Multivariate Gaussian Modeling in Blood-Based FTIR Spectroscopy: Framework and Applications

Niklas Leopold-Kerschbaumer[1,2,3,*], Flora B. Nemeth[1,2,3], Nico Feiler[1,2,3], and Kosmas V. Kepesidis[1,2,3,*]

[1] Center for Molecular Fingerprinting (CMF), Czuczor Street 2-10, 1093 Budapest, Hungary
[2] Faculty of Physics, Ludwig-Maximilians-Universität München (LMU), Am Coulombwall 1, 85748 Garching, Germany
[3] Laboratory for Attosecond Physics, Max Planck Institute of Quantum Optics (MPQ), Hans-Kopfermann-Straße 1, 85748 Garching, Germany
[*] Correspondence: kosmas.kepesidis@cmf.hu, niklas.leopoldkerschbaumer@cmf.hu

ABSTRACT

Infrared molecular fingerprinting of human blood samples offers a powerful and minimally invasive approach for disease detection and health monitoring. In this study, we investigate the use of multivariate Gaussian (MVG) modeling to represent blood-based infrared molecular fingerprints obtained from a large case–control study that targets various cancer types. We demonstrate that synthetic FTIR datasets generated using MVGs preserve the key statistical and physical characteristics of real molecular fingerprints, thereby enabling the development and validation of analytical methodologies without relying on sensitive clinical data. Using this approach, we address three major challenges in clinical FTIR spectroscopy. First, we evaluate the potential of longitudinal study designs for early disease detection, such as lung cancer, by adjusting covariance structures of the case-control study data using the index of individuality, estimated from two independent longitudinal studies. Second, we propose a domain adaptation method to harmonize classification models, trained on data from two spectrometers of the same type, using the covariance from a small calibration set measured on both devices. The method's performance is assessed by training and testing these models across devices. Third, we provide MVG models for each cohort to serve as a proxy for real data in methodological research, facilitating reproducibility and collaboration in biomedical spectroscopy. This approach offers a practical solution for overcoming ethical barriers in clinical data sharing in spectroscopic biomarker research.

Keywords: Infrared spectroscopy, Molecular fingerprinting, Synthetic data, Cancer diagnostics

BACKGROUND & SUMMARY

Advancements in spectroscopic techniques have opened new frontiers in minimally-invasive disease diagnostics and personalized health monitoring[1–4]. Among these, infrared molecular fingerprinting of blood-based samples has demonstrated significant potential for detecting and characterizing various pathological conditions, including cancer and metabolic disorders[5–16]. However, the full utility of such datasets is often hindered by strict ethical and legal restrictions that prevent their open sharing, thereby limiting collaborative research and methodological development.

To address this challenge, we generate synthetic blood-based infrared molecular fingerprints that retain the essential statistical and physical properties of real patient data while containing no identifiable information. Synthetic datasets are constructed using multivariate Gaussian models fitted on real-world measurements from a cross-sectional study investigating various cancer types. The resulting synthetic data records encompass matched case–control cohorts for lung, breast, bladder, and prostate cancer, including independent training and test splits. For lung cancer, additional stratification by stage (I–IV) is provided. All MVG parameters are publicly released, allowing researchers to generate arbitrarily large synthetic cohorts conditioned on biologically meaningful attributes. Technical validation demonstrates that simulated datasets reproduce key statistical properties of the real data, including differential spectral effect sizes, peak-ratio distributions, principal component structure, and classification performance across independent test sets.

Beyond providing static synthetic cohorts, we introduce two methodological extensions that leverage the MVG framework. First, we develop a variance-scaling approach based on the index of individuality to simulate

Data Science for Photonics and Biophotonics II, edited by Thomas Bocklitz
Proc. of SPIE Vol. 14099, 1409905 · 2026 Published by SPIE
0277-786X · doi: 10.1117/12.3098971

longitudinal self-referencing scenarios. By modifying the covariance structure of healthy controls to reflect reduced within-person variability, we quantify the potential diagnostic gains achievable in longitudinal study designs before experimental implementation. Second, we present AdaptFTIR, a domain-aware data augmentation strategy that incorporates structured cross-device variability into the MVG covariance matrix. This approach enables systematic evaluation and mitigation of domain shift effects in cross-device classification tasks, improving robustness and generalization across measurement sites.

Together, these components establish a unified statistical framework for privacy-preserving data sharing, longitudinal study planning, and cross-device robustness analysis in blood-based FTIR spectroscopy.

Methods

Study cohorts and design

This work is based on blood plasma samples collected within three clinical studies employing infrared molecular fingerprinting: the multi-center *Lasers4Life* (L4L) study (DRKS00013217), its longitudinal substudy (LG), and the large-scale longitudinal *Health for Hungary* (H4H) study (Study approval reference number: 2754-11/2020/EÜ IG).

The L4L study was conducted in the Munich area and approved by the Ethikkommission bei der LMU München (EK 20170820—Nr.:17-532). The study complied with Good Clinical Practice (ICH-GCP), the Declaration of Helsinki, and all applicable regulatory requirements. All participants provided written informed consent before inclusion.

The primary L4L cohort follows a cross-sectional case–control design with lung cancer, prostate cancer, breast cancer, and bladder cancer patients, as well as a non-symptomatic reference group. In addition, a longitudinal substudy (LG) followed 18 healthy individuals over time, with repeated blood sampling at predefined intervals of 2-19 days and 8-11 visits per individual.

The independent H4H study constitutes a large-scale longitudinal cohort including 9363 individuals with repeated blood sampling over four years.

Blood collection and sample handling protocols as well as FTIR measurement specifications are described in detail by Huber et. al.[6] for the L4L and LG study, and by Nemeth et.al.[17] for the H4H study. Importantly, all studies follow nearly identical study protocols.

Spectral preprocessing

All spectra from L4L, LG, and H4H were processed using an identical preprocessing pipeline rooted in literature[6,12], to ensure methodological consistency. In brief, the pipeline is as follows:

Outlier detection was performed separately for training and test sets using the Local Outlier Factor (LOF) algorithm implemented in scikit-learn (Python version 1.6.1). Spectra identified as technical artifacts or incomplete measurements were excluded before downstream statistical modeling. Afterwards, the spectral region was restricted to 1000–3000 cm^{-1}. Next, the spectra underwent l2-normalization and lastly, the silent region (1800 cm^{-1} to 2800 cm^{-1}) was removed.

Case–control construction and statistical matching

Within the cross-sectional L4L cohort, individuals were classified as either therapy-naïve patients diagnosed with lung, prostate, bladder, or breast cancer or as non-symptomatic reference individuals. Before measuring, the obtained blood samples were split into a training set and a test set. The measurements were carried out in a fully randomized manner over 19 weeks. After a 10-week gap, introduced to account for potential drifts in spectrometer performance and ensure robust testing, the test set was measured in randomized order over 2 weeks.

To construct balanced case–control datasets and reduce confounding, propensity score matching was performed based on age and sex. Matching was implemented in R (version 4.4.1).

Following matching, the final cohorts comprised 2,079 individuals, of whom 1,650 were assigned to training sets and 429 to test sets. Detailed demographic characteristics and sample distributions are provided in Table S1 in *Supplementary Information A*.

Longitudinal analysis framework

Both longitudinal studies were used to assess and compare the index of individuality (IOI) of infrared molecular fingerprints, which is further used in the "Longitudinal Study Prospects" section. Additionally, a subset of blood samples in the H4H study was measured on two different devices with identical construction specifications. This allowed for the assessment of domain shift due to different devices and is handled in the "ADAPT FTIR" section. Tables on the demographics of both longitudinal studies are provided in Table S2 in *Supplementary Information B* and Table S3 in *Supplementary Information C*, respectively.

Multivariate Gaussian modeling

Assuming that absorbance values at each wavenumber are approximately Gaussian distributed, synthetic infrared spectra can be generated by sampling from a multivariate Gaussian distribution (MVG) fitted to measured spectra. The MVG defines the probability density function

$$\mathcal{G}(x|\mu,\Sigma) = \frac{1}{(2\pi)^{N/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right), \tag{1}$$

where $\mu \in \mathbb{R}^N$ represents the mean spectrum, and $\Sigma \in \mathbb{R}^{N\times N}$ is the covariance matrix capturing spectral correlations. Both the mean vectors and covariance matrices are provided for each dataset split described in Section . This modeling approach builds on prior work that applied MVG distributions to generate synthetic FTIR spectroscopy datasets for purposes such as sample size planning and method development[17–21]. In addition to patient spectra, MVG models were also fit to quality control (QC) samples—constructed by pooling blood specimens—with 1,048 QC samples included in the training set and 131 in the test set.

Technical Validation

To validate our simulation approach, we simulated cohorts of the L4L study of equal sizes to the real data and compared effect sizes (Cohen's d) by analyzing both the full spectra and selected peak ratios, in line with prior work[5]. We also evaluated the performance of a logistic regression classifier with L2 regularization ($C = 10$) on both datasets, using ROC curves for comparison. Results for the lung cancer cohort, stratified by disease status and sex (fourth row in Figure 2), are summarized in Figure 1.

Figure 1a1 shows 10 representative spectra alongside a PCA scatter plot, illustrating the visual similarity between real and simulated data. As seen in Figure 1a2, the effect size between real and simulated datasets is close to zero across all wavenumbers. Similarly, a test comparing the means per wavenumber shows that there is no significant difference between real and simulated data. Figures 1a3 and 1a4 demonstrate that effect sizes with respect to disease label and sex in the simulated data closely mirror those in the real data.

Similarly, Figure 1b1 shows that the distribution of peak ratios in the simulated data closely matches that of the real data. Figures 1b2–b4 provide a corresponding effect size and t-test analysis on these peak ratios, analogous to panel A. Which peak ratios were assigned to which index is shown in Table S4 in *Supplementary Information D*.

In terms of classification performance, the ROC curves in Figure 1C and Figure 1D show comparable results for sex prediction on all healthy individuals and disease prediction on the lung cancer cohort across real and simulated data, as well as between the training and independent test sets. The results also reveal a lower AUC for classification in the test set compared to the training set, indicating a domain shift. Notably, classifiers trained on multivariate Gaussian (MVG) simulations generally slightly outperform those trained on real data, while performance decreases on the independent test set. A complete summary of AUC scores across all dataset splits using the logistic regression classifier is provided in *Table S5* in *Supplementary Information E*.

For training logistic regression models, we applied standard scaling to the preprocessed data. The ROC curves on the training set were calculated using repeated cross-validation with 10 splits and 5 repetitions, while for evaluating the test set, the entire training set was used. This setup, along with the preprocessing pipeline and classification model, follows established practices in blood-based FTIR spectroscopy tasks[6].

All technical validation experiments were conducted using Python 3.10.16 with NumPy 2.2.4 and scikit-learn 1.6.1.

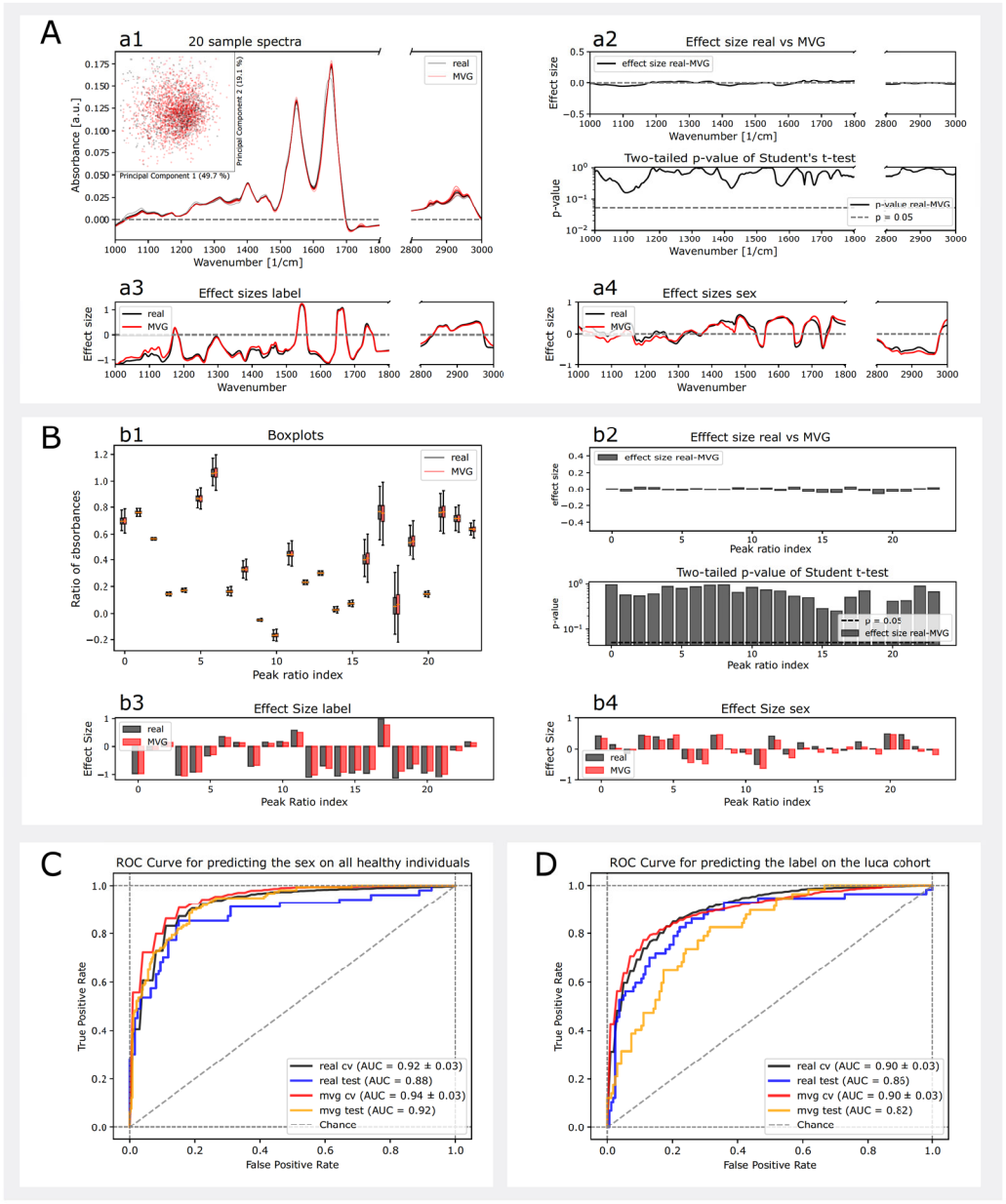

Figure 1: Comparison between real data and simulated data. **A)** Basic analysis on the spectra, including 1. 10 spectra of real data and 10 spectra of simulated data with PCA scatter plot of all spectra, 2. Effect size and t-test p-values between real and simulated data, 3. Effect size between the lung cancer label for real and simulated data, 4. Effect size between the sexes for real and simulated data. **B)** Basic analysis on the peak ratios, including 1. Boxplot for each peak ratio for real and simulated data, 2. Effect size and t-test p-values between real and simulated data, 3. Effect size between the lung cancer label for real and simulated data, 4. Effect size between the sexes for real and simulated data. **C)** ROC curve for predicting sex with the logistic regression classifier described above on all healthy individuals on the test set, as well as within the training set for both real and simulated data. **D)** ROC curves for predicting the disease with the logistic regression classifier described above on the lung cancer cohort on the test set, as well as within the training set for both real and simulated data.

SYNTHETIC DATA RECORDS

Using MVGs, we generated synthetic spectra for matched cohorts of healthy and diseased individuals across several cancer types: lung cancer (luca), breast cancer (brca), bladder cancer (blca), and prostate cancer (prca). For lung cancer, additional stratification was performed by disease stage, resulting in subcohorts for stages I through IV. Due to limited data for early-stage lung cancer, we also include a multivariate Gaussian (MVG) fit for a combined cohort representing stages I and II.

As previously described, each (sub-)cohort was divided into training and test sets before sample measurement. Within each set, samples were further grouped by disease status (healthy or diseased) and sex (male or female). A schematic overview of the full data simulation pipeline is shown in Figure 2A, while detailed data splits for each (sub-)cohort are presented in Figure 2B. Every data split highlighted in orange in Figure 2B corresponds to a fitted MVG model—characterized by a mean spectrum and covariance matrix—available in our GitHub repository [22].

These MVG models, defined by Equation 1, enable users to sample synthetic spectra conditioned on sex (male or female), disease status (healthy or diseased), and, for lung cancer, disease stage (I–IV). Additionally, we provide a reference cohort composed solely of healthy individuals, stratified by sex, from which MVG models for healthy male and female spectra were derived.

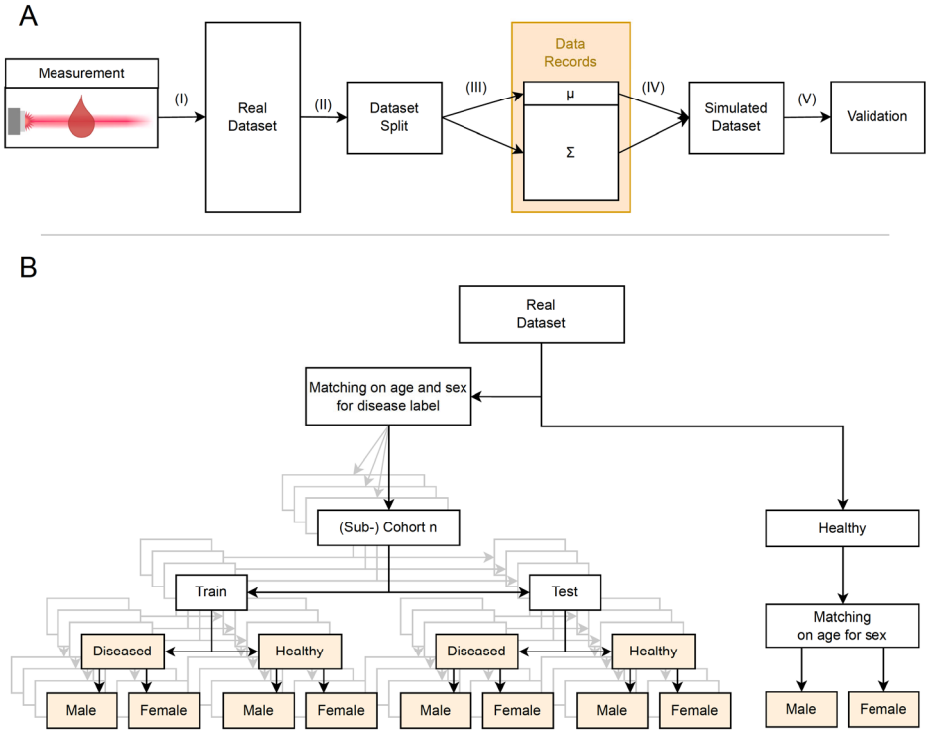

Figure 2: **A)** Diagram illustrating how the simulated datasets are obtained. (I) Samples of study participants are measured via FTIR, and measurements are stored in a dataset. (II) The real dataset is split into subsets based on disease status and/or sex according to Figure 2. (III) The mean and covariance matrix are calculated for each subset. (IV) Simulated spectra are sampled according to Equation 1. (V) The quality of simulated datasets is verified by comparing the ROC curves, differential fingerprints, and effect sizes to real data. **B)** Dataset splits visualized. Orange colored boxes indicate that the mean and covariance matrix are available for sampling new spectra using an MVG.

LONGITUDINAL STUDY PROSPECTS

Designing longitudinal studies—where subjects are measured repeatedly over time—can provide greater sensitivity for detecting disease-related changes compared to traditional case–control designs. Quantifying the potential benefit of a longitudinal approach, however, remains a key challenge. For classical clinical biomarkers, this assessment is often based on the index of individuality (ioi), which expresses how much an individual's biomarker values vary relative to the variability across a population. The ioi is defined as

$$\text{ioi} = \frac{\text{wpv}}{\text{bpv}}, \tag{2}$$

where wpv denotes the within-person variability and bpv the between-person variability. For multivariate datasets such as infrared (IR) spectra, this division is performed element-wise. Our formal definitions of wpv and bpv are provided in Section F in the *Supplementary Information*.

MVG self-referencing approach

Extending this concept to blood-based FTIR spectroscopy, we estimated the ioi using the two longitudinal datasets described in the methods section. The resulting ioi values across the spectral domain are shown in Figure 3A.

To model how a longitudinal design might improve disease detectability, we modified the covariance matrix of healthy control samples, Σ^{con}, to reflect reduced within-person variance expected in repeated measurements from the same individual. Specifically, we applied a variance-scaling transformation based on the estimated ioi values:

$$\Sigma' = \begin{pmatrix} \text{ioi}_1^2\Sigma_{11}^{\text{con}} & \dots & \text{ioi}_1\text{ioi}_D\Sigma_{1D}^{\text{con}} \\ \vdots & \ddots & \vdots \\ \text{ioi}_1\text{ioi}_D\Sigma_{D1}^{\text{con}} & \dots & \text{ioi}_D^2\Sigma_{DD}^{\text{con}} \end{pmatrix} = \Lambda\Sigma^{\text{con}}\Lambda, \quad \text{where} \quad \Lambda = \text{diag}(\text{ioi}), \tag{3}$$

where D denotes the number of spectral grid points and Σ^{con} the covariance matrix of the control group. This adjustment effectively scales the covariance structure of the healthy cohort, yielding a "self-referenced" population distribution with reduced variability (Figure 3B).

Using this modified covariance, we simulated synthetic longitudinal spectra from the healthy population while keeping the disease spectra unchanged. We then repeated the same cross-validation procedure as described previously, this time training classifiers on the variance-scaled healthy data and testing them against real diseased samples. The resulting ROC–AUC values (Figure 3C) quantify the potential gain in discriminative performance under a longitudinal design scenario.

Overall, our results indicate that incorporating longitudinal information—modeled here as a reduction in within-subject variance—can substantially improve classification performance in blood-based FTIR spectroscopy. This provides a quantitative framework for estimating the added diagnostic value of longitudinal sampling before such studies are conducted experimentally.

ADAPT FTIR: A CROSS-DEVICE GENERALIZATION TECHNIQUE

In medical research, measurements from the same device type at two different locations might exhibit a domain shift due to various reasons such as device imperfections, lab conditions such as humidity and air pressure, etc. We propose AdaptFTIR, a method in which MVG modeling is employed as a domain-aware data augmentation strategy to address systematic cross-device variability in FTIR spectroscopy. Each preprocessed FTIR spectrum is treated as a high-dimensional random vector $x \in \mathbb{R}^W$, where W denotes the number of retained wavenumber channels after truncation, normalization, and removal of non-informative spectral regions. The objective of the augmentation is to generate synthetic spectra that preserve biologically meaningful spectral structure while explicitly incorporating device-induced variability.

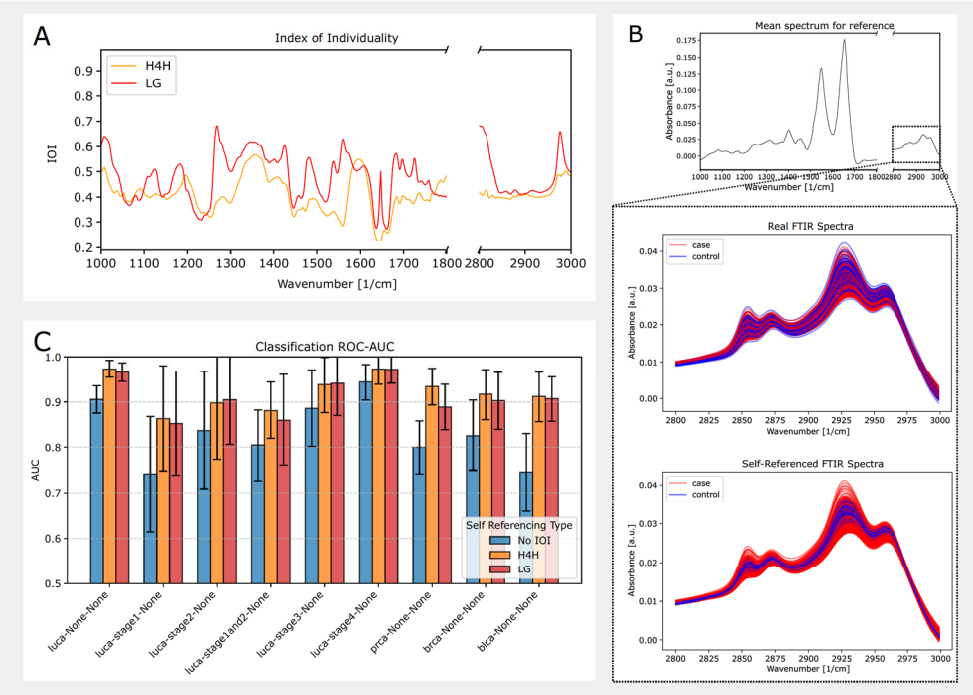

Figure 3: **A)** Index of individuality (IOI) derived from the H4H and LG longitudinal studies. **B)** Illustration of the variance scaling applied to healthy samples. **C)** ROC–AUC results comparing classification performance with and without longitudinal variance scaling.

MVG formulation and parameterization

Synthetic spectra are generated by sampling from a multivariate normal distribution

$$x \sim \mathcal{N}(\mu_i, \Sigma_{\text{cal}}), \tag{4}$$

where the mean vector μ_i corresponds to the subject-specific average spectrum computed from the measured training data, and the covariance matrix Σ_{cal} is shared across all subjects.

The covariance matrix Σ_{cal} is estimated from a calibration dataset comprising repeated measurements of the same individuals acquired on multiple FTIR devices. For each calibration subject, a covariance matrix is computed from their paired spectra across devices and visits, capturing wavelength-resolved correlations that arise from device-dependent effects rather than biological variation. These subject-specific covariance matrices are then averaged element-wise to obtain a single, stable covariance estimate:

$$[\Sigma_{\text{cal}}]_{mn} = \frac{1}{H} \sum_{i=1}^{H} [\Sigma(p_i)]_{mn}\,, \tag{5}$$

where H denotes the number of calibration subjects and p_i represents the matrix of spectra belonging to subject i.

This construction embeds structured cross-device spectral variability—including baseline distortions, amplitude scaling, and localized spectral deviations—directly into the generative model. Sampling from this MVG yields synthetic spectra that remain anchored to individual biochemical fingerprints via μ_i, while spanning the joint variability observed across devices. For each individual in the training set, multiple synthetic spectra are generated and appended to the original data, effectively simulating additional visits under heterogeneous measurement conditions.

Evaluation strategy

The impact of MVG-based augmentation on cross-device generalization is assessed using deliberately stringent classification protocols designed to isolate domain-shift effects. Two complementary prediction tasks are considered.

First, we compared both domains before and after data augmentation via principal component analysis (PCA) (see Figure 4A). Second, a binary classification task is formulated to predict biological sex, evaluated on individuals not seen during training in order to prevent subject-specific information leakage. In Figure 4B we see that the MVG augmentation approach successfully overcomes the difference in AUC scores. Third, a multi-class classification task is defined in which each individual constitutes a separate class. This task is highly sensitive to subtle spectral distortions and therefore serves as a stringent test of cross-device robustness. Figure 4C shows that also in this task the gap between within-site testing and cross-site testing was closed.

For both tasks, models are trained exclusively on spectra acquired from a single FTIR device and evaluated on data measured either on the same device (within-device evaluation) or on a different device (cross-device evaluation). Logistic regression with L2 regularization is used consistently across all experiments to ensure model interpretability and to avoid confounding effects related to model complexity.

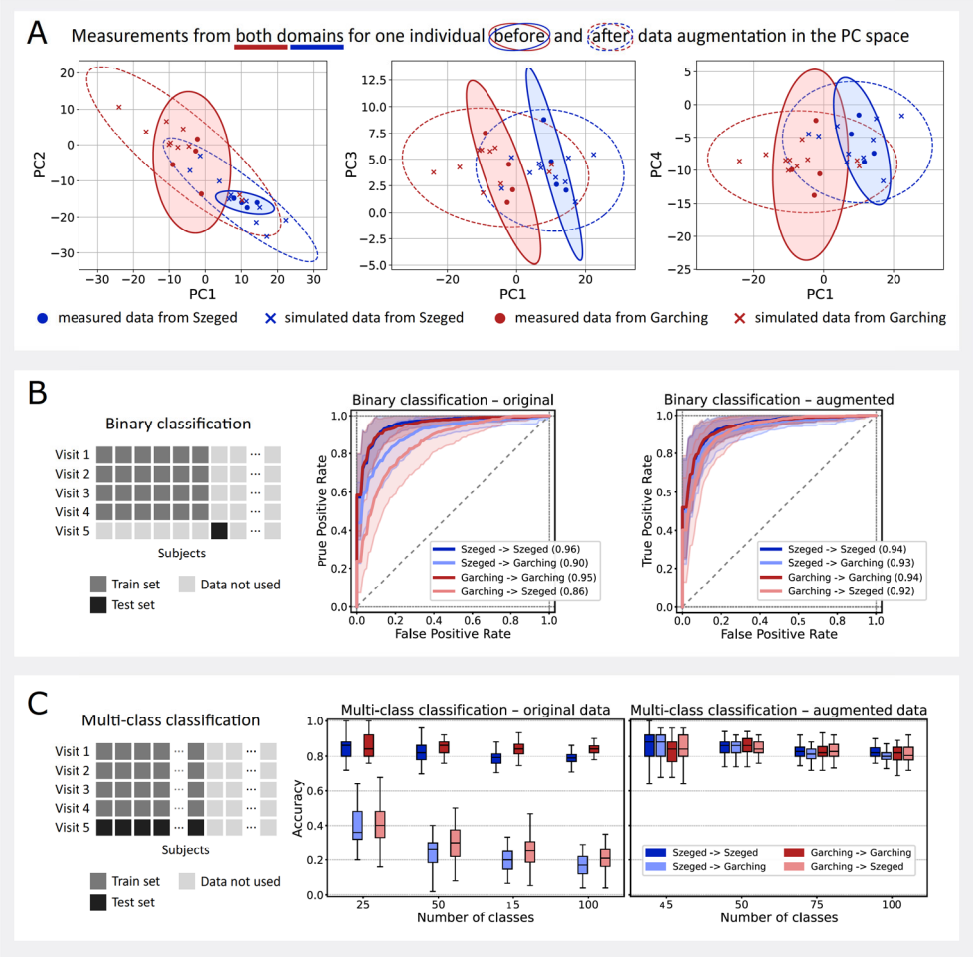

Figure 4: A) Principal component analysis for measurements from both domains. Dashed and solid lines show 2D kernel density estimates, B) Binary classification for predicting the sex, C) Multiclass classification for predicting the subject ID.

USAGE NOTES

Limitations

While simulations based on multivariate Gaussian fits provide highly interpretable approximations of data distributions, they come with important limitations. Most notably, MVG-based sampling does not support accurate modeling of individual-level variability. As a result, the generated spectra do not account for personal covariates such as age or BMI. Additionally, because the underlying dataset originates from a case-control study, it does not support longitudinal simulations or temporal analyses. Finally, MVG models assume linear relationships among features; any non-linear dependencies across wavenumbers present in real measurements may be lost in the fitted distributions.

DATA AVAILABILITY

All mean spectra and covariance matrices described in the "Synthetic Data Records" section are available in our GitHub repository[22]. Furthermore, the AdaptFTIR method is presented in a GitHub repository[23] and can be installed via pip.

CODE AVAILABILITY

Both GitHub repositories also include an example Python scripts demonstrating how to use these models.

ACKNOWLEDGMENTS

We thank Mihaela Žigman and all contributors involved in the design and execution of the *Lasers4Life* clinical study. We also express our gratitude to Ferenc Krausz for fostering the research environment that enabled the realization of this work.

This work was supported by the Center for Molecular Fingerprinting Research Nonprofit LLC (CMF), the Centre for Advanced Laser Applications (CALA) at LMU Munich, and the Max Planck Institute of Quantum Optics (MPQ). The work is part of Project no. 2020-2.1.1-ED-2022-00213 that has been implemented with the support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, financed under the 2020-2.1.1-ED funding scheme.

References

[1] Baker, M. J. *et al.* Using fourier transform ir spectroscopy to analyze biological materials. *Nature protocols* **9**, 1771–1791 (2014).

[2] Butler, H. J. *et al.* Development of high-throughput atr-ftir technology for rapid triage of brain cancer. *Nature communications* **10**, 4501 (2019).

[3] Voronina, L. *et al.* Molecular origin of blood-based infrared spectroscopic fingerprints. *Angewandte Chemie* **133**, 17197–17206 (2021).

[4] Paraskevaidi, M. *et al.* Clinical applications of infrared and raman spectroscopy in the fields of cancer and infectious diseases. *Applied Spectroscopy Reviews* **56**, 804–868 (2021).

[5] Huber, M. *et al.* Stability of person-specific blood-based infrared molecular fingerprints opens up prospects for health monitoring. *Nature communications* **12**, 1511 (2021).

[6] Huber, M. *et al.* Infrared molecular fingerprinting of blood-based liquid biopsies for the detection of cancer. *Elife* **10**, e68758 (2021).

[7] Martin, F. L. *et al.* Distinguishing cell types or populations based on the computational analysis of their infrared spectra. *Nature protocols* **5**, 1748–1760 (2010).

[8] Ghimire, H. *et al.* Protein conformational changes in breast cancer sera using infrared spectroscopic analysis. *Cancers* **12**, 1708 (2020).

[9] Ollesch, J. *et al.* An infrared spectroscopic blood test for non-small cell lung carcinoma and subtyping into pulmonary squamous cell carcinoma or adenocarcinoma. *Biomedical Spectroscopy and Imaging* **5**, 129–144 (2016).

[10] Eissa, T. *et al.* Plasma infrared fingerprinting with machine learning enables single-measurement multi-phenotype health screening. *Cell Reports Medicine* **5** (2024).

[11] Kepesidis, K. V. *et al.* Assessing lung cancer progression and survival with infrared spectroscopy of blood serum. *BMC medicine* **23**, 101 (2025).

[12] Kepesidis, K. V. *et al.* Breast-cancer detection using blood-based infrared molecular fingerprints. *BMC cancer* **21**, 1–9 (2021).

[13] Backhaus, J. *et al.* Diagnosis of breast cancer with infrared spectroscopy from serum samples. *Vibrational Spectroscopy* **52**, 173–177 (2010).

[14] Elmi, F., Movaghar, A. F., Elmi, M. M., Alinezhad, H. & Nikbakhsh, N. Application of ft-ir spectroscopy on breast cancer serum analysis. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **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. *Journal of biophotonics* **7**, 210–221 (2014).

[16] Anderson, D., Anderson, R., Moug, S. & Baker, M. Liquid biopsy for cancer diagnosis using vibrational spectroscopy: systematic review. *BJS open* **4**, 554–562 (2020).

[17] Nemeth, F. B. *et al.* Bridging spectral gaps: Cross-device model generalization in blood-based infrared spectroscopy. *Analytical Chemistry* (2025).

[18] Beleites, C., Neugebauer, U., Bocklitz, T., Krafft, C. & Popp, J. Sample size planning for classification models. *Analytica chimica acta* **760**, 25–33 (2013).

[19] Eissa, T., Kepesidis, K. V., Zigman, M. & Huber, M. Limits and prospects of molecular fingerprinting for phenotyping biological systems revealed through in silico modeling. *Analytical Chemistry* **95**, 6523–6532 (2023).

[20] Wegner, C. *et al.* Toward informative representations of blood-based infrared spectra via unsupervised deep learning. *Journal of Biophotonics* e70011 (2025).

[21] Schroeder, J. *et al.* Information-optimal measurement: From fixed sampling protocols to adaptive spectroscopy. *arXiv preprint arXiv:2505.14364* (2025).

[22] Leopold-Kerschbaumer, N. MVG-IMF. https://github.com/Attoworld-Data-Science/MVG-IMF (2025).

[23] Leopold-Kerschbaumer, N. AdaptFTIR. https://github.com/Niklas-LK/AdaptFTIR (2024).

[24] Blat, A. *et al.* Fourier transform infrared spectroscopic signature of blood plasma in the progression of breast cancer with simultaneous metastasis to lungs. *Journal of Biophotonics* **12**, e201900067 (2019).

[25] Ghimire, H., Jayaweera, P. & Perera, A. U. Longitudinal analysis of molecular alteration in serum samples of dextran sodium sulfate-induced colitis mice by using infrared spectroscopy. *Infrared Physics & Technology* **97**, 33–37 (2019).

[26] Ghimire, H., Venkataramani, M., Bian, Z., Liu, Y. & Perera, A. U. Atr-ftir spectral discrimination between normal and tumorous mouse models of lymphoma and melanoma from serum samples. *Scientific reports* **7**, 16993 (2017).

[27] Wang, X., Shen, X., Sheng, D., Chen, X. & Liu, X. Ftir spectroscopic comparison of serum from lung cancer patients and healthy persons. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **122**, 193–197 (2014).

[28] Lasch, P., Beekes, M., Fabian, H. & Naumann, D. Antemortem identification of transmissible spongiform encephalopathy (tse) from serum by mid-infrared spectroscopy. *Handbook of vibrational spectroscopy* (2006).

[29] Sahu, R. K. *et al.* Continuous monitoring of wbc (biochemistry) in an adult leukemia patient using advanced ftir-spectroscopy. *Leukemia research* **30**, 687–693 (2006).

[30] Staniszewska-Slezak, E., Mateuszuk, L., Chlopicki, S., Baranska, M. & Malek, K. Alterations in plasma biochemical composition in no deficiency induced by l-name in mice analysed by fourier transform infrared spectroscopy. *Journal of Biophotonics* **9**, 1098–1108 (2016).

[31] Sheng, D. *et al.* Comparison of serum from gastric cancer patients and from healthy persons using ftir spectroscopy. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **116**, 365–369 (2013).

[32] Staniszewska-Slezak, E. *et al.* A possible fourier transform infrared-based plasma fingerprint of angiotensin-converting enzyme inhibitor-induced reversal of endothelial dysfunction in diabetic mice. *Journal of Biophotonics* **11**, e201700044 (2018).

[33] Paraskevaidi, M., Martin-Hirsch, P. L. & Martin, F. L. Atr-ftir spectroscopy tools for medical diagnosis and disease investigation. In *Nanotechnology Characterization Tools for Biosensing and Medical Diagnosis*, 163–211 (Springer, 2018).

[34] Paraskevaidi, M. *et al.* Differential diagnosis of alzheimer's disease using spectrochemical analysis of blood. *Proceedings of the National Academy of Sciences* **114**, E7929–E7938 (2017).

[35] Movasaghi, Z., Rehman, S. & ur Rehman, D. I. Fourier transform infrared (ftir) spectroscopy of biological tissues. *Applied Spectroscopy Reviews* **43**, 134–179 (2008).

[36] Gazi, E. *et al.* Applications of fourier transform infrared microspectroscopy in studies of benign prostate and prostate cancer. a pilot study. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland* **201**, 99–108 (2003).

[37] Smith, B. R. *et al.* Combining random forest and 2d correlation analysis to identify serum spectral signatures for neuro-oncology. *Analyst* **141**, 3668–3678 (2016).

[38] Carmona, P., Molina, M., López-Tobar, E. & Toledano, A. Vibrational spectroscopic analysis of peripheral blood plasma of patients with alzheimer's disease. *Analytical and bioanalytical chemistry* **407**, 7747–7756 (2015).

Supplementary Information

A Demographics of L4L cohorts

Disease	Set	Size	Population (%)				Age		BMI	
			Healthy		Diseased					
			Male	Female	Male	Female	Avg	Std	Avg	Std
BLCA	Test	213	59.62	25.82	10.80	3.76	66.42	10.64	26.53	4.75
	Train	347	40.35	10.66	38.90	10.09	71.41	9.87	26.32	4.89
BRCA	Test	281	0.00	67.07	0.00	32.93	62.00	13.19	25.51	4.88
	Train	82	0.00	50.53	0.00	49.47	60.07	13.65	25.57	5.49
LUCA	Test	219	53.42	20.55	15.07	10.96	66.39	10.13	26.71	5.01
	Train	908	20.48	29.96	26.32	23.24	65.08	10.39	25.69	4.87
	Stage 1 Train	144	27.78	22.92	27.78	21.53	69.74	8.68	25.69	4.97
	Stage 1+2 Train	235	27.66	22.55	28.09	21.70	68.55	9.20	25.94	4.97
	Stage 2 Train	91	27.47	21.98	28.57	21.98	66.67	9.67	26.34	4.95
	Stage 3 Train	176	28.41	22.16	27.84	21.59	67.60	9.75	26.49	6.07
	Stage 4 Train	275	23.27	26.91	24.00	25.82	68.49	8.83	25.30	4.54
PRCA	Test	295	49.03	0.00	50.97	0.00	67.07	8.93	27.13	4.21
	Train	569	50.09	0.00	49.91	0.00	61.46	12.86	26.75	4.32

Table S1: Demographics across different diseases, stages, and train/test sets.

B Demographics of H4H dataset

Total Number of		
Samples	Individuals	Visits/Person
45497	9363	1-6

Demographics			
	Male—Female	Age	BMI
Range	3234—6129	39-88	14.5-91.3
Mean	-	53.1	27.3
STD	-	8.0	5.0

Visit Coverage	
Individuals with	
≥2 visits	8935
≥3 visits	8674
≥4 visits	8031
≥5 visits	6246
≥6 visits	3669

Visit Intervals	
Visit Interval	mean ± std
V1-V2	129.2±22.9
V2-V3	134.3±35.0
V3-V4	136.7±38.1
V4-V5	157.9±48.5
V5-V6	350.8±44.0

Table S2: H4H demographics and visit information summary.

C Demographics of LG dataset

Total Number of		
Samples	Individuals	Visits/Person
180	18	8-11

Demographics			
	Male—Female	Age	BMI
Range	7—11	20-59	18.2-38.8
Mean	-	40.5	23.6
STD	-	13.4	4.7

Visit Coverage	
Individuals with	
≥2 visits	18
≥3 visits	18
≥4 visits	18
≥5 visits	18
≥6 visits	18
≥7 visits	18
≥8 visits	18
≥9 visits	17
≥10 visits	14
≥11 visits	5

Visit Intervals [days]	
Visit Interval	mean ± std
V1-V2	3.4±1.7
V2-V3	2.4±0.8
V3-V4	6.0±3.8
V4-V5	3.3±1.0
V5-V6	4.8±2.0
V6-V7	4.1±2.1
V7-V8	5.2±4.7
V8-V9	4.6±3.6
V9-V10	4.0±1.4
V10-V11	5.4±2.2

Table S3: LG dataset demographics and cisit information summary

D Peak ratio indices

Index	Peak Ratio	Description
0	I_{1635}/I_{1654}	Ratio of β-sheet to α-helix secondary structures;[24–28]
1	I_{1546}/I_{1655}	Amide I to amide II ratio; secondary structure/fibrils[27,29–31]
2	$I_{1655}/(I_{1655}+I_{1548})$	α-helix to total proteins[30]
3	$I_{1684}/(I_{1655}+I_{1548})$	Antiparallel β-sheet to total proteins[30]
4	$I_{1515}/(I_{1655}+I_{1548})$	Tyrosine-rich proteins to total proteins[30]
5	I_{2959}/I_{2931}	$\nu_{as}(CH_3)/\nu_{as}(CH_2)$; lipid chain length[31]
6	$(I_{2855}+I_{2927})/(I_{2962}+I_{2871})$	Fatty-acid elongation[24,32]
7	$(I_{2851}+I_{2927})/(I_{1655}+I_{1548})$	Lipid-to-protein ratio[30]
8	$I_{1239}/(I_{2851}+I_{2927})$	Phospholipids to total lipids[30]
9	I_{1741}/I_{1640}	Lipid-to-protein ratio[33]
10	I_{1740}/I_{1400}	Lipid-to-protein ratio[29,33]
11	I_{2852}/I_{1400}	Lipid-to-protein ratio[29]
12	I_{1450}/I_{1539}	Lipid-to-protein ratio[34]
13	I_{1240}/I_{1517}	Tyrosine phosphorylation degree[30]
14	I_{1045}/I_{1545}	Phosphate-to-carbohydrate[34]
15	I_{1080}/I_{1550}	Phosphate-to-amide II[27,29,31]
16	I_{1060}/I_{1230}	$\nu_s(PO_2^-)/\nu_{as}(PO_2^-)$[34]
17	I_{1170}/I_{1080}	Relative nucleic-acid content[27]
18	I_{1030}/I_{1080}	Glycogen/phosphate; metabolic turnover[35–37]
19	I_{1080}/I_{1243}	$\nu_s(PO_2^-)/\nu_{as}(PO_2^-)$
20	$I_{1587}/(I_{1655}+I_{1548})$	Free amino acids to proteins[27]
21	I_{1156}/I_{1171}	Carbohydrate moieties in plasma globulins[38]
22	I_{1243}/I_{1314}	Protein/nucleic acid changes[31]
23	I_{1453}/I_{1400}	$\delta_{as}(CH_3)/\delta_s(CH_3)$[31]

Table S4: Peak ratios used to evaluate simulation quality (adapted from[5]).

E AUCs for all FTIR datasets

split	datatype	cohort	stage	testtype	metric	sex	auc	std auc
sex&disease	mvg	blca	-	cv	label	0&1	77.36	6.48
					sex	0&1	93.79	4.74
		brca	-	cv	label	0&1	86.22	6.55
					sex	0&1	-	-
		luca	-	cv	label	0&1	93.09	2.19
					sex	0&1	94.17	2.15
			1	cv	label	0&1	87.6	9.23
					sex	0&1	92.77	6.5
			1&2	cv	label	0&1	85.09	9.02
					sex	0&1	94.5	3.75
			2	cv	label	0&1	86.94	10.54
					sex	0&1	88.25	12.8
			3	cv	label	0&1	91.08	7.13
					sex	0&1	92.44	5.16
			4	cv	label	0&1	94.41	4.09
					sex	0&1	92.44	4.52
		prca	-	cv	label	0&1	83.21	5.78
					sex	0&1	-	-
		blca	-	test	label	0&1	61.43	-
					sex	0&1	78.84	-
		brca	-	test	label	0&1	50.17	-
					sex	0&1	-	-
		luca	-	test	label	0&1	78.14	-
					sex	0&1	90.11	-
		prca	-	test	label	0&1	60.27	-
					sex	0&1	-	-
	real	blca	-	cv	label	0&1	69.16	8.44
					sex	0&1	90.81	4.72
		brca	-	cv	label	0&1	76.49	7.45
					sex	0&1	-	-
		luca	-	cv	label	0&1	89.88	3.23
					sex	0&1	91.57	2.58
			1	cv	label	0&1	67.95	13.81
					sex	0&1	88.72	8.82
			1&2	cv	label	0&1	75.02	10.83
					sex	0&1	91.07	6.45
			2	cv	label	0&1	78.3	14.9
					sex	0&1	85.13	12.74
			3	cv	label	0&1	86.32	9.38
					sex	0&1	84.08	9.77
			4	cv	label	0&1	91.64	5.11
					sex	0&1	84.49	7.18
		prca	-	cv	label	0&1	74.45	6.07
					sex	0&1	-	-
		blca	-	test	label	0&1	61.5	-

Continued on next page

split	datatype	cohort	stage	testtype	metric	sex	auc	std auc
					sex	0&1	89.13	-
		brca	-	test	label	0&1	60.07	-
					sex	0&1	-	-
		luca	-	test	label	0&1	85.91	-
					sex	0&1	92.68	-
		prca	-	test	label	0&1	66.83	-
					sex	0&1	-	-
	mvg	blca	-	cv	label	0	80.21	7.18
		luca	-	cv	label	0	93.94	3.73
			1	cv	label	0	92.29	8.99
			1&2	cv	label	0	89.32	9.36
			2	cv	label	0	78.12	20.19
			3	cv	label	0	90.22	11.76
			4	cv	label	0	94.66	6.72
		prca	-	cv	label	0	83.92	4.4
		blca	-	test	label	0	63.06	-
		luca	-	test	label	0	76.17	-
		prca	-	test	label	0	60.27	-
	real	blca	-	cv	label	0	64.2	10.16
		luca	-	cv	label	0	89.19	4.92
			1	cv	label	0	66.48	19.98
			1&2	cv	label	0	72.92	12.19
			2	cv	label	0	68.47	21.93
			3	cv	label	0	84.06	10.4
			4	cv	label	0	92.11	6.32
		prca	-	cv	label	0	74.49	6.36
		blca	-	test	label	0	62.38	-
		luca	-	test	label	0	77.08	-
		prca	-	test	label	0	66.83	-
	mvg	blca	-	cv	label	1	90.25	11.79
		brca	-	cv	label	1	85.79	6.94
		luca	-	cv	label	1	91.85	3.24
			1	cv	label	1	77.56	20.74
			1&2	cv	label	1	82.6	11.67
			2	cv	label	1	96.03	10.01
			3	cv	label	1	88.35	14.56
			4	cv	label	1	92.08	7.56
		blca	-	test	label	1	47.05	-
		brca	-	test	label	1	50.17	-
		luca	-	test	label	1	93.7	-
	real	blca	-	cv	label	1	71.1	20.08
		brca	-	cv	label	1	77.08	8.01
		luca	-	cv	label	1	86.63	5.44
			1	cv	label	1	60.08	21.14
			1&2	cv	label	1	69.69	12.51
			2	cv	label	1	96.03	11.19
			3	cv	label	1	75.12	15.22
			4	cv	label	1	88.17	9.91

Continued on next page

split	datatype	cohort	stage	testtype	metric	sex	auc	std auc
		blca	-	test	label	1	49.09	-
		brca	-	test	label	1	60.07	-
		luca	-	test	label	1	90.83	-
disease	mvg	blca	-	cv	label	_	78.73	7.19
					sex	_	-	-
		brca	-	cv	label	_	87.17	6.58
					sex	_	-	-
		luca	-	cv	label	_	89.61	2.88
					sex	_	-	-
			1	cv	label	_	84.86	9.54
					sex	_	-	-
			1&2	cv	label	_	85.95	6.34
					sex	_	-	-
			2	cv	label	_	85.85	12.31
					sex	_	-	-
			3	cv	label	_	93.45	5.5
					sex	_	-	-
			4	cv	label	_	95.89	3.07
					sex	_	-	-
		prca	-	cv	label	_	83.91	5.28
					sex	_	-	-
		blca	-	test	label	_	60.14	-
					sex	_	-	-
		brca	-	test	label	_	54.48	-
					sex	_	-	-
		luca	-	test	label	_	81.61	-
					sex	_	-	-
		prca	-	test	label	_	58.97	-
					sex	_	-	-
	real	blca	-	cv	label	_	69.16	8.44
					sex	_	90.81	4.72
		brca	-	cv	label	_	76.49	7.45
					sex	_	-	-
		luca	-	cv	label	_	89.88	3.23
					sex	_	91.57	2.58
			1	cv	label	_	67.95	13.81
					sex	_	88.72	8.82
			1&2	cv	label	_	75.02	10.83
					sex	_	91.07	6.45
			2	cv	label	_	78.3	14.9
					sex	_	85.13	12.74
			3	cv	label	_	86.32	9.38
					sex	_	84.08	9.77
			4	cv	label	_	91.64	5.11
					sex	_	84.49	7.18
		prca	-	cv	label	_	74.45	6.07
					sex	_	-	-
		blca	-	test	label	_	61.5	-

Continued on next page

split	datatype	cohort	stage	testtype	metric	sex	auc	std auc
					sex	_	89.13	-
		brca	-	test	label	_	60.07	-
					sex	_	-	-
		luca	-	test	label	_	85.91	-
					sex	_	92.68	-
		prca	-	test	label	_	66.83	-
					sex	_	-	-

Table S5: AUCs for predicting the disease label and the sex on real and on simulated data on all cohorts.

F Formal definition of the index of individuality

Consider a longitudinal dataset of FTIR measurements $X \in \mathbb{R}^{I\times J\times K}$ with measurements $x_{ijk} \in X$, where i denotes the subject, j the visit and k the wavenumber. For simplicity we will only consider $J = \text{const.}$, however it is straightforward to generalize the definitions to differing numbers of visits per individual $J = J_i$.
For each subject i and wavenumber k, we define the subject-specific mean across visits as

$$\bar{x}_{ik} = \frac{1}{J}\sum_{j=1}^{J} x_{ijk}. \tag{6}$$

Following the standard variance-components formulation, the within-person variability (wpv) at wavenumber k is defined as the pooled within-subject variance:

$$\mathrm{wpv}_k^2 = \frac{1}{I(J-1)}\sum_{i=1}^{I}\sum_{j=1}^{J}\left(x_{ijk} - \bar{x}_{ik}\right)^2. \tag{7}$$

Next, we define the grand mean at wavenumber k across all subjects and visits as

$$\bar{x}_{..k} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} x_{ijk}. \tag{8}$$

The pooled variance at wavenumber k, reflecting the total variance across the entire cohort, is then given by

$$s^2_{\mathrm{pooled},k} = \frac{1}{IJ-1}\sum_{i=1}^{I}\sum_{j=1}^{J}\left(x_{ijk} - \bar{x}_{..k}\right)^2. \tag{9}$$

Finally, the index of individuality (ioi) at wavenumber k is defined as the ratio of within-person variability to pooled variance:

$$\mathrm{ioi}_k = \frac{\mathrm{wpv}_k}{s_{\mathrm{pooled},k}}. \tag{10}$$

The index of individuality therefore quantifies the relative contribution of intra-individual variability to the overall population variance at each spectral coordinate.