IFSCC 2025 full paper (IFSCC2025-1275)

"Protein signature of human epidermis aging from the IN-SPIRE-T cohort: A machine learning model for estimating biological age"

Patrick Bogdanowicz[1]*, Amaury Alves[2], Pedro Vásquez-Ocmín[1], Pascale Bianchi[1], Eléonore Gravier[1], Gwendal Josse[1], Luciana Bostan[3], Anna Kychygina[3], Michel Simon[3], Nicolas Gaudenzio[3], Nicola Coley[4], Sophie Guyonnet[4], Bruno Vellas[4], Katia Ravard[1], Hélène Duplan[1], Sandrine Bessou-Touya[1]

[1]Pierre Fabre Dermocosmetics & Personal Care, Toulouse, France ; [2]Ippon Innovation, Toulouse ; [3]INFINITY, Toulouse University, INSERM U1291, Toulouse, France ; [4]IHU HealthAge, CHU Toulouse, Toulouse, France.

***Presenting author:** Patrick Bogdanowicz (PhD), Pierre Fabre Dermo-Cosmétique et Personal Care, R&D, Toulouse; patrick.bogdanowicz@pierre-fabre.com

1. Introduction

Aging is defined as the progressive accumulation of deleterious changes over time, including molecular and cellular damages, leading to functional decline, chronic diseases, and ultimately, mortality[1]. Biological age, distinct from chronological age, provides a more nuanced perspective on the aging process. While chronological age, measured from the time of birth, is strongly correlated with declining health, morbidity, and mortality, biological age is inferred from the intricate interplay between cellular and biochemical processes. This offers an individual-specific reflection of physiological function and overall health status[2]. This distinction is particularly crucial for organs such as the skin. Indeed, skin aging is a complex process influenced by a multitude of biological and environmental factors.

Recently, the integration of artificial intelligence in aging and longevity research for the construction of predictive models has revolutionized our ability to analyze and interpret biological data from omics approaches[3]. By leveraging machine learning models to analyze these proteomic data, researchers can accurately predict the biological age of the skin and identify factors contributing to its aging. Consequently, understanding this process is crucial for the development of effective dermo-cosmetic products, which have the potential to slow down or even reverse the effects of skin aging[1,4].

The translational INSPIRE-T cohort, comprising 1,200 subjects aged 20 to 102, was established to investigate the novel concept of healthy aging trajectories based on integrated

clinical and biological markers. Its objectives include elucidating the biological mechanisms that sustain key functions and validating measures of biological age.

Our study aimed to employ machine learning to analyze a specific signature of proteins in the epidermis within the INSPIRE-T cohort, enabling us to train our model using the subjects' chronological skin age and estimate their biological ages.

2. Materials and Methods

INSPIRE-T cohort

The INSPIRE-T cohort comprised 77 subjects. Skin biopsies were obtained from the inner arm. Subjects were stratified into three age groups: Young (19 individuals, aged 22 to 37), Middle-aged (21 individuals, aged 40 to 58), and Aged (37 individuals, aged 61 to 87).

Dermal-Epidermal separation and sample Treatment

Dermis and epidermis were separated by heat treatment (1 min at 60°C) as previously described [5].

Each sample was disrupted by micro cavitation (Bioruptor Pico, Diagenode) with glass Diagenode protein extraction beads and heated (95°C 10 minutes) in 100 µL of LYSE 1X buffer (PreOmics, Gmbh). The protein concentration was determined by a BCA method compatible with reducing agents. After a centrifugation step, the supernatant was transferred into LoBind 96-well plate.

Proteins (between 50 and 100 µg) were then digested with a mix of Trypsin and LysC, following the PreOmics instructions for the use of iST kit (PreOmics, Gmbh). Peptides were purified using mixed mode reverse phase cation exchanger SPE column (PreOmics Gmbh), eluted, dried, and solubilized in 100 µL of 3% acetonitrile 0.1% formic acid aqueous solution. Then, peptide concentration in each sample was determined using BCA method.

Shotgun proteomic by LC-MS/MS

250 ng of peptides were injected once for each epidermis sample. Chromatography was performed using a Vanquish Neo system using PepMap100 C18 (75 µm x 50 cm, 2 µm material) column applying a 3 % to 45% acetonitrile 90 minutes gradient at a flow rate of 300 nl/min after a 3 min trapping step on precolumn. Data were acquired using an Exploris 480 (Thermo Scientific) mass spectrometer. MS/MS scan was performed on the most intense ions of each MS1 survey scan for a total cycle time of 2 sec.

Data analysis for protein identification and quantification

For protein identification, the data were processed by Proteome Discoverer 3.0. Proteins were identified by SEQUEST-HT search algorithms, comparing acquired MS/MS spectra to theoretical MS/MS peptide spectra derived in silico from a database of protein sequences. The database was composed with the Human reference proteome (UP000005640 download the 12/02/2024) and a database of frequently observed experimental contaminants in mass spectrometry without human contaminants (cRAP database, https://www.thegpm.org/crap/; downloaded 26/06/2024). INFERYS™, a deep learning prediction-based rescoring system developed by Thermo Scientific, was used to increase the number of MS/MS identifications. False Discovery Rate (FDR) determination was made using Percolator algorithm. All spectra reported with a confidence less than high by SEQUEST-HT, thus considered as not identified, were processed a second time with the same databases without the INFERYS™ node, and with the following modified parameters:

For protein relative quantification, the abundance was measured for each peptide, and these abundances summed for each protein, using the Minora Feature Detector and the Feature Mapper nodes from Proteome Discoverer 3.0.

Statistical analysis and model building

Proteins with more than 80% missing values, as well as those marked as "contaminants," were subsequently removed. After cleaning, 4054 proteins were obtained. A PCA was performed to show a separation of age classes. Imputation of missing data was performed using the missForest method. Relative abundance normalization was selected for this analysis

The heatmap was generated from the 50 proteins with the highest variability in terms of IQR (Interquartile Range). Hierarchical clustering using Ward's method is applied to the samples with Pearson distance and to the proteins with absolute Pearson distance.

Variable selection using Recursive Feature Elimination (RFE) method was performed on a training set consisting of 62 subjects to retain a maximum of 30 proteins[6,7]. After this selection, an XGBoost machine learning model was trained on the same training set and then evaluated on a test set comprising 15 subjects that had not been previously utilized (using the proteins retained RFE method)[8]. The model's hyperparameters were optimized through cross-validation (using 4-folds in this instance).

Pathway analysis and pharmacological interpretation

Ingenuity Pathway Analysis (IPA) was employed to analyze the top 50 most variables proteins identified in the heatmap, as well as proteins selected for the machine learning model.

3. Results

Predictive Model for the INSPIRE-T cohort

Following the data cleaning process mentioned above (with 4,054 proteins remaining), the PCA score plot derived from the first two principal components accounted for 42.8 % of the total variance. The PCA results demonstrated a clear separation of age classes, particularly visible for the younger individuals, while middle age group show a dispersion between the other two classes (**Fig. 1**).

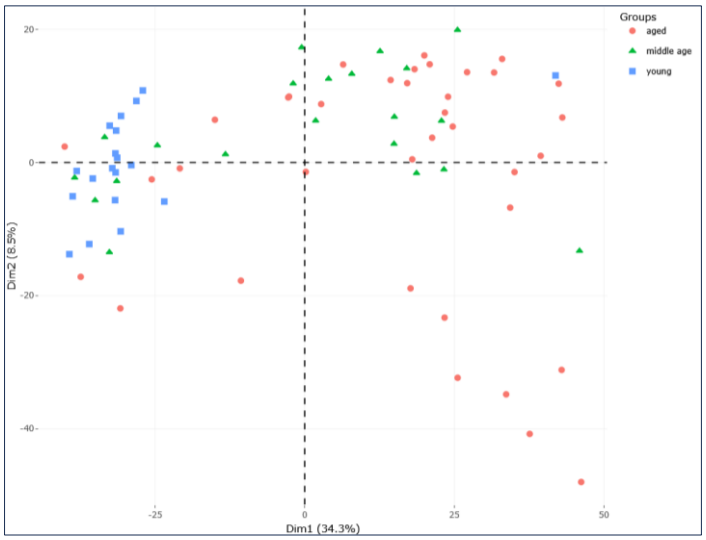

Figure 1. Global PCA of 77 subjects by age classes.

A complementary heatmap analysis was conducted using the top 50 most variables proteins (**Fig. 2**). Consistent with the PCA results, subjects from the young class generally differed from other classes, with the exception of one sample (from a 29-year-old individual) that clustered with the aged group. While more than 60% of the middle-aged subjects were dispersed between the young and aged groups. Surprisingly, two subjects from the aged group (61- and 80-years-old, respectively) clustered with the young group. The primary pathways identified for these proteins are associated with pro-inflammatory conditions, barrier integrity, autophagy and fatty acid synthesis modifications.

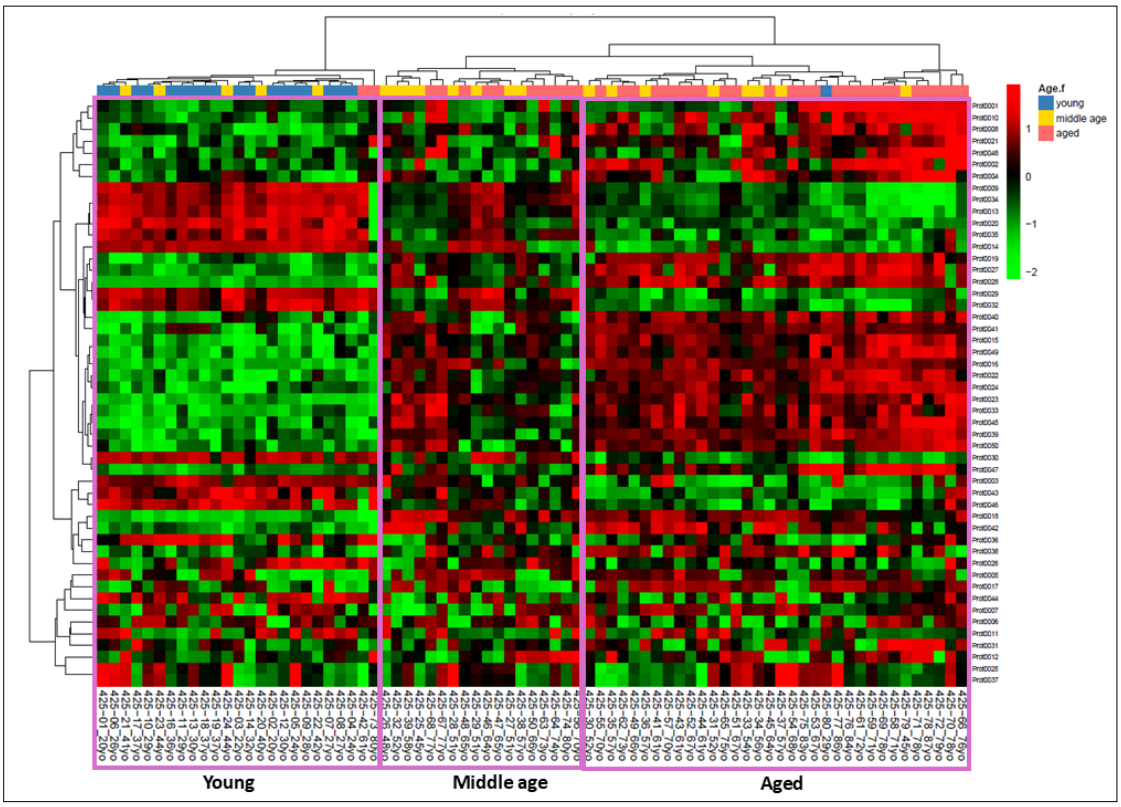

Figure 2. Heatmap of the top 50 most variables proteins by class.

The following results were obtained using the XGBoost model, yielding to the selection of 28 proteins. **Fig. 3** highlights the steps for our model building. We demonstrated the estimation of biological age with a Mean Absolute Error (MAE) of 8.1 years and an R^2 of 0.68 on the independent test set.

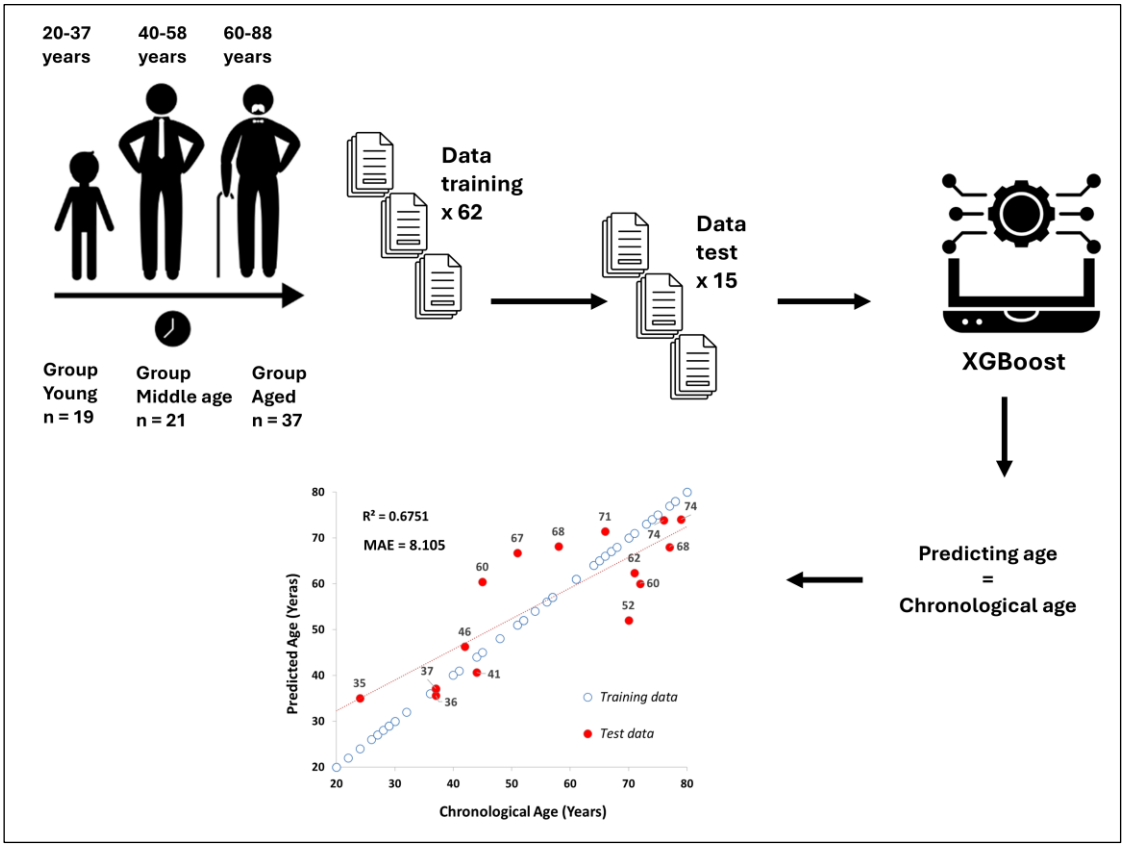

Figure 3. Steps for model building using INSPIRE-T cohort.

Pathways Analysis about proteins selected fro the machine learning model showed that these proteins belong to five canonical pathways: biotin-carboxyl carrier protein assembly, carnitine metabolism, fatty acid activation, phenylalanine degradation IV and unfolded protein response.

4. Discussion

The INSPIRE cohort enabled us to work with subjects from three age groups (young, middle age and aged). However, our proteomic analysis revealed that the chronological age of these individuals differs from their biological age. While young subjects clustered together based on both the total proteins and the top 50 proteins, middle-aged group (40–60 years), appears to be more of an artificial representation of the biological state of the individuals, positioning itself between the young and aged groups.

Our results showed that differentially expressed proteins are involved in inflammation, autophagy, skin barrier, and fatty acid synthesis pathways. These proteins regulate

inflammation by modulating the production of cytokines and other inflammatory mediators. Additionally, they alter fatty acid modifications by affecting the metabolic pathways responsible for lipid synthesis and degradation. Over time, these changes can contribute to the deterioration of skin structure and function.

While gold standard aging clocks like Hannum or Horvath's DNA methylation-based clocks exist[9,10], the recent integration of artificial intelligence in proteomic data analysis has led to the creation of protein clocks, principally for plasma samples[11,12]. Our machine learning model from epidermis samples allowed us to determine a predictive model of skin health from 28 proteins, with a Mean Absolute Error (MAE) of 8.1 years.

5. Conclusion

The development of our protein-clock approach holds significant promise for applications in the dermo-cosmetic industry, offering novel pathways for elucidating epidermal skin aging mechanisms and enabling targeted interventions and personalized anti-aging treatments.

6. References

1. Moqri, M. *et al.* Biomarkers of aging for the identification and evaluation of longevity interventions. *Cell* **186**, 3758–3775 (2023).
2. Chen, R. *et al.* Biomarkers of ageing: Current state-of-art, challenges, and opportunities. *MedComm – Future Medicine* **2**, e50 (2023).
3. Zhavoronkov, A. *et al.* Artificial intelligence for aging and longevity research: Recent advances and perspectives. *Ageing Research Reviews* **49**, 49–66 (2019).
4. Ma, J. *et al.* Quantitative proteomics analysis of young and elderly skin with DIA mass spectrometry reveals new skin aging-related proteins. *Aging* **12**, 13529–13554 (2020).
5. Jian, L., Cao, Y. & Zou, Y. Dermal-Epidermal Separation by Heat. in *Epidermal Cells: Methods and Protocols* (ed. Turksen, K.) 23–25 (Springer US, New York, NY, 2020). doi:10.1007/7651_2019_270.
6. Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A. & Vingron, M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18**, S96–S104 (2002).
7. Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning* **46**, 389–422 (2002).
8. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (Association for Computing Machinery, New York, NY, USA, 2016). doi:10.1145/2939672.2939785.
9. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biology* **14**, 3156 (2013).
10. Hannum, G. *et al.* Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. *Mol Cell* **49**, 359–367 (2013).
11. Johnson, A. A., Shokhirev, M. N. & Lehallier, B. The protein inputs of an ultra-predictive aging clock represent viable anti-aging drug targets. *Ageing Research Reviews* **70**, 101404 (2021).

12. Lehallier, B., Shokhirev, M. N., Wyss-Coray, T. & Johnson, A. A. Data mining of human plasma proteins generates a multitude of highly predictive aging clocks that reflect different aspects of aging. *Aging Cell* **19**, e13256 (2020).