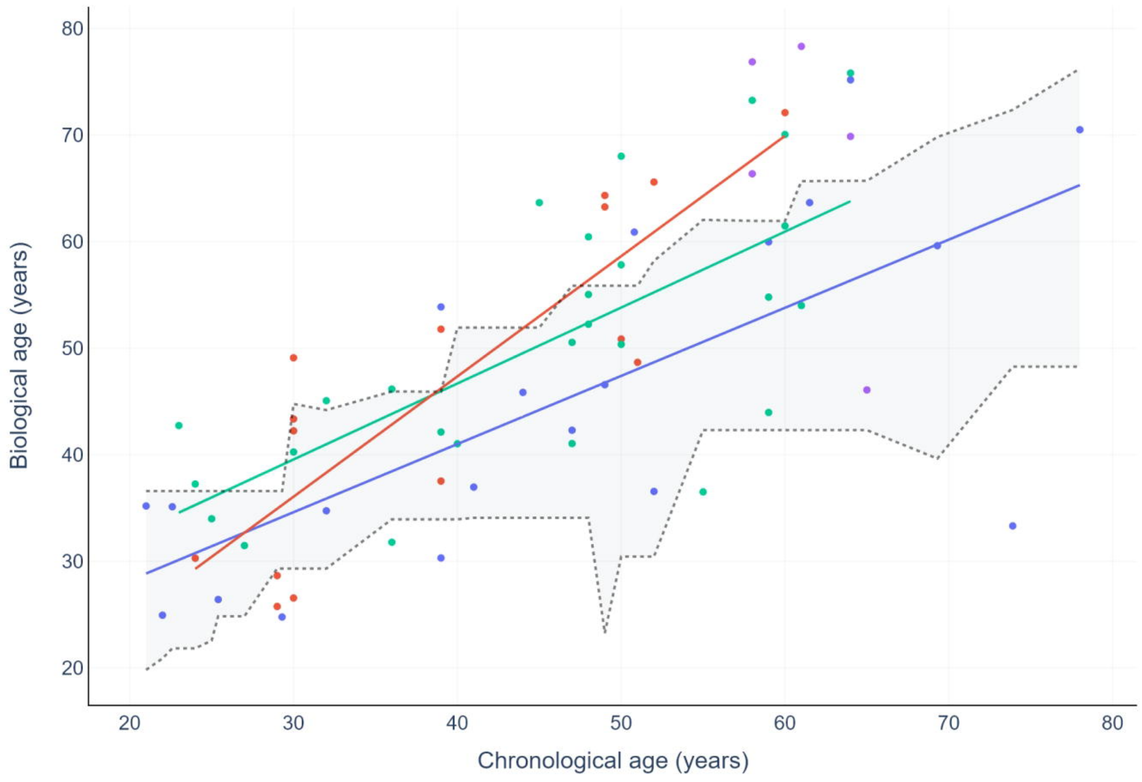
Supplementary Figure 3. The average train MAD vs test MAD after 100 iterations on the dataset. Boosting techniques and linear regression techniques are able to extract the signal smoothly. Although they overfit, regularization can be applied.

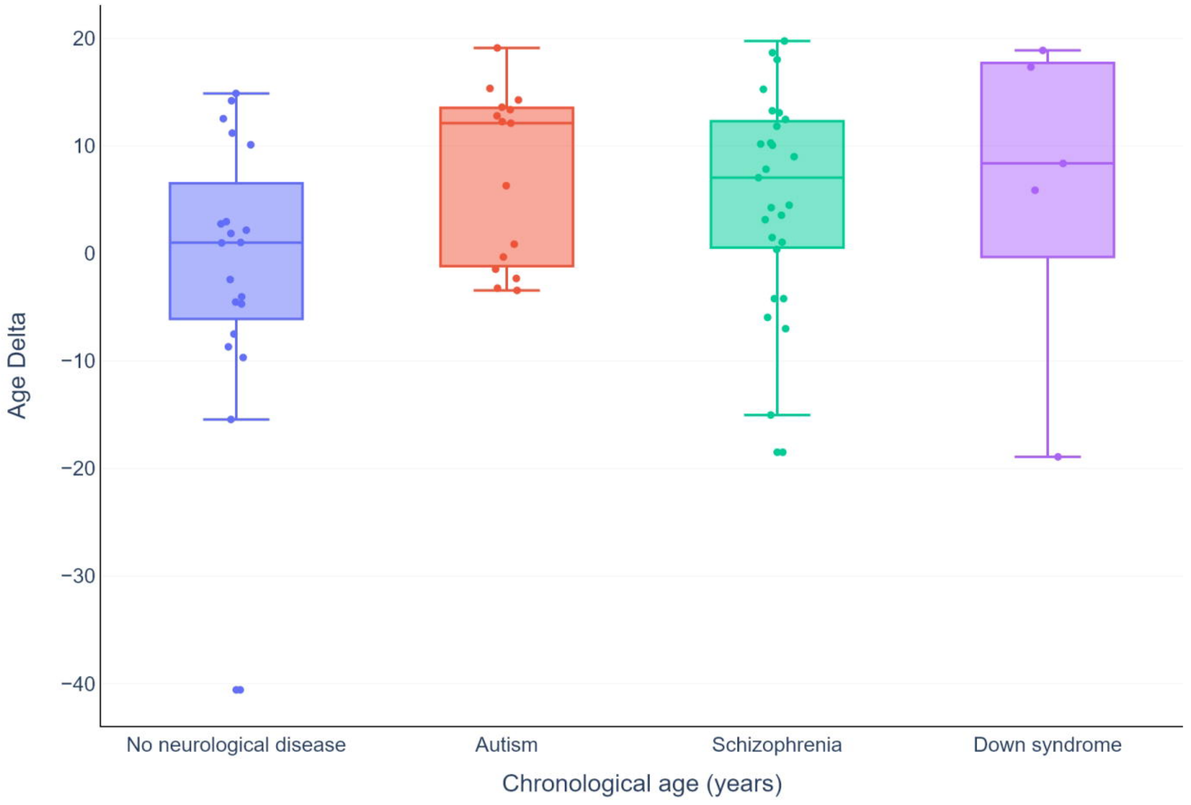
Supplementary Figure 4. Prediction of a XGBoost ensemble model including entropy and dolichol. A trendline was calculated for each diagnosis using OLS.

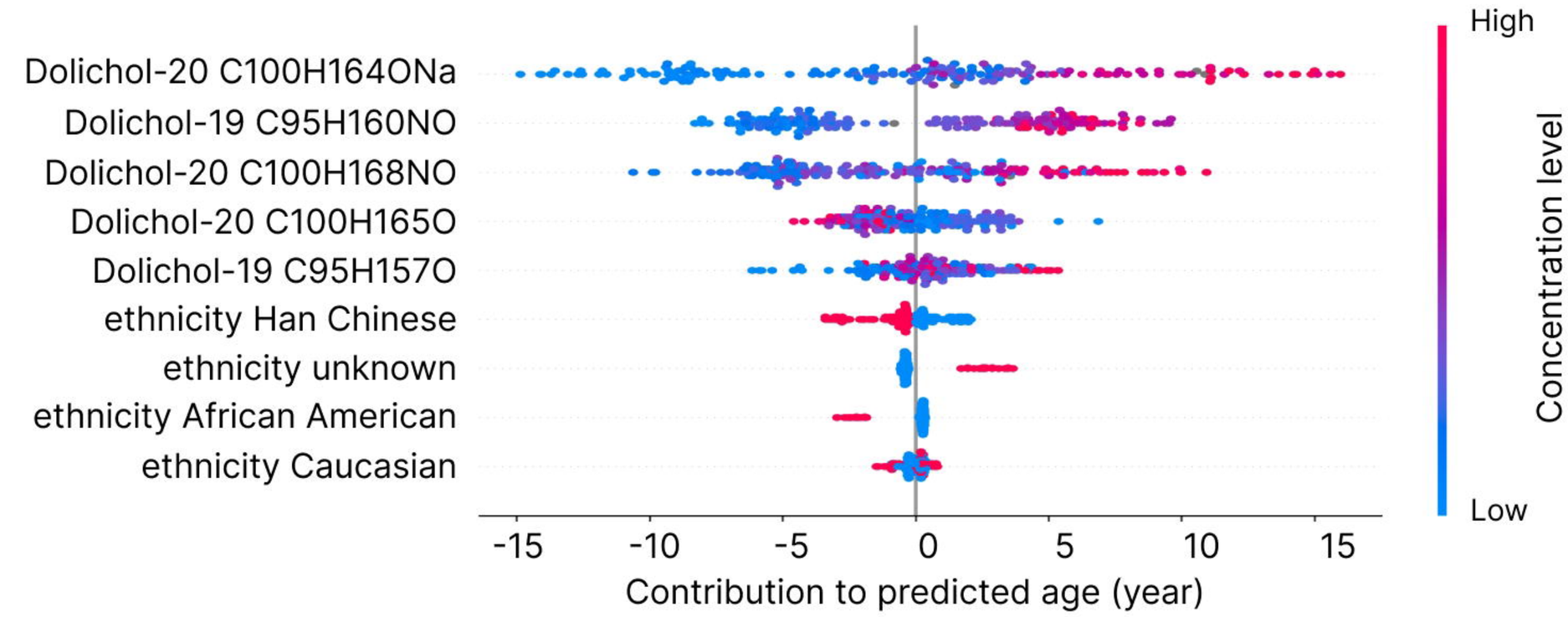

Lipids
Metadata
Samples
(A) Feature matrix
ML Model
Explainable AI
Age prediction of samples
Biological age (years)
Chronological age (years)
(B) Prediction
Global
f(x)
Dolichol-20_C100H164ONa
Dolichol-19_C95H160NO
Dolichol-20_C100H168NO
Dolichol-20_C100H165O
Dolichol-19_C95H157O
ethnicity_han_chinese
ethnicity_unknown
ethnicity_african_american
ethnicity_caucasian
Instances
SHAP value (impact on model output)
11.08
−11.08
(C) Global explanation
Individual
f(x) = 57.963
13566.975 = Dolichol-20_C100H164ONa
8229.926 = Dolichol-19_C95H160NO
4098.428 = Dolichol-20_C100H165O
7540.928 = Dolichol-20_C100H168NO
1 = ethnicity_han_chinese
0 = ethnicity_unknown
0 = ethnicity_caucasian
0 = ethnicity_african_american
8403.049 = Dolichol-19_C95H157O
+9.16
+3.97
−2.03
+1.61
−0.31
−0.25
+0.24
−0.17
E[f(X)] = 46.85
(D) Local eplanation


2.5
2
1.5
1
0.5
Entropy
(20, 30)
(30, 40)
(40, 50)
(50, 60)
(60, 70)
(70, 80)
Chronological age (years)

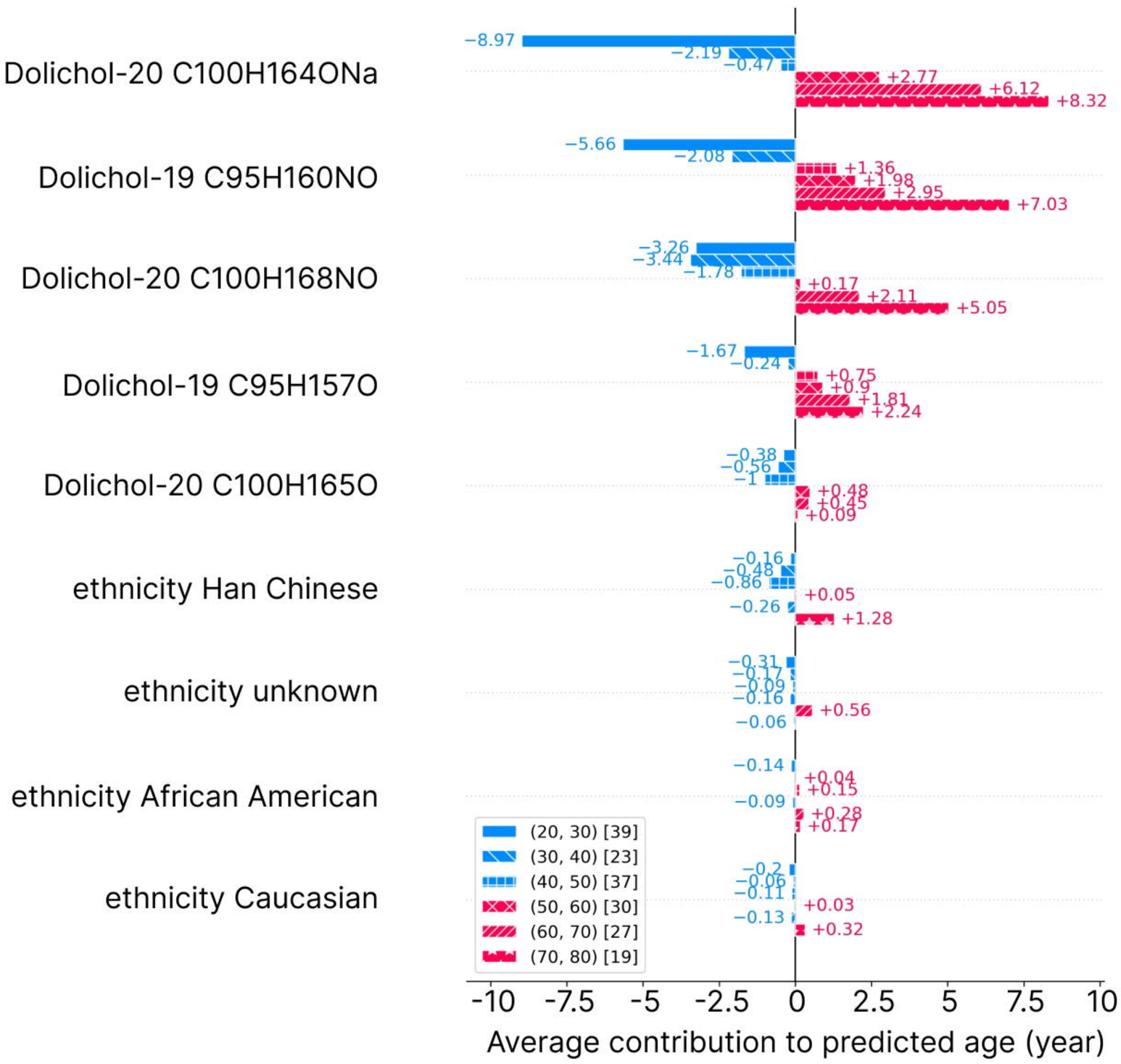

Dolichol concentration (cumulative)
100k
80k
60k
40k
20k
0
20
30
40
50
60
70
80
Chronological age (years)


Biological age (years)
Chronological age (years)
80
70
60
50
40
30
20
20
30
40
50
60
70
80

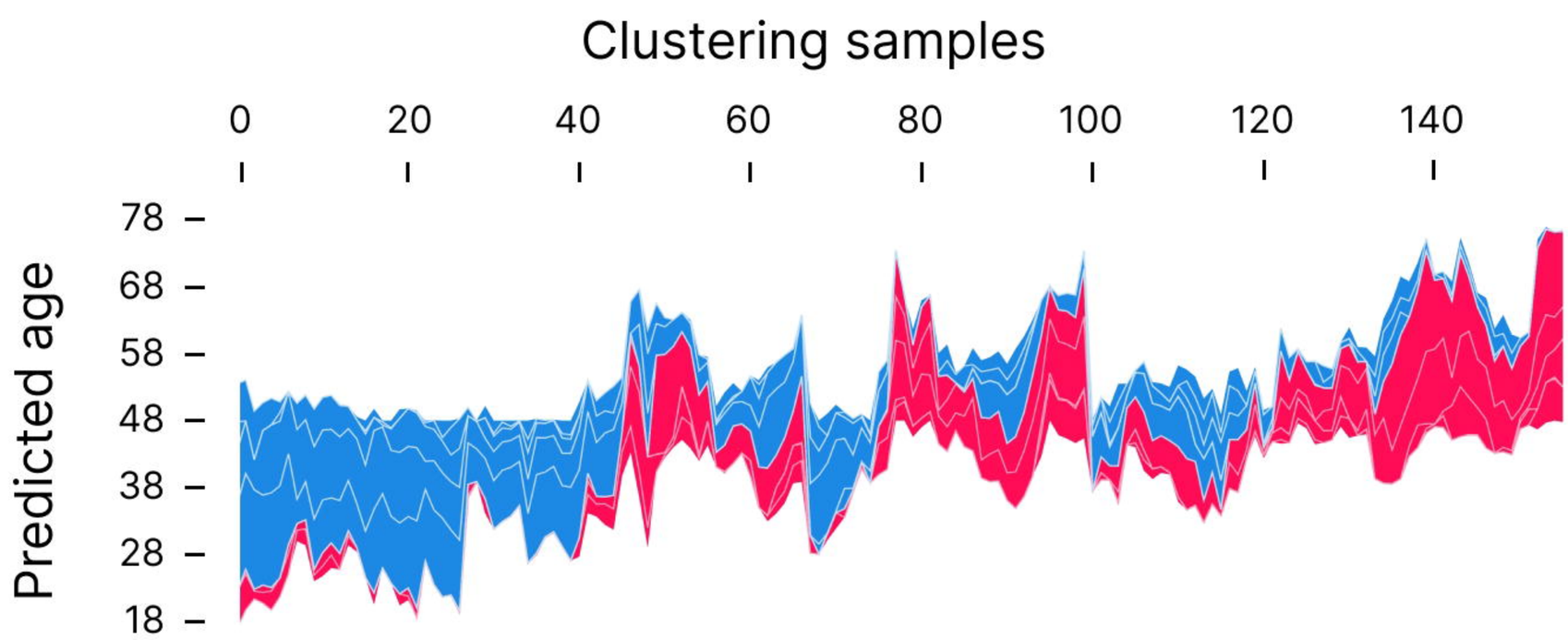

Age Delta
20
10
0
−10
−20
−30
−40
No neurological disease
Autism
Schizophrenia
Down syndrome
Chronological age (years)


Dolichol-20 C100H164ONa
Dolichol-19 C95H160NO
Dolichol-20 C100H168NO
Dolichol-20 C100H165O
Dolichol-19 C95H157O
ethnicity Han Chinese
ethnicity unknown
ethnicity African American
ethnicity Caucasian
-15
-10
-5
0
5
10
15
Contribution to predicted age (year)
High
Concentration level
Low

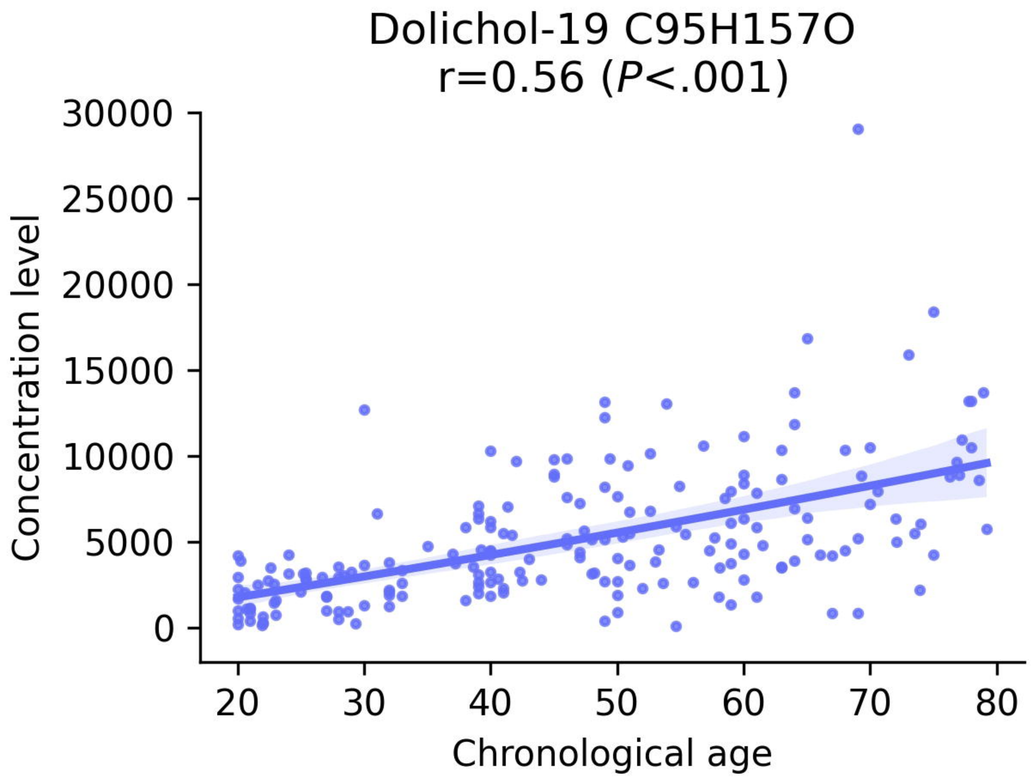
(a)


Dolichol-20 C100H164ONa
Dolichol-19 C95H160NO
Dolichol-20 C100H168NO
Dolichol-19 C95H157O
Dolichol-20 C100H165O
ethnicity Han Chinese
ethnicity unknown
ethnicity African American
ethnicity Caucasian
−8.97
−2.19
−0.47
+2.77
+6.12
+8.32
−5.66
−2.08
+1.36
+1.98
+2.95
+7.03
−3.26
−3.44
−1.78
+0.17
+2.11
+5.05
−1.67
−0.24
+0.75
+0.9
+1.81
+2.24
−0.38
−0.56
−1
+0.48
+0.45
+0.09
−0.16
−0.48
−0.86
+0.05
−0.26
+1.28
−0.31
−0.17
−0.09
−0.16
+0.56
−0.06
−0.14
+0.04
+0.15
−0.09
+0.28
+0.17
−0.2
−0.06
−0.11
+0.03
−0.13
+0.32
(20, 30) [39]
(30, 40) [23]
(40, 50) [37]
(50, 60) [30]
(60, 70) [27]
(70, 80) [19]
-10
-7.5
-5
-2.5
0
2.5
5
7.5
10
Average contribution to predicted age (year)

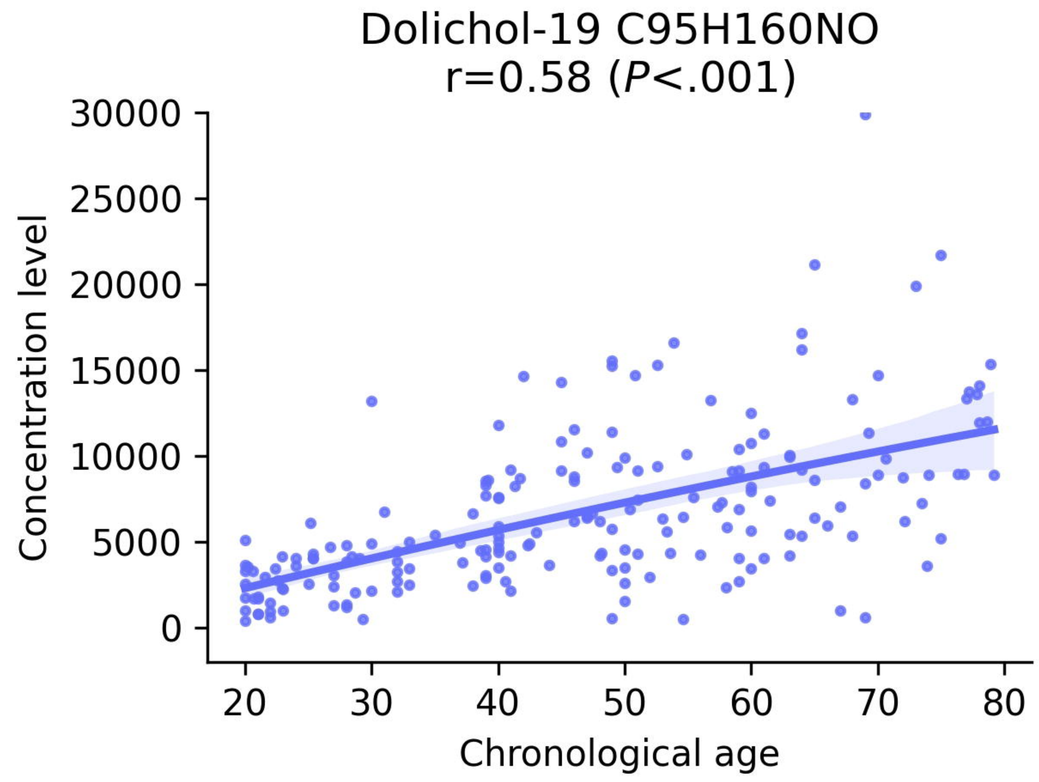
(b)


Clustering samples
0
20
40
60
80
100
120
140
Predicted age
78
68
58
48
38
28
18

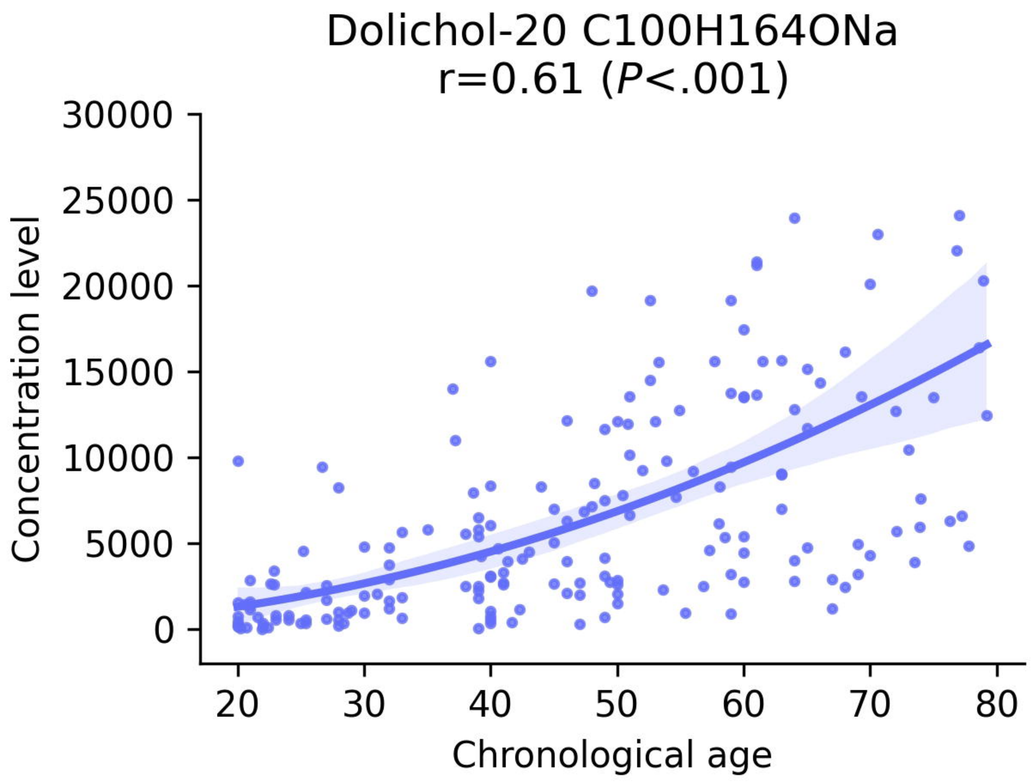
(c)


Dolichol-19 C95H157O
r=0.56 (P<.001)
Concentration level
Chronological age

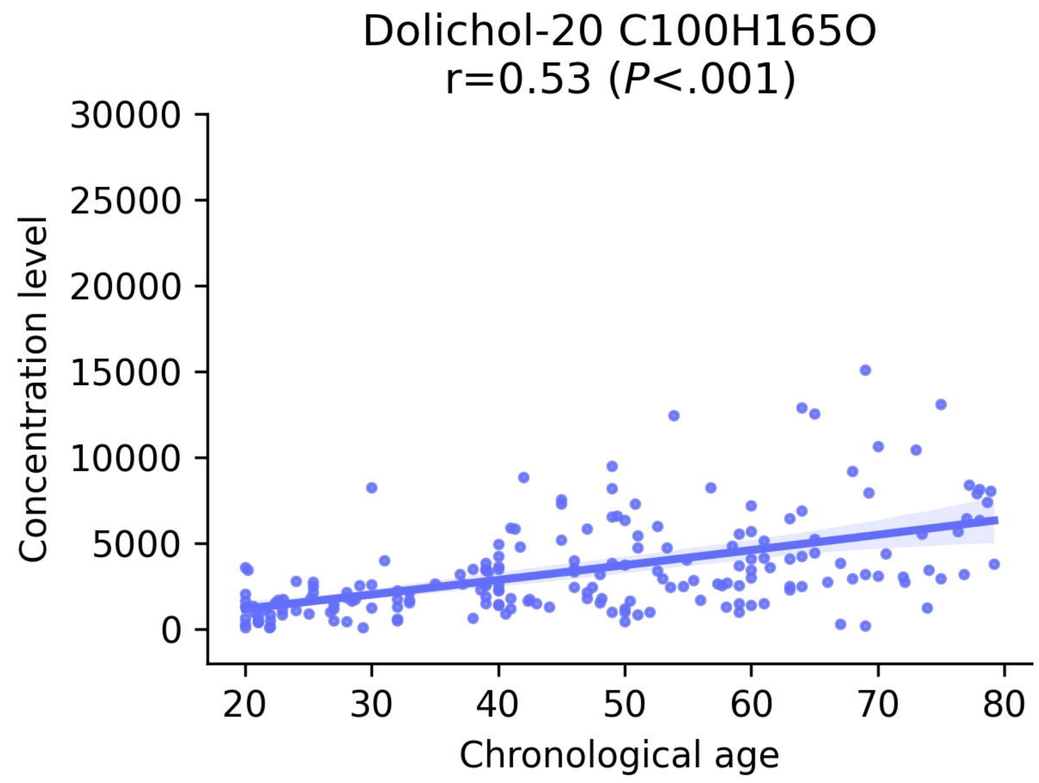
(a)


Dolichol-19 C95H160NO
r=0.58 (P<.001)
Concentration level
Chronological age

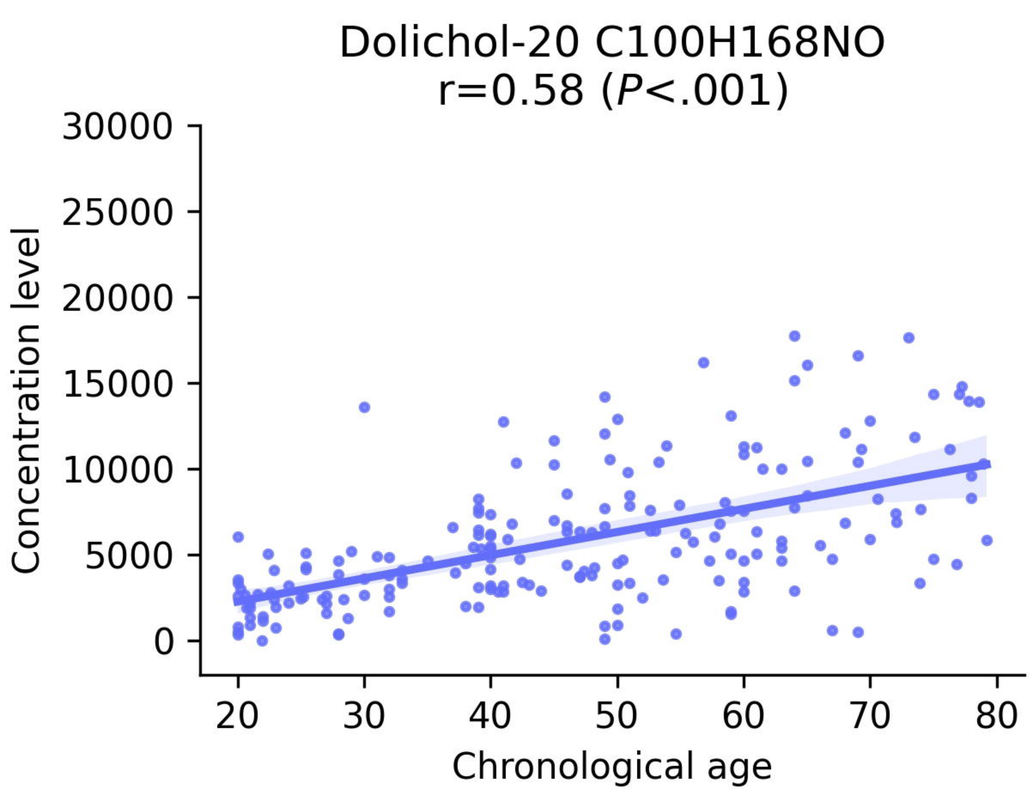
(b)


Dolichol-20 C100H164ONa
r=0.61 (P<.001)
Concentration level
Chronological age

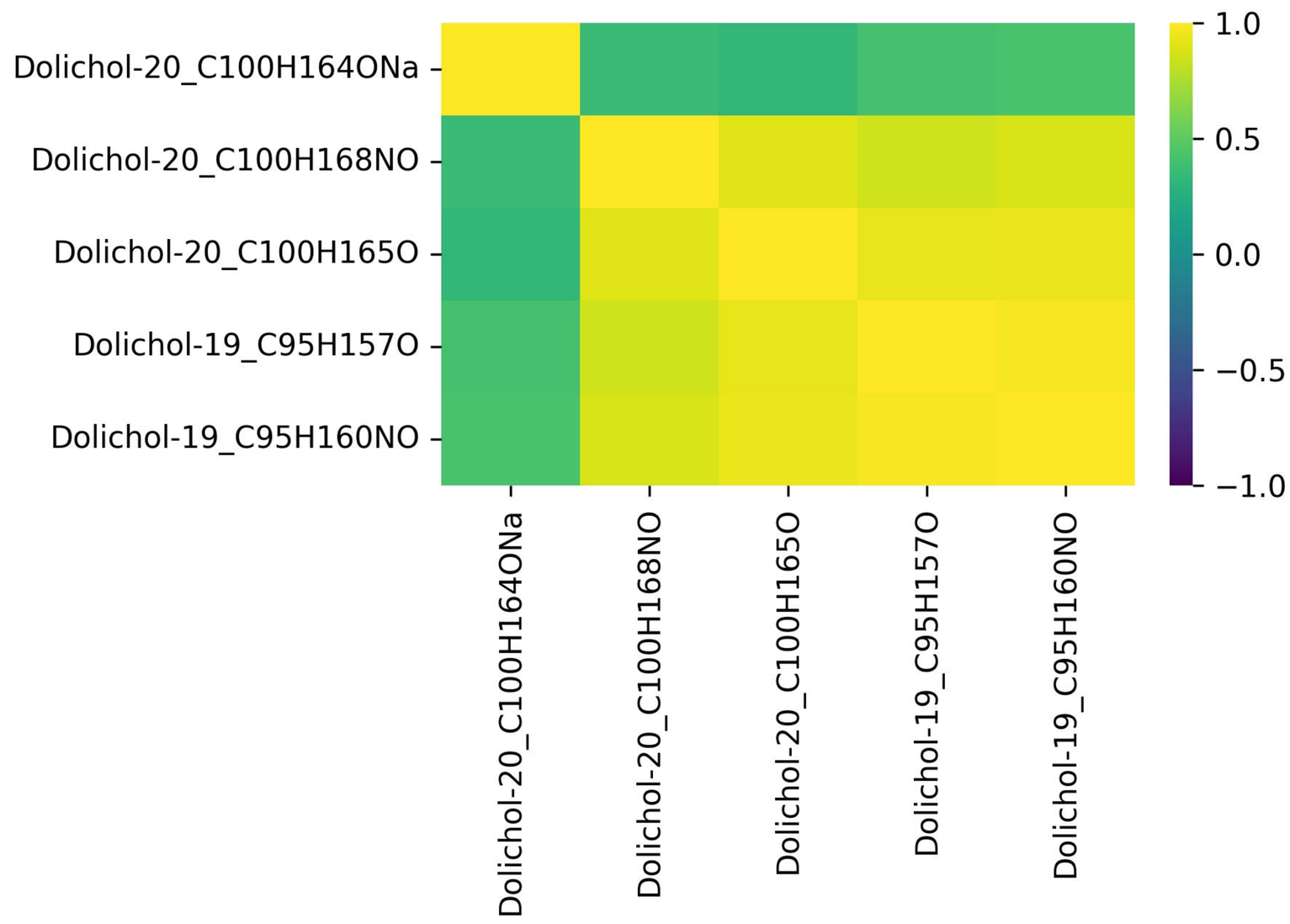
(c)


Dolichol-20 C100H165O
r=0.53 (P<.001)
Concentration level
Chronological age

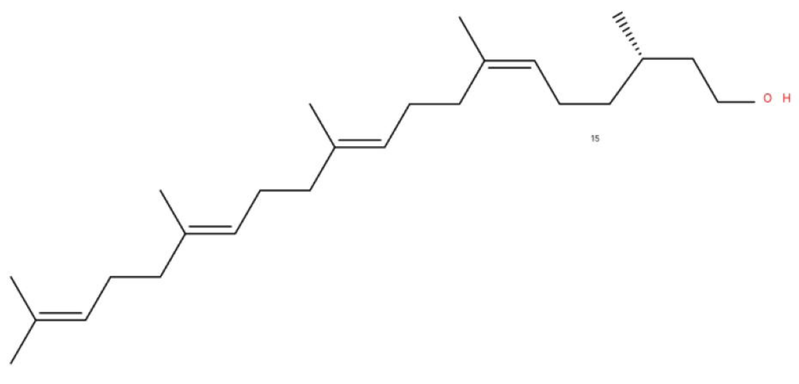
(d)


Dolichol-20 C100H168NO
r=0.58 (P<.001)
Concentration level
Chronological age

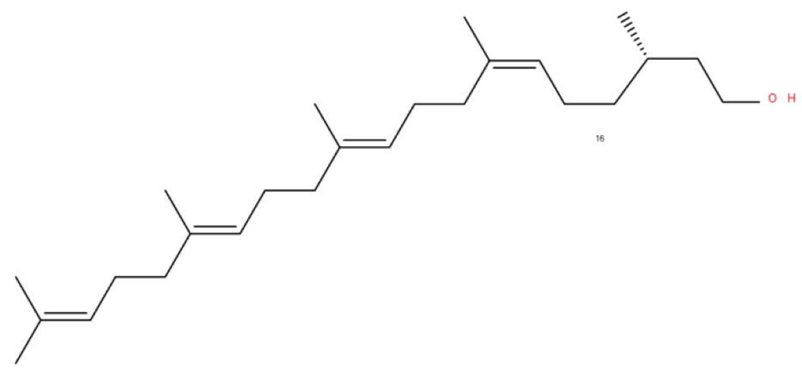
(e)


Dolichol-20_C100H164ONa
Dolichol-20_C100H168NO
Dolichol-20_C100H165O
Dolichol-19_C95H157O
Dolichol-19_C95H160NO
1.0
0.5
0.0
−0.5
−1.0

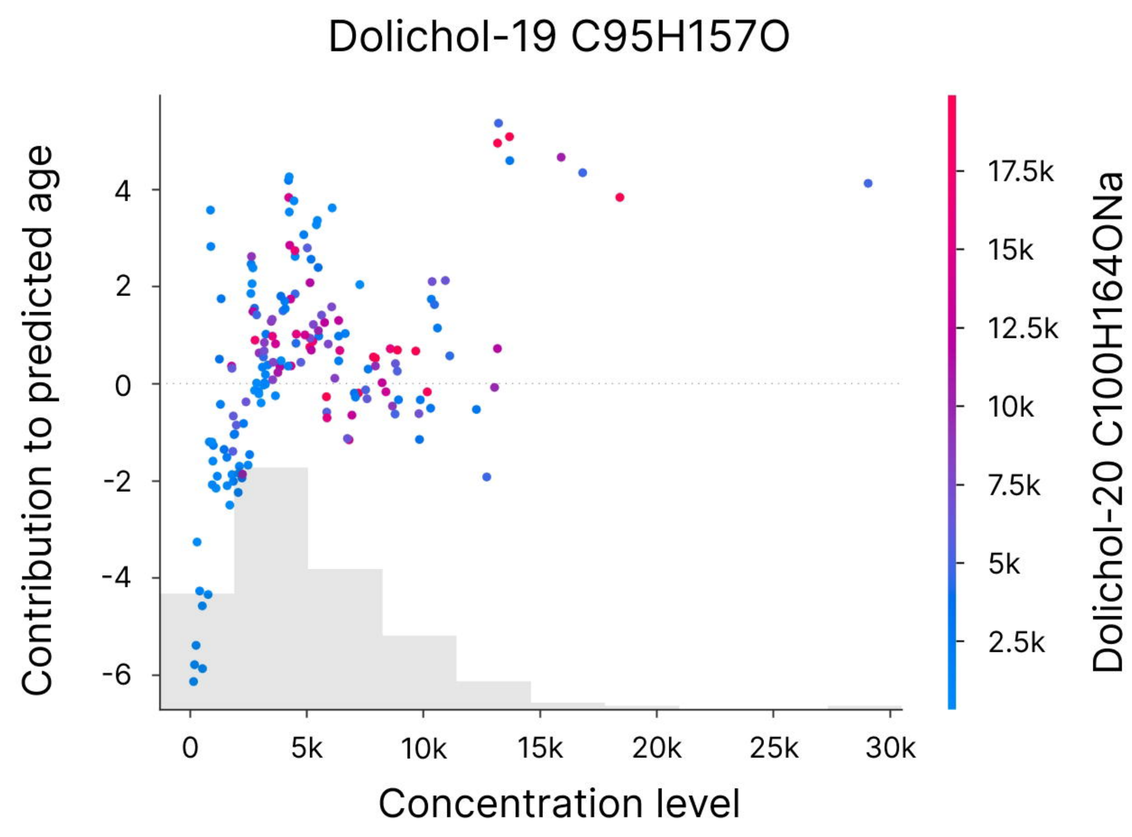
(f)

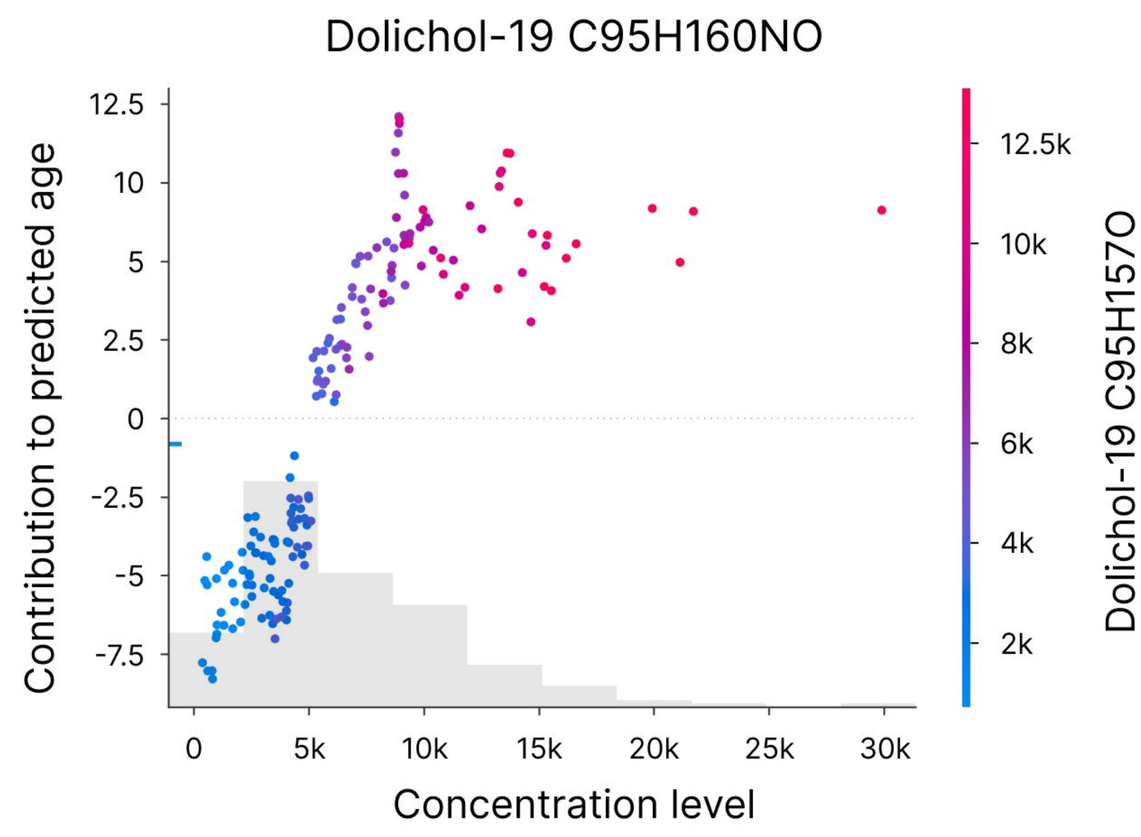


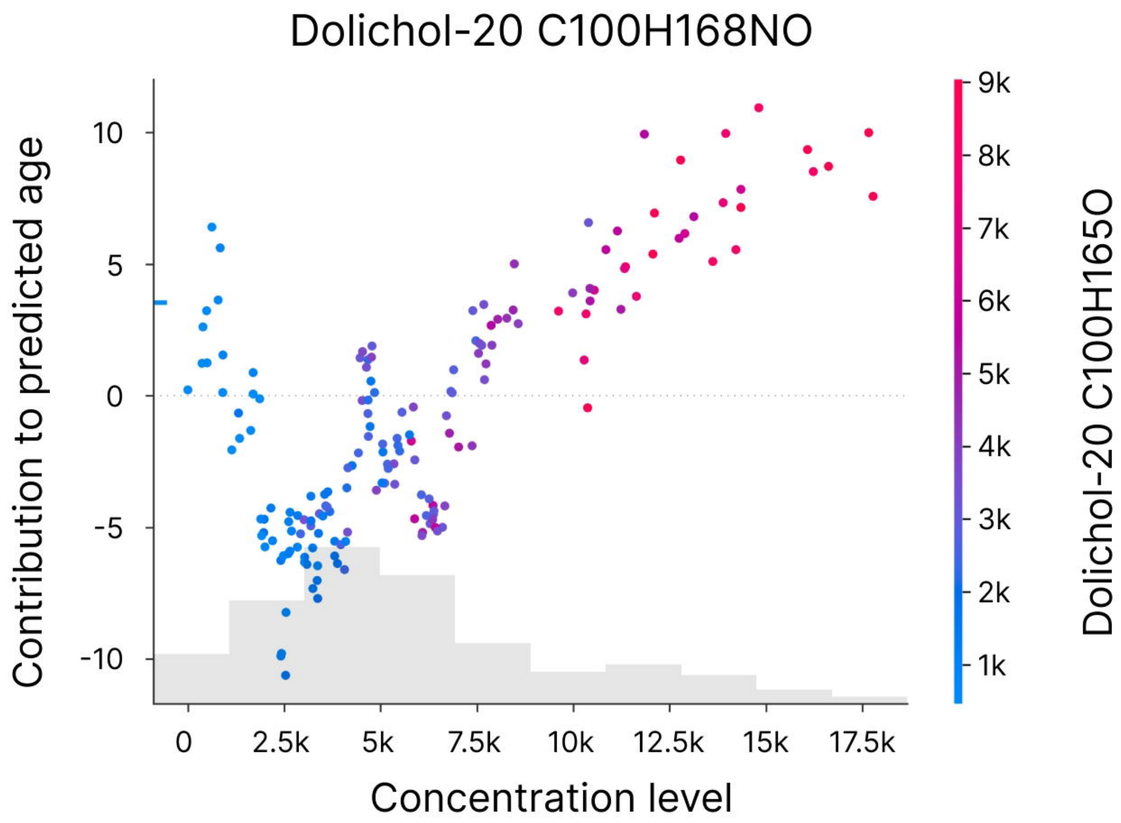
(g)

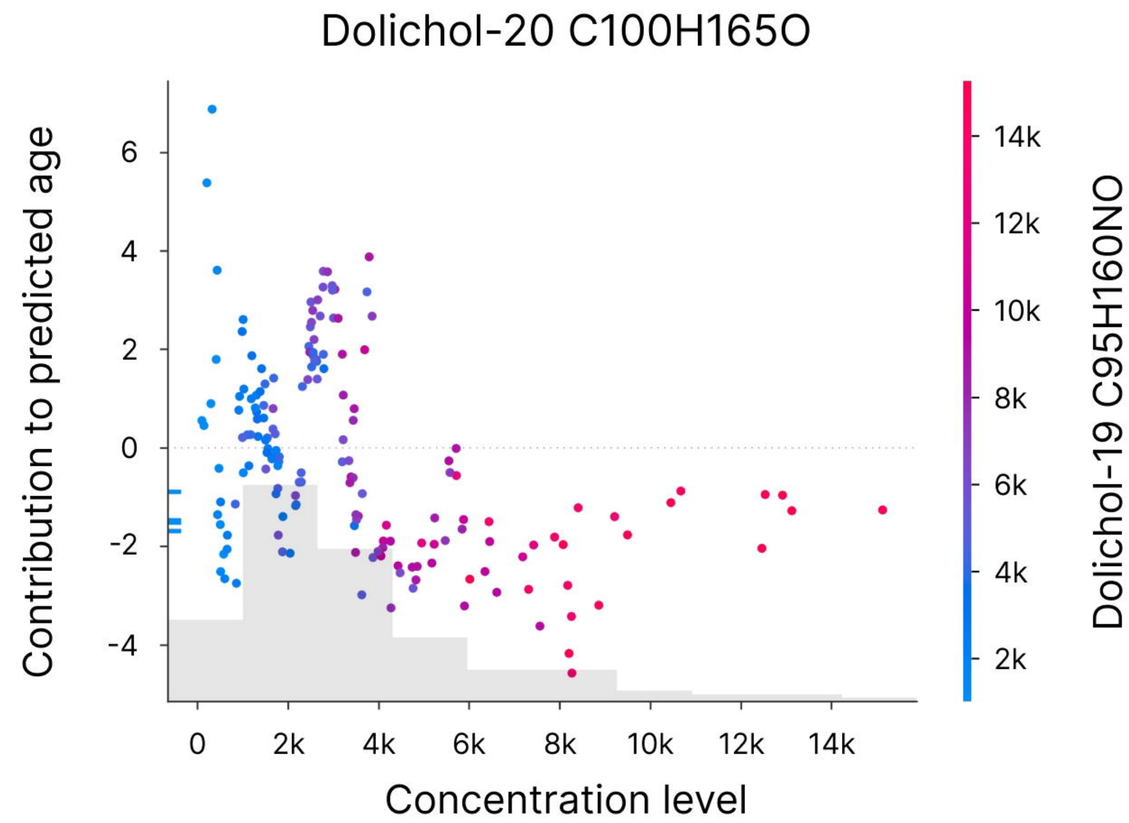


(h)


Dolichol-19 C95H157O
Contribution to predicted age
Concentration level
Dolichol-20 C100H164ONa
(a)


Dolichol-19 C95H160NO
Contribution to predicted age
Concentration level
Dolichol-19 C95H157O
(b)


Dolichol-20 C100H168NO
Contribution to predicted age
Concentration level
Dolichol-20 C100H165O
(c)


Dolichol-20 C100H165O
Contribution to predicted age
Concentration level
Dolichol-19 C95H160NO
(d)

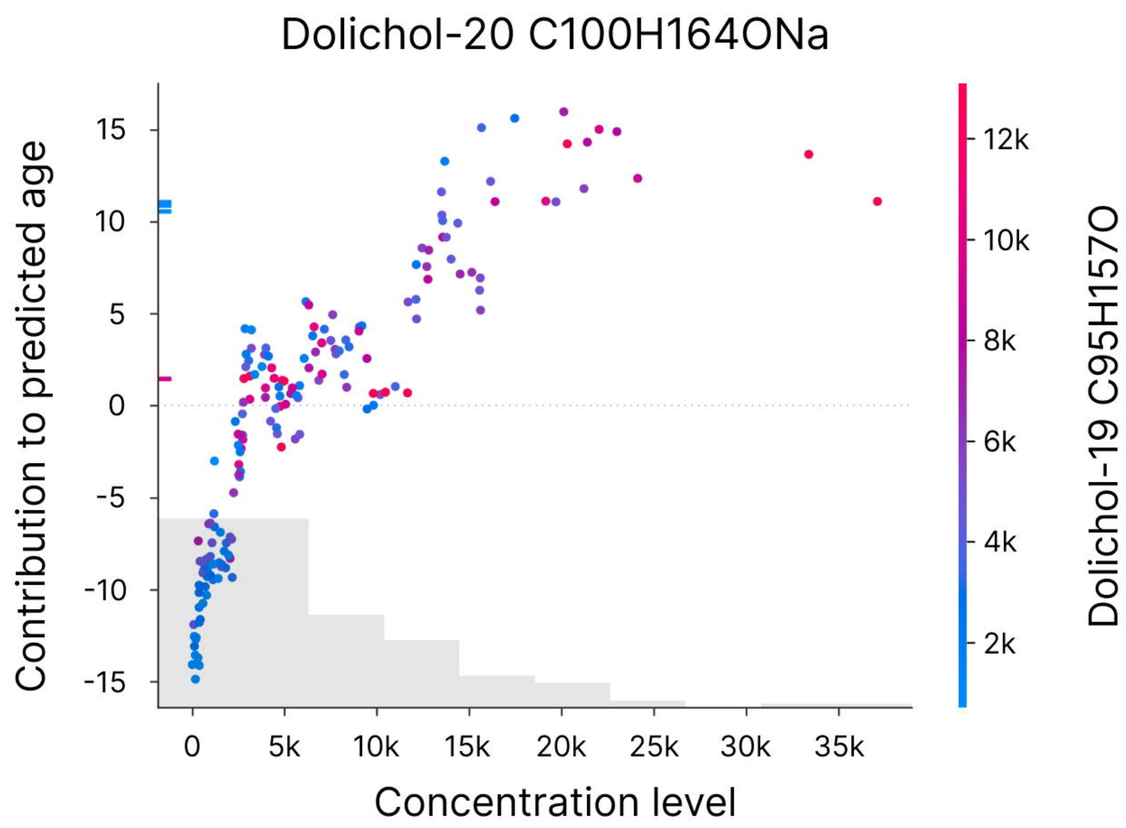

Dolichol-20 C100H164ONa
Contribution to predicted age
Concentration level
Dolichol-19 C95H157O
(e)

Table 1. Search space used by grid search for elastic net with PCA, and XGBoost.

Elastic PCA	
Parameter	**Search space**
alpha	[0.001, 0.01, 0.1, 1, 10]
l1_ratio	[0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8]
n components	[.95, .90, .85, .80]
XGBoost	
Parameter	**Search space**
n_estimators	[50, 70, 90, 110]
max depth	[2, 3, 4]
subsample	[0.5, 1]
colsample_bytree	[0.5, 1]
reg_lambda	[0, 1, 2, 3]
reg_alpha	[0, 1, 2, 3]

Table 2. Results of nested stratified cross-validation. The best-performing model is shown in bold. Standard error of the mean is shown in brackets. Maximum and minimum scores per model are provided. The discrepancy between maximum and minimum scores is caused by high levels of variability.

	MAD (SE)	**ME** (SE)	**r**	**RMSE** (SE)	**R2** (SE)
Elastic PCA					
average	**7.02 (0.38)**	**-0.05 (0.81)**	**0.76[a] (0.03)**	**11.48 (0.68)**	**0.54 (0.05)**
max	8.21	2.27	0.86	14.15	0.67
min	5.44	-3.35	0.67	9.81	0.32
XGBoost					
average	8.14 (0.34)	-0.53 (0.64)	0.67[a] (0.05)	12.87 (0.77)	0.42 (0.07)
max	9.08	1.32	0.79[b]	15.36	0.58
min	6.65	-3.07	0.51[a]	10.71	0.15

[a]$P<.001$ for Pearson correlation coefficient
[b]$P<.005$ for Pearson correlation coefficient

Table 3. Performance of the ensemble models trained on WND test samples. The best-performing model is shown in bold.

	MAD	**ME**	**r**	**RMSE**	**R2**
Elastic PCA					
Train	4.71	0	0.88	8.01	0.78
Test	9.61	-2.29	0.73[a]	11.92	0.52
XGBoost					
Train	0.87	0.33	0.99[a]	2.55	0.98
Test	6.62	0.01	0.72[a]	11.91	0.52

[a]$P<.001$ for Pearson correlation coefficient

Table 4. A comparison of the top 5 features identified by Elastic PCA ensemble and XGBoost ensemble with their corresponding feature ranking among all features.

	Elastic PCA ensemble rank	XGBoost ensemble rank
Dolichol-20 C100H164ONa	1	2
Dolichol-20 C100H168NO	2	6
Dolichol-19 C95H160NO	3	1
Dolichol-19 C95H157O	4	>10
Dolichol-20 C100H165O	5	5
PC(6:2(3E,5E)/1:2(11E,13E))	6	4
Sorbitan palmitate	>10	3

Table 5. Results of nested stratified cross-validation based on only dolichol and dolichol and ethnicity. The best-performing model is shown in bold. Standard error of the mean is shown in brackets. Maximum and minimum scores per model are provided. The discrepancy between maximum and minimum scores is caused by high levels of variability.

	MAD (SE)	**ME** (SE)	**r**	**RMSE** (SE)	**R2** (SE)
Elastic PCA Dolichol					
average	7.75 (0.66)	**-0.01 (0.60)**	0.71[a] (0.02)	12.17 (0.26)	**0.49 (0.03)**
max	8.87	1.61	0.75[a]	13.27	0.56
min	4.8	-2.7	0.64[a]	11.54	0.37
Elastic PCA Dolichol Ethnicity					
average	8.16 (0.68)	-0.48 (0.96)	**0.73[a] (0.02)**	**12.11 (0.26)**	0.49 (0.02)
max	10.12	2.09	0.78[a]	13.13	0.54
min	5.36	-3.78	0.65[a]	11.26	0.38
XGBoost Dolichol					
average	8.02 (0.89)	0.58 (0.78)	0.69[a] (0.03)	12.87 (0.78)	0.42 (0.07)
max	12.34	2.72	0.81[b]	15.8	0.66
min	6.57	-2.55	0.56[a]	10.08	0.16
XGBoost Dolichol Ethnicity					
average	**7.31 (0.62)**	1.01 (0.96)	0.69[a] (0.03)	12.84 (0.64)	0.42 (0.06)
max	9.23	4.36	0.77[a]	15.02	0.57
min	5.07	-2.78	0.57[a]	10.94	0.24

[a]$P<.001$ for Pearson correlation coefficient

[b]$P<.005$ for Pearson correlation coefficient

Table 6. Dolichol concentration levels x10^3 among disorders. Delta shows the average change between case and control.

Molecule	case ±std	control ±std	delta	significant difference
Down syndrome				
Dolichol-19 C95H157O	12.3 ±12.0	6.1 ±3.7	102%	*P*=0.05
Dolichol-19 C95H160NO	14.9 ±11.0	8.2 ±4.4	81%	*P*=0.04
Dolichol-20 C100H164ONa	14.3 ±14.0	9.5 ±6.3	52%	*P*=0.06
Dolichol-20 C100H165O	10.7 ±5.7	4.1 ±2.9	162%	*P*=0.05
Dolichol-20 C100H168NO	15.1 ±7.3	7.5 ±4.3	100%	*P*=0.04
Autism Spectrum Disorder				
Dolichol-19	2.9 ±3.4	3.3 ±3.0	-14%	*P*=0.64
Dolichol-19 C95H157O	4.5 ±4.4	4.5 ±3.7	0%	*P*=0.55
Dolichol-19 C95H160NO	2.8 ±4.2	2.9 ±4.8	-3%	*P*=0.42
Dolichol-20 C100H164ONa	2.2 ±2.6	2.2 ±2.2	2%	*P*=0.27
Dolichol-20 C100H165O	3.8 ±4.6	4.1 ±3.3	-7%	*P*=0.42
Schizophrenia				
Dolichol-19 C95H157O	4.6 ±4.6	4 ±3.1	14%	*P*=0.22
Dolichol-19 C95H160NO	6 ±5.5	5.4 ±3.8	11%	*P*=0.35
Dolichol-20 C100H164ONa	3.2 ±7.0	4.3 ±5.5	-26%	*P*=0.55
Dolichol-20 C100H165O	3.1 ±3.9	2.6 ±2.4	20%	*P*=0.27
Dolichol-20 C100H168NO	5.2 ±4.1	4.7 ±3.5	9%	*P*=0.28