Integration of Infrared Molecular Fingerprinting Data in a Longitudinal Health Profiling Cohort

Kosmas V. Kepesidis[1,2,3](✉), Zita I. Zarandy[1,2,3], Flora B. Nemeth[1,2,3], Lea Gigou[1,2,3], Mihaela Žigman[1,2,3], and Ferenc Krausz[1,2,3]

[1] Ludwig-Maximilians-Universität München (LMU), Chair of Experimental Physics - Laser Physics, Garching, Germany
kosmas.kepesidis@lmu.de

[2] Max Planck Institute of Quantum Optics (MPQ), Laboratory for Attosecond Physics, Garching, Germany

[3] Center for Molecular Fingerprinting (CMF), Budapest, Hungary

Abstract. This study explores the characteristics and interpretation of infrared molecular fingerprints (IMFs)—blood-based profiles that capture broad molecular information—for applications in precision medicine. Using data from 4,196 healthy individuals across five longitudinal visits, we integrated Fourier-transform infrared (FTIR) spectroscopy with routine clinical chemistry tests. IMF measurements showed high inter-individual and low intra-individual variability, indicating stable and unique molecular profiles over time. Machine learning models re-identified individuals with over 90% accuracy based solely on their IMF data, highlighting strong individual specificity. To enhance diagnostic resolution, we quantified within-person and between-person variability across the infrared spectrum. A tree-based optimization algorithm stratified individuals into sub-cohorts by maximizing the Index of Individuality, minimizing between-person variability to levels close to within-person. The algorithm was based on 27 blood parameters and three demographic variables, producing hierarchical splits based on averaged longitudinal values. We further modeled the relationships between IMFs and clinical parameters using linear regression, revealing robust, biologically interpretable associations. To uncover latent physiological structure, we applied Pareto Task Inference (ParTI), which identified a tetrahedral organization in the combined IMF-clinical data space, representing four archetypal physiological states. Individual trajectories within this space may serve as early indicators of health deviation. Archetypes were further characterized using demographic and health-related data, supporting hypotheses on systemic trade-offs in health maintenance.

Keywords: Precision medicine · Infrared molecular profiling · Health trajectories · Longitudinal study · High-dimensional data

R. Bellazzi et al. (Eds.): AIME 2025, LNAI 15734, pp. 180–190, 2025.
https://doi.org/10.1007/978-3-031-95838-0_18

1 Introduction

Advancements in precision medicine increasingly depend on comprehensive molecular profiling to capture the dynamic complexity of human health [1]. Among emerging technologies, infrared molecular fingerprinting (IMF) of human blood via Fourier-transform-infrared (FTIR) spectroscopy has shown promise as a blood-based tool for assessing physiological and biochemical states [2–5]. By detecting vibrational signatures of biomolecules, IMF provides a holistic molecular snapshot, offering potential for individualized health monitoring and disease prediction [6–16]. Recent studies have demonstrated the feasibility of IMF for cancer detection and multi-phenotype health screening using blood-based samples [7,9], highlighting its potential as a cost-effective, high-throughput tool for large-scale health monitoring. However, its integration into large-scale longitudinal health profiling studies remains largely unexplored.

In this study, we leverage an expanding prospective longitudinal health profiling cohort [17] to systematically evaluate the stability, variability, and interpretability of IMF data in healthy individuals. Analyzing blood samples from 4,196 participants with at least five visits, we assess intra- and inter-individual variations in infrared molecular profiles and their correlation with routine clinical chemistry markers.

Our findings demonstrate that IMF data exhibit high individuality and stability over time, reinforcing previous research [6]. Additionally, we show that IMF data encode information on the concentrations of various clinical analytes commonly measured in standard blood tests, corroborating prior findings from an independent cohort [9]. This shared molecular information enhances the interpretability of IMF data. Furthermore, sub-cohorts of molecularly similar individuals can be identified using the clinical chemistry blood panel. We show that, within these sub-cohorts, inter-individual variability approaches intra-individual levels, effectively reducing baseline variation. This stratification could enable more precise and personalized early disease detection and health monitoring.

The longitudinal nature of our study further allowed us to explore methods for extracting interpretable individual health trajectories. Specifically, we applied Pareto Task Inference (ParTI)—a framework for inferring biological tasks from high-dimensional data—to both the IMF and clinical chemistry panel data [18,19]. This analysis revealed a tetrahedral organization of interpretable states, suggesting a constrained space of molecular variability. We linked lifestyle, anthropometric, and health variables to this geometric structure by employing an enrichment method. Individuals thus navigate on an interpretable molecular landscape, enabling the characterization of individual health trajectories and potentially signaling transitions from wellness to disease before conventional biomarkers detect pathology.

Table 1. Characteristics of study cohort

	# Subjects	Age	BMI
Female	2,839	52.02 ± 0.18	26.61 ± 5.04
Male	1,357	52.02 ± 0.16	28.45 ± 4.29
Total	4,196	52.02 ± 0.17	27.21 ± 4.88

2 Longitudinal Study Overview

In this work, we utilize blood plasma samples from a longitudinal study involving healthy individuals [17], with study code H4H_HU_2020_Sample Collection (study approval reference number: 2754-11/2020/EÜIG). As part of this study, FTIR spectroscopy was employed to analyze the collected blood plasma samples. In addition, a clinical chemistry panel consisting of 27 blood parameters was collected, together with demographic, lifestyle, and health-related information on the subjects. The spectroscopic measurements were performed using a commercial FTIR device designed for liquid sample analysis (MIRA-Analyzer, CLADE GmbH) and pre-processed similarly to previous studies [4,6,7,9,11]. The pre-processed FTIR spectra are shown in panel Fig. 1(a).

The cohort consists of 4,196 subjects who had at least five visits. The first four visits occurred approximately 130 days apart, while the interval between the fourth and fifth visits was around 150 days, resulting in an average interval of approximately 1.5 years between the first and fifth visits. The basic characteristics of this cohort are summarized in Table 1. The study population consists of significantly more female than male participants. The age distribution across sexes is nearly identical, while the BMI of males is slightly higher than that of females.

3 Stability of Infrared Molecular Profiles

This section uses multi-class classification to examine the stability and person-specific nature of blood-based molecular profiles. Logistic regression models were trained using the Python package Scikit-learn [20] to identify individuals based on FTIR spectral data, clinical chemistry panel data, and their combination. This analysis was performed on 50 individuals. Figure 1(b) shows that the classification accuracy increases with the number of visits used for training, reaching 0.90 for FTIR, 0.96 for clinical chemistry, and 0.98 for the combined data when training on four visits and predicting the fifth.

To assess the robustness of the models, cross-validation was used with four visits for training and one for testing. Figure 1(c) displays the resulting confusion matrices. The mean classification accuracies were 0.924 ± 0.069 for FTIR, 0.908 ± 0.057 for blood panel data, and 0.960 ± 0.043 for their combination, with no statistically significant differences between models.

These findings demonstrate that both spectral and biochemical data capture stable, individual-specific signatures across repeated visits.

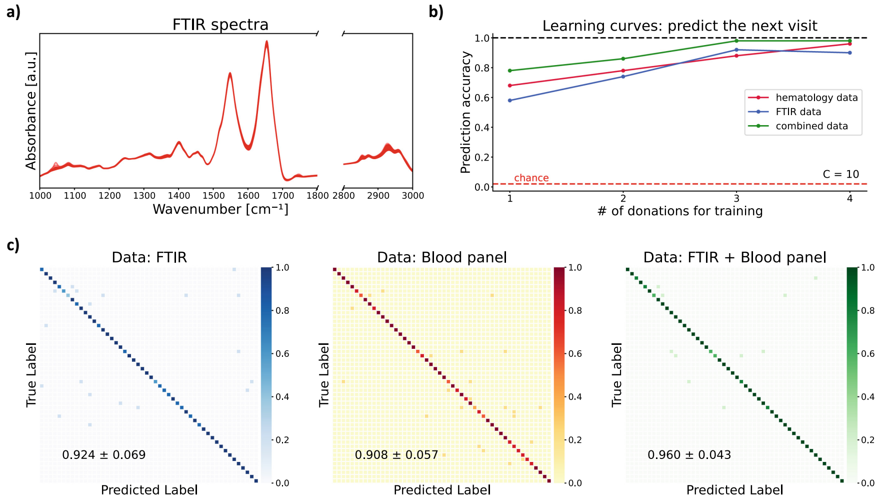


Fig. 1. Multi-class classification of 50 subjects with five visits, predicting individual identities using three different feature sets: FTIR data (519 spectral points), blood panel data (27 measured blood parameters), and their combination. **a)** The 250 FTIR spectra used for this analysis. **b)** Classification accuracy trends as additional visits are added to the training set, with predictions made on the next visit. **c)** Multi-class classification models trained on four visits and tested on the remaining visit using a 5-fold cross-validation procedure. Confusion matrices display results for the three feature sets in the order described above.

4 Blood Analyte Information Encoded Within Infrared Molecular Profiles

This section investigates shared information between the IMF data and the 27 clinical chemistry blood parameters. Specifically, we trained partial least squares regression models [20] to predict clinical chemistry variables from the IMF features. The analysis was performed using 5-fold cross-validation to ensure robust and reliable results. Table 2 summarizes the cross-validation performance for the ten best-performing blood parameters. Parameters are ranked by the models' predictive performance in terms of mean R^2, mean-square deviation (MSE), and Pearson correlation coefficient values between the predicted and the measured parameter values.

Strong predictive power was observed for glucose and lipid profile markers, all showing R^2 values above 0.80. Moderate predictability was found for hemoglobin, hematocrit, creatinine, albumin, and total protein. These findings confirm results presented in previous work based on a different cohort [9] and enhance the interpretability of IMF data.

Table 2. Cross-validation results for regression models

Analyte	R^2	Relative MSE	Pearson's r
Triglycerides	0.876 ± 0.024	0.125 ± 0.032	0.937 ± 0.013
Glucose	0.864 ± 0.023	0.135 ± 0.019	0.930 ± 0.012
Total Cholesterol	0.852 ± 0.005	0.148 ± 0.008	0.923 ± 0.003
HDL Cholesterol	0.843 ± 0.013	0.157 ± 0.016	0.918 ± 0.007
LDL Cholesterol	0.803 ± 0.001	0.197 ± 0.002	0.896 ± 0.001
Creatinine	0.600 ± 0.004	0.400 ± 0.007	0.775 ± 0.002
Hemoglobin	0.584 ± 0.013	0.416 ± 0.014	0.765 ± 0.008
Hematocrit	0.523 ± 0.015	0.477 ± 0.020	0.723 ± 0.010
Total Protein	0.462 ± 0.010	0.538 ± 0.011	0.680 ± 0.008
Albumin	0.472 ± 0.013	0.527 ± 0.012	0.688 ± 0.009

5 Individual Variation in Infrared Molecular Profiles

Personalized preventive medicine requires an understanding of an individual's molecular profile and its natural variability over time. Frequent molecular sampling could establish individualized biomarker baselines, allowing for to distinction of normal fluctuations from disease signals more sensitively. This may, e.g., be crucial for early disease detection. However, logistical and financial constraints make this approach impractical. An alternative is to group individu-

Table 3. Clinical chemistry and demographic parameters that resulted in the greatest increase in the Index of Individuality (IoI) when used for grouping similar individuals.

Covariate	IoI Increase
Total Cholesterol	0.0449
Triglycerides	0.0389
LDL Cholesterol	0.0270
Insulin	0.0164
Glucose	0.0103
Hba1C	0.0100
Albumin	0.0100
Total Protein	0.0091
CEA	0.0063
Calcium	0.0056
Age	0.0081
BMI	0.0061
Sex	0.0051
Combined	0.2112

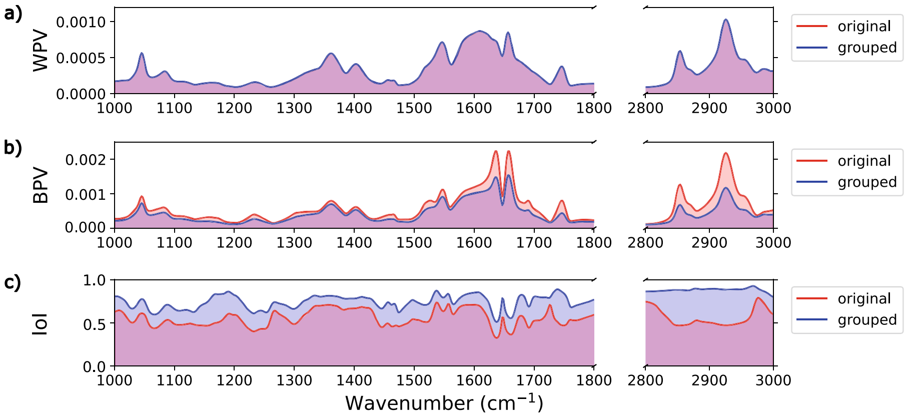


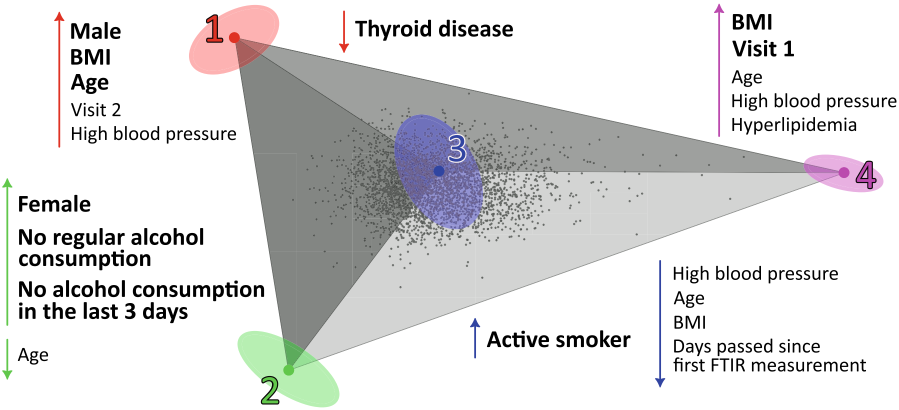
Fig. 2. The three plots show the within-person variability (a), between-person variability (b), and Index of Individuality (c). Grouped data reduces between-person variability compared to the original data, while within-person variability remains consistent, leading to an increased Index of Individuality.

als with similar molecular profiles, enabling broader but less frequent sampling across these sub-cohorts. If variability within such groups is reduced in comparison to between-person fluctuations, molecular data from a larger cohort can be integrated to create a sensitive setting, enhancing early disease detection while reducing the burden of continuous monitoring.

To construct such sub-cohorts for IMF data, individuals can be grouped based on measurable parameters such as anthropometric traits (e.g., age, sex, BMI) and blood clinical chemistry markers. The main challenge is identifying the combination of parameters that minimizes between-person variability (BPV), optimally to levels comparable to within-person variability (WPV).

Toward this goal, we evaluated BPV and WPV across individuals, aiming to form groups where BPV approaches WPV. The Index of Individuality (IoI), defined as the ratio of WPV/BPV, is increased in these subgroups in comparison to the overall population. To assess the influence of various factors on IoI, we examined 27 blood parameters and 3 demographic variables individually. Subjects were divided into evenly populated groups based on the average values of each parameter across all visits, and the resulting IoI changes were used to rank the parameters by their significance. The top 10 blood parameters and 3 demographic variables from this ranking were selected for further multivariate analysis, as detailed in Table 3. An optimization algorithm, available as an open-source Python package on GitHub [21], was then applied to the selected 13 variables, identifying optimal split points and iteratively refining group assignments to maximize IoI by minimizing BPV. The optimization algorithm consisted of a hierarchical tree structure to define the grouping strategy. A back-pruning algorithm was applied to eliminate unnecessary subdivisions and prevent overfitting. This resulted in 252 groups, with an IoI increase of

0.21, averaged across all groups and wavenumbers. The original and grouped IoI, BPV, and WPV distributions are visualized in Fig. 2. We show that, within these sub-cohorts, inter-individual variability approaches intra-individual levels. Due to the reduced variation in IMF data, such stratification could enable more precise disease detection and health monitoring.



	Discrete variables			Continuous variables		
archetype #	Feature Name: category	P value (Hypergeom.)	P maximal	Feature Name	P value (Mann-Whitney)	Is first bin maximal?
1	**sex: Male**	**1.89E-05**	**0.6245**	**bmi**	**1.79E-12**	**Yes**
1	**thyroid disease: No**	**4.11E-04**	**0.7151**	**age**	**1.08E-03**	**Yes**
1	high blood pressure: Yes	2.75E-03	0.0969			
1	visit: Visit 2	2.91E-03	0.2872			
2	**alcohol in last three days: No**	**2.48E-13**	**0.9997**	age	5.82E-09	No
2	**sex: Female**	**1.50E-08**	**0.9540**			
2	**alcohol consumption: max 4 times per month**	**1.19E-05**	**0.8625**			
3	**smoking status: Active smoker**	**2.81E-05**	**0.8094**	age	1.08E-09	No
3	high blood pressure: No	3.91E-03	0.0009	bmi	3.68E-15	No
3				ftir days passed	2.58E-04	No
4	**visit: Visit 1**	**1.49E-07**	**0.8938**	**bmi**	**8.94E-05**	**Yes**
4	high blood pressure: Yes	9.94E-04	0.3233	age	1.69E-02	No
4	hyperlipdaemia: Yes	1.31E-03	0.0469			

Fig. 3. Tetrahedral structure identified by the ParTI algorithm, with vertices representing four distinct archetypes. Colored ellipses indicate the uncertainty of archetype positions, estimated via bootstrap resampling, with shape and orientation determined by the covariance of the resampled positions. Results of enrichment analysis are shown adjacent to each archetype, listing variables significantly enriched near each vertex. Bold, enlarged entries mark variables most enriched in the bin closest to the archetype (P maximal > 0.5 for discrete variables or "Yes" for continuous ones). P-values for discrete variables were calculated using the hypergeometric test, and for continuous variables using the Mann-Whitney U-test. All variables shown remained significant after Benjamini-Hochberg correction.

6 Towards Interpretable Individual Trajectories

Originally developed in the context of evolutionary biology, ParTI was used to explain how phenotypes evolve under pressure to perform multiple conflicting tasks, such as in finch beak morphology, enzyme kinetics, or ant behavioral roles. In these systems, evolutionary trade-offs lead to phenotypic distributions forming low-dimensional geometric shapes, with archetypes corresponding to task specialists [18]. While these tasks are classically ecological or developmental, the same conceptual framework can be extended to human health data: physiological or metabolic states may be shaped by underlying constraints and trade-offs, such as immune function, metabolic regulation, or aging processes [19].

Thus, although ParTI was initially grounded in evolutionary theory, it offers a powerful lens for interpreting complex, high-dimensional datasets in biomedical settings. In this broader view, health states themselves can be seen as emergent outcomes of competing physiological demands, and archetypes as individuals or time points representing specialization in different health-related functions. This allows ParTI to bridge mechanistic understanding with clinical interpretation in longitudinal, multi-omics studies.

Here, we applied ParTI—using the available MATLAB implementation [22]—to IMF and clinical chemistry panel data collected longitudinally to determine whether high-dimensional blood-based measurements could be structured into a geometry reflecting distinct health states. To explore this, we fitted ParTI using three feature sets: (i) IMF data alone, (ii) blood panel data alone, and (iii) a combination of both, with and without standard scaling, resulting in six configurations. Among these, only the model trained on unscaled IMF data yielded a statistically significant ($p = 0.002$) tetrahedral geometry. To assess whether this arrangement arose from random variation, we performed a permutation test by shuffling feature values across samples while preserving their distribution. By comparing the volume ratio of the archetypal tetrahedron to the convex hull across 10,000 randomized datasets, we confirmed that the observed structure was unlikely to occur by chance. Additionally, bootstrapping estimated variability in archetype positions, providing confidence intervals for the identified extreme states.

To minimize systematic bias and ensure consistency in sample collection, processing, and IMF data analysis, only data from one clinical site were used. The following analysis is based on a cohort comprising 3,583 IMF and blood panel profiles across five visits. After identifying archetypes, we conducted enrichment analysis using discrete and continuous variables from the longitudinal study, such as sex, comorbidities (e.g., hypertension, malignancies, respiratory diseases), and lifestyle habits like smoking and alcohol consumption. We also examined whether individuals' positions within the tetrahedral space shifted over time, assessing both categorical (visit number) and continuous (days since the first visit) temporal variables.

In the case of unscaled IMF data, enrichment analysis revealed distinct demographic and physiological associations. For example, as illustrated in Fig. 3, Archetype 1 was linked to older males with higher BMI, while Archetype 2 was enriched for females with low alcohol consumption. Additionally, systematic time-dependent shifts were observed, possibly related to device inconsistencies, with subjects' first visits clustering near Archetype 4 and subsequent earlier measurements around Archetype 3. However, the lack of detailed lifestyle parameters (e.g., physical activity, mental well-being, nutrition) limits our ability to define archetypes associated with optimal health states. Furthermore, the rarity of certain comorbidities in our relatively healthy cohort potentially obscured disease-related archetype associations.

The identification of a robust tetrahedral structure in our dataset echoes findings from a previous application of ParTI [19], where a similar geometry emerged with archetypes corresponding to sex-specific physiological profiles, reinforcing the reproducibility of ParTI across diverse biological settings and data types. This consistency underscores the method's robustness and suggests its potential utility in future applications, such as tracking early deviations from typical health trajectories, stratifying populations based on physiological trade-offs, or monitoring personalized responses to interventions. As high-dimensional clinical data becomes increasingly accessible, ParTI offers a principled, interpretable framework for uncovering the latent structure of human health.

7 Discussion and Outlook

These findings highlight that IMF data is a scalable and personalized tool for health monitoring and early disease detection. The observed stability and individuality of IMF profiles underscore their potential as robust biomarkers for tracking physiological changes over time. Additionally, reducing variability in IMF profiles by grouping molecularly similar individuals enables the development of AI models that account for inter-individual differences while leveraging population-level insights. Furthermore, identifying structured health states through Pareto Task Inference (ParTI) provides a framework for categorizing and interpreting individual health trajectories. This could facilitate early identification of deviations that indicate the onset of non-communicable diseases before symptoms manifest, and open avenues for future applications such as population stratification, early disease prediction, and personalized monitoring. While our current dataset lacked detailed lifestyle data, future studies could leverage richer metadata to link molecular archetypes to behavioral or environmental factors, enhancing clinical interpretability and utility.

Acknowledgements. We thank all Center for Molecular Fingerprinting (CMF) employees for fostering the research environment that enabled the realization of this work. In particular, we thank Diána Debreceni's and her team for preparing the FTIR measurements. KVK acknowledges insightful discussions with Marija Marolt.

Disclosure of Interests. The authors declare no competing interests.

References

1. Tebani, A., et al.: Integration of molecular profiles in a longitudinal wellness profiling cohort. Nat. Commun. **11**(1), 4487 (2020)
2. Baker, M.J., et al.: Using Fourier transform IR spectroscopy to analyze biological materials. Nat. Protocols **9**(8), 1771–1791 (2014)
3. Butler, H.J., et al.: Development of high-throughput ATR-FTIR technology for rapid triage of brain cancer. Nat. Commun. **10**(1), 4501 (2019)
4. Voronina, L., et al.: Molecular Origin of Blood-Based Infrared Spectroscopic Fingerprints. Angewandte Chemie **133**(31), 17197–17206 (2021)
5. Paraskevaidi, M., et al.: Clinical applications of infrared and Raman spectroscopy in the fields of cancer and infectious diseases. Appl. Spectrosc. Rev. **56**(8-10), 804–868 (2021)
6. Huber, M., et al.: Stability of person-specific blood-based infrared molecular fingerprints opens up prospects for health monitoring. Nat. Commun. **12**(1), 1511 (2021)
7. Huber, M., et al.: Infrared molecular fingerprinting of blood-based liquid biopsies for the detection of cancer. Elife **10**, e68758 (2021)
8. Ghimire, H., et al.: Protein conformational changes in breast cancer sera using infrared spectroscopic analysis. Cancers **12**(7), 1708 (2020)
9. Eissa, T., et al.: Plasma infrared fingerprinting with machine learning enables single-measurement multi-phenotype health screening. Cell Reports Med. **5**(7), 101625 (2024)
10. Zigman, M., et al.: 90P Infrared molecular fingerprinting: a new in vitro diagnostic platform technology for cancer detection in blood-based liquid biopsies. Ann. Oncol. **33**, S580 (2022)
11. Kepesidis, K.V., et al.: Breast-cancer detection using blood-based infrared molecular fingerprints. BMC Cancer **21**, 1–9 (2021)
12. Kepesidis, K.V., et al.: Assessing lung cancer progression and survival with infrared spectroscopy of blood serum. BMC Med. **23**(1), 101 (2025)
13. Backhaus, J., et al.: Diagnosis of breast cancer with infrared spectroscopy from serum samples. Vib.l Spectrosc. **52**(2), 173–177 (2010)
14. Elmi, F., et al.: Application of FT-IR spectroscopy on breast cancer serum analysis. Spectrochimica Acta Part A: Mol. Biomol. Spectrosc. **187**, 87–91 (2017)
15. Ollesch, J., et al.: It's in your blood: spectral biomarker candidates for urinary bladder cancer from automated FTIR spectroscopy. J. Biophotonics **7**(3-4), 210–221 (2014)
16. Anderson, D.J., et al.: Liquid biopsy for cancer diagnosis using vibrational spectroscopy: systematic review. BJS open **4**(4), 554–562 (2020)
17. H4H study homepage. https://h4h.hu/en/. Accessed 30 Jan 2025
18. Hart, Y., et al.: Inferring biological tasks using pareto analysis of high-dimensional data. Nat. Methods **12**(3), 233–235 (2015)
19. Zimmer, A., et al.: The geometry of clinical labs and wellness states from deeply phenotyped humans. Nat. Commun. **12**(1), 3578 (2021)
20. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)

21. Zarandy, Z.I.: Stratified IOI subgrouping: a tree-based method to group subjects by index of individuality. GitHub repository (2025). https://github.com/center-for-molecular-fingerprinting/stratified-ioi-subgrouping
22. Hart, Y., et al.: A MATLAB package to perform pareto task inference (ParTI), which infers tasks from high-dimensional datasets. GitHub repository (2015). https://github.com/AlonLabWIS/ParTI