Analytical Methods

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d5ay02166a

Synthetic blood-based infrared molecular fingerprints: artificial cohorts for methodological research

Niklas Leopold-Kerschbaumer,*[ab] Nico Feiler[c] and Kosmas V. Kepesidis *[ac]

Infrared molecular fingerprinting of human blood samples provides a powerful, minimally invasive approach for disease detection and health monitoring. However, ethical and legal constraints often limit the sharing of real patient data collected from clinical studies. In this work, we present a synthetic dataset of blood-based infrared molecular fingerprints, generated using multivariate Gaussian models fitted on real measurements from a large case-control study targeting various cancer types. The synthetic dataset retains the statistical and physical properties of real molecular fingerprints, enabling the development and validation of analytical methodologies without compromising patient privacy. We demonstrate that the provided artificial dataset can 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.

Received 30th December 2025
Accepted 26th May 2026

DOI: 10.1039/d5ay02166a

rsc.li/methods

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.[5–17] 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. Synthetic datasets are constructed using multivariate Gaussian (MVG) models fit on real-world measurements from a cross-sectional study investigating various cancer types. By capturing and reproducing key spectral patterns in an artificial yet statistically valid manner, the synthetic cohorts provide a valuable resource for researchers aiming to develop and refine analytical techniques without direct access to sensitive patient data.

This paper presents the methodology used to generate the synthetic dataset, evaluates its fidelity to real-world spectral fingerprints, and discusses its potential applications in biomedical research. Importantly, the resulting synthetic cohorts are made publicly available through a dedicated open-access repository, enabling immediate and unrestricted use by the research community.[18] This approach facilitates ethical data sharing, reproducibility, and innovation in the field of biomedical spectroscopy. Our findings support the broader adoption of synthetic datasets as a viable solution for overcoming ethical barriers while advancing the study of blood-based molecular diagnostics.

Methods

Infrared fingerprinting of human blood plasma

Blood samples utilized in this work were collected in the framework of the multi-center *Lasers4Life* clinical study. This study was conducted in the Munich area and is registered (ID DRKS00013217) at the German Clinical Trials Register (DRKS). The study was reviewed and approved by "Ethikkommission bei der LMU München" (EK 20170820—Nr.: 17-532), and was conducted according to Good Clinical Practice (ICH-GCP), the principles of the Declaration of Helsinki, and all applicable legislations and regulations. Informed consent was obtained from all participants before blood collection. Blood plasma samples of study participants were obtained following previously established sample handling procedures.[5] Before measuring the samples using a commercially available FTIR device (MIRA-Analyzer, CLADE GmbH), 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

[a]Center for Molecular Fingerprinting (CMF), Frontiers Foundation, Budapest, Hungary. E-mail: niklas.leopoldkerschbaumer@cmf.hu; kosmas.kepesidis@cmf.hu
[b]Department of Mathematics, Ludwig-Maximilians-Universität München (LMU), Munich, Germany
[c]Faculty of Physics, Ludwig-Maximilians-Universität München (LMU), Garching, Germany

potential drifts in spectrometer performance and ensure robust testing, the test set was measured in randomized order over 2 weeks. The infrared molecular fingerprints collected span the spectral range of 930–3051 cm^{-1}, capturing characteristic absorption bands for proteins, carbohydrates, and lipids. Once all measurements were obtained, outliers were detected using the Local Outlier Factor method from the scikit-learn library in Python 1.6.1, separately for the training and the test set. This ensured that erroneous or incomplete measurements were removed before analysis.

Case-control designs

The health status of individuals in the study is categorized into either therapy-naïve patients with lung, prostate, bladder, or breast cancer or non-symptomatic reference individuals. Based on these classifications and available demographic information, case-control designs were constructed to ensure meaningful comparisons. To reduce potential confounding effects, statistical matching was performed using propensity scores, aligning cases with their closest control counterparts based on age and sex. Specifically, we performed optimal pair matching using the Mahalanobis distance within propensity score calipers,[19] implemented in R version 4.4.1. Following this procedure, the resulting matched cohorts comprise a total of 2079 individuals, with 1650 assigned to training sets and 429 to test sets. Detailed information on cohort demographics and sample sizes is provided in Table S1 in SI A. Using these well-balanced cohorts, we fitted statistical models capable of generating new synthetic infrared spectra that capture the key statistical and spectral properties of the original datasets.

Multivariate Gaussian modeling

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

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

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 the next 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.[20–24] In addition to blood plasma spectra, we fitted MVG models to spectra from quality-control serum samples consisting of pooled human serum.[6] The dataset included 1048 quality-control samples in the training set and 131 in the test set.

Synthetic data records

Using our approach, we generated synthetic spectra for matched cohorts of healthy and diseased individuals across four 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 an 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 Fig. 1A, while detailed data splits for each (sub-)cohort are presented in Fig. 1B. Every data split highlighted in orange in Fig. 1B corresponds to a fitted MVG model—characterized by a mean spectrum and covariance matrix—available in our GitHub repository.[18]

These MVG models, defined by eqn (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.

In the following section, we validate this simulation approach with use cases motivated by real-world applications such as disease diagnostics.

Technical validation

To validate our simulation approach, we simulated cohorts of equal size to the real data. We 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 Fig. 1), are summarized in Fig. 2.

Fig. 2a1 shows 10 representative spectra alongside a PCA scatter plot, illustrating the visual similarity between real and simulated data. As seen in Fig. 2a2, the effect size between real and simulated datasets is close to zero across all wavenumbers. Similarly, Fig. 2a3 compares the means per wavenumber, which shows that there is no significant difference between real and simulated data. Fig. 2a4 and a5 demonstrate that effect sizes with respect to disease label and sex in the simulated data closely mirror those in the real data.

Similarly, Fig. 2b1 shows that the distribution of peak ratios in the simulated data closely matches that of the real data. Fig. 2b2–b5 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 S2 in SI B. Effect sizes for all MVG cohorts are provided in Fig. S1, S2 in SI C and D.

In terms of classification performance, the ROC curves in Fig. 2C and D 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

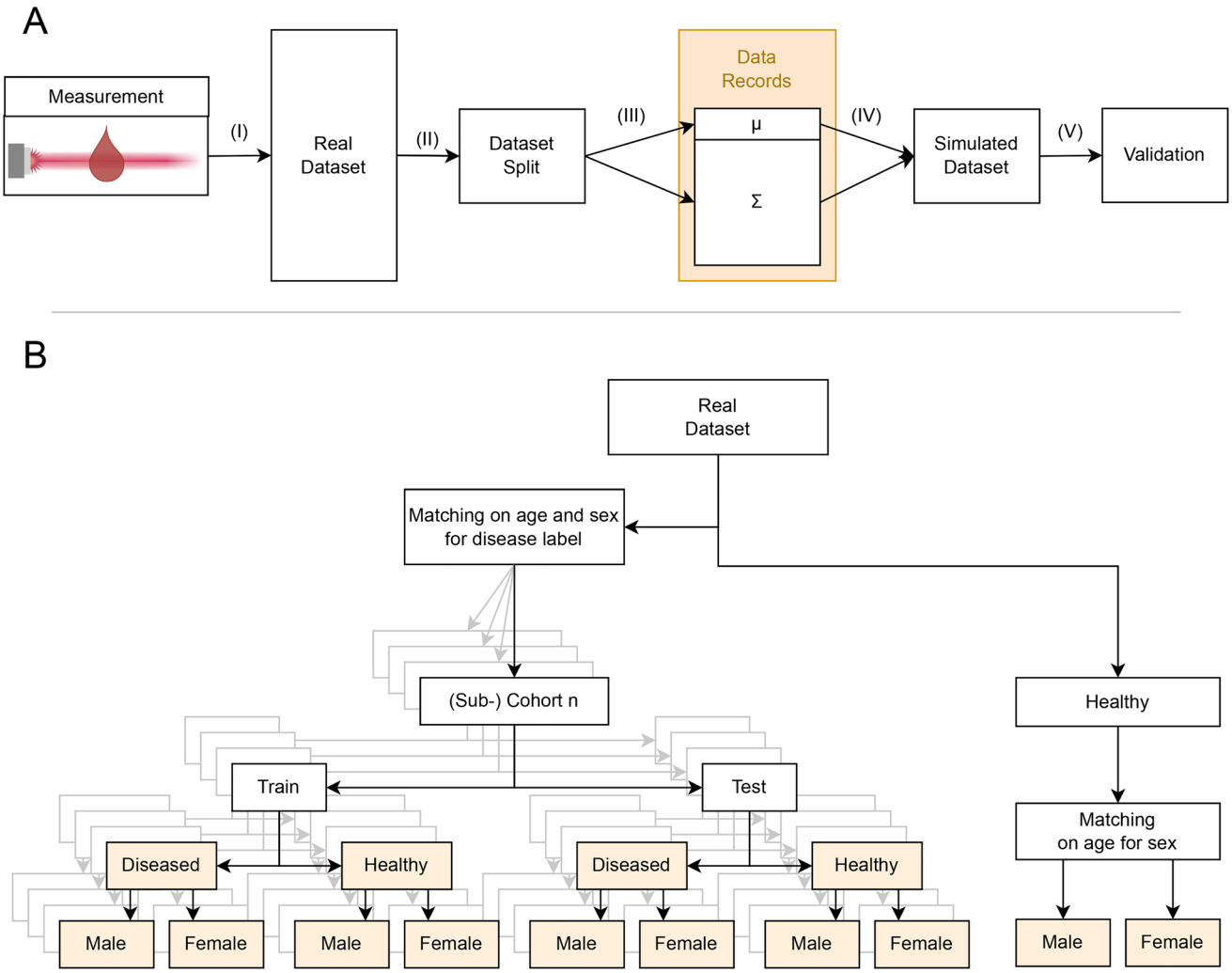

Fig. 1 (A) Diagram illustrating how the simulated datasets are obtained. (I) Blood plasma samples of study participants are measured *via* FTIR, and measurements are stored in a database. (II) The real dataset is split into subsets based on disease status and/or sex according to figure. (III) The mean and covariance matrix are calculated for each subset. (IV) Simulated spectra are sampled according to eqn (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.

training set, indicating a domain shift. A complete summary of AUC scores across all dataset splits using logistic regression is provided in Table S3 in SI E.

All data underwent the following preprocessing steps, consistent with established methods:[6,12]

(1) Truncation I: spectra were truncated to the range of 1000–3000 cm^{-1} to standardize measurements, excluding regions without peaks.

(2) L2 normalization: each spectrum was normalized by its L2 norm:

$$x^{\text{norm}} = \frac{x}{\|x\|} \quad \text{with} \quad \|x\| = \sqrt{\sum_{k=1}^{N} {x_k}^2} \tag{2}$$

(3) Truncation II: absorbance values and wavenumbers between 1800–2800 cm^{-1} were removed. This "silent region" is dominated by water absorbance and contains no molecular information.

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.

Usage notes

Applications

The synthetic dataset introduced here is designed to support methodological research in blood-based infrared molecular

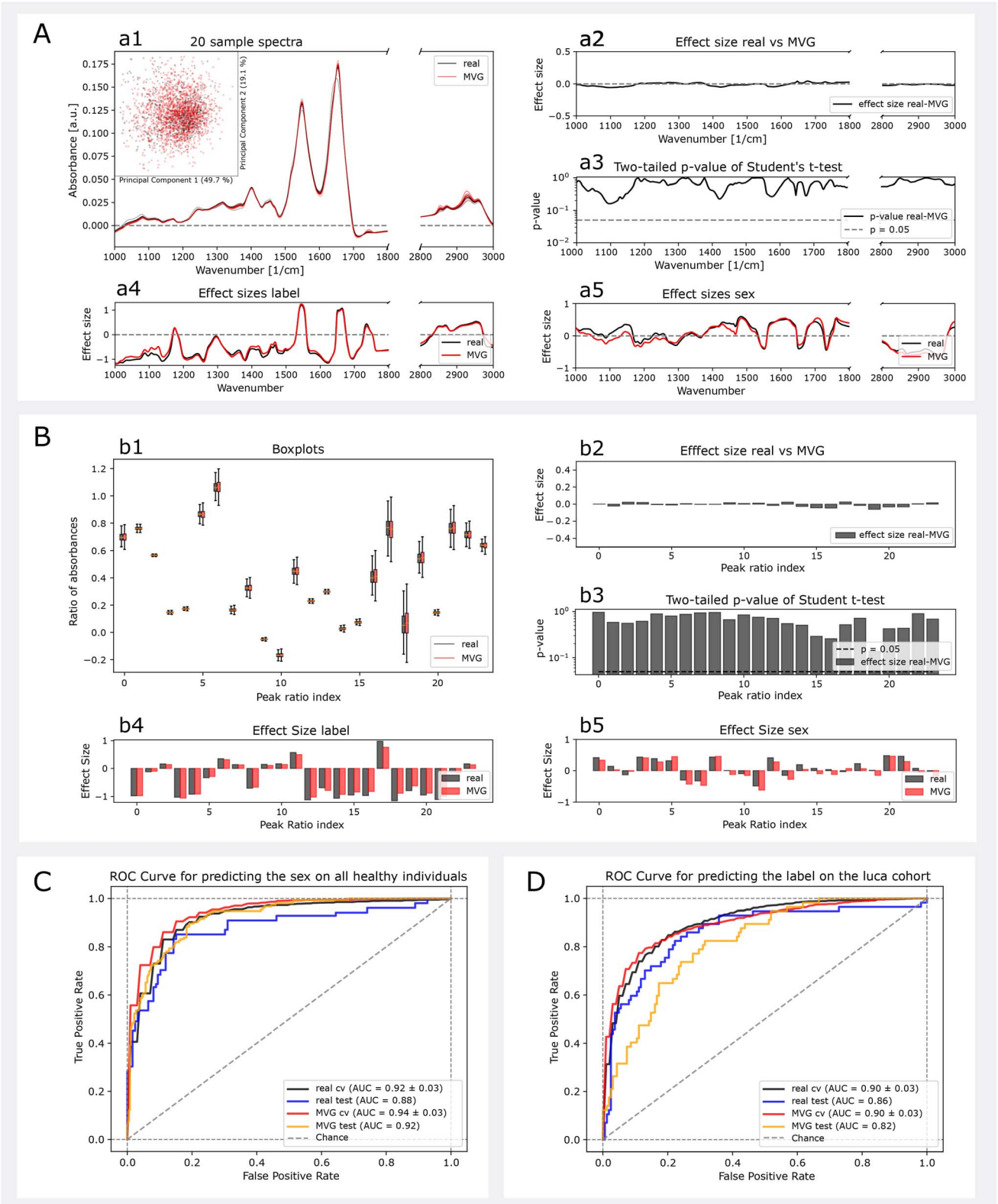

Fig. 2 Comparison between real data and simulated data. (A) Basic analysis on the spectra, including (1) 10 spectra of real data and 10 spectra of simulated data with the outlier removed, PCA scatter plot of all spectra, (2) effect size between real and simulated data, (3) *t*-test *p*-values between real and simulated data, (4) effect size between the lung cancer label for real and simulated data, (5) 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 between real and simulated data, *t*-test *p*-values between real and simulated data, (4) effect size between the lung cancer label for real and simulated data, (5) 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.

fingerprinting under controlled and reproducible conditions. In particular, it enables the development, benchmarking, and validation of statistical, machine-learning, and signal-processing methods for FTIR spectra, without requiring access to sensitive clinical data.

One class of applications is the evaluation of new inference and measurement methodologies. For example, Schröder *et al.*[24] used an early version of the dataset to develop and validate Bayesian autocorrelation spectroscopy, with the MVG models serving as priors for reconstructing infrared spectra from noisy measurements. This controlled setting enabled systematic assessment of reconstruction accuracy, uncertainty quantification, and sampling efficiency. More broadly, statistically well-characterized synthetic spectra provide a reproducible basis for benchmarking analytical pipelines, particularly when real-world datasets are inaccessible, restricted, or limited in size.

Another class of applications is enabled by the explicit availability of the underlying MVG parameters. Because both the mean vectors and covariance matrices are provided, users are not limited to sampling fixed synthetic cohorts; rather, they can directly manipulate the statistical structure of the data. This enables systematic analyses of how changes in variability affect downstream results. For example, increased or reduced variability may reflect higher or lower measurement noise, respectively, with reduced noise potentially arising from improved instrumentation and a higher signal-to-noise ratio. More generally, tuning variability provides a way to probe different medical settings, including precision diagnostics. In such settings, one could assess how disease-classification performance benefits from the reduced biological variability expected in more precisely defined patient populations. In addition, informed manipulation of the covariance matrix has been shown to support domain adaptation approaches.[22]

Limitations

While simulations based on MVG 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.

Prospects

Although the synthetic datasets presented in this work are derived from cancer-specific cohorts, the underlying methodology is not inherently limited to oncological applications. In principle, the MVG modeling framework can be applied to any condition for which appropriately designed and statistically matched cohorts of infrared molecular fingerprints are available, including metabolic and cardiovascular diseases. The key requirement is that the measured spectra exhibit sufficiently stable statistical structure for their distributions to be reasonably approximated by an MVG model. This assumption should be empirically validated for each new application, for example, by assessing distributional properties, covariance structure, and downstream task performance, such as effect sizes or classification metrics, in comparison with real data.

Author contributions

Niklas Leopold-Kerschbaumer: investigation; formal analysis; visualization; methodology; writing – original draft preparation; writing – review & editing. Nico Feiler: investigation; formal analysis; visualization; methodology; writing – original draft preparation; writing – review & editing. Kosmas V. Kepesidis: conceptualization; supervision; project administration; methodology; writing – original draft preparation; writing – review & editing.

Conflicts of interest

The authors declare no competing interests.

Code availability

The GitHub repository[18] also includes an example Python script demonstrating how to use these models to generate synthetic spectra.

Data availability

All mean spectra and covariance matrices described in this work are available in our GitHub repository.[18] If promising results are achieved using the simulated data, researchers are encouraged to contact the corresponding authors to explore validation opportunities on real datasets.

Supplementary information (SI) is available. See DOI: https://doi.org/10.1039/d5ay02166a.

Acknowledgements

This work was supported by the Centre for Advanced Laser Applications (CALA) at LMU Munich, the Center for Molecular Fingerprinting Research Nonprofit LLC (CMF), and the Frontiers Foundation. 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. 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.

References

1 M. J. Baker, *et al.*, Using fourier transform ir spectroscopy to analyze biological materials, *Nat. Protoc.*, 2014, **9**, 1771–1791.
2 H. J. Butler, *et al.*, Development of high-throughput atr-ftir technology for rapid triage of brain cancer, *Nat. Commun.*, 2019, **10**, 4501.
3 L. Voronina, *et al.*, Molecular origin of blood-based infrared spectroscopic fingerprints, *Angew. Chem.*, 2021, **133**, 17197–17206.
4 M. Paraskevaidi, *et al.*, Clinical applications of infrared and raman spectroscopy in the fields of cancer and infectious diseases, *Appl. Spectrosc. Rev.*, 2021, **56**, 804–868.
5 M. Huber, *et al.*, Stability of person-specific blood-based infrared molecular fingerprints opens up prospects for health monitoring, *Nat. Commun.*, 2021, **12**, 1511.
6 M. Huber, *et al.*, Infrared molecular fingerprinting of blood-based liquid biopsies for the detection of cancer, *eLife*, 2021, **10**, e68758.
7 F. L. Martin, *et al.*, Distinguishing cell types or populations based on the computational analysis of their infrared spectra, *Nat. Protoc.*, 2010, **5**, 1748–1760.
8 H. Ghimire, *et al.*, Protein conformational changes in breast cancer sera using infrared spectroscopic analysis, *Cancers*, 2020, **12**, 1708.
9 J. Ollesch, *et al.*, An infrared spectroscopic blood test for non-small cell lung carcinoma and subtyping into pulmonary squamous cell carcinoma or adenocarcinoma, *Biomed. Spectrosc. Imag.*, 2016, **5**, 129–144.
10 T. Eissa, *et al.*, Plasma infrared fingerprinting with machine learning enables single-measurement multi-phenotype health screening, *Cell Rep. Med.*, 2024, **5**, 7.
11 K. V. Kepesidis, *et al.*, Assessing lung cancer progression and survival with infrared spectroscopy of blood serum, *BMC Med.*, 2025, **23**, 101.
12 K. V. Kepesidis, *et al.*, Breast-cancer detection using blood-based infrared molecular fingerprints, *BMC Cancer*, 2021, **21**, 1–9.
13 J. Backhaus, *et al.*, Diagnosis of breast cancer with infrared spectroscopy from serum samples, *Vib. Spectrosc.*, 2010, **52**, 173–177.
14 F. Elmi, A. F. Movaghar, M. M. Elmi, H. Alinezhad and N. Nikbakhsh, Application of ft-ir spectroscopy on breast cancer serum analysis, *Spectrochim. Acta, Part A*, 2017, **187**, 87–91.
15 J. Ollesch, *et al.*, It's in your blood: spectral biomarker candidates for urinary bladder cancer from automated ftir spectroscopy, *J. Biophot.*, 2014, **7**, 210–221.
16 D. Anderson, R. Anderson, S. Moug and M. Baker, Liquid biopsy for cancer diagnosis using vibrational spectroscopy: systematic review, *BJS Open*, 2020, **4**, 554–562.
17 M. Zigman, *et al.*, 90p infrared molecular fingerprinting: A new in vitro diagnostic platform technology for cancer detection in blood-based liquid biopsies, *Ann. Oncol.*, 2022, **33**, S580.
18 N. Leopold-Kerschbaumer, MVG-IMF, 2025, https://github.com/Attoworld-Data-Science/MVG-IMF.
19 P. R. Rosenbaum, *Design of Observational Studies*, Springer, 2010, vol. 10.
20 C. Beleites, U. Neugebauer, T. Bocklitz, C. Krafft and J. Popp, Sample size planning for classification models, *Anal. Chim. Acta*, 2013, **760**, 25–33.
21 T. Eissa, K. V. Kepesidis, M. Zigman and M. Huber, Limits and prospects of molecular fingerprinting for phenotyping biological systems revealed through in silico modeling, *Anal. Chem.*, 2023, **95**, 6523–6532.
22 F. B. Nemeth, *et al.*, Bridging spectral gaps: Cross-device model generalization in blood-based infrared spectroscopy, *Anal. Chem.*, 2025, **97**(19), 10264–10272.
23 C. Wegner, *et al.*, Toward informative representations of blood-based infrared spectra via unsupervised deep learning, *J. Biophot.*, 2025, e70011.
24 J. Schroeder, *et al.*, Information-optimal measurement: From fixed sampling protocols to adaptive spectroscopy, *arXiv*, 2025, preprint, arXiv:2505.14364, DOI: 10.48550/arXiv.2505.14364.