

Infrared Molecular Fingerprinting to Probe Human Health

Tarek Abdelmaksoud Abdalla Eissa

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz:

Prof. Dr. Julien Gagneur

Prüfende der Dissertation:

1. Prof. Dr. Burkhard Rost
2. Prof. Dr. Ferenc Krausz

Die Dissertation wurde am 18.02.2025 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 20.05.2025 angenommen.

Abstract

Human blood plasma is a rich molecular reservoir, containing diverse proteins, lipids, nucleic acids, and metabolites that reflect the body’s health state. This biochemical complexity makes plasma an ideal target for molecular profiling, offering the possibility to develop minimally invasive diagnostics tools targeting a variety of systemic physiological conditions. Fourier transform infrared (FTIR) absorption spectroscopy is a particularly promising analytical tool due to its capacity to capture molecularly diverse information in a fast, cost-effective, and label-free measurement. By analyzing the vibrational signatures across different chemical bonds within a complex sample when interacting with infrared (IR) light, FTIR spectroscopy reveals a so-called ‘IR molecular fingerprint’ – a single snapshot measurement that characterizes the sample’s overall biochemical makeup in its native liquid state. When combined with machine learning (ML), IR fingerprinting holds excellent promise as a high-throughput blood-based *in vitro* diagnostic tool. Although successful results have been reported for various health-related applications, the approach has yet to be adopted in clinical diagnostic practice. This publication-based thesis comprises four complementary studies that collectively address key challenges and opportunities for establishing blood-based IR fingerprinting as a robust platform for clinical diagnostics.

A critical barrier to adoption has been the lack of large-scale studies validating the approach in clinically relevant contexts. The first article in this thesis addresses this gap by evaluating the feasibility of IR fingerprinting within a diverse, population-based cohort, involving the analysis of 5184 plasma samples from 3169 individuals. Developing a multi-task classification framework, we demonstrate that a single IR fingerprint can reliably identify multiple co-occurring physiological states. Moreover, we show that IR fingerprinting has the capacity to predict the onset of metabolic disorders years before their clinical manifestation, providing actionable medical insights and facilitating timely pharmacological/lifestyle interventions. This article underscores the potential of IR fingerprinting as a single-measurement, multi-phenotype screening platform, marking a significant advancement in the field.

Another challenge facing the IR fingerprinting domain is the widespread misinterpretation of spectral data when analyzing complex samples like blood plasma, where each spectral feature stems from the superposition of signals from diverse organic molecules. The second article in this thesis addresses the complexities of analyzing IR spectral data. Supported by new and available experimental evidence, we provide methodological and best-practice guidelines to improve data analysis practices. Rather than overemphasizing the interpretation of single spectral peaks and features, we highlight the value of ML and multivariate analysis to decode the complex molecular information in IR fingerprints. Overall, this perspective article aims to improve the broader understanding of the limitations and the feasible, fruitful applications of IR fingerprinting.

Abstract

In the third article, we develop and employ an *in silico* model to systematically explore the factors affecting the performance of IR molecular fingerprinting in clinical applications. These factors include the technical measurement noise, biological variability, chemical complexity of the studied sample, and cohort sizes used for model training – all of which critically impact the predictive accuracy in clinical settings. Our developed model enables the simulation of realistic IR spectra that model diverse experimental conditions, offering a tool to investigate the potential and limitations of IR fingerprinting under controlled conditions. This tool can guide experimental efforts and resource allocation, ultimately aiming to improve the diagnostic precision and clinical utility of IR fingerprinting.

The final article in this thesis tackles a key challenge in combining molecular profiling technologies with ML: out-of-distribution (OOD) measurements. Traditional supervised ML algorithms assume that the train and test data follow similar distributions. In practice, it is unlikely that any given training set is fully representative of the true data domain, and OOD data are inevitable. OOD data can arise from biological factors (e.g., varying age or ethnic groups) or (pre)analytical factors (e.g., changes in sample handling procedures or measurement devices). Such distributional shifts compromise a model’s capacity to generalize to measurement events encountered in practical applications. To address this, we develop a new method: Contextual Out-of-Distribution Integration (CODI). CODI is an approach that generates synthetic data which integrate unrepresented sources of variation into a given molecular fingerprint dataset. By augmenting a dataset with OOD variance, CODI enables an ML model to better generalize to samples beyond the original training data. We evaluate CODI across multiple IR fingerprinting applications, involving three longitudinal clinical studies and a case-control study for detecting four common cancers (lung, prostate, bladder, and breast). We practically demonstrate CODI as a tool to improve model generalization and predictive accuracy, marking a significant step toward reliable ML diagnostics in practical real-world conditions.

This thesis presents a roadmap – from practical demonstrations of IR-based multi-phenotyping to algorithmic developments – that advances the basic research foundation for using IR fingerprinting across diverse clinical diagnostic settings.

Zusammenfassung

Das menschliche Blutplasma ist ein reichhaltiges molekulares Reservoir, das verschiedene Proteine, Lipide, Nukleinsäuren und Metaboliten enthält, welche den Gesundheitszustand des Körpers widerspiegeln. Diese biochemische Komplexität macht das Plasma zu einem idealen Ziel zur Erstellung von Molekülprofilen. Zudem bietet es die Möglichkeit, minimal-invasive Diagnostikverfahren zu entwickeln, die auf eine Vielzahl verschiedener physiologischer Zustände abzielen. Die Fourier-Transformations-Infrarot-Absorptionsspektroskopie (FTIR) ist ein besonders vielversprechendes Analysewerkzeug, da sie in der Lage ist, vielfältige molekulare Informationen in einer schnellen, kostengünstigen und markierungsfreien Messung zu erfassen. Durch die Analyse der Schwingungssignaturen verschiedener chemischer Bindungen in einer komplexen Probe bei der Wechselwirkung mit einer Infrarot (IR)-Lichtquelle zeigt die FTIR-Spektroskopie einen sogenannten „molekularen IR-Fingerprint“ – eine einzelne Momentaufnahme, welche die gesamte biochemische Zusammensetzung der Probe in ihrem nativen, flüssigen Zustand charakterisiert. In Verbindung mit maschinellem Lernen (ML) ist das IR-Fingerprinting ein vielversprechendes Verfahren für die blutbasierte In-vitro-Diagnostik im Hochdurchsatz. Obwohl erfolgreiche Ergebnisse für verschiedene gesundheitsbezogene Anwendungen berichtet wurden, wird dieser Ansatz noch nicht in der Diagnostik praktiziert. Diese publikationsbasierte Dissertation umfasst vier komplementäre Studien, die sich gemeinsam mit den wichtigsten Herausforderungen und Möglichkeiten für die Etablierung des blutbasierten IR-Fingerprintings als robuste Plattform für die klinische Diagnostik befassen.

Ein kritisches Hindernis für die Einführung war das Fehlen von groß angelegten Studien, die den Ansatz in klinisch relevanten Kontexten validieren. Der erste Artikel dieser Dissertation adressiert diese Lücke, indem die Machbarkeit des IR-Fingerprintings innerhalb einer diversen, bevölkerungsbasierten Kohorte bewertet wird, welche die Analyse von 5184 Plasma-Proben von 3169 Individuen umfasst. Durch die Entwicklung eines Multi-Task-Klassifikationssystems demonstrieren wir, dass ein einzelner IR-Fingerprint mehrere gleichzeitig auftretende physiologische Zustände zuverlässig identifizieren kann. Darüber hinaus zeigen wir, dass IR-Fingerprinting die Fähigkeit besitzt, den Beginn von Stoffwechselstörungen Jahre vor deren klinischer Manifestation vorherzusagen, was zu verwertbaren medizinischen Erkenntnissen führt und rechtzeitige pharmakologische sowie lebensstilbezogene Interventionen erleichtert. Dieser Artikel unterstreicht das Potenzial von IR-Fingerprinting als Einzelmessungs-, Multi-Phänotyp-Screening-Plattform und stellt einen bedeutenden Fortschritt in diesem Bereich dar.

Eine weitere Herausforderung für den Bereich des IR-Fingerprintings ist die weit verbreitete Fehlinterpretation von Spektraldaten bei der Analyse komplexer Proben wie Blutplasma, bei denen jedes spektrale Merkmal aus der Überlagerung von Signalen verschiedener organischer Moleküle entsteht. Der zweite Artikel dieser Disser-

Zusammenfassung

tation befasst sich mit der Komplexität der Analyse von IR-Spektraldaten. Gestützt auf neue und verfügbare experimentelle Belege bieten wir methodologische und Best-Practice-Richtlinien zur Verbesserung der Datenauswertung. Anstatt die Interpretation einzelner spektraler Peaks und Merkmale überzubetonen, heben wir den Wert von maschinellem Lernen (ML) und multivariater Analyse hervor, um die komplexen molekularen Informationen in IR-Fingerprints zu entschlüsseln. Insgesamt zielt dieser Perspektivartikel darauf ab, das breitere Verständnis der Einschränkungen sowie der machbaren und fruchtbaren Anwendungen von IR-Fingerprinting zu verbessern.

Im dritten Artikel entwickeln und nutzen wir ein *in silico*-Modell, um systematisch die Faktoren zu untersuchen, welche die Leistungsfähigkeit des IR-Fingerprintings in klinischen Anwendungen beeinflussen. Zu diesen Faktoren gehören das technische Messrauschen, die biologische Variabilität, die chemische Komplexität der untersuchten Probe sowie die für das Modelltraining verwendeten Kohortengrößen – allesamt Faktoren, welche die prädiktive Genauigkeit in klinischen Kontext entscheidend beeinflussen. Unser entwickeltes Modell ermöglicht die Simulation realistischer IR-Spektren, die unterschiedliche experimentelle Bedingungen abbilden, und bietet ein Werkzeug zur Untersuchung des Potenzials und der Einschränkungen des IR-Fingerprintings unter kontrollierten Bedingungen. Dieses Werkzeug kann experimentelle Bemühungen und die Ressourcenallokation lenken, um letztendlich die diagnostische Präzision und den klinischen Nutzen des IR-Fingerprintings zu verbessern.

Der letzte Artikel dieser Dissertation behandelt eine zentrale Herausforderung bei der Kombination von molekularen Profilierungstechnologien mit maschinellem Lernen (ML): Out-of-Distribution (OOD) Messungen. Traditionelle überwachte ML-Algorithmen setzen voraus, dass die Trainings- und Testdaten ähnlichen Verteilungen folgen. In der Praxis ist es unwahrscheinlich, dass ein gegebenes Trainingsset die wahre Daten-Domäne vollständig repräsentiert, und OOD-Daten sind unvermeidlich. OOD-Daten können durch biologische Faktoren (z.B. unterschiedliche Alters- oder ethnische Gruppen) oder (prä)analytische Faktoren (z.B. Änderungen in der Probenverarbeitung oder an Messgeräten) entstehen. Solche Verteilungsunterschiede beeinträchtigen die Fähigkeit eines Modells, auf Messereignisse, die in praktischen Anwendungen auftreten, zu generalisieren. Um dem entgegenzuwirken, entwickeln wir eine neue Methode: Contextual Out-of-Distribution Integration (CODI). CODI ist ein Ansatz, der synthetische Daten erzeugt, welche nicht repräsentierte Variationssquellen in einen gegebenen molekularen Fingerprint-Datensatz integrieren. Durch die Erweiterung eines Datensatzes mit OOD-Varianz ermöglicht CODI einem ML-Modell, besser auf Proben zu generalisieren, die über die ursprünglichen Trainingsdaten hinausgehen. Wir evaluieren CODI in mehreren IR-Fingerprinting-Anwendungen, darunter drei longitudinale klinische Studien und eine Fall-Kontroll-Studie zur Erkennung von vier häufigen Krebsarten (Lunge, Prostata, Blase und Brust). In mehreren Anwendungen zeigen wir, dass CODI die Modellgeneralisierung und die Vorhersagegenauigkeit verbessert und somit einen bedeutenden Schritt hin zu einer zuverlässigen ML-Diagnostik unter realen Bedingungen darstellt.

Diese Dissertation stellt eine Roadmap vor – von praktischen Demonstrationen der IR-basierten Multi-Phänotypisierung bis hin zu algorithmischen Entwicklungen – welche die grundlegende Forschungsbasis für die Nutzung des IR-Fingerprintings in vielfältigen klinischen diagnostischen Bereichen vorantreibt.

Acknowledgments

My journey from Cairo to Munich – beginning with my M.Sc. to the completion of this Ph.D. dissertation – has been one of immense academic and personal growth. I have been fortunate to receive support and encouragement from many incredible individuals, to whom I am very grateful.

First and foremost, I wish to thank **Dr. Mihaela Žigman**, my mentor and direct supervisor. Mihaela, you created a lab with an outstanding, collaborative atmosphere. Your support in both the academic pursuits and personal growth of your team members, and your guidance whenever anyone needs help staying on track, are constant sources of motivation. From early on, you trusted me to take responsibility, shape and direct decisions – even in aspects beyond the scope of my project. This level of involvement is rare for a doctoral student, and had a lasting impact on my development. The way you lead – with both expertise and genuine care – has set an example I will always carry with me. I have been incredibly fortunate to be a part of your lab.

I am also grateful to **Prof. Dr. Ferenc Krausz**. Your vision created a well-resourced research environment that laid a strong foundation for this work and so many others. Even while managing several institutes and countless responsibilities, you always made time to engage, discuss my research, and provide valuable feedback. I am especially grateful for your forward-thinking approach – not just in science but in supporting the career growth of those around you.

My sincere thanks go to **Prof. Dr. Burkhard Rost**, my formal supervisor. Your willingness to supervise this Ph.D. was not for personal gain but out of a genuine desire to support and push others toward their goals. Our conversations, whether about research, career decisions, or broader perspectives, always left me with greater clarity and confidence. Just knowing that there was always someone to turn to – for advice, encouragement, or simply a reality check – made a profound difference. I also want to express my appreciation to **Dr. Lothar Richter** – you had a big role in the success of my M.Sc. thesis years ago, which pushed me to continue and transition into this Ph.D.

I have been very lucky to work alongside colleagues who showed me how exciting interdisciplinary research can be. **Dr. Marinus Huber**, for choosing to engage with me early on in the Ph.D., then working together throughout it, where you offered broad insights that extended well beyond your expertise in physics. Your active role in discussions and ability to keep the bigger picture and final outcome in mind was invaluable throughout our work together. I learned a great deal from you. **Dr. Frank Fleischmann**, you co-led our research group and took on much of the day-to-day organization, allowing many (post-)doctoral students to focus on their research. Your readiness to help – in discussions, experimental designs, biobanking, or countless other aspects – went above and beyond. Your support made a real difference, not just to me but to everyone around you. **Dr. Liudmila (Lucy) Voronina**, we had

Acknowledgments

very many engaging conversations, and I always learned something new each time we talked. Thank you for your thoughtful feedback, support, and collaborative spirit, which made working together a pleasure. **Dr. Kosmas V. Kepesidis**, thank you for our many discussions and your expert assistance with the data analysis. **Dr. Philip Jacob**, thank you for your keen eye for detail and constructive feedback on various aspects of my work. **Jacqueline Hermann, Carola Spindler, and Sabine Witzens**, thank you for your expertise in managing and carefully carrying out the clinical studies. Likewise, I appreciate **Daniel Meyer** and **Viola Zoka** for their incredible efforts in handling the many thousands of sample measurements. I wish you all the best in your current and future endeavors.

Finally, I owe a debt of gratitude to my family. To my mother, **Madiha**, and late father, **Abdelmaksoud**. Thank you for always pushing me to aim higher and instilling a deep belief in the power of education. To my siblings – **Noura, Eiman, Yehia**, and **Salwa** – and all of your children, I am proud to be part of a family whose talents and achievements span the arts and the sciences. Your success in your respective fields continues to inspire me every day.

List of included publications

The following core, lead-authored, and peer-reviewed works, form the basis of this publication-based dissertation. The publications are listed in their order of appearance.

1. **Eissa, T.**, Leonardo, C., Kepesidis, K. V., Fleischmann, F., Linkohr, B., Meyer, D., Zoka, V., Huber, M., Voronina, L., Richter, L., Peters, A., Žigman, M., “Plasma infrared fingerprinting with machine learning enables single-measurement multi-phenotype health screening,” *Cell Reports Medicine*, Volume 5, Issue 7, July 2024, 101625, DOI: 10.1016/j.xcrm.2024.101625.

Type: Research Article

Highlight: Featured on the cover

Included as: Appendix A

2. **Eissa, T.**, Voronina, L., Huber, M., Fleischmann, F., Žigman, M., “The perils of molecular interpretations from vibrational spectra of complex samples,” *Angewandte Chemie Int. Ed.*, Volume 63, Issue 50, November 2024, e202411596, DOI: 10.1002/anie.202411596.

Type: Scientific Perspective

Highlight: Also available in German (10.1002/ange.202411596)

Included as: Appendix B

3. **Eissa, T.**, Kepesidis, K. V., Žigman, M., Huber, M., “Limits and prospects of molecular fingerprinting for phenotyping biological systems revealed through *in silico* modeling,” *Analytical Chemistry*, Volume 95, Issue 16, April 2023, 6523-6532, DOI: 10.1021/acs.analchem.2c04711.

Type: Research Article

Included as: Appendix C

4. **Eissa, T.**, Huber, M., Obermayer-Pietsch, B., Linkohr, B., Peters, A., Fleischmann, F., Žigman, M., “CODI: Enhancing machine learning-based molecular profiling through contextual out-of-distribution integration,” *PNAS Nexus*, Volume 3, Issue 10, October 2024, pgae449, DOI: 10.1093/pnasnexus/pgae449.

Type: Research Article

Included as: Appendix D

Contents

Abstract	iii
Zusammenfassung	v
Acknowledgments	vii
List of included publications	ix
Contents	x
1 Introduction	1
1.1 Cell-free blood as a diagnostic reservoir	2
1.2 Infrared molecular fingerprinting	3
1.3 The role of machine learning and data quality	7
1.4 Challenges of clinical translation	12
1.5 Objectives and structure of thesis	13
2 Materials and methods	15
2.1 Clinical study design and cohorts	16
2.2 Definitions of clinical phenotypes	18
2.3 Fourier transform infrared spectroscopy	19
2.4 Machine learning and computational frameworks	21
2.5 Ethics statements	25
3 Summary of publications	26
3.1 Plasma infrared fingerprinting with machine learning enables single-measurement multi-phenotype health screening	27
3.2 The perils of molecular interpretations from vibrational spectra of complex samples	30
3.3 Limits and prospects of molecular fingerprinting for phenotyping biological systems revealed through <i>in silico</i> modeling	32
3.4 CODI: Enhancing machine learning-based molecular profiling through contextual out-of-distribution integration	34
4 Concluding remarks	36
Bibliography	39

CONTENTS

Appendix A

-  Plasma infrared fingerprinting with machine learning enables single-measurement multi-phenotype health screening 49

Appendix B

-  The perils of molecular interpretations from vibrational spectra of complex samples 84

Appendix C

-  Limits and prospects of molecular fingerprinting for phenotyping biological systems revealed through *in silico* modeling 109

Appendix D

-  CODI: Enhancing machine learning-based molecular profiling through contextual out-of-distribution integration 128

Appendix E

- Copyright notices 156

1 Introduction

1.1 Cell-free blood as a diagnostic reservoir

Living organisms are composed of complex networks of interacting organic molecules [1, 2]. These molecules – including diverse proteins, lipids, nucleic acids, and metabolites – provide a snapshot of an organism’s function at a point in time [3]. In the context of human health, analyzing these biochemical constituents provides a powerful means to quantitatively examine the molecular makeup of body compartments and identify deviations linked to disease and dysfunction. Among the diverse tissues available, blood is a particularly valuable analysis target. It serves as the primary medium for transporting oxygen, nutrients, hormones, and metabolic byproducts, and as it circulates throughout the body, its interactions with various organs and tissues allow it to reflect the body’s functional state [4, 5]. Its accessibility through minimally invasive sampling further makes it a strategic starting point for developing diagnostic tools to investigate human physiology at the molecular level.

The presence of cells in whole blood can overshadow the detection of other molecular components (e.g., proteins, lipids, and metabolites), which are often indicators of underlying physiological conditions [4]. Therefore, blood plasma, the cell- and platelet-free fraction that constitutes $\sim 54\%$ of blood volume, along with blood serum, produced by removing clotting factors from plasma, are the preferred mediums for clinical analysis and biodiagnostic research [6, 7].

The value of cell-free blood stems from the diverse molecular landscape it contains, which can reflect physiological changes and indicate various health disorders. For instance, protein biomarkers such as C-reactive protein (CRP) and serum amyloid A (SAA) are significantly elevated in lung cancer patients compared to healthy individuals, indicating an underlying inflammatory response [8]. Metabolites, such as glucose and lactate, signal the onset of type-2 diabetes [9]. Lipid profiles serve as early risk indicators for cardiovascular disease [10]. Shortly, the capacity to capture multiple physiological conditions from a single analysis medium/sample underscores the potential of blood-based molecular diagnostics [3, 11, 12].

In medical practice, the diagnostic potential of blood components has long been utilized in clinical chemistry laboratories through targeted assays that quantify the concentrations of select analytes [13, 14]. For example, a comprehensive metabolic panel includes a set of fourteen blood tests measuring analytes such as albumin, creatinine, potassium, and glucose [14]. Another commonly employed test is the lipidomic panel, which includes high-density lipoprotein (HDL) cholesterol, low-density lipoprotein (LDL) cholesterol, very low-density lipoprotein (VLDL) cholesterol, and triglycerides [14]. Clinical laboratory tests are employed in multiple medical contexts, including in-patient diagnosis, predicting disease progression, and monitoring treatment response [13, 15]. As a result, they are currently the most frequently performed medical activity that precedes more expensive downstream patient care [15, 16]. Nevertheless, these assays rely on several measurement instruments and require labor-intensive procedures from trained specialists, leading to extended turnaround times and a significant cost burden on healthcare [4, 13–15]. Furthermore, biomarkers were developed in candidate approaches – requiring decades of research to link any given analyte to a physiological state.

1 Introduction

Shotgun approaches that probe the human phenotype by simultaneously analyzing an ensemble of molecules, requiring no *a priori* information, thus carry the potential to forge the future of blood-based clinical testing. Untargeted mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectroscopy are two highly promising technologies on the forefront to fulfill this role [3, 9, 11, 12, 17]. The recent advances in profiling the ‘proteome,’ ‘lipidome,’ and ‘metabolome’ through MS and NMR have spawned a multi-omics revolution in biomedical research. Broadly, multi-omics have demonstrated their potential in screening for multiple diseases, disease subtyping, risk stratification, and biomarker discovery [3, 9, 17, 18]. Still, concerns relating to low measurement throughput, run-to-run and between-lab variability, and involved experimental handling have hindered their applications in the diagnostic clinical setting, beyond niche applications [9, 19–21].

Taken together, there is an extensive unmet need to develop and validate new approaches for comprehensively profiling the molecular composition of cell-free blood in a high-throughput, reproducible manner. Such approaches could enable more efficient and cost-effective clinical diagnostics.

1.2 Infrared molecular fingerprinting

The interaction of light with matter has long been recognized as a powerful means of probing the physical and chemical properties of diverse materials without necessarily requiring direct contact. Whether analyzing gases, liquids, or solids, photons can provide information about the spatial structure, molecular concentrations, and dynamics (e.g., motion) of the material under study. These characteristics make optical-based techniques appealing in a broad range of domains – including biomedical applications, where non- or minimally-invasive methods are highly desirable.

One of the main strengths of photon-based techniques stems from the broad range of energies available across the electromagnetic spectrum (Figure 1.1). From high-energy gamma rays and X-rays to lower-energy infrared (IR) and radio waves, each spectral region interacts with matter in distinct ways. For example, X-ray radiation can penetrate thick or dense materials, enabling imaging of internal structures – an approach exploited in classical clinical X-ray imaging for bones and, more recently, phase-contrast imaging for soft tissues [22]. Ultraviolet (UV) spectroscopy is frequently used to quantify nucleic acid content and study protein folding [23, 24]. The different energies of X-ray and UV photons give rise to different phenomena – X-rays often ionize core electrons or induce scattering, whereas UV light typically excites valence electrons to higher energy levels.

In contrast, IR light is lower in energy than X-ray or UV light and is therefore less likely to induce chemical modifications in a sample. IR spectroscopy directly probes molecular vibrational modes, which correspond to quantized oscillations of atoms within a molecule around their equilibrium positions (Figure 1.1) [25, 26]. These oscillations include non-destructive modes of bond deformations, such as stretching (changes in bond length) and bending (changes in bond angle), that occur at characteristic frequencies [25]. When broadband IR radiation interacts with a sample, certain photons are absorbed, corresponding to the energy required to excite these vibrational modes and the presence of the bond in the sample. This excitation, known

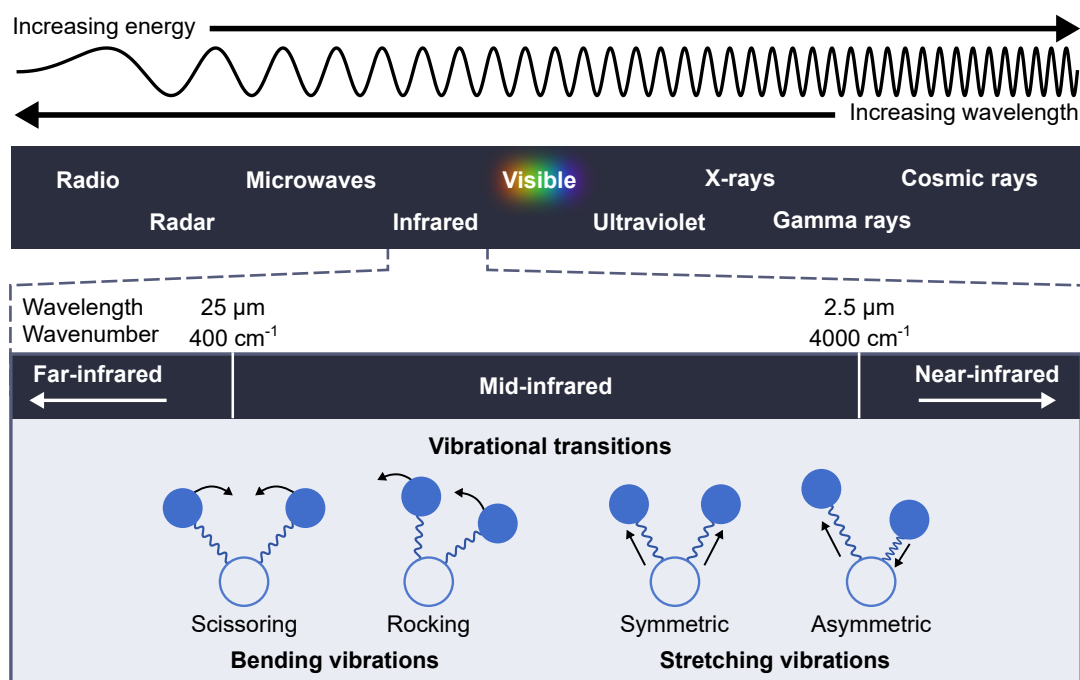


Figure 1.1: Overview of the electromagnetic spectrum focusing on the IR region. Molecules can be modeled as atoms bonded by springs (e.g., a water molecule, as shown). When interacting with an IR light source, these chemical bonds can undergo vibrational motions. Absorption occurs when the energy of an IR photon matches the transition between a molecule's ground and excited vibrational states. This absorption provides information about the chemical structure of the molecule. In the mid-IR region (400-4000 cm^{-1}), bending and stretching vibrations are prominent. This region is particularly valuable for analyzing biological samples, as it can provide molecular 'fingerprints' of functional groups within proteins, lipids, carbohydrates, and other organic molecules.

as a vibrational transition, requires a change in the molecular dipole moment. As a result, bonds or functional groups with significant dipole variations exhibit strong, characteristic absorption bands. For instance, complex molecules like proteins strongly absorb in multiple regions of the mid-IR (Figure 1.2 A), while nonpolar molecules like oxygen and nitrogen (lacking a dipole moment) are largely IR-inactive.

The mid-IR region (roughly 400-4000 cm^{-1} , but more predominately between 950-1800 cm^{-1} and 2800-3000 cm^{-1}) is particularly valuable for analyzing biological matrices, as it captures the vibrational modes of many covalent bonds in organic molecules [25, 26]. Proteins, lipoproteins, nucleic acids, and diverse metabolites exhibit distinct absorbance signatures within the mid-IR (Figure 1.2 A-C), enabling their identification from the recorded spectrum, depending on the analysis context. The intensity of absorbance at each wavenumber is proportional to the concentration of the functional group within the sample [25, 26]. In the analysis of complex mixtures such as plasma, the combined spectral response of all molecular constituents is recorded in the IR spectrum, providing a comprehensive molecular overview of the sample (Figure 1.2 D). This enables the simultaneous quantitative assessment of various molecules in a single measurement, capturing both the overall concentrations of functional groups and the variability between a set of analyzed samples.

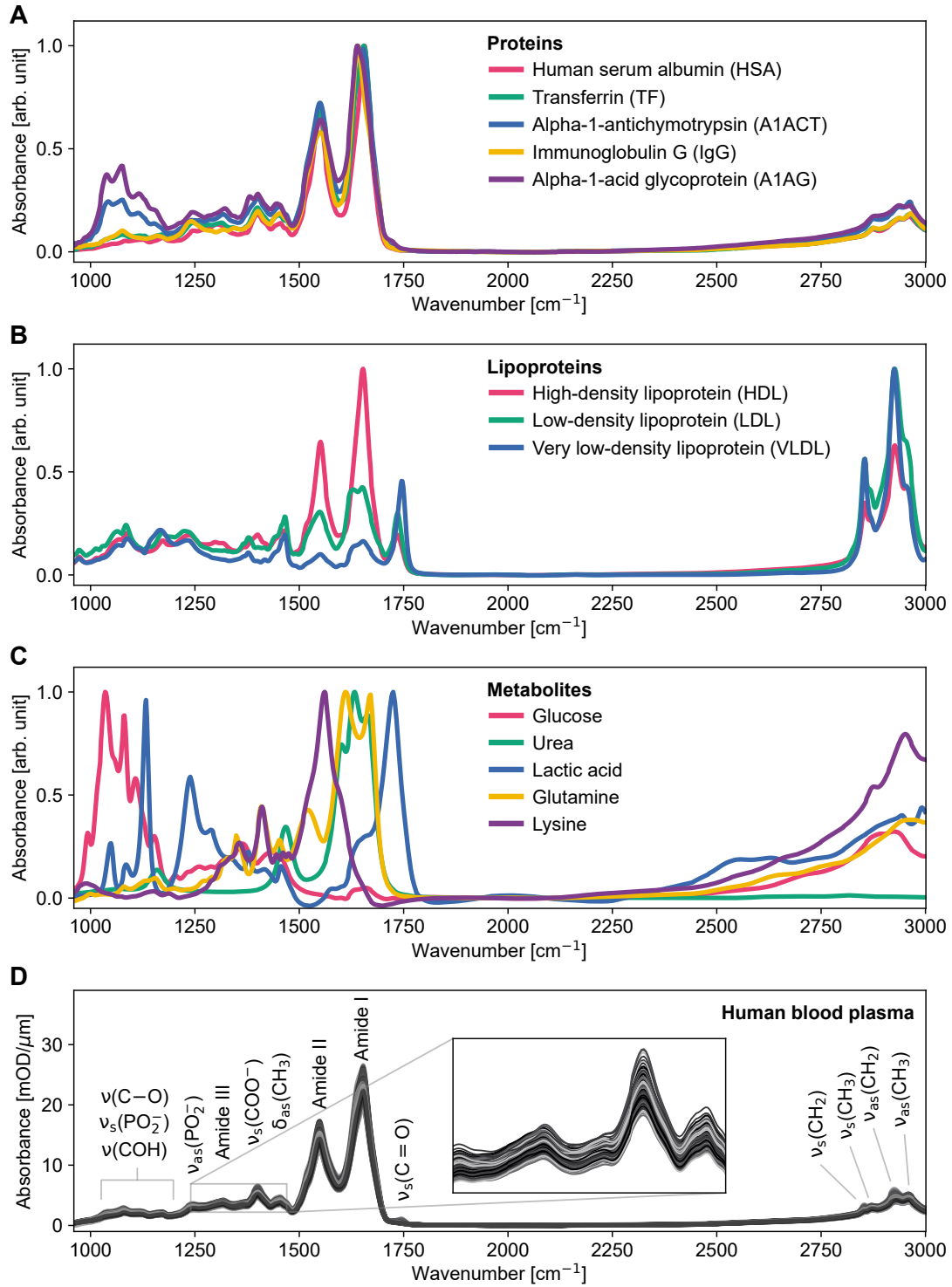


Figure 1.2: IR spectra of plasma molecular components. (A-C) IR absorption spectra of selected proteins, lipid particles, and metabolites, illustrating their characteristic peaks and unique molecular structures. IR spectra of individual components were normalized to their highest peak. (D) Composite nature of blood plasma IR spectra, representing the superposition of multiple molecular signatures within the plasma matrix. Plasma IR spectra were measured from healthy individuals, and are depicted without normalization. Inset depicts a close-up of the spectra in a selected region. Absorption peaks correspond to molecular vibrational modes: ν – stretching; δ – bending; s – symmetric; as – asymmetric.

1 Introduction

Broadly speaking, IR spectroscopy refers to any technique in which IR light is used to measure the absorption or emission properties of a sample. In a biologically focused context, IR *fingerprinting* refers to a more application-focused approach, wherein the IR spectrum of a biological specimen is recorded, and its overall unique absorption pattern – the “fingerprint” – is analyzed in relation to phenotypes of interest (Figure 1.3) [27–29]. Thus, rather than studying the chemical structure of a given sample, IR fingerprinting is applied to a collection of samples to identify patterns that differentiate classes (e.g., healthy vs. diseased) or correlate with quantitative parameters of interest (e.g., glucose content).

One of the most compelling advantages of IR fingerprinting is its minimal sample preparation requirements and label-free nature of measurement [26, 30]. Unlike techniques that require fluorescent tagging [31], radioactive labeling [32], or sample purification [33], IR spectra are recorded in a molecule-agnostic manner. Furthermore, samples can be measured in their native state – e.g., no requirement of chemical modifications or drying of liquid samples. In blood plasma analysis, for instance, the primary preparatory step involves isolating plasma from whole blood – a straightforward and widely established procedure. IR spectra can be recorded in relatively short timeframes, with widely available spectrometers yielding measurements on the scale of a few minutes, significantly reducing the risk of sample degradation over the course of a measurement [25, 27]. Moreover, small amounts of sample are typically needed to record a reliable IR spectrum ($\sim 50\ \mu\text{l}$ of plasma is sufficient [27]). These attributes enable the acquisition of reproducible data while supporting high-throughput workflows.

Fourier transform infrared (FTIR) spectrometers are widely regarded as the gold standard for IR spectroscopy [29, 30, 34–37]. Based on decades of advancements and refinements, modern commercial FTIR instruments can capture high-resolution broadband IR spectra, can be integrated with microfluidic systems for (semi-)automated sample delivery, and are combined with software to process and easily extract measurements (Materials and Methods). FTIR-based fingerprinting has been employed in various proof-of-principle applications in medical contexts, including the detection of multiple cancer types from plasma and serum [38, 39], acute chest pain from serum [40], metabolic conditions from plasma [27], glycosylation patterns in multiple plasma proteins [41], viral infections from blood serum [42], and diabetes from urine [43]. Beyond biofluid fingerprinting, FTIR has also been employed in spatially-resolved tissue analysis in histopathologic examinations [44].

Despite its merits, IR fingerprinting is not without limitations. When analyzing complex samples, such as plasma, one major challenge stems from the overlap of absorption peaks arising from numerous molecular constituents [25, 28, 45, 46]. For instance, the IR spectra of different proteins share many of their vibrational responses, making it virtually impossible to reliably link an observed signal to a single protein (Figure 1.2 A). This challenge is particularly pronounced in realistic medical screening applications, where the concentrations of several molecular species are changing between different classes of samples, complicating the identification of the molecular origins of observed spectral patterns. Furthermore, FTIR-based fingerprinting has a relatively modest capacity to detect the signals of low-abundance molecules compared to more targeted methods, or without additional sample preparation (e.g., enrichment

techniques). When biomarkers of interest are present at very low concentrations (e.g., circulating plasma DNA, RNA, CRP, CA-125, mannose, or PSA), their vibrational signatures are too weak to be detectable, being overshadowed by both the technical noise of the measurement (with a detection limit ~ 0.01 mg/ml [28, 34, 47]) and the molecular variability of more abundant substances.

Despite the mentioned challenges, the merits of IR molecular fingerprinting remain highly compelling. The approach offers an unparalleled capacity for providing high-throughput, cost-effective, and label-free biochemical “snapshots” of patient samples. With ongoing research advancements and efforts toward integration and acceptance, IR fingerprinting has the potential to complement or even replace current diagnostic practices in the near future.

1.3 The role of machine learning and data quality

The wealth of molecular information captured through IR fingerprinting offers tremendous potential for a multitude of applications in medical diagnostics and beyond. However, realizing this potential is inherently tied to the ability to collect high-quality datasets and extract meaningful insights from them – i.e., correlating information from a set of spectra to a physiological characteristic. In the context of blood-based *in vitro* diagnostics, a challenge lies in analyzing spectra derived from samples that share a broadly similar molecular composition. Unlike applications that analyze a set of different materials or identify clear contaminants in relatively homogeneous samples, blood-based spectra exhibit largely similar shapes and characteristics (Figure 1.2 D). Thus, the task is often not to identify visually-obvious differences, such as the emergence of new peaks, but to detect more subtle changes superimposed on a largely consistent baseline. These subtle changes are further masked by sources of data errors – e.g., noise in data annotations – as well as variations that should not correlate with the characteristic of interest – e.g., intrinsic biological variability and (pre-)analytical inconsistencies. In other terms, we are often operating in a low signal-to-noise ratio environment.

Practically, a broadband IR spectrum is a vector of continuous numbers ($\mathbf{x} \in \mathbb{R}^m$), with each dimension representing the total IR absorbance at a particular wavenumber. The number of dimensions in this vector can range from several hundred to few thousand features, depending on the spectral resolution and the analyzed spectral range. This translates to a matrix spanning a multidimensional space when multiple samples (n) are analyzed ($\mathbf{X} \in \mathbb{R}^{n \times m}$). Within this dataset, the signal of interest – often a combination of several features linked to a specific outcome – constitutes only a small fraction of the total variability.

Machine learning (ML), and more specifically multivariate analysis, plays a pivotal role in this context. Given a spectral dataset, ML algorithms can learn feature weights, accounting for all feature dependencies, to identify subtle signals linked to annotated outcomes. These outcomes, derived from participant characteristics, can take various forms, including:

- **Binary:** one of two exclusive classes (e.g., case or control group);

1 Introduction

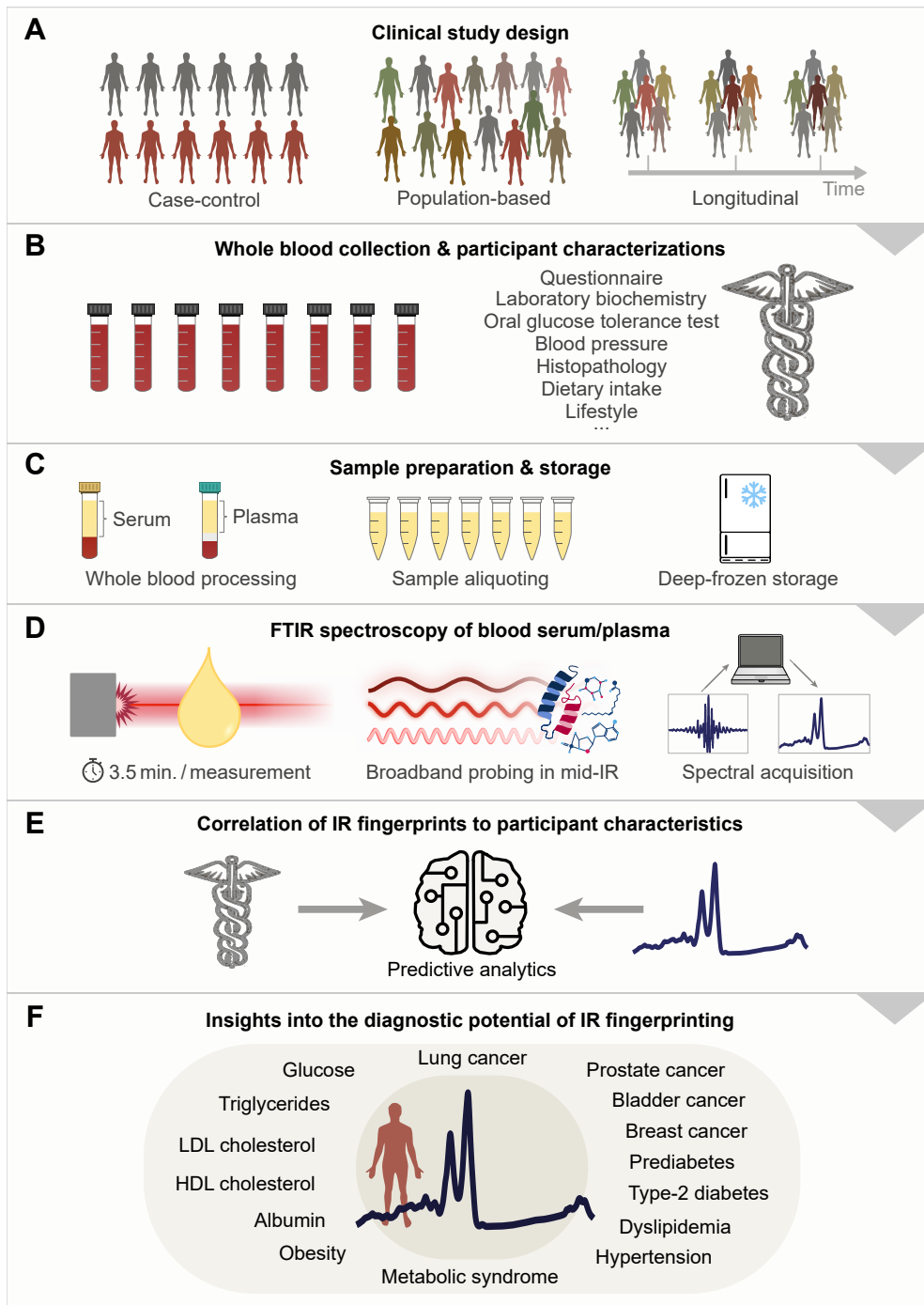


Figure 1.3: Overview of blood-based IR fingerprinting workflow. (A) Clinical studies are designed as case-control (comparing two groups), population-based (capturing heterogeneity across a population), longitudinal (monitoring changes over time), or mixed designs (e.g., longitudinal population-based). (B) Whole blood is collected and participants are comprehensively characterized through questionnaires, clinical chemistry, and medical assessments. (C) Blood is processed into serum or plasma, aliquoted into smaller volumes, and stored for long-term use. (D) FTIR spectroscopy enables high-throughput molecular probing of serum or plasma. (E) Predictive analytics correlate the IR spectral measurements with participant characteristics, extracting meaningful patterns. (F) A snapshot of diagnostic insights is provided through this workflow, revealing the potential of IR fingerprinting to quantify biomarker concentrations and detect phenotypic conditions from a single measurement.

1 Introduction

- **Multi-class:** one of three or more mutually exclusive classes (e.g., type-1 diabetes vs type-2 diabetes vs no diabetes);
- **Multi-label:** two or more non-exclusive classes (e.g., co-occurring conditions);
- **Continuous:** real-valued outcomes (e.g., molecular concentration); or
- **Ordinal:** ranked outcomes (e.g., disease stage).

As the target is often to map a given spectrum to one of the above label formats, *supervised* classification or regression algorithms are of instrumental use. During the training phase, the algorithm is exposed to a set of spectra together with the known associated outcomes. The algorithm’s objective is to learn *patterns* within the given training set that correlate to the outcome of interest. The success of this learning is judged by the capacity of the trained model to make predictions on new, unseen samples. This *generalization* capacity – which can be judged by several performance metrics (Materials and methods) – is the vital aspect on which to assess the success of the whole IR fingerprinting workflow and the diagnostic potential it possesses.

Past studies have shown that ML can effectively decode the informational content of spectral data to predict a variety of phenotypic outcomes using different learning algorithms (e.g., support vector machines, ridge regression, logistic regression, random forest, discriminant analysis, and neural networks) [27, 38, 42, 48–51]. The process of selecting and tuning the learning algorithms is an important aspect that should be done with care. However, established methodologies offer clear guidance in this area [52, 53]. Other critical and more time-consuming considerations lie in the broader aspects of the IR fingerprinting workflow (Figure 1.3), such as the technicalities of study design, data collection, and validation strategies. Ensuring robust model performance requires careful attention to potential biases, sources of variability in the data, and rigorous validation to avoid overly optimistic assessments of performance. In this context, adopting a data-centric approach to ML, which prioritizes the quality, representativeness, and integrity of the dataset, can be the decisive factor in the success or failure of the whole workflow.

For one, supervised learning relies on high-quality annotations (i.e., labels) that accurately represent the phenomena of interest, serving as the ground truth essential for model training and validation. However, label noise – manifesting as misclassified samples (e.g., diseased individuals labeled as healthy) or unreliable quantitative measurements (e.g., assays with high error margins) – can undermine this process. The reliability of these labels varies depending on the measured outcome and the assessment method employed. For instance, study participants often complete questionnaires at the time of blood donation, providing self-reported health information. However, self-reported health data are notoriously prone to inaccuracies. Consider the diagnosis of diabetes: in the Southern German adult population (aged 55 and above), approximately 40% are in a (pre)diabetic state [54]. Yet, half of these cases remain undiagnosed when relying solely on self-reported data [54]. This discrepancy results in a significant source of label noise, where a large portion of individuals self-reporting as healthy may in fact be in a (pre)diabetic state. Even clinically established diagnostic assays, such as an oral glucose tolerance test (OGTT), are not immune to discrepancies. OGTT results are influenced by physiological variances, sometimes leading to inconsistent diagnoses depending on the timing and conditions of adminis-

1 Introduction

tration [55, 56]. Consequently, the diagnostic accuracy of any newly developed model is inherently constrained by the reliability of the ground truth on which it is assessed. Ensuring that annotations are as accurate and reliable as possible is therefore a critical step prior to model training.

Another challenge stems from the presence of potential confounding factors that bias the training data. Confounders are variables that correlate with the outcome of interest but are not causally related [57]. Consider a case-control study designed to distinguish a specific cancer from healthy controls. If the cancer patients are systematically older than the healthy control individuals, the model may inadvertently learn to associate molecular signals related to age rather than those specific to cancer. Such confounding effects are not confined to biological variables but also extend to analytical aspects of data collection. For example, suppose the measurements of case samples are conducted sequentially, followed by those of control samples. In this instance, the predictive algorithm may learn to distinguish samples based on their time of measurement, or measurement batch. Similarly, site-specific confounders can occur when samples are collected from different locations and operators, leading ML algorithms to detect variations associated with the collection site instead of the disease-induced variations [58]. These confounders can significantly undermine the model’s capacity to accurately identify the true relationship between the spectral input and the physiological condition being studied. It is therefore essential to implement practical measures that ensure comparability between the studied groups, such as matching cases to controls based on relevant characteristics (e.g., age, gender, body mass index, and comorbidities). Furthermore, randomizing the order of measurements and ensuring that samples are collected under standardized procedures are necessary to mitigate the impact of confounding factors.

The representativeness of the collected training dataset is another critical consideration. Supervised ML models operate under the assumption that the training and test data are independent and identically distributed (i.i.d.) [59]. To ensure adequate model generalization, the training dataset must therefore adequately reflect the statistical distribution of data encountered in real-world applications. This requirement is particularly challenging in the domain of multi-molecular fingerprinting, where numerous factors can influence the measurements obtained. These factors can be linked to biochemical changes in the sample or technical variations in data acquisition, among them:

- **Biological variability:** variances observed at a within-person level (e.g., recycling, turnover, rhythmic diurnal/annual oscillations), between-person level (e.g., genetics, lifestyle factors, environmental exposures), or population-level variances (e.g., ethnicity, genetic background) [1, 49, 60–62];
- **Pathophysiological variability:** variances arising from the stages or severity of health conditions, the presence of comorbidities, or subclinical states that may alter the biochemical profile [60, 62];
- **Preanalytical variability:** variances introduced during sample handling, such as differences in the site of needle puncture for blood collection, variations in the tube filling volumes, delays in sample processing (e.g., time at room temperature), or inconsistencies in centrifugation protocols or tube types used to prepare plasma or serum [60, 63, 64]; and

1 Introduction

- **Analytical variability:** technical variances related to data acquisition, including wear and tear of spectrometer components, component replacements, inter-spectrometer differences, or external operating conditions (e.g., temperature or humidity during analysis) [65, 66].

Each of these factors can impact the data, introducing systematic drifts, or variations of a more random nature. For accurate assessments of diagnostic capacity and to achieve robust generalization, it is essential that the collected training dataset captures a wide range of potential data variances and measurement conditions that may be encountered in practice.

Finally, a critical perspective is needed when devising the approach of testing models on unseen data – to estimate their predictive performance under model deployment settings. In biondiagnostic applications, where collecting large amounts of biological specimens and data points is difficult, a cross-validation (CV) procedure is commonly applied to estimate true (external) predictive performance. Briefly, in CV, the available dataset is typically shuffled and split into a training set (e.g., $\sim 90\%$ of the data) and a test set (the remaining portion). This procedure is then repeated, holding-out a different portion of the data for testing in each iteration. Compared to performing a single random train-test split of a limited dataset, CV offers a more robust performance estimate as it allows the model to be both trained and tested on a larger sample pool and provides a level of confidence in the measured predictive efficacy. However, if confounding factors are present in the data, e.g., caused by measurement batch effects, a CV with a random shuffle may produce overly optimistic performance estimates that fail to translate to true external test settings [67]. Similarly, if the dataset used for CV does not adequately capture the range of real-world variability, the developed model may struggle with out-of-distribution (OOD) data when deployed in practice [68].

A more rigorous approach to evaluating diagnostic capacity requires testing procedures that explicitly address realistic sources of data variability. For example, to evaluate the model’s robustness to measurement instrument drift, test samples should be collected and measured at a later time point than the training set, ensuring that any temporal or operational changes are reflected. If the trained model generalizes well to these temporally shifted measurements, it indicates resilience to (pre)analytical drift on the time scale that was examined. Even more ideally, these temporally shifted test samples should be collected from independent clinical studies. This offers a more stringent setting to assess the model’s capacity to also generalize across diverse populations and data collection environments [69, 70]. However, obtaining external, independent validation samples is often impractical as it is resource-intensive and requires access to multiple clinical studies targeting the same phenotypic outcomes. Nevertheless, carefully designed validation strategies – tailored to mimic realistic application settings – offer a more robust assessment of generalization capacity than traditional random train-test splits, which may not capture the potential decline in predictive capacity in actual clinical deployment.

Ultimately, predictive ML and multivariate analysis play an important role in the domain of IR fingerprinting. They provide a means to transform a difficult-to-interpret spectral measurement into an actionable health outcome. However, their effective application requires careful attention to study design, data collection, and model validation strategies.

1.4 Challenges of clinical translation

As societies face aging populations and a rising prevalence of chronic conditions, the demand for high-throughput screening tools is increasing. Newly developed tools should prioritize simplicity, automation, and affordability to increase their likelihood of widespread adoption in clinical practice. While IR spectroscopy is not a new technology – FTIR spectrometers have been commercially available since the late 1960s [26] and the potential for biomedical applications was recognized early on [71] – recent years have seen advancements in its methodologies and applications. Innovations such as tunable quantum cascade lasers (QCLs) [72], field-resolved spectroscopy (FRS) [34, 73], improved FTIR spectral resolution and measurement reproducibility [26, 30], and the integration of ML for data analysis have expanded the scope of potential applications [26, 74]. These developments have contributed to a growing interest in employing IR fingerprinting for biomedical research, producing a wealth of evidence supporting its utility in health diagnostics [27, 38, 41, 44, 49, 75–83]. Despite this progress, the transition from research to practice still faces barriers, and IR spectroscopy has yet to be translated to the clinical setting [74, 84–86].

In 2004, Jackson offered a critique on the slow adoption of optical fingerprinting in clinical settings [87]. In one part, he noted:

“... we must ask ourselves why we have failed to demonstrate clinical utility. The answer may be simple but unpleasant: perhaps biological optical spectroscopy is not advancing. ... Much of the research being presented orally and in the literature may best be summarised as ‘more of the same’ with little truly new information being presented. ... [A]t times much current work seems merely derivative. ... In short, at this point a lack of imagination would seem to be a major problem.”

While his words are over two decades old, as I reflect on them, the sentiment still resonates with the current state of the IR fingerprinting research domain.

Byrne et al. [84] and Cameron et al. [74] have addressed similar themes in their perspectives. Both emphasize the value of screening body fluids as they represent an achievable target. Both cite technical challenges associated with pre-analytical standardization and its implications for clinical adoption. Indeed, the quality of an IR spectrum from a blood-derived sample can be significantly affected by sample collection methods (e.g., tube types, filling volumes, mixing procedures), subsequent sample handling (e.g., deep-frozen storage temperatures and durations, number of freeze-thaw cycles, aliquoting procedures, time at room temperature), and even intra- and inter-operator differences. While pre-analytical standardization is an essential prerequisite for clinical adoption, these challenges are not unique to IR fingerprinting. They are inherent to biofluid analytics already established in clinical settings and affect other molecular profiling technologies, including nuclear magnetic resonance (NMR), mass spectrometry (MS), and genetic sequencing methods [88–90]. As such, standard operating procedures (SOPs) for IR fingerprinting can and have been developed [49, 91], analogous to those used in other technologies, to yield reproducible data.

Another recurrent reason often cited as a barrier to clinical adoption is the need for large-scale clinical studies [74, 84, 85, 87]. Most previous proof-of-concept studies relied on relatively small patient cohorts (often below 100 individuals) to demonstrate

1 Introduction

clinical utility. The use of small cohorts brings into question the reliability of the results and their capacity to generalize to larger populations in practical implementation [74, 84, 85]. Especially with the growing use of ML in the domain, the statistical power of such studies becomes even more uncertain, as ML algorithms often require large sample sets to train the model and validate its diagnostic capacity on unseen, held-out data with high confidence [92, 93]. Beyond pure sample size counts, the adequate representativeness of training data is essential to avoid biases that can compromise diagnostic performance. Limited heterogeneity in training datasets – whether due to unrepresentative demographics, phenotypic stages, dynamic human physiologies (e.g., rhythmic oscillations, aging), sample collection technicalities (e.g., inevitable SOP deviations), or technical measurement drifts (e.g., different spectrometers) – can lead to overestimation of diagnostic efficacy and reduced generalizability to real-world settings when more heterogeneous data is encountered [50, 59, 94–96]. Addressing these challenges will require considerable resource and time investments to enroll subjects representative of a potential patient population, perform multi-centric sample collections, and account for inter-laboratory and inter-spectrometer variations.

Overall, the clinical translation of IR fingerprinting requires advancement on multiple fronts. Beyond the need for standardization and large-scale studies, closer collaboration with clinicians is crucial to ensure that research efforts target clinically meaningful questions. Once proof-of-concept is established, the path to clinical adoption involves commercialization and obtaining regulatory approvals, both of which pose added challenges [74]. Nevertheless, recurring themes in the literature over the years that address the challenges of clinical adoption [74, 84–87] suggest that the field remains largely stagnant. Jackson’s earlier observation, quoted above, still resonates powerfully here: the field has struggled to turn its early promise into real-world impact. Breaking this pattern will require prioritizing investments in large-scale, representative clinical studies that address relevant clinical questions and generate robust, generalizable findings.

1.5 Objectives and structure of thesis

The work presented in this thesis is unified by a central research objective: **To evaluate the feasibility of IR molecular fingerprinting as a robust blood-based *in vitro* analytical tool for high-throughput phenotype screening and analysis.**

To address this objective, four complementary studies are presented, each guided by the following primary research questions (RQs):

- RQ #1: Can IR fingerprinting, combined with ML, reliably detect and predict complex physiological conditions within a large, naturally heterogeneous patient population?
- RQ #2: What are the key considerations and best practices for interpreting IR spectra of molecularly complex samples, particularly those derived from blood-based media and their molecular components?
- RQ #3: How do technical measurement, biological, and experimental study design factors affect the diagnostic capacity of IR fingerprinting, and what

1 Introduction

approaches can be adopted to refine them for improved diagnostic performance?

RQ #4: What strategies can address the challenges of standardization, out-of-distribution data, and generalization in ML modeling to improve model transfer across naturally variable molecular fingerprint datasets?

These RQs, tackled in the included studies, address key areas of basic and translational research, aiming to identify opportunities and challenges, validate methods, and inform future applications and efforts of establishing IR fingerprinting in biomedical diagnostics. However, they do not cover the broader aspects of real-world deployment, such as regulatory approvals or commercialization.

The subsequent chapters provide the methodological background (**Chapter 2**), summarize each study’s objectives and findings (**Chapter 3**), and conclude with final remarks and future directions (**Chapter 4**). The four RQs are investigated in publications included in the appendices: RQ #1 – **Appendix A**; RQ #2 – **Appendix B**; RQ #3 – **Appendix C**; RQ #4 – **Appendix D**.

2 Materials and methods

2.1 Clinical study design and cohorts

One of the first considerations in establishing an IR fingerprinting medical application is selecting an appropriate clinical study design (Figure 1.3). At the heart of this design lies a central question: **which specific problem is being addressed, and what considerations are required to enable its evaluation?** In this thesis, four independent clinical studies involving the collection of human blood samples were used to address different research questions (RQs #1–4 in Section 1.5). Collectively, these studies enabled an assessment of the feasibility and technical robustness of IR fingerprinting for probing human health, while also facilitating the development of new data analysis tools to improve diagnostic capacities.

Blood samples and corresponding medical data from the following clinical studies were utilized in the publications presented in this thesis:

1. **KORA** study [27, 28, 50];
2. **Lasers4Life-Cancer** study [28, 47, 50];
3. **Lasers4Life-LG** study [50]; and
4. **BioPersMed** study [50].

The **KORA** (Cooperative Health Research in the Region of Augsburg) study was performed within the framework of a project established by Helmholtz Zentrum München and was designed as a research platform open to external project partners [97]. KORA is a population-based cohort in Southern Germany, comprising an age- and gender-stratified sample of adult participants randomly drawn from the resident registration offices within the study area. The study features a longitudinal setup, with several rounds of blood donations and medical characterizations over time. The first baseline characterization (KORA-S4) took place in 1999-2001, followed by KORA-F4 in 2006-2008 and KORA-FF4 in 2013-2014.

Upon successful project application, the analyses presented in this thesis were conducted in collaboration with Helmholtz Zentrum München, using retrospectively collected samples from the KORA cohort. Specifically, we obtained medical examination data and blood plasma samples from KORA-F4 ($n = 3044$) and KORA-FF4 ($n = 2140$), among which $n = 2015$ individuals participated in both. Being population-based, the KORA study captures a realistic degree of biological variability across a wide range of adult individuals, as well as reflecting a realistic estimate of the prevalence of physiological conditions in the population. The KORA study enabled us to conduct the largest study to date examining IR fingerprinting as a modality to probe human health, covering common health aberrations and their potential combinations. Moreover, the longitudinal design supported investigations into predicting the future onset of health aberrations [27] and assessing the stability of individual IR fingerprints over a 6.5 year follow-up period [50].

The **Lasers4Life-Cancer** clinical study was established by Ludwig-Maximilians-Universität München to prospectively collect samples for the detection of four common cancer phenotypes: lung, prostate, bladder, and breast cancers [38]. The study was set up as a case-control design to evaluate disease-specific biochemical differences that can be captured by IR fingerprinting. The study is multi-centric, enrolling individuals diagnosed with one of the four malignant carcinomas, benign pathologies affecting

2 Materials and methods

the same organs, or non-symptomatic healthy controls. Individuals with suspicion of having one of the four listed cancers were enrolled in the study. Thus, all cancer cases were enrolled prior to any cancer treatment or therapy (i.e., therapy naïve) at the time of blood collection. This design feature eliminated the possibility of capturing the signal of treatment, as a confounder, rather than the presence of the disease in the case-control setup. Blood plasma and serum were collected from all participants satisfying the inclusion criteria.

Within this thesis, samples from the Lasers4Life-Cancer study were utilized in several analyses [28, 47, 50]. Samples from individuals with lung cancer ($n = 536$), prostate cancer ($n = 449$), bladder cancer ($n = 205$), and breast cancer ($n = 163$) were utilized [50]. Samples from healthy individuals pair-matched to each cancer case individual – based on age, gender, and body mass index – of the same corresponding sample sizes were included as appropriate controls. The study enabled the evaluation of IR fingerprinting as a modality to detect several common cancers [50]. The study also facilitated the development of a data simulation tool that enabled deeper investigations into the limiting factors hindering the diagnostic capacity of IR fingerprinting [47], as well as a new method of training ML models in a manner that increases their likelihood of generalizing to realistic data collection settings [50].

The **Lasers4Life-LG** (LG: longitudinal) study was established by Ludwig Maximilians Universität München [49]. The study cohort comprised of 31 nominally healthy individuals, initially sampled up to 13 times over a 7-week period, with an additional follow-up after 6 months as detailed in a previous publication [49]. Since this initial publication, the same individuals were invited to participate in two additional sampling sessions at 3.5 and 4.5 years post their initial involvement. In both the latter two samplings, 8 individuals participated. In total, 335 whole blood samples were collected within the framework of the study. Both blood plasma and serum were processed from all samples.

Within this thesis, samples from all participants and all visits spanning the 4.5 follow-up period were considered. This setup enabled the investigation of the stability of person-specific molecular signatures obtained through IR fingerprinting and developing methods that further enhance the stability of the IR fingerprinting approach over multi-year periods [50].

The **BioPersMed** (Biomarkers of Personalised Medicine) study was established by the Medical University of Graz, Austria [98]. The study focused on enrolling asymptomatic individuals from the regional communities at risk for cardiovascular or metabolic disease to evaluate the predictive value of various traditional and novel biomarkers and observe disease courses starting in the preclinical, asymptomatic phase. The study incorporated a longitudinal setup, wherein repetitive blood donations and examinations of participants were conducted in 2-year intervals.

Upon successful project application, we collaborated with the Medical University of Graz. Within this thesis, we utilized blood plasma samples and medical data from a subset of 44 nominally healthy individuals out of 1022 participants. A total of 44 individuals participated in five visits, each separated by 2 years, spanning up to 8 years from the baseline visit to the last follow-up considered. BioPersMed served as an additional independent cohort to examine the long-term stability of person-specific

IR fingerprints, demonstrating this stability over 8 years, the most extended follow-up time interval investigated to date [50].

2.2 Definitions of clinical phenotypes

As discussed in Section 1.3, the annotations associated with each IR spectrum provide detailed characterizations of study participants (e.g., age, gender, height, weight, etc.) as well as the ground truth basis for assessing diagnostic efficacy for conditions of interest. Such participant characterizations (i.e., the ‘metadata’) are collected in various forms, depending on the nature of the variables collected and the targets of the clinical study. General characteristics of study participants are relatively easy to collect reliably. All participants are asked to complete a standardized questionnaire at the time of their blood draw, providing information such as their age, gender, smoking status, and current medications. Additionally, it is relatively easy to get all participants to undergo simple physical measurements to determine characteristics such as their height, weight, and waist circumference. The more challenging data to obtain in a reliable manner is the deeper clinical characterizations – e.g., confirmed diagnoses, disease staging, and biomarker profiles. Within the scope of this thesis, samples from four independent clinical studies were utilized (Section 2.1), each targeting different outcomes.

The **KORA** study provided comprehensive medical phenotypic outcomes of the individuals derived from medical records, diagnostic tests, and laboratory biochemistry [97]. For the investigations presented in this thesis utilizing KORA samples [27, 28, 50], several outcomes were considered and were defined as follows:

- Dyslipidemia was defined based on abnormally elevated levels of non-HDL cholesterol with the commonly used cut-off point at 4.1 mmol/l [99, 100].
- Hypertension was defined based on a blood pressure test with a reading ≥ 140 and/or 90 mmHg or by known use of antihypertensive medication.
- Prediabetic and type-2 diabetic states were determined by an oral glucose tolerance test (OGTT) or a validated self-report. Self-reported diabetes was confirmed by contacting the responsible physicians, reviewing medical charts, or by taking self-reported glucose-lowering medications. An OGTT was performed on individuals with no prior diagnosis of diabetes, and the diabetic state was determined according to the 1999 WHO criteria [101]. Prediabetes included individuals with impaired fasting glucose, impaired glucose tolerance, or both.
- Metabolic syndrome (MetS) was defined according to the harmonized criteria proposed by the International Diabetes Federation (IDF) and American Heart Association/National Heart, Lung, and Blood Institute (AHA/NHLBI) [102]. MetS was present if ≥ 3 of the following 5 criteria were satisfied: (1) waist circumference ≥ 94 cm for males, ≥ 80 cm for females; (2) systolic blood pressure ≥ 130 mmHg or diastolic blood pressure ≥ 85 mmHg or treatment with antihypertensive medication; (3) fasting serum glucose ≥ 100 mg/dl or intake of antidiabetic medication; (4) serum HDL cholesterol < 40 mg/dl for males, < 50 mg/dl for females, or drug treatment for reduced HDL (fibrates); and (5) fasting serum triglycerides ≥ 150 mg/dl or drug treatment for elevated triglycerides (fibrates).

2 Materials and methods

- Blood glucose levels were determined using a hexokinase method from serum.
- HbA1c was determined using a reverse-phase cation-exchange High-Pressure Liquid Chromatography (HPLC) analysis.
- Total cholesterol, HDL cholesterol, and LDL cholesterol were determined using enzymatic, colorimetric routines (CHOD-PAP) from serum.
- Triglycerides were determined using an enzymatic, colorimetric routine (GPO-PAP) from serum.
- Creatinine was determined using a modified kinetic Jaffe reaction from serum.
- Hematologic parameters (leukocytes, erythrocytes, hematocrit, platelets) were determined by impedance measurements from EDTA.
- Albumin was determined using immunonephelometry from serum.

The **Lasers4Life-Cancer** study enrolled individuals that were at suspicion of having one of the four common cancers, comparable control participants with benign conditions affecting the same organs, and non-symptomatic references. The phenotype groups were defined based on the following:

- Presence of malignancies (cancer cases): Individuals diagnosed with lung, prostate, bladder, or breast cancer – clinically diagnosed using medical imaging methods such as X-ray and MRI, and pathologically confirmed through immunohistochemical tissue biopsy – and who had not received any cancer-related therapies within the past two years, were included in the analysis.
- Presence of benign conditions (control individuals): Individuals diagnosed with chronic obstructive pulmonary disease (COPD) or pulmonary hamartoma for lung cancer, and benign prostate hyperplasia (BPH) for prostate cancer – clinically defined based on self-reporting and corroborated by reviewing medical records.
- Non-symptomatic references (control individuals): Individuals not diagnosed with malignancies or benign conditions, as defined above.

For the remaining two longitudinal clinical studies – **BioPersMed** and **Lasers4Life-LG** – no clinical endpoints were utilized. For the investigations presented in this thesis utilizing samples from these two studies [50], the task was to examine the stability of person-specific measurements over time rather than investigating medical endpoints.

2.3 Fourier transform infrared spectroscopy

As discussed in Section 1.2, Fourier transform infrared (FTIR) spectrometers are the most widely commercialized and utilized instruments for IR spectroscopy. In this thesis, FTIR spectroscopy serves as the primary measurement modality. To provide additional context, a brief overview of the fundamental principles underlying FTIR spectrometers is discussed below (Figure 2.1).

Every spectrometer system includes four core components: (1) a light source; (2) an interferometer to modulate the light into an interference pattern; (3) a detector to record light intensity; and (4) an external or embedded computer for signal processing.

2 Materials and methods

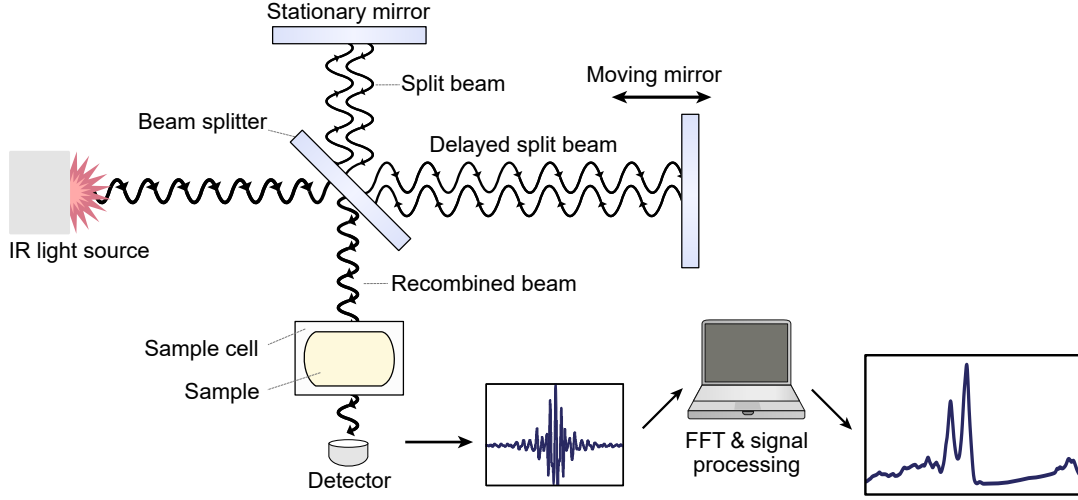


Figure 2.1: Schematic of a basic FTIR setup featuring a Michelson interferometer.

The process begins with the generation of broadband IR radiation. In the mid-IR, the most common light source is a Globar – a silicon carbide rod [26]. The Globar is heated to high temperatures (~ 1300 K), continuously emitting a broad spectrum of IR wavelengths. The IR radiation is then directed into an interferometric setup, typically a Michelson interferometer, which is at the heart of the FTIR technique [26]. In this setup, the incoming IR beam is split into two paths: one path is reflected off a stationary mirror while the other off a (motorized) movable mirror, and then recombined. The controlled motion of the motorized mirror modulates the optical path difference between the split beams, creating an interference pattern in the combined beam that encodes wavelength information. This enables the wavelengths to be resolved at a later stage.

After leaving the interferometer, the recombined beam reaches and interacts with the sample. In transmission FTIR, the sample is injected into a flow-through sample cell, or cuvette, with windows made from an IR-transparent material. Calcium fluoride (CaF_2) is a commonly used and appropriate material for the cell when analyzing aqueous samples – albeit it is slightly soluble in water and will need replacement after prolonged use. The pathlength for the sample cell – the distance over which the IR beam passes through the sample – must be carefully chosen. Longer pathlengths increase the amount of sample that interacts with the light, which can improve the molecular detection sensitivity. However, strongly absorbing samples, such as those containing water, require shorter pathlengths to avoid complete absorption of the IR signal. For aqueous samples like blood plasma or serum, which are made of approximately 92% water, very small pathlengths (below $10\ \mu\text{m}$) are required to ensure measurable transmission given the intensity of light in FTIR spectrometers [34, 103, 104].

Within the interactions of IR radiation with the sample, wavelengths corresponding to the vibrational modes of molecular bonds in the sample are absorbed, while unabsorbed wavelengths pass through and are recorded by a detector. IR detectors can be thermal detectors or photodetectors [25, 26]. Thermal detectors respond to changes in temperature of an absorbing element, while photodetectors rely on a semiconductor whose electrical properties change in response to incoming IR photons. Both types

2 Materials and methods

of detectors convert the transmitted IR signal into an electrical output, known as an interferogram. The interferogram is a time-domain measurement that represents the signal intensity as a function of the changing optical path difference introduced by the movable mirror. To convert the interferogram into a more usable measurement, a fast Fourier transform (FFT) is applied to yield the IR spectrum – a plot of absorbance (or transmittance) as a function of wavenumber (the reciprocal of wavelength, measured in cm^{-1}). The measured absorbance at each wavenumber is proportional to the total concentration of the molecular components that react to the corresponding energy level.

When analyzing aqueous samples, and the signal of interest is not water, another water reference spectrum can be recorded immediately before or after the sample of interest is measured. The spectrum of water can then be subtracted from the corresponding measured spectrum of the sample. In automated systems, the reference water measurement can be automatically recorded whenever any sample is measured, directly producing the water-subtracted sample spectrum.

Additional automation steps within the measurement setup can be taken to enable high-throughput analysis and more reproducible measurements. Most widely used experimental spectrometer systems require the sample to be manually injected into the sample cell. While low cost, this procedure requires more manual labor and can introduce added sources of measurement variability – e.g., inter- and intra-operator variation. Automated spectrometers can be equipped with microfluidic sample delivery systems, which feature a robotic arm designed for handling batch measurements. The operator places a set of tubes on a rack (e.g., 60 tubes). The robotic arm processes the sample rack, sequentially injecting the sample contents into the sample cell system. Cleaning procedures can also be automatized, where solutions pass through the system between sample measurements to eliminate residual contamination and cross-contamination in the measurement cell. These automated features facilitate more reproducible measurements and support large-scale analytical workflows.

Within this thesis, all FTIR spectra were measured in native liquid phase using a commercially available spectrometer with automated microfluidic sample handling (MIRA Analyzer, Clade.io). The system features a flow-through transmission cell made of CaF_2 and a pathlength of approximately $8\text{ }\mu\text{m}$. Microfluidic sample delivery, sample exchange, and cleaning procedures are automatized, based on previously established workflows. The IR spectra were acquired with a resolution of 4 cm^{-1} . On average, one measurement within a batch lasts 3.5 minutes, including the time taken for sample injection and cleaning procedures in between samples.

2.4 Machine learning and computational frameworks

As described in Section 1.3, ML and multivariate methods play an important role in the domain of IR fingerprinting. They enable the transformation of the information-rich absorption spectra into meaningful characteristics that can be linked to phenotypic outcomes. In this section, a general overview of ML techniques, evaluation metrics, validation methods, and computational frameworks is provided. More detailed descriptions of the specific methods used in each study are provided within the respective publications included in the appendix.

2 Materials and methods

For context on data representation, the IR measurement output from the spectrometer consists of wavenumber-absorption value pairs. The wavenumber axis can be treated as implicit, allowing the IR spectrum of a single sample to be represented as a vector $\mathbf{x}$:

$$\mathbf{x} = [x_1, x_2, \dots, x_m] \in \mathbb{R}^m,$$

where m corresponds to the number of discrete absorption values (features) across the available wavenumber range. When analyzing a collection of n samples, the data can be represented as a matrix $\mathbf{X}$:

$$\mathbf{X} = \begin{bmatrix} -\mathbf{x}_1- \\ -\mathbf{x}_2- \\ \vdots \\ -\mathbf{x}_n- \end{bmatrix} \in \mathbb{R}^{n \times m}.$$

Each sample $\mathbf{x}_i$ is associated with one or more outputs of interest, denoted by y_i . These outputs are typically scalar values. However, in the case of multi-label outputs, it is represented as a vector $\mathbf{y}_i$. The outputs investigated within this thesis can be categorized as follows:

- $y_i \in \mathbb{R}$: a continuous value;
- $y_i \in \{0, 1\}$: a binary value;
- $y_i \in \{1, 2, \dots, k\}$: a multi-class value, where one out of k classes is selected;
- $\mathbf{y}_i \in \{0, 1\}^k$: multi-label values, represented as a vector of binary values where each element corresponds to one of k classes.

For a set of samples $\mathbf{X}$, the output is represented as a vector $\mathbf{y}$, or as a matrix $\mathbf{Y}$ in the case of multi-label outputs.

In supervised learning, the objective is to learn a mapping from an individual spectrum to the corresponding outcome $f : \mathbf{x}_i \rightarrow y_i$, or $f : \mathbf{x}_i \rightarrow \mathbf{y}_i$ in the case of a multi-label output. This mapping is derived from a *training dataset* of input-output pairs. After training, the learned model f is evaluated on unseen data to estimate its predictive performance and assess how well it generalizes to samples where the outcomes are unknown. For each input $\mathbf{x}_i$, the model generates a prediction $\hat{y}_i = f(\mathbf{x}_i)$.

This map can be derived through several different learning algorithms. Among the diverse learning algorithms, linear regression is one of the simpler methods of deriving the map to predict a continuous output. Linear regression has been employed in several analyses presented in this thesis [27, 28], albeit with a modification over ‘vanilla’ linear regression – *regularization*. IR spectral data contain noise that should not correlate to the true signal of interest, the spectral features observe a high degree of multicollinearity, are relatively high-dimensional (hundreds or thousands of features), and the sample size n in biological applications is typically small relative to the number of features m . These challenges often cause the learning procedure to overfit the training dataset or fail to properly learn the map. Regularization addresses this by introducing a penalty term that controls the complexity of the model, effectively balancing a trade-off between the model’s complexity and its fit to the training data [105]. In other terms, regularization helps the model utilize information from the most relevant features while reducing the influence of noise and redundant information that can cause model instability.

2 Materials and methods

In addition to linear regression, logistic regression has also been employed in various analyses presented in this thesis [27, 47, 50]. Logistic regression extends linear regression to predict binary outcomes by modeling the probability that a given input belongs to one of two classes. It uses the logistic function to ensure that predicted probabilities remain within the $[0, 1]$ interval. Similar to linear regression, logistic regression can be regularized to reduce the complexity of the model and combat overfitting.

Linear and logistic regressions are relatively simple learning algorithms compared to neural networks, ensemble methods (e.g., random forest), or kernel-based methods. However, in the context of deriving relationships between IR spectra and phenotypic outcomes, linear and logistic regression had consistently proven to be effective and appropriate, as the relationships observed in the data were predominantly linear. When more complex, non-linear methods (e.g., kernel-based approaches or ensembles) were employed, they typically approximated linear solutions without offering additional insights or significant improvements in predictive performance. As a result, simpler learning algorithms were preferred within the investigations performed in this thesis. Nevertheless, it should be noted that neural networks have found a potential fruitful application in the analysis of data from IR-based micro-spectroscopy – where the spectrally-derived chemical information is captured across the spatial distribution of a tissue, resulting in hyperspectral image data that benefit from the capacity of neural networks to model non-linear relationships [51, 106].

While logistic regression was initially developed to model binary outcomes, extensions of logistic regression to model multi-class outputs are possible. One common extension is multinomial logistic regression [107], which generalizes the binary framework to handle multiple categories by modeling the probability of each class relative to a baseline category. This is achieved through separate sets of coefficients for each non-baseline class, allowing the model to capture class-specific relationships between the predictors and outcomes. Multinomial logistic regression is well-suited for problems where the outcome is categorical with no inherent order among the classes. In one of the studies presented in this thesis [50], multinomial logistic regression was effectively applied to investigate multi-class outcomes.

Further extension of binary classifiers to multi-label outputs can be achieved by creating several binary classifiers, each tasked with predicting the presence or absence of a specific label. This approach, often referred to as problem transformation, includes methods like binary relevance, which treats each label independently, and more sophisticated techniques like classifier chains. Classifier chains model dependencies between labels by training a sequence of binary classifiers, where the predictions of previous classifiers are included as additional features for subsequent ones [108]. This approach accounts for correlations between labels, improving predictive performance compared to treating each label independently. Classifier chaining has been successfully employed in one of the studies presented in this thesis [27].

Model testing and the calculation of prediction error metrics are critical aspects of the predictive workflow. As discussed in Section 1.3, error metrics must be computed on fully unseen samples – those not used during model training and not involved in model selection or hyperparameter tuning – to ensure an unbiased estimate of the model’s performance. Ideally, these test samples should account for domain-specific variances, such as time-aware data splits in cases where data exhibit temporal

2 Materials and methods

shifts (e.g., based on sample collection time or measurement time) or, if possible, use external datasets (e.g., samples from independent clinical studies). Furthermore, if the task is to examine the diagnostic performance of a phenotype detection model on an unseen participant population, care must be taken to ensure that the test data does not include individuals overlapping with the training data. This consideration is relevant when multiple measurements from the same individual are present in the dataset – e.g., due to sample re-measurements or multiple sample collections from the same individuals. These strategies aim to better reflect the model’s capacity to generalize to measurements collected under real-world deployment settings. In all studies included in this thesis, all phenotype classification models were tested exclusively on individuals not used for training. Furthermore, the model selection process, including the tuning of the learning models (namely the strength of the regularization parameter), was always performed using internal data validation splits. This process involved splitting the training set into cross-validation folds to tune a learning model, then retraining a new model on the entire available training split of the data using the tuning parameters that performed best on the internal validation splits. Finally, the trained model was tested on held-out test sets. Temporal test splits were also considered in the classification models when possible to ensure robustness to potential measurement drifts [27, 50].

The prediction error metrics used to quantify a model’s performance on unseen samples depend on the nature of the prediction task. Across the investigations in this thesis, a range of metrics have been applied. For binary outputs, the receiver operating characteristic (ROC) curve and the area under the curve (AUC) were heavily utilized [27, 47, 50]. These metrics allowed for the examination of both sensitivity and specificity, with the AUC serving as a single summarizing value to assess the model’s overall discriminative capacity. For multi-class outputs, accuracy was the primary evaluation metric, noting that the class distributions in these investigations were well-balanced [50]. For multi-label outputs, several metrics were employed to capture the complexity of predictions, including the exact match ratio, Hamming score, and Hamming loss [27]. For continuous outputs, regression metrics such as the coefficient of determination (R^2) and root mean squared error (RMSE) were used [27, 28].

All data analyses in the studies presented in this thesis were performed in a Python environment (version 3.8.8). The open-source Scikit-learn package (version 0.24.1) was used for several implementations of the above-discussed concepts.

Beyond the predictive modeling, the two tools developed within this thesis have been demonstrated to aid the analysis of spectral data. The first is an *in silico* spectral simulation tool, which allows for the generation of simulated IR spectra tailored to different experimental designs [47]. This tool is accessible through a dedicated repository (<https://github.com/tarek-eissa/in-silico-spectral-generator>). The second tool, CODI (contextual out-of-distribution integration), is a general-purpose framework designed to take a seed training dataset and augment it with additional data variance [50]. CODI has demonstrated that this method of training improves the generalization capacity of predictive models when the initial seed training dataset is limited in size and representativeness. CODI is available both as a standalone repository (<https://github.com/tarek-eissa/codi>) and as a Python package (pycodi).

Descriptions of these two tools, their use cases, and theoretical background can be found in the associated publications and GitHub repositories [47, 50].

2.5 Ethics statements

All samples collected within the framework of the four clinical studies (Section 2.1) utilized in this thesis adhere to the relevant ethical regulations.

The Lasers4Life-LG study was approved by the Ethics Committee of the Ludwig-Maximilian-University (LMU) of Munich under research study protocol #17-532. The BioPersMed study was approved by the Ethics Committee of the Medical University of Graz, Austria (EC Nr. 24-224 ex 11/12; project application number 4008_22). The KORA-F4 and KORA-FF4 study methods were approved by the ethics committee of the Bavarian Chamber of Physicians, Munich (EC No. 06068). The Lasers4Life-Cancer study was approved by the Ethics Committee of the LMU of Munich under research study protocols #17-141 and #17-182, and is registered at the German Clinical Trials Register (ID DRKS00013217). All analyzed participants enrolled in the four above-listed studies provided written informed consent.

3 Summary of publications

3.1 Plasma infrared fingerprinting with machine learning enables single-measurement multi-phenotype health screening

Authors:

Tarek Eissa, Cristina Leonardo, Kosmas V. Kepesidis, Frank Fleischmann, Birgit Linkohr, Daniel Meyer, Viola Zoka, Marinus Huber, Liudmila Voronina, Lothar Richter, Annette Peters, and Mihaela Žigman.

Published:

Cell Reports Medicine, Volume 5, Issue 7, July 2024.

Full text version:

Enclosed as Appendix A.

Preface:

This publication was guided by the following research question (RQ #1):

Can IR fingerprinting, combined with ML, reliably detect and predict complex physiological conditions within a large, naturally heterogeneous patient population?

Prior to this work, several proof-of-concept studies had demonstrated the potential of IR fingerprinting for health diagnostics. However, these studies were typically limited in sample size (often fewer than 100 individuals), focused on a single condition versus a narrow control group, and were conducted over short timeframes without accounting for real-world technical variabilities. While valuable for initial validation, such controlled setups do not reflect the complexities of practical applications. Human populations are inherently diverse – marked by physiological variability, co-occurring health conditions, and broad age ranges. Moreover, real-world studies must contend with unavoidable variations in sample handling, as samples are collected and measured over multiple years. Motivated by these gaps and the promise of IR fingerprinting, we conducted the largest study of its kind within a population-based cohort, analyzing thousands of plasma samples collected across multiple years. Our goal was to determine whether IR fingerprinting could move beyond isolated demonstrations and serve as a robust, scalable tool for broad screening of common health conditions.

Summary:

Minimally invasive diagnostic tools have long been desirable for efficiently identifying health risks and chronic conditions across diverse human populations. Although laboratory clinical chemistry panels are widely used for this purpose, they require separate assays for each biomarker, making them time- and resource-intensive and limiting their utility to the targeted biomarkers. Biofluid-based IR fingerprinting has emerged as a promising alternative, offering a single-measurement approach capable of providing health-related insights. Previous proof-of-concept studies examining IR fingerprinting were generally limited by small sample sizes, were performed in controlled case-control study setups, and did not account for the complexity of real-world populations, where multimorbid conditions are especially common among older adults. In

3 Summary of publications

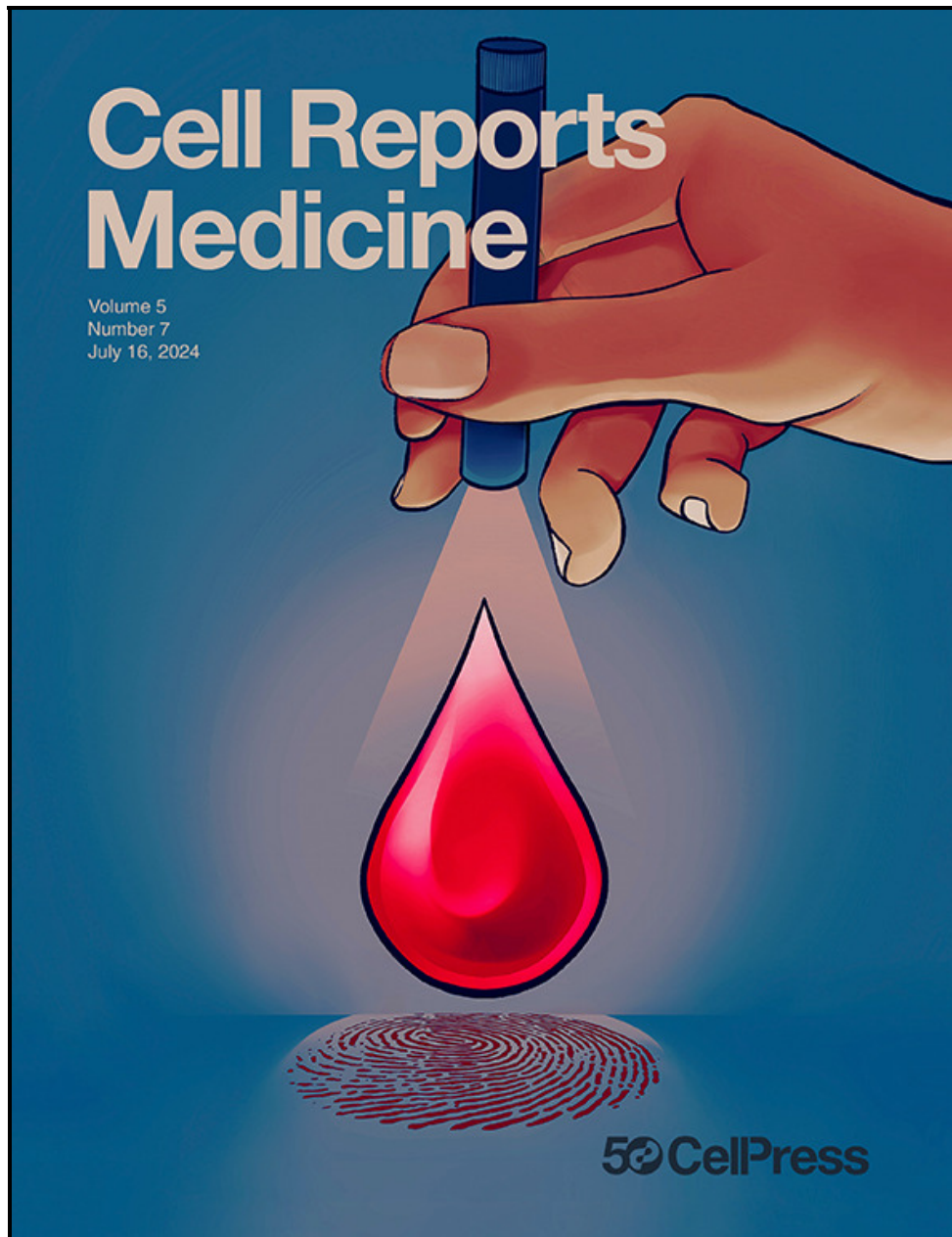
this study, we analyzed 5184 blood plasma samples collected from 3169 individuals within the population-based Cooperative Health Research in the Region of Augsburg (KORA) longitudinal cohort – a phenotypically diverse potential patient population. This represents the largest cohort profiled with IR fingerprinting to date. Our goal was to evaluate the capacity of this approach to comprehensively detect several medically relevant human health phenotypes, with a particular focus on metabolic disorders and associated risk factors. Our methodology involved capturing molecular information provided by FTIR spectroscopy, combined with ML analyses – including binary, multilabel, and regression models – to translate IR fingerprints into actionable medical insights. The study design included samples collected and measured years apart, ensuring robust model predictions despite variations in sample handling, sample storage times, and measurement intervals.

Our results demonstrate that IR molecular fingerprints provide a robust foundation for simultaneous health screening. Using a multilabel classification model designed to differentiate between dyslipidemia, hypertension, prediabetes, type 2 diabetes, healthy states (i.e., the absence of these conditions), and their potential combinations, we found that the approach can effectively characterize several multimorbid states. Impressive individual binary classification efficacies were observed across all conditions, with ROC-AUC values of 0.76 for prediabetes, 0.94 for type 2 diabetes, 0.77 for hypertension, 0.97 for dyslipidemia, and 0.90 for identifying healthy individuals. Additionally, we discovered that IR fingerprints can predict the future onset of metabolic syndrome (MetS), a progressive condition marked by multiple metabolic risk factors, with an ROC-AUC of 0.77, forecasting its occurrence approximately 6.5 years prior to clinical diagnosis. Moreover, we demonstrated that IR fingerprints encode anthropometric parameters such as gender, BMI, and age, yielding an ROC-AUC of 0.96 for gender classification and R^2 values of 0.43 and 0.58 for BMI and age estimation, respectively. Strong correlations ($R^2 \geq 0.91$) were observed between IR spectral features and key clinical analytes such as triglycerides, HDL cholesterol, LDL cholesterol, and glucose, underscoring the physiological and diagnostic significance of this approach.

Overall, this study establishes IR fingerprinting as a minimally invasive diagnostic tool for tackling complex multimorbidities in adult populations. The capacity to detect multiple chronic conditions, forecast their onset, and link spectral data to known biomarkers position the approach as a feasible alternative to traditional diagnostics. Its affordability, speed (~ 3.5 min per measurement), and lack of pre-analytical sample preparation make the approach particularly appealing for widespread use. Notably, the power to identify preclinical states like prediabetes and pre-MetS underscores its potential for early intervention and preventing or delaying severe diseases (e.g., cardiovascular conditions). While further external validation across diverse populations can bring the approach closer to application, our study provides a strong foundation for integrating IR fingerprinting into routine diagnostic practices.

Highlight:

This publication was chosen to be featured on the journal issue’s front cover.



Cover art by Sama Aljarhi (www.samaalj.com).

Copied with permission (Appendix E).

Contributions:

I formulated the main outline of the manuscript. I designed and performed all data analyses, visualizations, and interpretations. Mihaela Žigman and I wrote the manuscript, with all authors contributing to revising. Mihaela Žigman and I are corresponding authors.

3.2 The perils of molecular interpretations from vibrational spectra of complex samples

Authors:

Tarek Eissa, Liudmila Voronina, Marinus Huber, Frank Fleischmann, and Mihaela Žigman.

Published:

Angewandte Chemie Int. Ed., Volume 63, Issue 50, November 2024.

Full text version:

Enclosed as Appendix B.

Preface:

This publication was guided by the following research question (RQ #2):

What are the key considerations and best practices for interpreting IR spectra of molecularly complex samples, particularly those derived from blood-based media and their molecular components?

Within our research group, we often found ourselves repeatedly facing the same dilemma: colleagues, referees, and fellow spectroscopists would ask for a straightforward mapping between the IR signals we observe and their precise molecular origins. Yet each time, we realized that this question – though natural – was often misguided, as it demands oversimplifications that neither reflect the intricacies of complex biological matrices nor the real strengths of IR fingerprinting. Our collective drive to clarify these recurring misunderstandings led us to bring together the insights and expertise of the authors listed above. We wrote this *Scientific Perspective* article with the goal to clarify common misconceptions in the field and offer a more rigorous and pragmatic framework for IR spectral analysis.

Summary:

We have demonstrated that IR fingerprinting exhibits remarkable efficacy in characterizing the holistic biochemical makeup of complex samples in a single measurement, enabling innovative applications in biomedicine. This already marks a significant advancement beyond the traditional uses of IR spectroscopy, which were primarily limited to analyzing the structure of individual organic compounds or identifying contaminant chemicals in purified substances. Despite this advancement, significant challenges persist in interpreting IR spectra of complex multi-molecular mixtures, such as blood derivatives. Unfortunately, overlapping absorption peaks and interferences between molecular components often hinder the attribution of spectral signals to specific molecules or even broad molecular classes (e.g., carbohydrates, lipids, or proteins). Thus, while IR spectroscopy excels in identifying chemical functional groups and can sensitively capture physiological states, its capacity for molecular identification is limited when analyzing complex biological specimens. Nevertheless, there remains a pervasive desire to link spectral signals to specific molecular substances

3 Summary of publications

in clinical applications, and misinterpretations of IR spectra are widespread across peer-reviewed literature.

In this perspective, we critically address the challenges of analyzing multi-molecular spectra, offering evidence-based guidelines and best practices to improve the broader understanding and application of IR spectroscopy in complex systems. We demonstrate that IR spectral regions commonly attributed to specific molecular classes, such as lipids or carbohydrates, are often dominated by signals from abundant proteins like human serum albumin (HSA). Computational reconstructions further reveal that spectral signals frequently arise from combinations of multiple, functionally-independent molecules, further hindering the capacity to attribute observed signals to specific biomarkers. Another common misconception in the field is the tendency to link measured spectral signals to low-abundance molecules (e.g., DNA or RNA) despite their concentrations typically being far below the detection limits of current IR technology. Moreover, we demonstrate how chemical modifications that molecules undergo in their life-cycle (e.g., formation of covalent bonds) and common measurement preprocessing steps (e.g., normalization) can distort spectral signals and compromise molecular interpretations.

Overall, our findings underscore the need for rigorous methodological approaches and caution in assigning spectral features to molecular substances. To enhance the interpretability of IR spectra, we recommend integrating orthogonal techniques like mass spectrometry or nuclear magnetic resonance (NMR) spectroscopy or reducing sample complexity through chemical fractionation. Multivariate statistical approaches, such as ridge regression, can also effectively decode the dependencies between spectral features and prove highly effective in extracting relevant molecular information, even when individual spectral features are heavily overshadowed. By addressing these limitations, we aim to improve the reliability and applicability of IR fingerprinting in biomedical applications, positioning it as a high-throughput diagnostic or screening tool rather than a molecularly-specific method.

Highlight:

The publication is also available in German (10.1002/ange.202411596).

Contributions:

I significantly contributed to formulating the outline of the manuscript. I designed and performed the majority of the formal analysis and data visualizations. I wrote the larger portion of the manuscript, with all authors contributing to writing and revising.

3.3 Limits and prospects of molecular fingerprinting for phenotyping biological systems revealed through *in silico* modeling

Authors:

Tarek Eissa, Kosmas V. Kepesidis, Mihaela Zigman, and Marinus Huber.

Published:

Analytical Chemistry, Volume 95, Issue 16, April 2023.

Full text version:

Enclosed as Appendix C.

Preface:

This publication was guided by the following research question (RQ #3):

How do technical measurement, biological, and experimental study design factors affect the diagnostic capacity of IR fingerprinting, and what approaches can be adopted to refine them for improved diagnostic performance?

In our ongoing efforts to employ IR fingerprinting as a practical diagnostic tool, we found ourselves returning to a fundamental question: what truly restricts the method’s diagnostic performance and how severe are those limitations? We already recognized that the biological variability between different individuals was a major hindrance, but to what extent? And could improvements in spectrometer instrumentation make a meaningful difference, or is the technology already approaching its theoretical limits? To address these questions, we turned to computational simulations, developing an *in silico* model to systematically understand the impact of the key factors that limit the technology’s diagnostic potential.

Summary:

The strengths – and limitations – of IR fingerprinting have become apparent by now. The technique yields a wealth of molecular information, which can nicely be uncovered through ML and multivariate analysis to enable high-throughput health diagnostics. However, the interplay between factors like the biological variability of the analyzed cohort of samples, technical noise inherent to the IR measurement, biochemical complexity of the analyzed specimens, and cohort sizes used to train ML models can all collectively impact the diagnostic potential of the approach. These factors can be addressed through controlled experimental designs, improved instrumentation, and focused applications in carefully planned studies. Yet, collecting large, physical datasets to investigate the impact of these parameters presents a limitation, as it requires a large amount of time and resources.

In this study, we developed an *in silico* model capable of simulating realistic IR molecular fingerprints under controlled conditions, providing a systematic means to explore the effects of biological variability, sample chemical complexity, and measurement noise. Using our *in silico* model, we targeted the application of detecting lung

3 Summary of publications

and prostate cancers through IR fingerprinting as an experimental paradigm. The model revealed that inter-individual biological variability is a dominant factor limiting the cancer classification performance, with even slight reductions leading to substantial classification improvements. For instance, reducing the biological variability from an inter-person to an intra-person level, as observed in longitudinal self-referencing scenarios, increased the binary classification efficiency of detecting lung cancer from an ROC-AUC of 0.88 to 0.98. This provided the first quantitative evidence demonstrating the potential IR fingerprinting in personalized, longitudinal health-monitoring settings to improve diagnostic capacity. Reducing technical measurement noise had a secondary yet significant impact, demonstrating that advancements in IR spectroscopic technology can further improve classification efficiency, especially under conditions of high biological variability. Our study also examined the hindering impact of the molecular complexity of cell-free blood, demonstrating that chemically fractionating the samples could mitigate some of the limitations posed by analyzing biological matrices.

This study underscores the utility of computational modeling in advancing molecular fingerprinting for health diagnostics. The developed model provides a platform to simulate experimental conditions and guide study designs without requiring large-scale physical sampling to model diverse experimental settings.

Additional material:

<https://www.github.com/tarek-eissa/in-silico-spectral-generator>

Contributions:

I implemented the idea, performed the research, data analyses and visualizations with assistance from Marinus Huber. Marinus Huber and I wrote the manuscript, with all authors contributing to revising.

3.4 CODI: Enhancing machine learning-based molecular profiling through contextual out-of-distribution integration

Authors:

Tarek Eissa, Marinus Huber, Barbara Obermayer-Pietsch, Birgit Linkohr, Annette Peters, Frank Fleischmann, and Mihaela Žigman.

Published:

PNAS Nexus, Volume 3, Issue 10, October 2024.

Full text version:

Enclosed as Appendix D.

Preface:

This publication was guided by the following research question (RQ #4):

What strategies can address the challenges of standardization, out-of-distribution data, and generalization in ML modeling to improve model transfer across naturally variable molecular fingerprint datasets?

Scientific progress is often driven by unexpected findings, and this study is no exception. It is well known that data noise, variability, and distributional shifts hinder the performance of supervised ML models. This is why considerable effort goes into standardizing data – through both careful experimental design and computational methods aimed at minimizing variability. Here, we took a different, somewhat counterintuitive approach: what if, instead of further refining standardization techniques, we deliberately introduced *more* variation into a given training dataset using external data sources? To our surprise, this strategy worked far better than we anticipated. Even more exciting, we are now seeing that this strategy extends beyond IR fingerprinting, showing promise for other molecular fingerprinting techniques. This unexpected success highlights a broader potential for improving ML generalization in molecular fingerprinting applications, and we hope it inspires further exploration of this strategy.

Summary:

Our assessments thus far reveal the significant potential of combining IR fingerprinting with ML analysis to advance health diagnostics. However, as with essentially any ML-driven analyses, success fundamentally depends on the quality and representativeness of the training data. Specifically, supervised ML algorithms typically operate under the assumption that the train and test data are independent and *identically distributed* (i.i.d). Yet, this assumption is frequently violated in real-world applications. Out-of-distribution (OOD) measurements are almost inevitable and can stem from biological factors such as variability across age or ethnic groups, as well as analytical factors like sample handling inconsistencies or measurement device operating conditions. These distributional shifts hinder a trained model’s capacity to generalize to certain measurement events encountered in practice. This challenge is particularly pronounced in molecular profiling applications, where collecting biological specimens

3 Summary of publications

and representative experimental datasets is highly resource-intensive. In this study, we develop a new method: Contextual Out-of-Distribution Integration (CODI). The concept of CODI is to incorporate information about possible OOD variations into a given seed training dataset. This procedure augments the training dataset with an arbitrarily large number of simulated samples that better capture the true data distribution. By modeling and integrating diverse sources of data variability, CODI increases the likelihood that a trained ML model generalizes to samples beyond the seed training data, minimizing the need for extensive experimental data collection.

To assess CODI’s efficacy, we applied the newly developed framework to IR molecular fingerprints obtained from three independent longitudinal clinical studies with follow-up periods of up to eight years, as well as a cancer case-control study detecting four common cancers (lung, prostate, bladder, and breast). By incorporating sources of variability such as biological variations, sampling handling operators, sample freezer-storage effects, and instances of measurement device maintenance, the framework enriched training datasets with realistic distribution shifts. In longitudinal molecular monitoring, CODI improved the accuracy of identifying individuals from their molecular profiles over multi-year follow-ups, achieving up to a 50% improvement in classification accuracy compared to models trained on unrepresentative experimental data. Similarly, across different cancer diagnostic settings, CODI improved the capacity of ML generalization, enabling a classifier to maintain its accuracy when applied to measurements collected years after the training measurements. In another proof-of-concept application, CODI enabled the training of a classifier on plasma-based fingerprints, which was successfully applied to serum-based fingerprints, demonstrating the approach’s potential for domain adaptation despite considerable distributional shifts between train and test data.

Overall, the method developed and validated in this study showcases CODI as a framework to address generalization and model transfer challenges in molecular profiling applications. These findings highlight the importance of utilizing domain-specific knowledge to simulate realistic data distributions. Importantly, the concept of CODI is highly extensible, enabling the same approach (and code implementation) to be applied to a variety of data modalities beyond IR fingerprinting.

Additional material:

<https://www.github.com/tarek-eissa/codi>

Contributions:

I conceived the project, designed and performed the research, implemented the idea, and analyzed the data with assistance from Marinus Huber and Mihaela Žigman. I formulated the outline of the manuscript. Mihaela Žigman and I wrote the manuscript, with all authors contributing to revising. Mihaela Žigman and I are corresponding authors.

4 Concluding remarks

4 Concluding remarks

This thesis has explored four interrelated research questions, each advancing IR fingerprinting as a diagnostic tool. We have demonstrated that (1) IR fingerprinting combined with ML can robustly detect and predict complex physiological conditions in large, diverse populations; (2) accurate interpretation of IR spectra in biological samples requires a rigorous methodological framework to avoid common misconceptions; (3) computational modeling offers crucial insights into factors influencing diagnostic performance – revealing strategies to mitigate biological variability, technical noise, and sample complexity; and (4) integrating simulated variability through the CODI framework improves model generalization, enabling reliable ML-driven diagnostics across diverse datasets and real-world conditions. Collectively, these findings establish that a single, molecule-agnostic measurement can encode extensive biochemical information relevant to multiple physiological states. Beyond establishing feasibility, this work provides a well-defined technological paradigm for how IR fingerprinting can enable efficient, high-throughput phenotype screening and diagnostics. By validating our approach across several large-scale clinical studies and addressing key methodological challenges, this work brings IR fingerprinting closer to clinical implementation. The evidence presented suggests that this technology is no longer just a promising exploratory concept but a viable diagnostic technological platform ready for prospective testing in clinical settings.

A critical step in implementing this technology is determining which clinical applications would benefit most from its capacities. We revealed that IR fingerprinting is effective for both acute, high-mortality diseases, such as lung cancer, as well as physiologically milder chronic conditions (e.g., type-2 diabetes, hypertension, and dyslipidemia) that often precede severe health complications. In oncology, IR fingerprinting could serve as a rapid and minimally invasive pre-screening tool for triaging, thus reducing the need for costly and painful tissue biopsies that involve time-consuming histopathological analyses. Meanwhile, the ability of IR spectroscopic profiling to detect multiple metabolic risk factors in a single measurement suggests a fruitful role in routine health checkups, enabling early intervention before the onset of severe disease (e.g., cardiovascular complications). This dual capacity – supporting both specialized diagnostics and general preventive screening – positions IR fingerprinting as a valuable addition to modern molecular healthcare.

To sustain the momentum established in this work, future research must address further technical and logistical challenges. The collection of larger and more representative experimental datasets will be needed to ensure that any application of this technology is accurate and reliable across different patient populations and clinical settings. Equally important is the development of computational strategies that can help mitigate and handle challenges of real-world variability in measurements and clinical conditions. Techniques like CODI, developed in this thesis, provide a pathway toward addressing the challenges of data variability, helping to maintain the predictive efficacy of trained ML models across diverse, practical ‘measurement events.’ Ultimately, it is becoming increasingly apparent that realizing the vision of IR fingerprinting as a new platform technology for human health profiling requires collaborative efforts beyond the scope of any single discipline. Bringing together expertise in chemistry, physics, bioinformatics, systems biology, and molecular medicine will certainly be the key to

4 Concluding remarks

refining analytical physicochemical methods and pushing the technology beyond the basic research setting.

Looking ahead, I believe that a particularly promising direction for IR fingerprinting is its application in resource-limited healthcare settings. The ability to rapidly and affordably screen for multiple health conditions with a single measurement could profoundly impact regions where access to medical diagnostics remains limited. By offering a scalable solution for preventive healthcare and stratifying at-risk populations for developing certain diseases, this technology and framework have the potential to improve health equity and contribute to more data-driven, proactive medical care on a global scale. Future efforts should focus on translating these findings into practical deployment strategies for various healthcare infrastructures, ensuring that this promising diagnostic approach reaches the populations that need it most.

Bibliography

- [1] C. López-Otín and G. Kroemer, “Hallmarks of health,” *Cell*, vol. 184, p. 33–63, Jan. 2021.
- [2] M. Civelek and A. J. Lusis, “Systems genetics approaches to understand complex traits,” *Nature Reviews Genetics*, vol. 15, p. 34–48, Dec. 2013.
- [3] M. Babu and M. Snyder, “Multi-omics profiling for health,” *Molecular and Cellular Proteomics*, vol. 22, p. 100561, June 2023.
- [4] Y. Song, Y.-Y. Huang, X. Liu, X. Zhang, M. Ferrari, and L. Qin, “Point-of-care technologies for molecular diagnostics using a drop of blood,” *Trends in Biotechnology*, vol. 32, p. 132–139, Mar. 2014.
- [5] E. W. Deutsch, G. S. Omenn, Z. Sun, M. Maes, M. Pernemalm, K. K. Palaniappan, N. Letunica, Y. Vandenbrouck, V. Brun, S.-c. Tao, X. Yu, P. E. Geyer, V. Ignjatovic, R. L. Moritz, and J. M. Schwenk, “Advances and utility of the human plasma proteome,” *Journal of Proteome Research*, vol. 20, p. 5241–5263, Oct. 2021.
- [6] H. J. Issaq, Z. Xiao, and T. D. Veenstra, “Serum and plasma proteomics,” *Chemical Reviews*, vol. 107, p. 3601–3620, July 2007.
- [7] M. Plebani, G. Banfi, S. Bernardini, F. Bondanini, L. Conti, R. Dorizzi, F. E. Ferrara, R. Mancini, and T. Trenti, “Serum or plasma? an old question looking for new answers,” *Clinical Chemistry and Laboratory Medicine (CCLM)*, vol. 58, p. 178–187, Sept. 2019.
- [8] W.-M. Gao, R. Kuick, R. P. Orzechowski, D. E. Misek, J. Qiu, A. K. Greenberg, W. N. Rom, D. E. Brenner, G. S. Omenn, B. B. Haab, and S. M. Hanash, “Distinctive serum protein profiles involving abundant proteins in lung cancer patients based upon antibody microarray analysis,” *BMC Cancer*, vol. 5, Aug. 2005.
- [9] T. Buergel, J. Steinfeldt, G. Ruyoga, M. Pietzner, D. Bizzarri, D. Vojinovic, J. Upmeyer zu Belzen, L. Loock, P. Kittner, L. Christmann, *et al.*, “Metabolomic profiles predict individual multidisease outcomes,” *Nature medicine*, vol. 28, no. 11, pp. 2309–2320, 2022.
- [10] M. J. Chapman, H. N. Ginsberg, P. Amarenco, F. Andreotti, J. Borén, A. L. Catapano, O. S. Descamps, E. Fisher, P. T. Kovanen, J. A. Kuivenhoven, P. Lesnik, L. Masana, B. G. Nordestgaard, K. K. Ray, Z. Reiner, M.-R. Taskinen, L. Tokgözoğlu, A. Tybjaerg-Hansen, G. F. Watts, and for the European Atherosclerosis Society Consensus Panel, “Triglyceride-rich lipoproteins and high-density

BIBLIOGRAPHY

- lipoprotein cholesterol in patients at high risk of cardiovascular disease: evidence and guidance for management,” *European Heart Journal*, vol. 32, pp. 1345–1361, Apr. 2011.
- [11] E. J. Topol, “The revolution in high-throughput proteomics and AI,” *Science*, vol. 385, Sept. 2024.
 - [12] Y. Hasin, M. Seldin, and A. Lusis, “Multi-omics approaches to disease,” *Genome Biology*, vol. 18, May 2017.
 - [13] U.-P. Rohr, C. Binder, T. Dieterle, F. Giusti, C. G. M. Messina, E. Toerien, H. Moch, and H. H. Schäfer, “The value of in vitro diagnostic testing in medical practice: A status report,” *PLOS ONE*, vol. 11, p. e0149856, Mar. 2016.
 - [14] R. McPherson and M. Pincus, *Henry’s Clinical Diagnosis and Management by Laboratory Methods*. Elsevier Health Sciences, 2017.
 - [15] M. Zhi, E. L. Ding, J. Theisen-Toupal, J. Whelan, and R. Arnaout, “The landscape of inappropriate laboratory testing: A 15-year meta-analysis,” *PLoS ONE*, vol. 8, p. e78962, Nov. 2013.
 - [16] B. Vrijzen, C. Naaktgeboren, L. Vos, W. van Solinge, H. Kaasjager, and M. ten Berg, “Inappropriate laboratory testing in internal medicine inpatients: Prevalence, causes and interventions,” *Annals of Medicine and Surgery*, vol. 51, p. 48–53, Mar. 2020.
 - [17] J. Carrasco-Zanini, M. Pietzner, J. Davitte, P. Surendran, D. C. Croteau-Chonka, C. Robins, A. Torralbo, C. Tomlinson, F. Grünschläger, N. Fitzpatrick, C. Ytsma, T. Kanno, S. Gade, D. Freitag, F. Ziebell, S. Haas, S. Denaxas, J. C. Betts, N. J. Wareham, H. Hemingway, R. A. Scott, and C. Langenberg, “Proteomic signatures improve risk prediction for common and rare diseases,” *Nature Medicine*, vol. 30, p. 2489–2498, July 2024.
 - [18] T. G. Meikle, K. Huynh, C. Giles, and P. J. Meikle, “Clinical lipidomics: realizing the potential of lipid profiling,” *Journal of Lipid Research*, vol. 62, p. 100127, 2021.
 - [19] C. Seger and L. Salzmann, “After another decade: LC–MS/MS became routine in clinical diagnostics,” *Clinical Biochemistry*, vol. 82, p. 2–11, Aug. 2020.
 - [20] A.-H. M. Emwas, “The strengths and weaknesses of NMR spectroscopy and mass spectrometry with particular focus on metabolomics research,” *Metabonomics: Methods and protocols*, pp. 161–193, 2015.
 - [21] P. E. Geyer, E. Voytik, P. V. Treit, S. Doll, A. Kleinhempel, L. Niu, J. B. Müller, M.-L. Buchholtz, J. M. Bader, D. Teupser, *et al.*, “Plasma proteome profiling to detect and avoid sample-related biases in biomarker studies,” *EMBO molecular medicine*, vol. 11, no. 11, p. e10427, 2019.
 - [22] A. Bravin, P. Coan, and P. Suortti, “X-ray phase-contrast imaging: from pre-clinical applications towards clinics,” *Physics in Medicine and Biology*, vol. 58, p. R1–R35, Dec. 2012.

BIBLIOGRAPHY

- [23] H. R. Dash, P. Shrivastava, and S. Das, *Quantification of DNA by Using UV-Visible Spectrophotometer*, p. 127–132. Springer US, 2020.
- [24] A. B. Biter, J. Pollet, W.-H. Chen, U. Strych, P. J. Hotez, and M. E. Bottazzi, “A method to probe protein structure from UV absorbance spectra,” *Analytical biochemistry*, vol. 587, p. 113450, 2019.
- [25] J. Green, *Handbook of Spectroscopy*, ch. 17, pp. 279–296. John Wiley & Sons, Ltd, 2003.
- [26] P. R. Griffiths, J. A. De Haseth, and J. D. Winefordner, *Fourier Transform Infrared Spectrometry*. Chemical Analysis: A Series of Monographs on Analytical Chemistry and Its Applications, Chichester, England: Wiley-Blackwell, 2 ed., Apr. 2007.
- [27] T. Eissa, C. Leonardo, K. V. Kepesidis, F. Fleischmann, B. Linkohr, D. Meyer, V. Zoka, M. Huber, L. Voronina, L. Richter, A. Peters, and M. Žigman, “Plasma infrared fingerprinting with machine learning enables single-measurement multi-phenotype health screening,” *Cell Reports Medicine*, vol. 5, p. 101625, July 2024.
- [28] T. Eissa, L. Voronina, M. Huber, F. Fleischmann, and M. Žigman, “The perils of molecular interpretations from vibrational spectra of complex samples,” *Angew. Chem. Int. Ed Engl.*, vol. 63, p. e202411596, Dec. 2024.
- [29] M. J. Baker, S. R. Hussain, L. Lovergne, V. Untereiner, C. Hughes, R. A. Lukaszewski, G. Thiéfin, and G. D. Sockalingum, “Developing and understanding biofluid vibrational spectroscopy: a critical review,” *Chem. Soc. Rev.*, vol. 45, pp. 1803–1818, Apr. 2016.
- [30] A. Barth and P. Haris, *Biological and Biomedical Infrared Spectroscopy*. Advances in biomedical spectroscopy, IOS Press, 2009.
- [31] K. M. McKinnon, “Flow cytometry: An overview,” *Current Protocols in Immunology*, vol. 120, Jan. 2018.
- [32] R. Grange, J. Thompson, and D. Lambert, “Radioimmunoassay, enzyme and non-enzyme-based immunoassays,” *British Journal of Anaesthesia*, vol. 112, p. 213–216, Feb. 2014.
- [33] K. Shah and P. Maghsoudlou, “Enzyme-linked immunosorbent assay (ELISA): the basics,” *British Journal of Hospital Medicine*, vol. 77, p. C98–C101, July 2016.
- [34] I. Pupeza, M. Huber, M. Trubetskov, W. Schweinberger, S. A. Hussain, C. Hofer, K. Fritsch, M. Poetzlberger, L. Vamos, E. Fill, T. Amotchkina, K. V. Kepesidis, A. Apolonski, N. Karpowicz, V. Pervak, O. Pronin, F. Fleischmann, A. Azzeer, M. Žigman, and F. Krausz, “Field-resolved infrared spectroscopy of biological systems,” *Nature*, vol. 577, p. 52–59, Jan. 2020.
- [35] A. Schwaighofer, M. Montemurro, S. Freitag, C. Kristament, M. J. Culzoni, and B. Lendl, “Beyond fourier transform infrared spectroscopy: External cavity quantum cascade laser-based mid-infrared transmission spectroscopy of proteins

BIBLIOGRAPHY

- in the amide i and amide ii region,” *Analytical Chemistry*, vol. 90, p. 7072–7079, May 2018.
- [36] R. Bhargava, “Infrared spectroscopic imaging: the next generation,” *Applied spectroscopy*, vol. 66, no. 10, pp. 1091–1120, 2012.
- [37] D. Rohleder, G. Kocherscheidt, K. Gerber, W. Kiefer, W. Köhler, J. Möcks, and W. Petrich, “Comparison of mid-infrared and raman spectroscopy in the quantitative analysis of serum,” *Journal of Biomedical Optics*, vol. 10, no. 3, pp. 031108–031108, 2005.
- [38] M. Huber, K. V. Kepesidis, L. Voronina, F. Fleischmann, E. Fill, J. Hermann, I. Koch, K. Milger-Kneidinger, T. Kolben, G. B. Schulz, F. Jokisch, J. Behr, N. Harbeck, M. Reiser, C. Stief, F. Krausz, and M. Zigman, “Infrared molecular fingerprinting of blood-based liquid biopsies for the detection of cancer,” *eLife*, vol. 10, Oct. 2021.
- [39] H. J. Butler, P. M. Brennan, J. M. Cameron, D. Finlayson, M. G. Hegarty, M. D. Jenkinson, D. S. Palmer, B. R. Smith, and M. J. Baker, “Development of high-throughput ATR-FTIR technology for rapid triage of brain cancer,” *Nature Communications*, vol. 10, Oct. 2019.
- [40] W. Petrich, K. B. Lewandrowski, J. B. Muhlestein, M. E. H. Hammond, J. L. Januzzi, E. L. Lewandrowski, R. R. Pearson, B. Dolenko, J. Früh, M. Haass, M. M. Hirschl, W. Köhler, R. Mischler, J. Möcks, J. Ordóñez-Llanos, O. Quarder, R. Somorjai, A. Staib, C. Sylven, G. Werner, and R. Zerback, “Potential of mid-infrared spectroscopy to aid the triage of patients with acute chest pain,” *The Analyst*, vol. 134, no. 6, p. 1092, 2009.
- [41] L. Voronina, F. Fleischmann, J. Šimunović, C. Ludwig, M. Novokmet, and M. Žigman, “Probing blood plasma protein glycosylation with infrared spectroscopy,” *Analytical Chemistry*, Feb. 2024.
- [42] S. Roy, D. Perez-Guaita, S. Bowden, P. Heraud, and B. R. Wood, “Spectroscopy goes viral: Diagnosis of hepatitis B and C virus infection from human sera using ATR-FTIR spectroscopy,” *Clinical Spectroscopy*, vol. 1, p. 100001, Dec. 2019.
- [43] M. S. Nogueira, A. L. Barreto, M. Furukawa, E. S. Rovai, A. Bastos, G. Bertin-cello, and L. F. d. C. e. S. d. Carvalho, “FTIR spectroscopy as a point of care diagnostic tool for diabetes and periodontitis: A saliva analysis approach,” *Photodiagnosis and Photodynamic Therapy*, vol. 40, p. 103036, Dec. 2022.
- [44] R. Bhargava, “Digital histopathology by infrared spectroscopic imaging,” *Annual Review of Analytical Chemistry*, vol. 16, p. 205–230, June 2023.
- [45] M. Diem, “Comments on recent reports on infrared spectral detection of disease markers in blood components,” *Journal of biophotonics*, vol. 11, no. 7, p. e201800064, 2018.
- [46] J. Coates *et al.*, “Interpretation of infrared spectra, a practical approach,” *Encyclopedia of analytical chemistry*, vol. 12, pp. 10815–10837, 2000.

BIBLIOGRAPHY

- [47] 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,” *Analytical Chemistry*, vol. 95, p. 6523–6532, Apr. 2023.
- [48] C. L. Song, M. Z. Vardaki, R. D. Goldin, and S. G. Kazarian, “Fourier transform infrared spectroscopic imaging of colon tissues: evaluating the significance of amide I and C–H stretching bands in diagnostic applications with machine learning,” *Analytical and Bioanalytical Chemistry*, vol. 411, p. 6969–6981, Aug. 2019.
- [49] M. Huber, K. V. Kepesidis, L. Voronina, M. Božić, M. Trubetskov, N. Harbeck, F. Krausz, and M. Žigman, “Stability of person-specific blood-based infrared molecular fingerprints opens up prospects for health monitoring,” *Nature Communications*, vol. 12, Mar. 2021.
- [50] T. Eissa, M. Huber, B. Obermayer-Pietsch, B. Linkohr, A. Peters, F. Fleischmann, and M. Žigman, “CODI: Enhancing machine learning-based molecular profiling through contextual out-of-distribution integration,” *PNAS Nexus*, vol. 3, Oct. 2024.
- [51] S. Berisha, M. Lotfollahi, J. Jahanipour, I. Gurcan, M. Walsh, R. Bhargava, H. Van Nguyen, and D. Mayerich, “Deep learning for FTIR histology: leveraging spatial and spectral features with convolutional neural networks,” *The Analyst*, vol. 144, no. 5, p. 1642–1653, 2019.
- [52] A. Geron, *Hands-on machine learning with scikit-learn, keras, and TensorFlow*. Sebastopol, CA: O’Reilly Media, 2 ed., Oct. 2019.
- [53] A. Burkov, *The hundred-page machine learning book*. Andriy Burkov, Jan. 2019.
- [54] W. Rathmann, B. Haastert, A. Icks, H. Löwel, C. Meisinger, R. Holle, and G. Giani, “High prevalence of undiagnosed diabetes mellitus in southern germany: Target populations for efficient screening. the KORA survey 2000,” *Diabetologia*, vol. 46, p. 182–189, Feb. 2003.
- [55] D. Bogdanet, P. O’Shea, C. Lyons, A. Shafat, and F. Dunne, “The oral glucose tolerance test—is it time for a change?—a literature review with an emphasis on pregnancy,” *Journal of clinical medicine*, vol. 9, no. 11, p. 3451, 2020.
- [56] S. Shilo, A. Keshet, H. Rossman, A. Godneva, Y. Talmor-Barkan, Y. Aviv, and E. Segal, “Continuous glucose monitoring and intrapersonal variability in fasting glucose,” *Nature Medicine*, vol. 30, p. 1424–1431, Apr. 2024.
- [57] R. H. Keogh and D. R. Cox, *Case-control studies*, vol. 4. Cambridge University Press, 2014.
- [58] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, “Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study,” *PLoS medicine*, vol. 15, no. 11, p. e1002683, 2018.

BIBLIOGRAPHY

- [59] J. P. Cohen, T. Cao, J. D. Viviano, C.-W. Huang, M. Fralick, M. Ghassemi, M. Mamdani, R. Greiner, and Y. Bengio, “Problems in the deployment of machine-learned models in health care,” *Cmaj*, vol. 193, no. 35, pp. E1391–E1394, 2021.
- [60] S. M. Batool, T. Hsia, A. Beecroft, B. Lewis, E. Ekanayake, Y. Rosenfeld, A. K. Escobedo, A. S. Gamblin, S. Rawal, R. J. Cote, M. Watson, D. T. Wong, A. A. Patel, J. Skog, N. Papadopoulos, C. Bettegowda, C. M. Castro, H. Lee, S. Srivastava, B. S. Carter, and L. Balaj, “Extrinsic and intrinsic preanalytical variables affecting liquid biopsy in cancer,” *Cell Reports Medicine*, vol. 4, p. 101196, Oct. 2023.
- [61] A. M. Hawkrigde and D. C. Muddiman, “Mass spectrometry-based biomarker discovery: toward a global proteome index of individuality,” *Annual Review of Analytical Chemistry*, vol. 2, no. 1, pp. 265–277, 2009.
- [62] N. Eling, M. D. Morgan, and J. C. Marioni, “Challenges in measuring and understanding biological noise,” *Nature Reviews Genetics*, vol. 20, p. 536–548, May 2019.
- [63] R. A. Bowen and A. T. Remaley, “Interferences from blood collection tube components on clinical chemistry assays,” *Biochemia medica*, vol. 24, no. 1, pp. 31–44, 2014.
- [64] H. Dvinge, R. E. Ries, J. O. Ilagan, D. L. Stirewalt, S. Meshinchi, and R. K. Bradley, “Sample processing obscures cancer-specific alterations in leukemic transcriptomes,” *Proceedings of the National Academy of Sciences*, vol. 111, no. 47, pp. 16802–16807, 2014.
- [65] C. L. Morais, K. M. Lima, M. Singh, and F. L. Martin, “Tutorial: multivariate classification for vibrational spectroscopy in biological samples,” *Nature Protocols*, vol. 15, no. 7, pp. 2143–2162, 2020.
- [66] S. Guo, C. Beleites, U. Neugebauer, S. Abalde-Cela, N. K. Afseth, F. Alsamad, S. Anand, C. Araujo-Andrade, S. Askrabic, E. Avci, *et al.*, “Comparability of raman spectroscopic configurations: a large scale cross-laboratory study,” *Analytical chemistry*, vol. 92, no. 24, pp. 15745–15756, 2020.
- [67] C. Soneson, S. Gerster, and M. Delorenzi, “Batch effect confounding leads to strong bias in performance estimates obtained by cross-validation,” *PloS one*, vol. 9, no. 6, p. e100335, 2014.
- [68] J. Liu, Z. Shen, Y. He, X. Zhang, R. Xu, H. Yu, and P. Cui, “Towards out-of-distribution generalization: A survey,” *arXiv*, 2021.
- [69] L. Adlung, Y. Cohen, U. Mor, and E. Elinav, “Machine learning in clinical decision making,” *Med*, vol. 2, p. 642–665, June 2021.
- [70] C. L. Ramspek, K. J. Jager, F. W. Dekker, C. Zoccali, and M. van Diepen, “External validation of prognostic models: what, why, how, when and where?,” *Clinical Kidney Journal*, vol. 14, p. 49–58, Nov. 2020.

BIBLIOGRAPHY

- [71] P. R. Griffiths, “Fourier transform infrared spectrometry,” *Science*, vol. 222, p. 297–302, Oct. 1983.
- [72] M. Brandstetter, L. Volgger, A. Genner, C. Jungbauer, and B. Lendl, “Direct determination of glucose, lactate and triglycerides in blood serum by a tunable quantum cascade laser-based mid-IR sensor,” *Applied Physics B*, vol. 110, pp. 233–239, 2013.
- [73] A. Weigel, P. Jacob, W. Schweinberger, M. Huber, M. Trubetskov, P. Karandušovský, C. Hofer, T. Buberl, T. Amotchkina, M. Högner, D. Hahner, P. Sulzer, A. Leitenstorfer, V. Pervak, F. Krausz, and I. Pupeza, “Dual-oscillator infrared electro-optic sampling with attosecond precision,” *Optica*, vol. 11, p. 726, May 2024.
- [74] J. M. Cameron, C. Rinaldi, S. H. Rutherford, A. Sala, A. G. Theakstone, and M. J. Baker, “Clinical spectroscopy: Lost in translation?,” *Applied Spectroscopy*, vol. 76, p. 393–415, Sept. 2021.
- [75] M. J. Baker, J. Trevisan, P. Bassan, R. Bhargava, H. J. Butler, K. M. Dorling, P. R. Fielden, S. W. Fogarty, N. J. Fullwood, K. a Heys, C. Hughes, P. Lasch, P. L. Martin-Hirsch, B. Obinaju, G. D. Sockalingum, J. Sulé-Suso, R. J. Strong, M. J. Walsh, B. R. Wood, P. Gardner, and F. L. Martin, “Using Fourier transform IR spectroscopy to analyze biological materials,” *Nature Protocols*, vol. 9, pp. 1771–91, 2014.
- [76] S. De Bruyne, M. M. Speeckaert, and J. R. Delanghe, “Applications of mid-infrared spectroscopy in the clinical laboratory setting,” *Critical Reviews in Clinical Laboratory Sciences*, vol. 55, p. 1–20, Dec. 2017.
- [77] D. J. Anderson, R. G. Anderson, S. J. Moug, and M. J. Baker, “Liquid biopsy for cancer diagnosis using vibrational spectroscopy: systematic review,” *BJS Open*, vol. 4, pp. 554–562, May 2020.
- [78] F. L. Martin, J. G. Kelly, V. Llabjani, P. L. Martin-Hirsch, I. I. Patel, J. Trevisan, N. J. Fullwood, and M. J. Walsh, “Distinguishing cell types or populations based on the computational analysis of their infrared spectra,” *Nature Protocols*, vol. 5, pp. 1748–1760, 2010.
- [79] Y. Li, L. Guo, L. Li, C. Yang, P. Guang, F. Huang, Z. Chen, L. Wang, and J. Hu, “Early diagnosis of type 2 diabetes based on near-infrared spectroscopy combined with machine learning and aquaphotomics,” *Frontiers in Chemistry*, vol. 8, Dec. 2020.
- [80] D. Townsend, M. Miljković, B. Bird, K. Lenau, O. Old, M. Almond, C. Kendall, G. Lloyd, N. Shepherd, H. Barr, N. Stone, and M. Diem, “Infrared microspectroscopy for cyto-pathological classification of esophageal cells,” *The Analyst*, vol. 140, no. 7, p. 2215–2223, 2015.
- [81] L. Voronina, C. Leonardo, J. B. Mueller-Reif, P. E. Geyer, M. Huber, M. Trubetskov, K. V. Kepesidis, J. Behr, M. Mann, F. Krausz, and M. Žigman, “Molecular origin of blood-based infrared spectroscopic fingerprints,” *Angewandte Chemie*, vol. 133, p. 17197–17206, May 2021.

BIBLIOGRAPHY

- [82] H. J. Byrne, I. Behl, G. Calado, O. Ibrahim, M. Toner, S. Galvin, C. M. Healy, S. Flint, and F. M. Lyng, “Biomedical applications of vibrational spectroscopy: Oral cancer diagnostics,” *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 252, p. 119470, May 2021.
- [83] J. M. Cameron, A. Sala, G. Antoniou, P. M. Brennan, H. J. Butler, J. J. A. Conn, S. Connal, T. Curran, M. G. Hegarty, R. G. McHardy, D. Orringer, D. S. Palmer, B. R. Smith, and M. J. Baker, “A spectroscopic liquid biopsy for the earlier detection of multiple cancer types,” *British Journal of Cancer*, vol. 129, p. 1658–1666, Sept. 2023.
- [84] H. J. Byrne, M. Baranska, G. J. Puppels, N. Stone, B. Wood, K. M. Gough, P. Lasch, P. Heraud, J. Sulé-Suso, and G. D. Sockalingum, “Spectropathology for the next generation: Quo vadis?,” *The Analyst*, vol. 140, no. 7, p. 2066–2073, 2015.
- [85] D. Finlayson, C. Rinaldi, and M. J. Baker, “Is infrared spectroscopy ready for the clinic?,” *Analytical Chemistry*, vol. 91, pp. 12117–12128, Sept. 2019.
- [86] A. L. Mitchell, K. B. Gajjar, G. Theophilou, F. L. Martin, and P. L. Martin-Hirsch, “Vibrational spectroscopy of biofluids for disease screening or diagnosis: translation from the laboratory to a clinical setting,” *Journal of Biophotonics*, vol. 7, p. 153–165, Mar. 2014.
- [87] M. Jackson, “From biomolecules to biodiagnostics: Spectroscopy does it all,” *Faraday Discussions*, vol. 126, p. 1, 2004.
- [88] O. Beckonert, H. C. Keun, T. M. D. Ebbels, J. Bundy, E. Holmes, J. C. Lindon, and J. K. Nicholson, “Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts,” *Nature Protocols*, vol. 2, p. 2692–2703, Oct. 2007.
- [89] M. K. Tuck, D. W. Chan, D. Chia, A. K. Godwin, W. E. Grizzle, K. E. Krueger, W. Rom, M. Sanda, L. Sorbara, S. Stass, W. Wang, and D. E. Brenner, “Standard operating procedures for serum and plasma collection: Early detection research network consensus statement standard operating procedure integration working group,” *Journal of Proteome Research*, vol. 8, p. 113–117, Dec. 2008.
- [90] H. L. Rehm, S. J. Bale, P. Bayrak-Toydemir, J. S. Berg, K. K. Brown, J. L. Deignan, M. J. Friez, B. H. Funke, M. R. Hegde, and E. Lyon, “ACMG clinical laboratory standards for next-generation sequencing,” *Genetics in Medicine*, vol. 15, p. 733–747, Sept. 2013.
- [91] L. Lovergne, J. Lovergne, P. Bouzy, V. Untereiner, M. Offroy, R. Garnotel, G. Thiéfin, M. J. Baker, and G. D. Sockalingum, “Investigating pre-analytical requirements for serum and plasma based infrared spectro-diagnostic,” *Journal of Biophotonics*, vol. 12, Aug. 2019.
- [92] A. Hackshaw, “Small studies: strengths and limitations,” *European Respiratory Journal*, vol. 32, p. 1141–1143, Oct. 2008.

BIBLIOGRAPHY

- [93] R. C. Deo, “Machine learning in medicine,” *Circulation*, vol. 132, no. 20, pp. 1920–1930, 2015.
- [94] L. Goetz, N. Seedat, R. Vandersluis, and M. van der Schaar, “Generalization—a key challenge for responsible AI in patient-facing clinical applications,” *npj Digital Medicine*, vol. 7, no. 1, p. 126, 2024.
- [95] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera, “A unifying view on dataset shift in classification,” *Pattern recognition*, vol. 45, no. 1, pp. 521–530, 2012.
- [96] Z. Obermeyer and E. J. Emanuel, “Predicting the future—big data, machine learning, and clinical medicine,” *New England Journal of Medicine*, vol. 375, no. 13, pp. 1216–1219, 2016.
- [97] R. Holle, M. Happich, H. Löwel, and H. Wichmann, “KORA - a research platform for population based health research,” *Das Gesundheitswesen*, vol. 67, p. 19–25, July 2005.
- [98] C. W. Haudum, E. Kolesnik, C. Colantonio, I. Mursic, M. Url-Michitsch, A. Tomaschitz, T. Glantschnig, B. Hutz, A. Lind, N. Schweighofer, C. Reiter, K. Ablasser, M. Wallner, N. J. Tripolt, E. Pieske-Kraigher, T. Madl, A. Springer, G. Seidel, A. Wedrich, A. Zirlik, T. Krahn, R. Stauber, B. Pieske, T. R. Pieber, N. Verheyen, B. Obermayer-Pietsch, and A. Schmidt, “Cohort profile: ‘biomarkers of personalised medicine’ (BioPersMed): a single-centre prospective observational cohort study in Graz/Austria to evaluate novel biomarkers in cardiovascular and metabolic diseases,” *BMJ Open*, vol. 12, p. e058890, Apr. 2022.
- [99] P. S. Jellinger, Y. Handelsman, P. D. Rosenblit, Z. T. Bloomgarden, V. A. Fonseca, A. J. Garber, G. Grunberger, C. K. Guerin, D. S. Bell, J. I. Mechanick, R. Pessah-Pollack, K. Wyne, D. Smith, E. A. Brinton, S. Fazio, M. Davidson, P. S. Jellinger, Y. Handelsman, D. S. Bell, Z. T. Bloomgarden, E. A. Brinton, M. H. Davidson, S. Fazio, V. A. Fonseca, A. J. Garber, G. Grunberger, C. K. Guerin, J. I. Mechanick, R. Pessah-Pollack, P. D. Rosenblit, D. A. Smith, K. Wyne, M. Bush, F. Zangeneh, Y. Handelsman, D. S. Bell, Z. T. Bloomgarden, E. A. Brinton, S. Fazio, V. A. Fonseca, A. J. Garber, G. Grunberger, C. K. Guerin, P. S. Jellinger, P. D. Rosenblit, D. A. Smith, K. Wyne, and M. H. Davidson, “American association of clinical endocrinologists and american college of endocrinology guidelines for management of dyslipidemia and prevention of cardiovascular disease,” *Endocrine Practice*, vol. 23, pp. 1–87, Apr. 2017.
- [100] N. Nantsupawat, A. Booncharoen, A. Wisetborisut, W. Jiraporncharoen, K. Pinyopornpanish, L. Chutarattanakul, and C. Angkurawaranon, “Appropriate total cholesterol cut-offs for detection of abnormal LDL cholesterol and non-HDL cholesterol among low cardiovascular risk population,” *Lipids in Health and Disease*, vol. 18, Jan. 2019.
- [101] World Health Organization, “Definition, diagnosis and classification of diabetes mellitus and its complications: report of a WHO consultation. part 1, diagnosis and classification of diabetes mellitus,” 1999.

BIBLIOGRAPHY

- [102] K. Alberti, R. H. Eckel, S. M. Grundy, P. Z. Zimmet, J. I. Cleeman, K. A. Donato, J.-C. Fruchart, W. P. T. James, C. M. Loria, and S. C. Smith, “Harmonizing the metabolic syndrome,” *Circulation*, vol. 120, pp. 1640–1645, Oct. 2009.
- [103] M. Huber, M. Trubetskov, S. A. Hussain, W. Schweinberger, C. Hofer, and I. Pupeza, “Optimum sample thickness for trace analyte detection with field-resolved infrared spectroscopy,” *Analytical Chemistry*, vol. 92, p. 7508–7514, Apr. 2020.
- [104] V. A. Lorenz-Fonfria, “Infrared difference spectroscopy of proteins: From bands to bonds,” *Chemical Reviews*, vol. 120, p. 3466–3576, Mar. 2020.
- [105] D. W. Marquardt and R. D. Snee, “Ridge regression in practice,” *The American Statistician*, vol. 29, no. 1, pp. 3–20, 1975.
- [106] P. Lasch, M. Diem, W. Hänsch, and D. Naumann, “Artificial neural networks as supervised techniques for FT-IR microspectroscopic imaging,” *Journal of Chemometrics*, vol. 20, p. 209–220, May 2006.
- [107] R. Christensen, *Log-linear models and logistic regression*. Springer Science & Business Media, 2006.
- [108] J. Read, B. Pfahringer, G. Holmes, and E. Frank, “Classifier chains for multi-label classification,” *Machine Learning*, vol. 85, p. 333–359, June 2011.

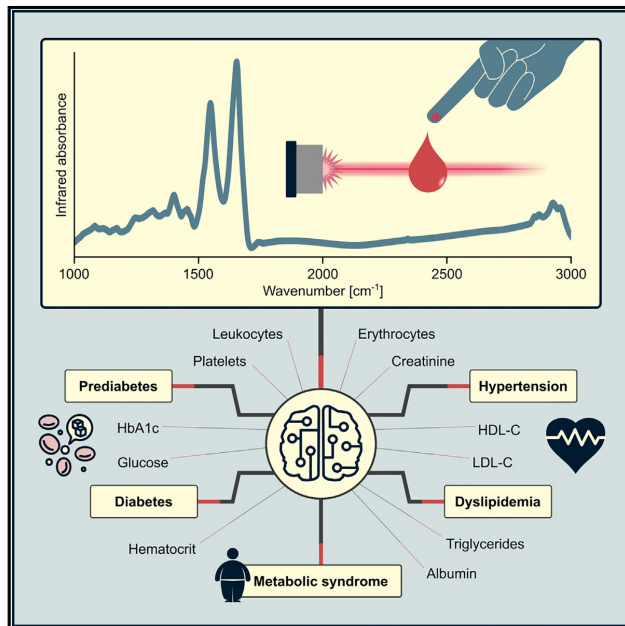
Appendix A

Plasma infrared fingerprinting with machine learning enables single-measurement multi-phenotype health screening

Copyright notice: © 2024 The Author(s). This is an open access article distributed under the terms of the Creative Commons CC-BY license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The CC-BY 4.0 license is printed in Appendix E.

Plasma infrared fingerprinting with machine learning enables single-measurement multi-phenotype health screening

Graphical abstract



Authors

Tarek Eissa, Cristina Leonardo, Kosmas V. Kepesidis, ..., Lothar Richter, Annette Peters, Mihaela Žigman

Correspondence

tarek.eissa@mpq.mpg.de (T.E.),
mihaela.zigman@mpq.mpg.de (M.Ž.)

In brief

Eissa et al. put forward a cost-effective diagnostic platform for high-throughput health profiling via infrared spectroscopy of blood plasma. Using a population-based study, they develop machine learning models to detect chronic health conditions and multimorbidities, forecast the development of metabolic syndrome, and quantify the concentrations of classical clinical chemistry analytes.

Highlights

- Multimorbidities are detectable through infrared molecular fingerprinting of plasma
- Infrared fingerprints forecast metabolic syndrome within 6.5 years pre-onset
- Concentrations of clinical chemistry analytes are reflected in infrared fingerprints
- Infrared fingerprinting enables cost-effective, high-throughput clinical diagnostics



Eissa et al., 2024, Cell Reports Medicine 5, 101625
July 16, 2024 © 2024 The Authors. Published by Elsevier Inc.
<https://doi.org/10.1016/j.xcrm.2024.101625>



Article

Plasma infrared fingerprinting with machine learning enables single-measurement multi-phenotype health screening

Tarek Eissa,^{1,2,3,*} Cristina Leonardo,¹ Kosmas V. Kepesidis,^{1,2,4} Frank Fleischmann,^{1,2} Birgit Linkohr,⁵ Daniel Meyer,^{2,4} Viola Zoka,^{1,4} Marinus Huber,^{1,2} Liudmila Voronina,¹ Lothar Richter,³ Annette Peters,^{5,6,7,8} and Mihaela Zigman^{1,2,9,*}¹Department of Laser Physics, Ludwig Maximilian University of Munich (LMU), Garching, Germany²Laboratory for Attosecond Physics, Max Planck Institute of Quantum Optics (MPQ), Garching, Germany³School of Computation, Information and Technology, Technical University of Munich (TUM), Garching, Germany⁴Center for Molecular Fingerprinting (CMF), Budapest, Hungary⁵Institute of Epidemiology, Helmholtz Zentrum München, Neuherberg, Germany⁶School of Public Health, Institute for Medical Information Processing, Biometry, and Epidemiology, Pettenkofer, Ludwig Maximilian University of Munich (LMU), Munich, Germany⁷German Center for Diabetes Research (DZD), Neuherberg, Germany⁸German Centre for Cardiovascular Research (DZHK), Partner Site Munich, Munich, Germany⁹Lead contact*Correspondence: tarek.eissa@mpq.mpg.de (T.E.), mihaela.zigman@mpq.mpg.de (M.Ž.)
<https://doi.org/10.1016/j.xcrm.2024.101625>

SUMMARY

Infrared spectroscopy is a powerful technique for probing the molecular profiles of complex biofluids, offering a promising avenue for high-throughput *in vitro* diagnostics. While several studies showcased its potential in detecting health conditions, a large-scale analysis of a naturally heterogeneous potential patient population has not been attempted. Using a population-based cohort, here we analyze 5,184 blood plasma samples from 3,169 individuals using Fourier transform infrared (FTIR) spectroscopy. Applying a multi-task classification to distinguish between dyslipidemia, hypertension, prediabetes, type 2 diabetes, and healthy states, we find that the approach can accurately single out healthy individuals and characterize chronic multimorbid states. We further identify the capacity to forecast the development of metabolic syndrome years in advance of onset. Dataset-independent testing confirms the robustness of infrared signatures against variations in sample handling, storage time, and measurement regimes. This study provides the framework that establishes infrared molecular fingerprinting as an efficient modality for populational health diagnostics.

INTRODUCTION

The pursuit of molecular gateways that describe the composition of chemically complex media in a minimally invasive, cost-effective, and high-throughput manner is a long-standing challenge.^{1,2} As physiological phenotypes often manifest in the molecular profiles of systemic biofluids (e.g., blood), gateways to quantitatively profile their compositions are invaluable for detecting and characterizing changing health states.^{3,4}

As minimally invasive screening and risk assessment tools, blood-based laboratory test panels are often employed in clinical diagnostics.^{5,6} Despite their foundational role in modern medicine, the process of measuring several biomarkers can be time-consuming and resource-intensive.⁷ Furthermore, biomarkers were developed in candidate approaches—requiring decades of research to link any given molecule to a physiological state.

“Shotgun” approaches that simultaneously probe an ensemble of molecules, requiring no *a priori* information, thus carry the

potential to forge the future of efficient clinical testing.^{8–10} Untargeted mass spectrometry and nuclear magnetic resonance spectroscopy are two highly promising technologies that may fulfill this role. Still, concerns relating to low measurement throughput, run-to-run and between-lab variability, and involved experimental handling have hindered their applications in high-throughput health diagnostics beyond clinical studies.^{8–11}

Infrared (IR) spectroscopy is a powerful top-down approach that enables the label-free profiling of a given sample by interrogating molecular fragments (i.e., functional groups) and measuring their resonant vibrational response to IR excitation.^{12–14} This technique can employ various light sources, including globars,¹⁵ few-cycle laser-based excitation,^{16,17} and discrete wavelength sources such as quantum cascade lasers.^{18,19} In broadband IR absorption spectroscopy, absorption occurs when the IR source emits frequencies that align with the vibrational frequency of a molecular fragment. When analyzing complex samples, the presence of similar chemical bonds and functional groups in different biomolecules leads to overlapping spectral response signals that



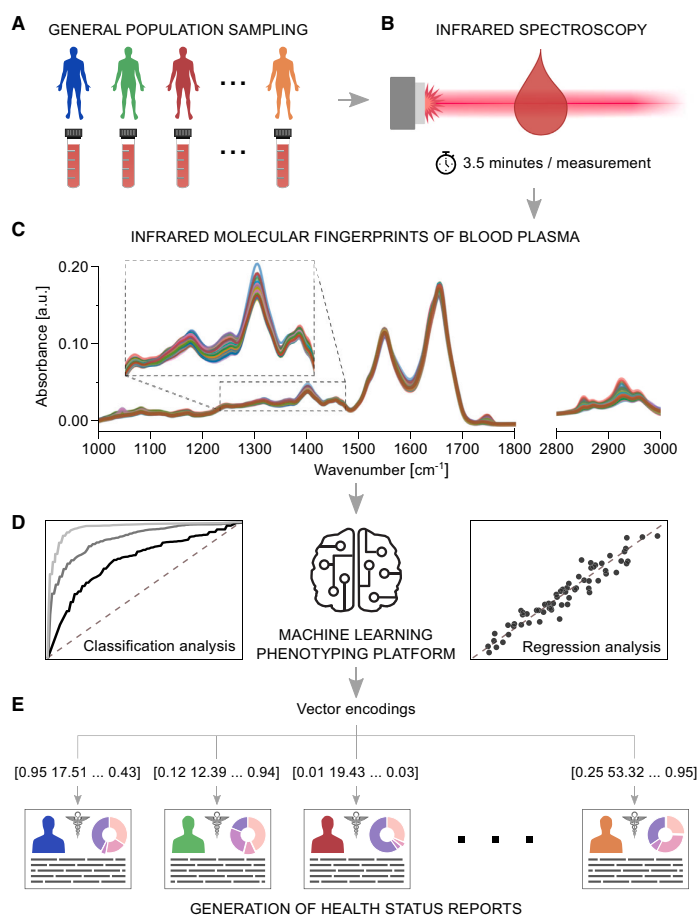


Figure 1. Proposed IR molecular fingerprinting strategy to aid clinical phenotyping and decision-making

(A) Blood samples are drawn, and blood plasma is prepared from individuals with unknown health states, or phenotypes.

(B) Collected plasma samples are measured using an automated FTIR spectrometer.

(C) Preprocessed absorption spectra of liquid blood plasma are captured. The inset depicts a close-up of the spectra in a truncated spectral region.

(D) Machine learning methods are applied to decode the measured spectra into numerical vectors, modeling diverse biological outcomes.

(E) Health reports are generated from a framework established on the measured IR spectra and machine learning models, which quantify unknown health-related properties of each individual to describe their physiology.

the approach been introduced into the canon of healthcare practices.^{36,37} To test its capacity—as an *in vitro* phenotype medical diagnostic tool—large-scale clinical studies involving deeply parametrized individuals whose physiology reflects a diverse patient population are required. Moreover, covariates potentially affecting IR fingerprints (e.g., anthropometric parameters or comorbidities) are yet to be established. Furthermore, trained phenotype detection models must be robustly validated in ways that closely align with real-world scenarios. Hindered by the scale of efforts required, progress toward incorporating IR-based phenotyping in medical diagnostic routines has been limited.³⁸

To assess whether an alliance between IR fingerprinting and machine learning

could fuel a single-measurement multi-phenotype medical diagnostics platform, we applied Fourier transform IR (FTIR) spectroscopy to profile blood plasma in a large phenotyped cohort. Experimentally, we measured 5,184 samples from 3,169 individuals within the population-based Cooperative Health Research in the Region of Augsburg (KORA) longitudinal cohort,³⁹ representing a potential patient population. Profiling the largest population with IR fingerprinting to our knowledge, our study aimed to examine the capacity of the approach to comprehensively detect several medically relevant human health phenotypes. With an age range of 32–88 years and the corresponding frequent occurrence of multimorbidity in the older age groups, we focused on the capacity to simultaneously detect a set of commonly occurring phenotypes—dyslipidemia, hypertension, prediabetes, type 2 diabetes, interrelated conditions of the metabolic syndrome (MetS), and healthy states. We employed different methods of validating how such a parallel, multilabel

hinder the identification of individual constituents. The measured IR absorption spectrum is characteristic of a sample's overall molecular composition, representing a superposition of the responses of all fragments. The corresponding IR absorption spectrum is thus referred to as a molecular fingerprint—simultaneously integrating the entire set of *omes* across biomolecules (e.g., proteins, lipids, and metabolites).²⁰ When analyzing systemic biofluids, such as plasma, physiological phenotypes may be encoded within the molecular fingerprints as spectral patterns, and machine learning algorithms can be trained to detect them—linking IR fingerprints to multiple phenotypes (Figure 1).

Applications of IR fingerprinting to address medical problems have been previously examined, including numerous studies applying different approaches of IR spectroscopy for *in vitro* clinical diagnostics.^{20–35} However, the concept of biofluid IR fingerprinting for medical health diagnostics or screening has not been evaluated in a large-scale naturally variable population, nor has

could fuel a single-measurement multi-phenotype medical diagnostics platform, we applied Fourier transform IR (FTIR) spectroscopy to profile blood plasma in a large phenotyped cohort. Experimentally, we measured 5,184 samples from 3,169 individuals within the population-based Cooperative Health Research in the Region of Augsburg (KORA) longitudinal cohort,³⁹ representing a potential patient population. Profiling the largest population with IR fingerprinting to our knowledge, our study aimed to examine the capacity of the approach to comprehensively detect several medically relevant human health phenotypes. With an age range of 32–88 years and the corresponding frequent occurrence of multimorbidity in the older age groups, we focused on the capacity to simultaneously detect a set of commonly occurring phenotypes—dyslipidemia, hypertension, prediabetes, type 2 diabetes, interrelated conditions of the metabolic syndrome (MetS), and healthy states. We employed different methods of validating how such a parallel, multilabel

phenotype classification model performed to examine the utility of the proposed approach. Furthermore, we examined the correlations between clinical laboratory test analytes and the informational content of IR spectra, facilitating their broader interpretability in a known clinical context. Our results provide a strategy to enable a robust and efficient multi-phenotyping platform.

RESULTS

Study characteristics for testing IR fingerprinting in populational screening

To test the performance of IR fingerprinting, we analyzed the KORA population-based longitudinal cohort.³⁹ The study cohort comprised a random sample from the general adult population, aged 32 to 88 years, in Southern Germany (STAR Methods; Table S1). Our analysis included samples from two time points within the cohort: KORA-F4, sample set #1 (conducted between 2006 and 2008) and KORA-FF4, sample set #2 (conducted between 2013 and 2014), both of which are follow-ups of the KORA-S4 study (conducted between 1999 and 2001). The first sample set (KORA-F4) included 3,044 participants, while the second (KORA-FF4) included 2,140 participants. In total, 5,184 blood plasma samples from 3,169 unique individuals were considered. Among these, 2,015 individuals participated in both sample sets. Extensive medical examinations were performed on the participants using standardized protocols (STAR Methods). FTIR spectroscopy in transmission mode was performed on all 5,184 samples in two independent measurement campaigns—one for each sample set—where the measurement times were separated by an average of 2.7 years. The two sample sets were thus collected at different times, freezer-stored for varying periods, and measured independently—all on a multi-year scale (STAR Methods).

Combining IR spectroscopy of human blood plasma with machine learning analysis, we put forward a rapid approach for describing several characteristics of human health at a populational level (Figure 1). Our study aimed to investigate the potential of IR molecular fingerprints to simultaneously detect commonly occurring chronic conditions and non-communicable diseases (NCDs) through a single, cost-efficient sample measurement.

As proof of concept, we focused on conditions that were highly prevalent in multimorbid individuals (Figure 2A). Dyslipidemia was the most frequently occurring condition among the population, accounting for a total of 49% of the first sample set and 42% of the second sample set. Hypertension was the second most commonly occurring condition, observed in 38% of the individuals in the first sample set and 39% of the individuals in the second. Prediabetic and type 2 diabetic individuals accounted for 28% of the first sample set and 33% of the second when combined. Corresponding to these terms, we defined an additional category of individuals as “healthy”—encompassing individuals negative for each of the four aforementioned phenotypes. Across the 5 phenotypes, we observed that the investigated conditions did not occur randomly in individuals but in certain natural combinations, such that 12 different co-occurrence patterns existed with varying prevalence (Figure 2A). The healthy group was found to be the largest sub-category of individuals in both sample sets.

IR fingerprinting enables simultaneous detection of multiple phenotypes

The capacity to capture and distinguish if a single phenotype or any combination of phenotypes existed in an individual would be essential for health screening and monitoring.^{40,41} To evaluate whether IR molecular fingerprinting has any capacity herein, we first examined the study cohort in a cross-sectional fashion. Specifically, we tested whether the approach was sensitive and robust enough to simultaneously detect the presence of dyslipidemia, hypertension, prediabetes, and type 2 diabetes and the absence of these conditions (the healthy group).

We developed a multilabel machine learning classifier capable of simultaneously detecting and distinguishing between the conditions of interest, with the output being a five-dimensional vector describing the phenotypic state of each individual (Figure 2B). The classifier incorporated a procedure in which the task was broken down into five binary classifications linked through a chain of decisions (STAR Methods). This allowed the multilabel classifier to consider correlations between the outcomes, leading to improved detection performance, as well as consider mutual exclusivities (e.g., diabetic individuals cannot be prediabetic and an individual with any chronic condition cannot be in the healthy group).

The inclusion of two sample sets from the examined study population allowed us to test how the classification performed with several validation methods (STAR Methods; Figure S1). First, we applied a 10-fold cross-validation, independently, on each of the two sample sets. This resulted in two classification performance estimates (one for the sample set #1 and another one for the sample set #2), whereby the two sample sets were collected as well as measured years apart. Yet, since a significant portion of individuals were sampled at both time points, and their samples contributed to both sample sets, the cross-validated metrics are not truly independent of each other in the initial examination here. To overcome this and ensure a robust evaluation unbiased by any possible overlap of the involved individuals, we applied a further validation strategy to examine how the classification performed when tested on samples from entirely different individuals. Specifically, the classifier was repeatedly trained on a portion of individuals from sample set #1 and tested on a held-out portion of independent individuals from sample set #2 (STAR Methods). This strategy was designed to test the classification's generalizability over sample storage time, measurement time, as well as across different individuals.

The complex multi-output nature of this classification, which encompasses the occurrence of 12 clinically feasible phenotype combinations (Figure 2A), requires a more thoughtful evaluation as predictions may be partially correct. We therefore considered three established metrics that measure multilabel prediction efficacy in different ways: the exact match ratio, Hamming score, and Hamming loss (metric definitions further explained in STAR Methods).^{42–45} Briefly, the exact match ratio rewards predictions that accurately identify the physiological state of a sample across all possible outcomes—scoring each 5-dimensional prediction with either a 0 or 1. The metric is strict as it treats partially correct predictions as wholly incorrect. In contrast, the Hamming score considers partially correct predictions as having value. For a given sample, it is defined as the proportion of

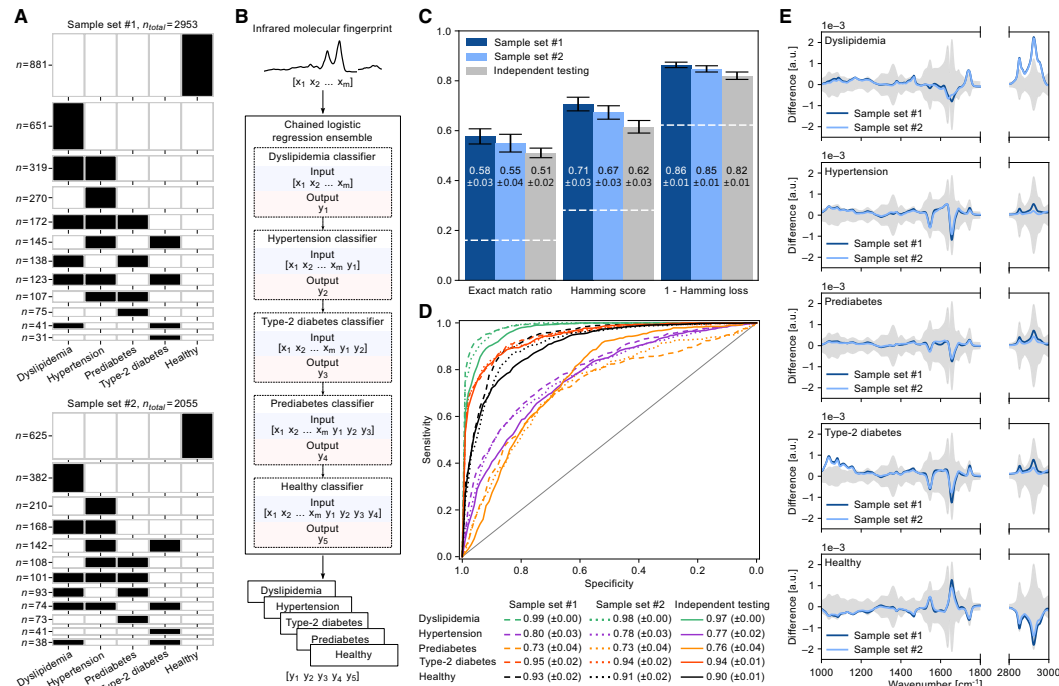


Figure 2. Detection of highly prevalent and co-occurring phenotypes using IR molecular fingerprints

(A) Distributions of all possible combinations of the investigated phenotypes in sample set #1 (upper panel) and sample set #2 (lower panel), encompassing the study cohort. Due to unknown/missing values in the medical records relevant to the listed phenotypes, 91 samples (out of 3,044) were excluded from sample set #1 and 85 samples (out of 2,140) were excluded from sample set #2. Individuals with listed phenotypes are highlighted in black, with each row representing a group of combined phenotypes and its height scaled by the number of individuals in that group.

(B) Model architecture of a multilabel classifier trained to describe the phenotypic state of each input IR spectrum, giving out a five-dimensional vector. A chained binary classification modeling method was applied where each model classifies one outcome against all others using the depicted classification order.

(C) Multilabel classification performance estimates for models trained to simultaneously predict all outcomes of a given IR fingerprint. Classification performance was estimated on unseen test samples by independently cross-validating on each sample set (sample set #1 in dark blue bar and sample set #2 in light blue bar). For independent testing (gray bar), classification performance was estimated by training on a portion of sample set #1 and testing the trained classifier on entirely independent individuals from sample set #2. This provides an estimate of classification generalization across measurements performed years apart, encompassing different training-test individuals. The white dashed lines are a reference that depicts estimates for random chance multilabel predictions for each classification metric (STAR Methods).

(D) Receiver operating characteristic (ROC) curves for binary classifications modeling each listed outcome against the remaining of the population. For each classification, the mean cross-validated ROC curves of the test sets are depicted and the area under the curve is listed below, along with its standard deviation.

(E) Difference between the mean IR fingerprints of samples positive and negative for each phenotype on each sample set (sample set #1 in dark blue and sample set #2 in light blue). Gray-shaded areas depict the standard deviation of IR fingerprints negative for each corresponding phenotype.

correctly predicted positive outcomes to the total number of predicted/actually positive outcomes for the sample. The Hamming loss is defined as the proportion of incorrectly predicted labels across all samples and the 5 labels. All three metrics range from 0 to 1, where higher values for the exact match ratio and Hamming score indicate better predictive performance, while a lower Hamming loss indicates better performance. To provide a consistent interpretation, we calculate $1 - \text{Hamming loss}$, where higher values also signify better predictive performance. Unlike binary predictions, where a threshold of 0.50 is often assumed to gauge classifiers making random predictions, a con-

textually relevant point of comparison for multilabel classification is more complex and dependent on the definition of the evaluation metric (details on chance predictions in STAR Methods).

Applying the multilabel classification, we found that the approach was capable of predicting all the exact phenotypic outcomes of the probed individuals with exact match ratios of 0.58 and 0.55 when independently cross-validated on each of the two sample sets (Figure 2C, light/dark blue bars). Importantly, the performance of the classification was stable between sample set #1 and #2, with differences falling within the standard deviation of the validation splits. When the classification was

repeatedly tested on the sample-set-independent measurements, we observed an exact match ratio of 0.51, which significantly exceeds a chance model (Figure 2C, gray bars). The Hamming loss ranged between 0.82 and 0.86 across the three validation methods—revealing that, in the worst case, only 18% of the individual phenotypes were misclassified across all the test samples. In light of the simplicity and cost-effectiveness of the single-measurement IR fingerprinting method, combined with the medical challenge to assess several phenotypes simultaneously, it was surprising to find such high multilabel prediction rates across the three metrics and validation methods. This underscores that the approach is not only sensitive to the diversity of phenotypes but also specific to medical aberrations and their combinations.

This investigation also revealed that the classification performance estimates derived through cross-validation were fairly consistent with how the classifier performed when tested on dataset-independent samples—a validation approach that more closely mirrors practical applications. However, it is important to highlight that a minor drop in predictive efficacy was observed. This drop underscores the importance of validating how such classifications perform when tested on samples that were collected and measured by different operators at considerably different time points from the training dataset. Such dataset-independent validation helps prevent the potential of overestimating how a classification would perform in more realistic applications. Despite this drop, it was surprising to observe that the phenotypes were encoded in IR molecular fingerprints to an extent that enabled their robust detection, even though variations between the sample sets and measurement conditions had significant effects on the spectra (Figure S2).

Phenotype-dependent detection capacity

Since the multilabel classification was built on a series of binary classifiers, where the predictions were driven exclusively by the information encoded in an input IR spectrum, we further examined the diagnostic capacity of each individual phenotype-specific classifier (Figure 2D). This analysis provides a quantitative assessment of the degree to which each individual phenotype was detectable by the IR spectra in a case-control setting—where the control class here combined the set of the remaining four phenotypes (i.e., one-vs-rest analysis). To assess predictive performance, we computed the receiver operating characteristic (ROC) curves of each binary classification, following the same validation procedure as previously described in the multilabel analysis. To combine the measures of sensitivity and specificity, we calculated the area under the ROC curve (AUC) as a summarizing metric of performance.

The AUCs ranged from 0.76 for prediabetes to 0.97 for dyslipidemia when independently tested, which shows that all 5 phenotypes were detectable with a high degree of confidence (Figure 2D). The highest prediction efficacies were observed for type 2 diabetes and dyslipidemia, underscoring the approach's potential for these conditions. Intriguingly, the classifier detecting whether an individual was healthy achieved an AUC of 0.90 when employing the sample set-independent testing method. This revealed that the IR-based approach was capable of both predicting each phenotype individually and providing a metric

of an overall healthy physiology. Importantly, the classification performance remained stable across the three testing methods for all the investigated phenotypes in this binary classification setting.

To further assess the efficacy of detecting each individual phenotype, we repeated the aforementioned analysis but removed the chain mechanism such that each phenotype was classified with an independent binary classifier (Figure S3). Without the chain, all classifications perform comparably well (Figure 2D vs. Figure S3). This similarity in predictive efficacy was expected for the binary classifications, as all predictions—whether chained or not—relied exclusively on information in the IR spectra. Hence, the classifiers did not require information about the predictions of previous classifiers as they were able to leverage all the information encoded in the IR spectrum. The only noticeable (albeit minor) drop in prediction efficacy was observed for prediabetes detection. This likely arose from the fact that prediabetes is an intermediate condition, and the control class included both type 2 diabetic and normal glucose tolerant (NGT) individuals. Therefore, we found that a single (linear) classifier struggled to single out the intermediate prediabetic cases, compared to combining multiple linear classifiers that have the capacity to single out NGT and type 2 diabetic individuals as well.

Phenotype-specific IR spectral signatures

Investigating the capacity to detect medically relevant phenotypes—both individually and simultaneously—raised the question of whether the molecular signature of individuals positive for a phenotype was characteristically distinct from those negative for it. To examine this, we calculated the difference between the mean IR molecular fingerprint of individuals positive for a phenotype and that of the remaining individuals' combined set of fingerprints encompassing the four remaining phenotypes (Figure 2E).

Differences in the molecular IR signatures of individuals bearing dyslipidemia, hypertension, and diabetic states showed very characteristic profiles with clear distinctions between one another. The signature of prediabetic individuals was revealed to share a very similar shape to that of individuals bearing type 2 diabetes—with only smaller variations in the magnitude of the differences. The latter finding is both promising and reassuring, given that prediabetes is an intermediate condition at a high risk of progressing into type 2 diabetes. Moreover, it was very reassuring to observe that the region with the most significant differences in the IR fingerprints of type 2 diabetic cases and controls was $1,000\text{--}1,180\text{ cm}^{-1}$, with the peaks aligning with known spectral signatures of glucose.⁴⁶ When the fingerprints of the healthy individuals were compared to all the other individuals in the population, the differences revealed a shape that is similar to the inverse differences of all the other conditions studied—thereby supporting the aforementioned results. Crucially, the signatures of all phenotypes remained stable between the two sample sets, given the heavy overlap between the two curves depicted for each phenotype. Summarizing, these comparisons of spectral differences further corroborate the distinctive properties and reproducibility of the plasma-based IR fingerprinting approach.

Comparing the predictive value of IR fingerprints to clinical analytes

To examine the predictive value of IR molecular fingerprints in relation to standard clinical analytes, we repeated the multilabel classification but in a modified fashion. Rather than using IR fingerprints, we used a set of 12 commonly measured analytes from blood-based clinical chemistry and hematology analysis as the input features to make predictions (Figure S4). These analytes are widely used in blood cell counts, along with metabolic and lipidomic panels.

Using the 12 clinical analytes as exclusive predictors of the studied phenotypes, this multilabel classification gained a slight advantage over the classification based only on IR fingerprints—with an exact match ratio of 0.57 and Hamming score of 0.70, compared to the previously mentioned exact match ratio of 0.51 and Hamming score of 0.68 (Figure 2C vs. Figure S4A). As the panels of clinical analytes included fasting glucose, hemoglobin A1c (HbA1c), low-density lipoprotein (LDL) cholesterol, high-density lipoprotein (HDL) cholesterol, and triglycerides, it was not surprising that an improvement in predictive capacity was observed when predicting type 2 diabetes and dyslipidemia as such clinical parameters are often used to clinically diagnose these conditions (Figure 2D vs. Figure S4B). For the condition that is not closely defined by such analytes, hypertension, the IR-based model outperformed the clinical-based model.

The involved costs of instrumentation and reagents, along with turnaround times, are a limiting factor for blood-based clinical laboratory testing. In contrast, the IR-based approach offers a reagent-free, one-shot measurement of bulk plasma, providing a comprehensive view of the sample's cross-molecular profile. This provides an efficient means of predicting the presence of health conditions, requiring minimal sample preparation, small sample volume, and short measurement time. A further highlight is that the IR-based approach is not limited to health conditions closely defined by conventional clinical laboratory parameters.

IR fingerprints are sensitive to anthropometric parameters

Gender, age, and body mass index (BMI) are risk factors for NCDs and chronic conditions.^{47,48} Should plasma-based IR fingerprints encode the contributions of such anthropometric parameters, their effects on the IR fingerprints would potentially influence the predictions of models trained to detect clinical phenotypes.

To investigate the extent of their contributions, we first examined the capacity to predict the gender, age, and BMI of the sampled individuals using their IR molecular fingerprints (Figures S5A–S5C). Classifying gender was highly successful, achieving an AUC of 0.96. Estimating age with multivariate regression modeling was possible with an R^2 value of 0.58 and a root-mean-square error (RMSE) of 8.39 years. BMI was estimated with an R^2 value of 0.43 and an RMSE of 3.66 kg/m². These results demonstrate the unambiguous impact of gender, age, and BMI on blood-based IR molecular fingerprints, affecting any future diagnostic applications.

To further investigate their contributions independently of any possible health deviation, we performed the same analysis, but

only on the healthy subset of individuals (Figures S5D–S5F). In this setting, classifying gender remained stable, with an AUC of 0.96. The efficacies of estimating age and BMI, however, were reduced to R^2 values of 0.52 and 0.37, respectively. Given that it was more challenging to estimate age and BMI in the healthy subset, this analysis suggested that NCDs or chronic conditions had an influence on the models estimating these anthropometric parameters. This supported the notion that the signals of aging and increased BMI are reflected in declining health states.

Prompted by these results, we further examined the capacity to detect each phenotype in a case-control setting—but classifying on samples of similar anthropometric distributions, rather than considering the populational distribution as previously. For each phenotype, we pair-matched a case individual to a control individual of similar gender, age, and BMI to minimize their possible contributions (Table S2). Using independent classification models, the AUCs ranged from 0.60 for prediabetes to 0.97 for dyslipidemia (Figure S6)—newly revealing that the contribution of these anthropometric parameters was higher for phenotypes with weaker disease signals.

Overall, we identified anthropometric parameters that significantly influence IR molecular fingerprints and, thereby, affect the phenotype classification models. These findings shall directly inform any future IR fingerprinting medical case evaluations as well as case-control clinical study designs.

IR fingerprints uncover shared pathophysiologies of MetS

MetS is a complex progressive condition characterized by a manifestation of interconnected metabolic abnormalities involving visceral obesity, hyperglycemia, raised blood pressure, raised triglycerides, and lowered HDL cholesterol (Figure 3A). Although its definition and pathogenesis are controversial,^{49,50} MetS identifies individuals with shared pathophysiology—ones who are at risk of developing further health conditions, such as atherosclerotic cardiovascular disease, type 2 diabetes,⁵¹ cerebrovascular events,⁵² liver cancer,⁵³ and colorectal cancers.⁵⁴ Robust yet unsophisticated methods of identifying individuals with MetS would not only facilitate a better understanding of its pathophysiology but also enable efficient methods of stratifying populations into different levels of risk. This approach could further facilitate individualized pharmacologic and/or lifestyle interventions.^{55,56}

Following the harmonized criteria of the International Diabetes Federation and the American Heart Association/National Heart, Lung, and Blood Institute,⁵⁷ MetS is present when at least three out of five aforementioned metabolic risk factors are present. As an early sign of MetS, one or two risk factors qualify as pre-MetS, a precursor condition that may represent the critical intermediate transition phase.^{58,59} Breaking down the number of risk factors we observed in our two sample sets, we found that most individuals were at a pre-MetS stage (Figure 3B). At the pre-MetS stage, visceral obesity was the most prevalent risk factor, followed by hypertension and hyperglycemia (Table S3).

Employing the IR fingerprinting approach, we first systematically evaluated the extent to which we can single out individuals with varying numbers of metabolic risk factors from those with none (Figure 3C). We found that the classification efficacy increased when the number of metabolic risk factors increased

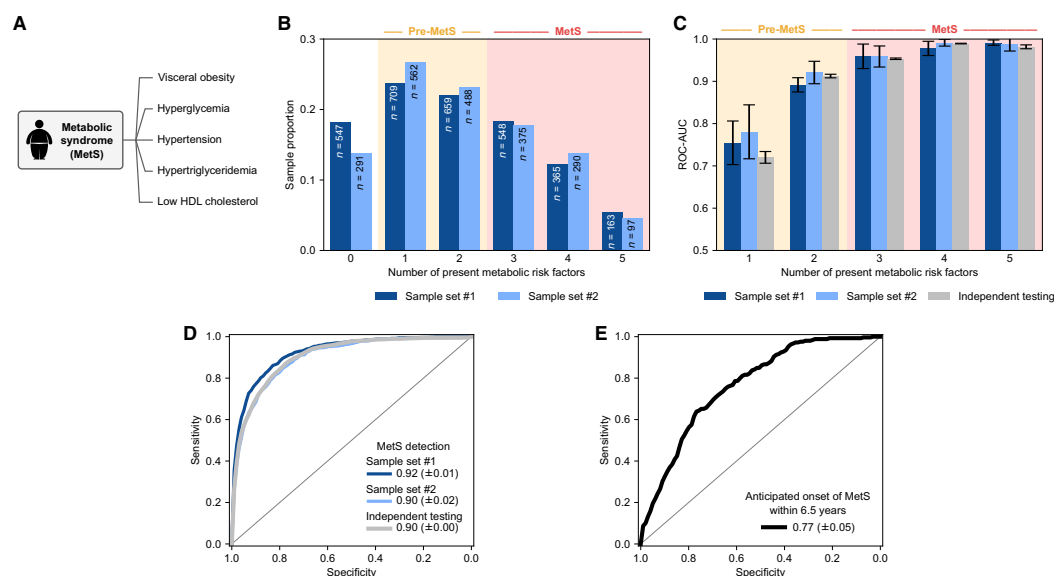


Figure 3. Detecting shared pathophysiology through the metabolic syndrome

(A) Metabolic syndrome (MetS) was defined according to the harmonized criteria of the International Diabetes Federation (IDF) and American Heart Association/National Heart, Lung, and Blood Institute (AHA/NHLBI)—requiring individuals to have at least three of the five listed metabolic risk factors.

(B) Proportion of samples with different numbers of concurrent risk factors. The pure sample counts are listed on each corresponding bar. Samples with unknown values for any of the risk factors were excluded. Distributions of sample set #1 (dark blue bars) and sample set #2 (light blue bars) are depicted separately. See also Table S3 for a breakdown of risk factor prevalence.

(C) Binary classifiers were trained on the IR molecular fingerprints to distinguish individuals with one, two, three, and so on risk factors from those with none of the risk factors. Classifier validation was carried out as previously described—by cross-validating on sample set #1, cross-validating on sample set #2, and repeatedly training on a portion of sample set #1, testing on independent individuals from sample set #2. The mean area under the ROC curve is depicted for classifying each group, with the error bars showing the standard deviation of the AUCs across the validation sets.

(D) Binary classifiers were trained on the IR molecular fingerprints of the two population-based sample sets to distinguish those with MetS (i.e., having at least three risk factors) and those without MetS (i.e., having two or fewer risk factors). ROC curves are depicted on unseen test samples where classifier validation was carried out as previously described. The mean AUC is listed for each validation type, along with the standard deviation across the validation sets.

(E) Binary classification to forecast the future onset of MetS using IR molecular fingerprints. The classifier was trained on individuals from two groups: those who did not have MetS at baseline and developed it by the follow-up ($n = 233$), and those who did not have MetS at baseline and remained without it by the follow-up ($n = 1,154$). Classifier training and testing (via a 10-fold cross-validation) was performed exclusively on measurements from the baseline visit (sample set #1), while the medical records of the follow-up were examined to set the outcomes that describe whether the syndrome was developed (within an average of 6.5 follow-up years). The mean test AUC is listed along with its standard deviation.

within the group. Intriguingly, the approach had the capacity to single out pre-MetS cases with a high efficacy—achieving an AUC near 0.75 when identifying individuals with one risk factor and an AUC near 0.90 when identifying individuals with two concurrent risk factors. This observation underscored the classifier's ability to stratify individuals depending on their metabolic burden.

We further discovered the high capacity of identifying individuals with MetS, achieving an AUC near 0.90 (Figure 3D). This further highlighted the high efficacy of identifying those at a greater risk of developing more serious health conditions, as previously mentioned.

IR fingerprints forecast the development of MetS

In contrast to detecting the presence of MetS and metabolic risk factors, the longitudinal aspect in our dataset enabled us to also

assess the capacity of forecasting the possible future development of MetS (Figure 3E). Within the study population, a subpopulation of 2,015 individuals participated in both baseline (sample set #1) and follow-up (sample set #2) samplings. Of these individuals, 1,387 did not have MetS at baseline, and among them, 233 developed MetS during the 6.5-year follow-up period. We constructed a predictive model that aimed to identify the individuals who transitioned from not having MetS at baseline but developed it in the follow-up sampling. As the control group, we used the 1,154 individuals who remained free of MetS throughout both samplings. Classifier training and testing were performed exclusively on the IR spectra of the baseline sampling (via 10-fold cross-validation). The binary outcomes, which denote whether each individual developed MetS during the follow-up, were determined by reviewing medical records from the follow-up. We found that the baseline measurements could effectively forecast

the future onset of MetS within the 6.5-year follow-up period, achieving an impressive AUC of 0.77 (Figure 3E). This AUC is comparable to a previously reported proteomics-based approach using 11 serum proteins, albeit over a 10-year follow-up period.⁶⁰

Altogether, these results further highlighted the IR fingerprinting modality in screening scenarios as an efficient means of early detection to reduce the risk of developing further health conditions.

Interpreting IR spectral features through clinical analytes

To facilitate the medical utility of the approach, the nature of IR molecular fingerprints must be described in the context of known substances—e.g., clinical chemistry analytes. Since different biomolecules are composed of different atoms and chemical bonds, the concentrations of distinct molecular groups may correlate to different spectral regions at varying degrees. We therefore investigated how each spectral feature correlated with the concentrations of 12 commonly measured clinical lab analytes (Figure 4A).

Triglycerides and LDL cholesterol heavily overlapped in the regions of highest correlation with absorbance—with the regions of 1,725–1,750 cm^{-1} (C=O stretching modes of fatty acids⁶¹) and 2,800–2,975 cm^{-1} (CH_2 and CH_3 stretching modes of lipid⁶¹) reflecting the highest levels of positive correlations. Triglycerides also observed high levels of correlation with absorbance in the region of 1,100–1,275 cm^{-1} (C–O and C–O–C stretching of carbohydrates⁶¹). As LDL is the main source of cholesterol build-up, the total cholesterol also showed a very similar trend in correlation across the spectral range. HDL cholesterol, on the other hand, revealed a relatively weak correlation pattern compared to the aforementioned parameters. The remaining analytes revealed different patterns from that of the lipid profile. As HbA1c is a glycosylated form of hemoglobin localized within erythrocytes, it was reassuring to find that HbA1c and glucose concentrations correlated to similar wave-numbers. It was also expected to find that the measures of hematocrit (representing the volume of erythrocytes) and the number of erythrocytes had very similar patterns. Albumin, the most abundant protein in blood plasma, showed its strongest correlations at the so-called amide-I and -II bands, generally specific to proteins (between 1,550 and 1,700 cm^{-1}).⁶² The remaining parameters observed relatively weak correlations across the spectral range.

These analyses were performed on the plasma IR fingerprints to broadly interpret the information encoded in the spectral positions by relating them to the concentrations of known clinical lab analytes. Although univariate analyses provide some insights by examining individual spectral features, they do not capture the full scope of molecular information in the IR fingerprints. Plasma, a highly complex biofluid, contains numerous molecular constituents that produce overlapping absorbance signals across different spectral features, potentially obscuring one another. To comprehensively understand the relationships between these clinical analyte concentrations and the information embedded in the IR fingerprints, it is crucial to consider the dependencies among spectral features.

In the following subsection, we employ multivariate analyses to further examine relations between the clinical lab analytes and plasma IR fingerprints.

Information is shared between clinical analytes and IR fingerprints

To further investigate the extent to which the studied clinical analytes were encoded within the IR fingerprints, we applied multivariate regression analysis to quantify their individual concentrations using the plasma IR spectra (Figure 4B). Unlike univariate analysis, the interactions and dependencies among all spectral features are considered here, potentially capturing relations that may be obscured by molecular absorbance overlap.

We found that analytes relating to the lipid profile and glucose level were most accurately estimated, all with R^2 values ≥ 0.91 (Figure 4B). Surprisingly, the concentration of HDL cholesterol was estimated with a high efficacy despite its relatively weak correlations with individual spectral features as depicted in Figure 4A. This highlighted that, despite the molecular overshadowing effect that likely concealed its molecular contributions at individual spectral features, the HDL cholesterol content had sufficient contributions that were distributed across several spectral features—thereby enabling the analyte to be accurately estimated when spectral dependencies were considered.

Albumin, creatinine, and HbA1c revealed relatively weaker yet still significantly strong associations with the informational content encoded in the fingerprints. For albumin, the relatively weaker R^2 value of 0.58 sparked intriguing considerations—particularly due to its high abundance and its integral role in shaping the IR spectrum of blood plasma.⁶³ This prompted us to re-perform the analysis depicted in Figure 4B, but on non-normalized spectra (Figure S7B). Without the normalization, no statistically significant change was observed in the concentration prediction efficacy—neither for albumin nor for any other analyte. Although it was previously reported that IR spectroscopy using quantum cascade lasers has the means to quantify albumin,⁶⁴ our FTIR evaluations of plasma show differences here. This may be attributed to the fact that, in our setting, the clinical chemistry laboratory albumin concentrations were measured in serum (Methods), which may have contributed noise to the task of quantifying albumin from the plasma-based IR spectra. In addition, in this molecularly complex milieu, however, the presence of numerous coexisting plasma molecules might conceivably eclipse the distinctive spectral contributions of albumin—especially given that several other highly abundant proteins overlap in their regions of absorbance.⁶³

As nearly half of the blood volume is made of erythrocytes, and as they serve as a reservoir of factors secreted in plasma, it was not surprising to find that their numbers could be estimated, to an extent, using IR molecular fingerprints of plasma. The parameters that were least reflected in the IR molecular fingerprints related to other types of blood cell counts—platelets and leukocytes. This indicated that neither their intracellular molecules nor their metabolic products strongly contribute to the extracellular composition of plasma and that the approach is not capable of specifically addressing their impacts on the chemistry of plasma (without any cellular enrichment and more dedicated analysis—e.g., immune cell profiling).

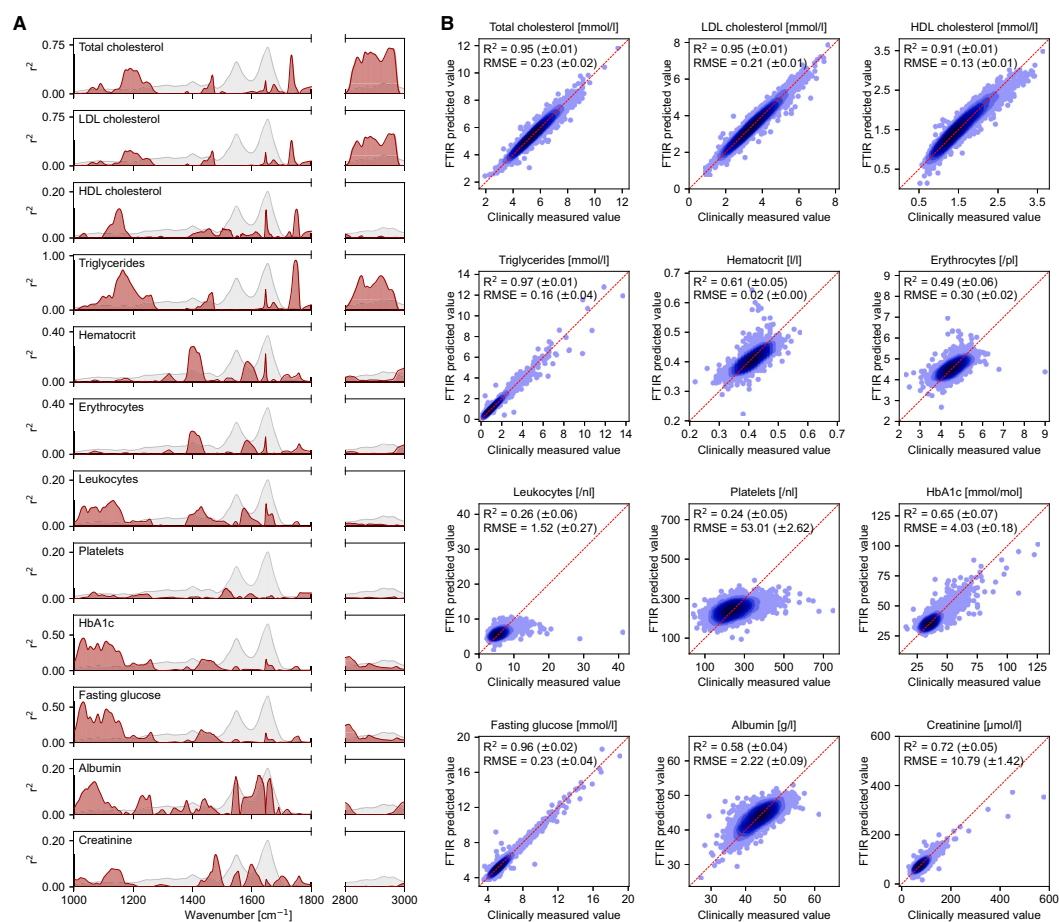


Figure 4. Classical clinical analytes reflected in IR molecular fingerprints

(A) Pearson correlation coefficient (red curves) between the concentrations of each clinical analyte and the absorbance at each wavenumber between 1,000 and 3,000 cm⁻¹ for preprocessed spectra. The mean absorbance spectrum of all measured IR spectra (5,184 samples) is depicted in gray on each panel as a visual reference for the shape of the spectrum.

(B) Performance of quantitatively predicting clinical analytes using the preprocessed IR spectra of the measured population. Regression algorithms were trained for each parameter to capture the relations between the clinically measured values and the multivariate spectral features. The predicted values follow from the test sets of 10-fold cross-validations. Each point represents a measurement. The mean coefficient of determination (R²) and the root-mean-square error (RMSE) are listed for each parameter, along with their standard deviations across the test splits. The diagonal (dashed red line) is a visual reference for a perfect fit. This same investigation was carried out on non-normalized spectra (Figure S7), where the univariate correlations (A) expectedly showed shifts in correlation trends while the efficacy of quantifying the concentration of each analyte (B) remained relatively unchanged.

In some instances, ratios between clinical analytes are utilized in medical diagnostics. For instance, ratios of triglycerides, LDL cholesterol, and total cholesterol to HDL cholesterol are often used for cardiovascular disease (CVD) risk assessments and have proven to be superior risk markers over individual analyte concentrations.⁶⁵ Since the IR molecular fingerprints are capable of estimating the concentrations of different parameters simultaneously, they are also informative to estimate ratios be-

tween them. To assess this, we calculated the ratios of triglycerides, LDL cholesterol, and total cholesterol to HDL cholesterol and studied the capacity to estimate their values (Figure S8). Thereby, we revealed that all three ratios were well estimated, each with an R² value near 0.92. This highlighted the potential of applying the approach to CVD risk assessments.

Altogether, it was evident that several distinct molecules and molecular groups correlate to the informational content of IR

fingerprints to a high degree despite the high complexity of blood plasma. IR fingerprint-based phenotype classification models can thus be linked to several relevant clinical lab parameters, providing a level of interpreting IR-based classifications in a known clinical context. Moreover, IR fingerprints present a viable alternative to traditional clinical laboratory methods for directly quantifying key clinical analytes, such as cholesterol and glucose levels, in routine clinical lab analysis. This provides an efficient, reagent-free means of simultaneously quantifying several analytes in a one-shot measurement.

DISCUSSION

The value of high-throughput molecular profiling is highest when molecular alterations have the capacity to define several medical phenotypes. The framework for IR fingerprinting to possibly fuel medical screening targeting the diversity of common conditions and their combinations within a naturally variable population has not been provided before.^{37,38,66} It was yet to be comprehensively proven whether IR fingerprinting could facilitate the robustness required for population-wide screening or whether it carries any capacity to capture the reality of complex multimorbid prevalence.

Combining IR spectroscopy with machine learning, our study demonstrated that IR molecular fingerprints (1) have the capacity to simultaneously detect and distinguish between common health phenotypes, providing a basis for multi-task phenotype detection; (2) can detect health-to-disease intermediate medical phenotypes (e.g., prediabetes and (pre-)MetS) that are of great value to screen and stratify populations for risk of developing further health conditions and intervene timely; (3) can forecast the future onset of clinically relevant health disorders, showcasing the approach's potential as an early detection approach; (4) strongly correlate to the concentrations of several clinical chemistry analytes; (5) reflect the physiological contributions of anthropometric parameters (gender, age, and BMI); and (6) are robust to variations in sample handling, storage time, and measurement regime differences on a multi-year scale. The approach is powerful for high-throughput population-scale screening. Specifically, for describing common, chronic health state phenotypes with a single cost-effective measurement that requires no preanalytical sample preparation or a *priori* molecular knowledge.

Health is multidimensional and individually heterogeneous, and parametrizing it is non-trivial. As complex clinical traits may be effectively probed systemically,⁶⁷ system-level cross-molecular IR fingerprints provide a new *in vitro* diagnostic advantage here. The impact of the presented paradigm, beyond clinical study scenarios, is manifold. Especially as, in any adult population, the majority of individuals with identified health aberrations will likely have multiple coexisting conditions, and various shared disease mechanisms can ideally be targeted via lifestyle or pharmacological interventions.^{40,55,68} Thus, screening avenues to efficiently discern between different multimorbid states are essential. Multimorbid states have been initially thought of as a collection of independent ailments and are still often measured separately. Given the direct coupling between many phenotypes (e.g., metabolic abnormalities and cardiovascular conditions),

informing measures on multimorbidity could advance therapeutic strategies and help assess and stratify risk levels. Importantly, the classifier detecting healthy individuals provided a very high prediction capacity (AUC of 0.90 when independently tested). This demonstrates that the approach is well suited to single out individuals in their healthspan (the period of life free from disease).

Beyond the phenotypes investigated here, the approach can serve as a platform that can be extended to screen and stratify for a variety of other health conditions. It may also be suitable for conditions of either unknown molecular origins or ones without knowledge of the underlying multi-molecular changes. To better understand and capture sub-clinical, intermediate phenotypes, we showcased how IR fingerprinting could facilitate the isolation of actionable at-risk phenotypes. Specifically, reporting on prediabetes could present an intervention that may prevent or delay the onset of type 2 diabetes that would, in turn, reduce the incidence of atherosclerotic CVD, kidney disease, vascular dementia, and Alzheimer's disease.^{69–72} Although methods for diagnosing prediabetes and type 2 diabetes are available,^{73,74} tools to identify individuals at risk for type 2 diabetes that could guide interventions are still lacking.⁷⁵ The second example is the capacity of IR fingerprinting to detect pre-MetS, a precondition and risk factor that may contribute to the evolution of further diseases. Thus, the presented framework is valuable for dynamic risk predictions across a variety of phenotypes as well as to capture combined conditions and potentially forecast their future onset—especially given its minimally invasive nature, ease of measurement, and low cost of the approach.

Previous clinical spectroscopy studies typically relied on rather small patient cohorts, hindering the potential of developing IR-based diagnostic models that adeptly generalize to larger populations.³⁸ Involving two datasets, from samples collected at two different times as well as experimentally measured years apart, we not only cross-validated the outcomes of each sample set separately but also further validated the robustness of our findings in a more realistic validation setting. Specifically, we were capable of training a classifier on samples from individuals collected at a point in time and testing this classifier on a different group of individuals that were measured years apart. Such a validation approach limited the potential to overestimate the clinical efficacy of our developed detection models. Additionally, the analyzed study paradigm reduced the possibility of patient selection bias since the individuals assessed were from a naturally heterogeneous population-based study involving a variety of potential conditions and health states—thus reflecting a realistic demographic distribution and true disease prevalence.

Limitations of the study

Critically seen, the generalizations of our machine learning applications may be limited to the studied population—where many individuals likely share similar genetic backgrounds and lifestyles.³⁹ To investigate the extent of this limitation, the models must be further tested on independent sets of ethnically diverse individuals. Secondly, the presented metrics of model validation considered the ground truth to be the health characterizations determined using existing clinical evaluations and may have been limited by the efficiency of the clinical analytical procedures

performed. For instance, diagnosing prediabetes and type 2 diabetes relied on an oral glucose tolerance test. Although commonly applied, this diagnostic approach is influenced by a multitude of physiological fluctuations.^{74,76} Finally, some of our analyses treated two time-separated samplings of partially overlapping individuals as two cross-sectional sample sets. This may partially explain the agreement between the cross-validations performed on the two sample sets. Additional testing on cohorts from fully independent clinical studies would provide an even more definitive verification of our findings.

Conclusions

The human body is a dynamic and complex system—with a multitude of ongoing biochemical processes driven by genetics, individual lifestyles, and aging.⁷⁷ The role of IR fingerprints to predict future disease onset and regression will need to be further assessed in additional longitudinal analyses that will be important to even more accurately capture and model (sub)clinical phenotypes, especially in light of the heterogeneity of person-specific healthspans. The concept of within-person stability of IR fingerprints has been revealed,⁷⁸ and we also demonstrated the underlying concept that time-tracking of IR fingerprints enhances phenotype detection.⁷⁹ However, it still remains to be proven whether decades-long health state trajectories can be decoded with IR fingerprinting.

Altogether, this study sets a framework with analytical validity and clinical utility that could reduce and streamline clinical operations, improve sample turnarounds, accelerate time to treatment for a variety of medical conditions, and risk stratify populations. We identify the value of IR fingerprinting when combined with machine learning as a multi-phenotyping modality able to fuel medical screening at very affordable costs and demonstrated potential to complement health checkups. If further developed and independently validated, the approach has the potential to provide actionable information to support and empower reliable clinical decisions while conserving resources.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- **KEY RESOURCES TABLE**
- **RESOURCE AVAILABILITY**
 - Lead contact
 - Materials availability
 - Data and code availability
- **EXPERIMENTAL MODEL AND SUBJECT DETAILS**
 - Study design and sample collection
 - Phenotype definitions, biochemical and hematological analysis
 - Exclusion of samples
- **METHOD DETAILS**
 - Sample preparation and FTIR measurements
 - Measurement preprocessing
- **QUANTIFICATION AND STATISTICAL ANALYSIS**
 - Classification analysis
 - Validation of multilabel and binary classifications
 - Multilabel classification metrics
 - Regression analysis
 - Feature selection analysis
 - Analysis software

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.xcrm.2024.101625>.

ACKNOWLEDGMENTS

This work was supported by LMU Munich, Centre for Advanced Laser Applications (CALA). The KORA study was initiated and financed by the Helmholtz Zentrum München – German Research Center for Environmental Health, which is funded by the German Federal Ministry of Education and Research (BMBF) and the State of Bavaria. Data collection in the KORA study is done in cooperation with the University Hospital of Augsburg. We thank Ferenc Krausz for the research environment, Sigrd Auweter for her support in initiating this study, and Jacqueline Hermann for her support throughout the study.

AUTHOR CONTRIBUTIONS

M.Ž. designed, initiated, and supervised the study; T.E., C.L., K.V.K., and M.Ž. designed the analytical framework; T.E. defined the data analysis methodology and performed all computational analyses, data visualizations, and data interpretations; T.E. and C.L. coordinated the data acquisition from the KORA study; B.L. and A.P. oversaw, organized, and coordinated the KORA study; C.L., D.M., and V.Z. performed the experimental measurements; F.F. supervised and contributed to the experimental measurements; K.V.K., L.R., M.H., and L.V. provided methodological and statistical advice on the data analysis; T.E. and M.Ž. wrote the manuscript; all authors contributed to revising the manuscript; and all authors approve the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 11, 2023

Revised: April 19, 2024

Accepted: June 7, 2024

Published: June 28, 2024

REFERENCES

1. Shilo, S., Rossman, H., and Segal, E. (2020). Axes of a revolution: challenges and promises of big data in healthcare. *Nat. Med.* 26, 29–38.
2. Hasin, Y., Seldin, M., and Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biol.* 18, 83.
3. Wright, J.T., and Herzberg, M.C. (2021). Science for the next century: Deep phenotyping. *J. Dent. Res.* 100, 785–789.
4. Duarte, T.T., and Spencer, C.T. (2016). Personalized proteomics: The future of precision medicine. *Proteomes* 4, 29.
5. Zhi, M., Ding, E.L., Theisen-Toupal, J., Whelan, J., and Arnaout, R. (2013). The landscape of inappropriate laboratory testing: A 15-year meta-analysis. *PLoS One* 8, e78962.
6. Kogan, A., Peleg, M., Tu, S.W., Allon, R., Khaitov, N., and Hochberg, I. (2020). Towards a goal-oriented methodology for clinical-guideline-based management recommendations for patients with multimorbidity: Gocom and its preliminary evaluation. *J. Biomed. Inf.* 112, 103587.
7. Shen, X., Kellogg, R., Panyard, D.J., Bararpour, N., Castillo, K.E., Lee-McMullen, B., Delfarah, A., Ubellacker, J., Ahadi, S., Rosenberg-Hasson, Y., et al. (2023). Multi-omics microsampling for the profiling of lifestyle-associated changes in health. *Nat. Biomed. Eng.* 8, 11–29.
8. Emwas, A.-H.M. (2015). The strengths and weaknesses of NMR spectroscopy and mass spectrometry with particular focus on metabolomics research. In *Methods in Molecular Biology* (Springer), pp. 161–193.
9. Seger, C., and Salzmann, L. (2020). After another decade: LC–MS/MS became routine in clinical diagnostics. *Clin. Biochem.* 82, 2–11.

10. Geyer, P.E., Voytik, E., Treit, P.V., Doll, S., Kleinhempel, A., Niu, L., Müller, J.B., Buchholtz, M.-L., Bader, J.M., Teupser, D., et al. (2019). Plasma proteome profiling to detect and avoid sample-related biases in biomarker studies. *EMBO Mol. Med.* **11**, e10427.
11. Buerge, T., Steinfeldt, J., Ruyoga, G., Pietzner, M., Bizzarri, D., Vojinovic, D., Upmeyer zu Belzen, J., Looch, L., Kittner, P., Christmann, L., et al. (2022). Metabolomic profiles predict individual multidisease outcomes. *Nat. Med.* **28**, 2309–2320.
12. Griffiths, P.R., De Haseth, J.A., and Winefordner, J.D. (2007). *Fourier Transform Infrared Spectrometry. Chemical Analysis: A Series of Monographs on Analytical Chemistry and its Applications*, Second Edition (Wiley-Blackwell).
13. Movasaghi, Z., Rehman, S., and ur Rehman, I. (2008). Fourier transform infrared (FTIR) spectroscopy of biological tissues. *Appl. Spectrosc. Rev.* **43**, 134–179.
14. Debska, B., and Guzowska-Swider, B. (2000). Fuzzy definition of molecular fragments in chemical structures. *J. Chem. Inf. Comput. Sci.* **40**, 325–329.
15. Miller, L.M., and Smith, R.J. (2005). Synchrotrons versus globars, point-detectors versus focal plane arrays: Selecting the best source and detector for specific infrared microspectroscopy and imaging applications. *Vib. Spectrosc.* **38**, 237–240.
16. Pupeza, I., Huber, M., Trubetskoy, M., Schweinberger, W., Hussain, S.A., Hofer, C., Fritsch, K., Poetzelberger, M., Vámos, L., Fill, E., et al. (2020). Field-resolved infrared spectroscopy of biological systems. *Nature* **577**, 52–59.
17. Kowligy, A.S., Timmers, H., Lind, A.J., Elu, U., Cruz, F.C., Schunemann, P.G., Biegert, J., and Diddams, S.A. (2019). Infrared electric field sampled frequency comb spectroscopy. *Sci. Adv.* **5**, eaaw8794.
18. Villares, G., Hugl, A., Blaser, S., and Faist, J. (2014). Dual-comb spectroscopy based on quantum-cascade-laser frequency combs. *Nat. Commun.* **5**, 5192.
19. Schwaighofer, A., Montemurro, M., Freitag, S., Kristament, C., Culzoni, M.J., and Lendl, B. (2018). Beyond Fourier transform infrared spectroscopy: External cavity quantum cascade laser-based mid-infrared transmission spectroscopy of proteins in the amide I and amide II region. *Anal. Chem.* **90**, 7072–7079.
20. Baker, M.J., Trevisan, J., Bassan, P., Bhargava, R., Butler, H.J., Dorling, K.M., Fielden, P.R., Fogarty, S.W., Fullwood, N.J., Heys, K.A., et al. (2014). Using Fourier transform IR spectroscopy to analyze biological materials. *Nat. Protoc.* **9**, 1771–1791.
21. Huber, M., Kepesidis, K.V., Voronina, L., Fleischmann, F., Fill, E., Hermann, J., Koch, I., Milger-Kneidinger, K., Kolben, T., Schulz, G.B., et al. (2021). Infrared molecular fingerprinting of blood-based liquid biopsies for the detection of cancer. *eLife* **10**, e68758.
22. Baker, M.J., Hussain, S.R., Lovergne, L., Untereiner, V., Hughes, C., Lukaszewski, R.A., Thieffn, G., and Sockalingum, G.D. (2016). Developing and understanding biofluid vibrational spectroscopy: a critical review. *Chem. Soc. Rev.* **45**, 1803–1818.
23. Martin, F.L., Kelly, J.G., Llabjani, V., Martin-Hirsch, P.L., Patel, I.I., Trevisan, J., Fullwood, N.J., and Walsh, M.J. (2010). Distinguishing cell types or populations based on the computational analysis of their infrared spectra. *Nat. Protoc.* **5**, 1748–1760.
24. Petrich, W., Lewandrowski, K.B., Muhlestein, J.B., Hammond, M.E.H., Januzzi, J.L., Lewandrowski, E.L., Pearson, R.R., Dolenko, B., Früh, J., Haass, M., et al. (2009). Potential of mid-infrared spectroscopy to aid the triage of patients with acute chest pain. *Analyst* **134**, 1092–1098.
25. Li, Y., Guo, L., Li, L., Yang, C., Guang, P., Huang, F., Chen, Z., Wang, L., and Hu, J. (2020). Early diagnosis of type 2 diabetes based on near-infrared spectroscopy combined with machine learning and aquaphotomics. *Front. Chem.* **8**, 580489.
26. de Souza, N.M.P., Machado, B.H., Koche, A., da Silva Furtado, L.B.F., Becker, D., Corbellini, V.A., and Rieger, A. (2023). Discrimination of dyslipidemia types with ATR-FTIR spectroscopy and chemometrics associated with multivariate analysis of the lipid profile, anthropometric, and pro-inflammatory biomarkers. *Clin. Chim. Acta* **540**, 117231.
27. Bhargava, R. (2023). Digital histopathology by infrared spectroscopic imaging. *Annu. Rev. Anal. Chem.* **16**, 205–230.
28. Bellisola, G., and Sorio, C. (2012). Infrared spectroscopy and microscopy in cancer research and diagnosis. *Am. J. Cancer Res.* **2**, 1–21.
29. Kazarian, S.G., and Chan, K.L.A. (2013). ATR-FTIR spectroscopic imaging: recent advances and applications to biological systems. *Analyst* **138**, 1940.
30. Townsend, D., Mijlković, M., Bird, B., Lenau, K., Old, O., Almond, M., Kendall, C., Lloyd, G., Shepherd, N., Barr, H., et al. (2015). Infrared micro-spectroscopy for cyto-pathological classification of esophageal cells. *Analyst* **140**, 2215–2223.
31. De Bruyne, S., Speckaert, M.M., and Delanghe, J.R. (2017). Applications of Mid-infrared Spectroscopy in the Clinical Laboratory Setting. *Crit. Rev. Clin. Lab. Sci.* **55**, 1–20.
32. Chen, X., Shu, W., Zhao, L., and Wan, J. (2022). Advanced mass spectrometric and spectroscopic methods coupled with machine learning for in vitro diagnosis. *View* **4**, 20220038.
33. Silveira, L., Borges, R.d.C.F., Navarro, R.S., Giana, H.E., Zangaro, R.A., Pacheco, M.T.T., and Fernandes, A.B. (2017). Quantifying glucose and lipid components in human serum by Raman spectroscopy and multivariate statistics. *Laser Med. Sci.* **32**, 787–795.
34. Perez-Guaita, D., Ventura-Gayete, J., Pérez-Rambla, C., Sancho-Andreu, M., Garrigues, S., and de la Guardia, M. (2012). Protein determination in serum and whole blood by attenuated total reflectance infrared spectroscopy. *Anal. Bioanal. Chem.* **404**, 649–656.
35. Perez-Guaita, D., Ventura-Gayete, J., Pérez-Rambla, C., Sancho-Andreu, M., Garrigues, S., and de la Guardia, M. (2013). Evaluation of infrared spectroscopy as a screening tool for serum analysis. *Microchem. J.* **106**, 202–211.
36. Anderson, D.J., Anderson, R.G., Moug, S.J., and Baker, M.J. (2020). Liquid biopsy for cancer diagnosis using vibrational spectroscopy: systematic review. *BJS Open* **4**, 554–562.
37. Finlayson, D., Rinaldi, C., and Baker, M.J. (2019). Is infrared spectroscopy ready for the clinic? *Anal. Chem.* **91**, 12117–12128.
38. Cameron, J.M., Rinaldi, C., Rutherford, S.H., Sala, A., G Theakstone, A., and Baker, M.J. (2022). Clinical spectroscopy: Lost in translation? *Appl. Spectrosc.* **76**, 393–415.
39. Holle, R., Happich, M., Löwel, H., and Wichmann, H. (2005). KORA - a research platform for population based health research. *Gesundheitswesen* **67**, 19–25.
40. Smith, S.M., Wallace, E., O'Dowd, T., and Fortin, M. (2021). Interventions for improving outcomes in patients with multimorbidity in primary care and community settings: a systematic review. *Syst. Rev.* **10**, 271.
41. Jakovljević, M., and Ostojić, L. (2013). Comorbidity and multimorbidity in medicine today: challenges and opportunities for bringing separated branches of medicine closer to each other. *Psychiatr. Danub.* **25**, 18–28.
42. Read, J., Pfahringer, B., Holmes, G., and Frank, E. (2011). Classifier chains for multi-label classification. *Mach. Learn.* **85**, 333–359.
43. Tsoumakas, G., and Katakis, I. (2007). Multi-label classification. *Int. J. Data Warehous. Min.* **3**, 1–13.
44. Godbole, S., and Sarawagi, S. (2004). Discriminative methods for multi-labeled classification. In *Advances in Knowledge Discovery and Data Mining* (Springer Berlin Heidelberg), pp. 22–30.
45. Sorower, M.S. (2010). *A Literature Survey on Algorithms for Multi-Label Learning* (Oregon State University).
46. Kasahara, R., Kino, S., Soyama, S., and Matsuura, Y. (2018). Noninvasive glucose monitoring using mid-infrared absorption spectroscopy based on a few wavenumbers. *Biomed. Opt. Express* **9**, 289–302.

Cell Reports Medicine

Article



47. Ujil, A., Koudstaal, S., Direk, K., Denaxas, S., Groenewold, R.H.H., Banejee, A., Hoes, A.W., Hemingway, H., and Asselbergs, F.W. (2019). Risk factors for incident heart failure in age- and sex-specific strata: a population-based cohort using linked electronic health records. *Eur. J. Heart Fail.* 21, 1197–1206.
48. Meisinger, C., Thorand, B., Schneider, A., Stieber, J., Döring, A., and Löwel, H. (2002). Sex differences in risk factors for incident type 2 diabetes mellitus. *Arch. Intern. Med.* 162, 82.
49. Grundy, S.M., Cleeman, J.I., Daniels, S.R., Donato, K.A., Eckel, R.H., Franklin, B.A., Gordon, D.J., Krauss, R.M., Savage, P.J., Smith, S.C., et al. (2005). Diagnosis and management of the metabolic syndrome. *Circulation* 112, 2735–2752.
50. Grundy, S.M. (2016). Metabolic syndrome update. *Trends Cardiovasc. Med.* 26, 364–373.
51. Sattar, N., McConnachie, A., Shaper, A.G., Blauw, G.J., Buckley, B.M., de Craen, A.J., Ford, I., Forouhi, N.G., Freeman, D.J., Jukema, J.W., et al. (2008). Can metabolic syndrome usefully predict cardiovascular disease and diabetes? outcome data from two prospective studies. *Lancet* 371, 1927–1935.
52. Wang, Y., Li, J., Zheng, X., Jiang, Z., Hu, S., Wadhera, R.K., Bai, X., Lu, J., Wang, Q., Li, Y., et al. (2018). Risk factors associated with major cardiovascular events 1 year after acute myocardial infarction. *JAMA Netw. Open* 1, e181079.
53. Esposito, K., Chiodini, P., Colao, A., Lenzi, A., and Giugliano, D. (2012). Metabolic syndrome and risk of cancer. *Diabetes Care* 35, 2402–2411.
54. Chen, H., Zheng, X., Zong, X., Li, Z., Li, N., Hur, J., Fritz, C.D., Chapman, W., Jr., Nickel, K.B., Tipping, A., et al. (2021). Metabolic syndrome, metabolic comorbid conditions and risk of early-onset colorectal cancer. *Gut* 70, 1147–1154.
55. Huang, P.L. (2009). A comprehensive definition for metabolic syndrome. *Dis. Model. Mech.* 2, 231–237.
56. Monnerie, S., Comte, B., Ziegler, D., Morais, J.A., Pujos-Guillot, E., and Gaudreau, P. (2020). Metabolomic and lipidomic signatures of metabolic syndrome and its physiological components in adults: A systematic review. *Sci. Rep.* 10, 669.
57. Alberti, K., Eckel, R.H., Grundy, S.M., Zimmet, P.Z., Cleeman, J.I., Donato, K.A., Fruchart, J.-C., James, W.P.T., Loria, C.M., and Smith, S.C. (2009). Harmonizing the metabolic syndrome. *Circulation* 120, 1640–1645.
58. de las Fuentes, L., Brown, A.L., Mathews, S.J., Waggoner, A.D., Soto, P.F., Gropler, R.J., and Davila-Roman, V.G. (2006). Metabolic syndrome is associated with abnormal left ventricular diastolic function independent of left ventricular mass. *Eur. Heart J.* 28, 553–559.
59. Chen, Y., Xu, W., Zhang, W., Tong, R., Yuan, A., Li, Z., Jiang, H., Hu, L., Huang, L., Xu, Y., et al. (2023). Plasma metabolic fingerprints for large-scale screening and personalized risk stratification of metabolic syndrome. *Cell Rep. Med.* 4, 101109.
60. Cai, X., Xue, Z., Zeng, F.-F., Tang, J., Yue, L., Wang, B., Ge, W., Xie, Y., Miao, Z., Gou, W., et al. (2023). Population serum proteomics uncovers a prognostic protein classifier for metabolic syndrome. *Cell Rep. Med.* 4, 101172.
61. Lacombe, C., Untereiner, V., Gobinet, C., Zater, M., Sockalingum, G.D., and Garnotel, R. (2015). Rapid screening of classic galactosemia patients: a proof-of-concept study using high-throughput FTIR analysis of plasma. *Analyst* 140, 2280–2286.
62. Barth, A. (2007). Infrared spectroscopy of proteins. *Biochim. Biophys. Acta* 1767, 1073–1101.
63. Voronina, L., Leonardo, C., Mueller-Reif, J.B., Geyer, P.E., Huber, M., Trubetskoy, M., Kepesidis, K.V., Behr, J., Mann, M., Krausz, F., and Žigman, M. (2021). Molecular origin of blood-based infrared spectroscopic fingerprints. *Angew. Chem., Int. Ed. Engl.* 60, 17060–17069.
64. Brandstetter, M., Sumalowitsch, T., Genner, A., Posch, A.E., Herwig, C., Drolz, A., Fuhrmann, V., Perkmann, T., and Lendl, B. (2013). Reagent-free monitoring of multiple clinically relevant parameters in human blood plasma using a mid-infrared quantum cascade laser based sensor system. *Analyst* 138, 4022–4028.
65. Millán, J., Pintó, X., Muñoz, A., Zúñiga, M., Rubiés-Prat, J., Pallardo, L.F., Masana, L., Mangas, A., Hernández-Mijares, A., González-Santos, P., et al. (2009). Lipoprotein ratios: physiological significance and clinical usefulness in cardiovascular prevention. *Vasc. Health Risk Manag.* 5, 757–765.
66. Su, K.-Y., and Lee, W.-L. (2020). Fourier transform infrared spectroscopy as a cancer screening and diagnostic tool: A review and prospects. *Cancers* 12, 115.
67. Civelek, M., and Lusis, A.J. (2013). Systems genetics approaches to understand complex traits. *Nat. Rev. Genet.* 15, 34–48.
68. Li, X., Ploner, A., Wang, Y., Zhan, Y., Pedersen, N.L., Magnusson, P.K., Jylhävä, J., and Hägg, S. (2021). Clinical biomarkers and associations with healthspan and lifespan: Evidence from observational and genetic data. *EBioMedicine* 66, 103318.
69. Afkarian, M., Sachs, M.C., Kestenbaum, B., Hirsch, I.B., Tuttle, K.R., Himmelfarb, J., and de Zeeuw, D.L. (2013). Kidney disease and increased mortality risk in type 2 diabetes. *J. Am. Soc. Nephrol.* 24, 302–308.
70. Peila, R., Rodriguez, B.L., and Launer, L.J.; Honolulu-Asia Aging Study (2002). Type 2 diabetes, APOE gene, and the risk for dementia and related pathologies: The Honolulu-Asia Aging Study. *Diabetes* 51, 1256–1262.
71. Poznyak, A., Grechko, A.V., Poggio, P., Myasoedova, V.A., Alfieri, V., and Orekhov, A.N. (2020). The diabetes mellitus-atherosclerosis connection: The role of lipid and glucose metabolism and chronic inflammation. *Int. J. Mol. Sci.* 21, 1835.
72. Xu, W., Qiu, C., Winblad, B., and Fratiglioni, L. (2007). The effect of borderline diabetes on the risk of dementia and Alzheimer's disease. *Diabetes* 56, 211–216.
73. US Preventive Services Task Force (2021). Screening for Prediabetes and Type 2 Diabetes: US Preventive Services Task Force Recommendation Statement. *JAMA* 326, 736–743.
74. Bogdanet, D., O'Shea, P., Lyons, C., Shafat, A., and Dunne, F. (2020). The oral glucose tolerance test—is it time for a change?—a literature review with an emphasis on pregnancy. *J. Clin. Med.* 9, 3451.
75. Edlitz, Y., and Segal, E. (2022). Prediction of type 2 diabetes mellitus onset using logistic regression-based scorecards. *eLife* 11, e71862.
76. Shilo, S., Keshet, A., Rossman, H., Godneva, A., Talmor-Barkan, Y., Aviv, Y., and Segal, E. (2024). Continuous glucose monitoring and intrapersonal variability in fasting glucose. *Nat. Med.* 30, 1424–1431.
77. Rutledge, J., Oh, H., and Wyss-Coray, T. (2022). Measuring biological age using omics data. *Nat. Rev. Genet.* 23, 715–727.
78. Huber, M., Kepesidis, K.V., Voronina, L., Božić, M., Trubetskoy, M., Harbeck, N., Krausz, F., and Žigman, M. (2021). Stability of person-specific blood-based infrared molecular fingerprints opens up prospects for health monitoring. *Nat. Commun.* 12, 1511.
79. Eissa, T., Kepesidis, K.V., Žigman, M., and Huber, M. (2023). Limits and prospects of molecular fingerprinting for phenotyping biological systems revealed through in silico modeling. *Anal. Chem.* 95, 6523–6532.
80. Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., et al. (2020). Array programming with NumPy. *Nature* 585, 357–362.
81. McKinney, W. (2010). Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference (SciPy)*, pp. 56–61.
82. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* 17, 261–272.
83. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011).

- Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830.
84. Hunter, J.D. (2007). Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95.
 85. Sujana, C., Seissler, J., Jordan, J., Rathmann, W., Koenig, W., Roden, M., Mansmann, U., Herder, C., Peters, A., Thorand, B., and Then, C. (2020). Associations of cardiac stress biomarkers with incident type 2 diabetes and changes in glucose metabolism: KORA F4/FF4 study. *Cardiovasc. Diabetol.* **19**, 178.
 86. Wichmann, H.-E., Gieger, C., and Illig, T. (2005). KORA-gen - resource for population genetics, controls and a broad spectrum of disease phenotypes. *Gesundheitswesen* **67**, 26–30.
 87. Rathmann, W., Haastert, B., Icks, A., Löwel, H., Meisinger, C., Holle, R., and Giani, G. (2003). High prevalence of undiagnosed diabetes mellitus in southern germany: Target populations for efficient screening. the KORA survey 2000. *Diabetologia* **46**, 182–189.
 88. Jellinger, P.S., Handelsman, Y., Rosenblit, P.D., Bloomgarden, Z.T., Fonseca, V.A., Garber, A.J., Grunberger, G., Guerin, C.K., Bell, D.S., Mechanick, J.I., et al. (2017). American association of clinical endocrinologists and American college of endocrinology guidelines for management of dyslipidemia and prevention of cardiovascular disease. *Endocr. Pract.* **23**, 1–87.
 89. Nantsupawat, N., Booncharoen, A., Wisetborisut, A., Jiraporncharoen, W., Pinyopornpanish, K., Chutarattanukul, L., and Angkurawaranon, C. (2019). Appropriate total cholesterol cut-offs for detection of abnormal LDL cholesterol and non-HDL cholesterol among low cardiovascular risk population. *Lipids Health Dis.* **18**, 28.
 90. Meisinger, C., Heier, M., Völzke, H., Löwel, H., Mitusch, R., Hense, H.-W., and Lüdemann, J. (2006). Regional disparities of hypertension prevalence and management within germany. *J. Hypertens.* **24**, 293–299.
 91. World Health Organization (1999). Definition, diagnosis and classification of diabetes mellitus and its complications: report of a WHO consultation. part 1, diagnosis and classification of diabetes mellitus.
 92. Elhadad, M.A., Wilson, R., Zaghloul, S.B., Huth, C., Gieger, C., Grallert, H., Graumann, J., Rathmann, W., Koenig, W., Sinner, M.F., et al. (2021). Metabolic syndrome and the plasma proteome: from association to causation. *Cardiovasc. Diabetol.* **20**, 111.
 93. Meisinger, C., Rückert, I., Stöckl, D., Thorand, B., Peters, A., Kowall, B., and Rathmann, W. (2014). Hematological parameters and prediabetes and diabetes in adults from the general population: A cross-sectional study. *J. Diabetes Metabol.* **05**.
 94. Riedl, A., Wawro, N., Gieger, C., Meisinger, C., Peters, A., Roden, M., Kronenberg, F., Herder, C., Rathmann, W., Völzke, H., et al. (2018). Identification of comprehensive metabolotypes associated with cardiometabolic diseases in the population-based KORA study. *Mol. Nutr. Food Res.* **62**, 1800117.
 95. Lau, L.H.Y., Nano, J., Cecil, A., Schederecker, F., Rathmann, W., Prehn, C., Zeller, T., Lechner, A., Adamski, J., Peters, A., and Thorand, B. (2021). Cross-sectional and prospective relationships of endogenous progesterone and estrogens with glucose metabolism in men and women: a KORA F4/FF4 study. *BMJ Open Diabetes Res. Care* **9**, e001951.
 96. Sangster, T., Major, H., Plumb, R., Wilson, A.J., and Wilson, I.D. (2006). A pragmatic and readily implemented quality control strategy for HPLC-MS and GC-MS-based metabolomic analysis. *Analyst* **131**, 1075–1078.
 97. Yang, H., Yang, S., Kong, J., Dong, A., and Yu, S. (2015). Obtaining information about protein secondary structures in aqueous solution using Fourier transform IR spectroscopy. *Nat. Protoc.* **10**, 382–396.
 98. Morais, C.L.M., Lima, K.M.G., Singh, M., and Martin, F.L. (2020). Tutorial: multivariate classification for vibrational spectroscopy in biological samples. *Nat. Protoc.* **15**, 2143–2162.
 99. Zhang, M.-L., and Zhou, Z.-H. (2014). A review on multi-label learning algorithms. *IEEE Trans. Knowl. Data Eng.* **26**, 1819–1837.
 100. Read, J., and Hollmén, J. (2014). A deep interpretation of classifier chains. In *Advances in Intelligent Data Analysis XIII* (Springer International Publishing), pp. 251–262.
 101. Marquardt, D.W., and Snee, R.D. (1975). Ridge regression in practice. *Am. Statistician* **29**, 3–20.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Human blood plasma of KORA cohort	Helmholtz Biobank	https://helmholtz-munich.de/en/epi/cohort/kora
Pooled human blood serum	BioWest, Nuaille, France	Cat# S4200-100; Lot# S1559454200
Deposited data		
Medical examination records of KORA cohort	KORA study, Holle et al. ³⁹	https://helmholtz-muenchen.managed-otrs.com/external
FTIR measurements of KORA cohort	This study	https://helmholtz-muenchen.managed-otrs.com/external
Software and algorithms		
Python (version 3.8.8)	Python Software Foundation	https://python.org
NumPy (version 1.21.2)	Harris et al. ⁸⁰	https://numpy.org
Pandas (version 1.2.4)	McKinney ⁸¹	https://pandas.pydata.org
SciPy (version 1.6.2)	Virtanen et al. ⁸²	https://scipy.org
Scikit-learn (version 0.24.1)	Pedregosa et al. ⁸³	https://scikit-learn.org
Matplotlib (version 3.5.1)	Hunter ⁸⁴	https://matplotlib.org
Other		
MIRA Analyzer (MA6) mid-infrared analyzer	Clade GmbH, Germany	https://clade.io

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Mihaela Žigman (mihaela.zigman@mpq.mpg.de).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The data of this study are available upon request by means of a project agreement from KORA (<https://helmholtz-muenchen.managed-otrs.com/external>). Requests should be sent to kora.passt@helmholtz-muenchen.de and are subject to approval by the KORA Board.
- This study does not report original code. Details on the software and algorithms used in this study are listed in the [key resources table](#) and [STAR Methods](#).
- Any additional information required to reanalyze the data reported in this study is available from the [lead contact](#) upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Study design and sample collection

The KORA study is a population-based cohort in Southern Germany.³⁹ It served as the basis for analysis in this work. The study comprised of an age- and gender-stratified sample of participants randomly drawn from the resident registration offices within the study area. The KORA S4 baseline was conducted in 1999–2001 (n = 4261) and was followed up in 2006–2008 (n = 3080) and 2013–2014 (n = 2279) — named KORA F4 and KORA FF4, respectively.⁸⁵ Medical examination data and blood plasma samples from the available subset of the KORA F4 (n = 3044) and FF4 samplings (n = 2140) were included in this work and were denoted by sample set #1 and sample set #2, respectively. Within these sample sets, some individuals participated in only one sample donation, while others participated in both. Across both sample sets, 2015 individuals participated in both donations, while 1154 individuals participated in only one donation. This results in a total of 5184 blood plasma donations from 3169 unique individuals. Data collection methods and standardized sample collections have been described in detail elsewhere.^{39,85–87} [Table S1](#) provides an overview of the cohort distributions. The KORA F4 and FF4 study methods were approved by the ethics committee of the Bavarian Chamber of Physicians, Munich (EC No. 06068).

Phenotype definitions, biochemical and hematological analysis

Individuals were identified as dyslipidemic based on abnormally elevated levels of non-HDL cholesterol with the commonly used cut-off point at 4.1 mmol/L.^{88,89} Hypertension was defined based on a blood pressure test with a reading ≥ 140 and/or 90 mmHg or by known use of antihypertensive medication.⁹⁰ The diabetic state was determined by an oral glucose tolerance test (OGTT) or a validated self-report.^{85,87} Self-reported diabetes was confirmed by contacting the responsible physicians, medical chart reviews, or by taking self-reported glucose-lowering medications.⁸⁵ For individuals with no prior diagnosis of diabetes, an OGTT test was performed and the diabetic state was determined according to the 1999 WHO diagnostic criteria.^{87,91} Prediabetes included individuals with impaired fasting glucose, impaired glucose tolerance, or both. As previously applied to the KORA F4 and FF4 samples,⁹² metabolic syndrome was defined according to the harmonized criteria proposed by the International Diabetes Federation (IDF) and American Heart Association/National Heart, Lung, and Blood Institute (AHA/NHLBI) in 2009.⁵⁷ Metabolic syndrome was present if ≥ 3 of the following 5 criteria were satisfied: (1) waist circumference ≥ 94 cm for males, ≥ 80 cm for females; (2) systolic blood pressure ≥ 130 mmHg or diastolic blood pressure ≥ 85 mmHg or treatment with antihypertensive medication; (3) fasting serum glucose ≥ 100 mg/dL or intake of antidiabetic medication; (4) serum HDL cholesterol < 40 mg/dL for males, < 50 mg/dL for females, or drug treatment for reduced HDL (fibrates); and (5) fasting serum triglycerides ≥ 150 mg/dL or drug treatment for elevated triglycerides (fibrates). Blood samples were collected from the cubital vein without stasis of all participants and immediately measured for the biochemical analysis using standardized protocols described elsewhere.^{93–95} Blood glucose levels were determined using a hexokinase method from serum.^{93,94} HbA1c was determined using a reverse-phase cation-exchange High-Pressure Liquid Chromatography (HPLC) analysis.⁹³ Total cholesterol, HDL cholesterol, and LDL cholesterol were determined using enzymatic, colorimetric routines (CHOD-PAP) from serum.^{93,94} Triglycerides were determined using an enzymatic, colorimetric routine (GPO-PAP) from serum.^{93,94} Creatinine was determined using a modified kinetic Jaffe reaction from serum. Hematological parameters (leukocytes, erythrocytes, hematocrit, platelets) were determined by impedance measurements from EDTA.^{93,94} Albumin was determined using immunonephelometry from serum.⁹⁵

Exclusion of samples

Due to missing values in the medical examination records, some samples were excluded from analysis whenever the missing values were relevant to the investigated question. The sample sizes used for each investigated question was mentioned in each corresponding figure.

METHOD DETAILS

Sample preparation and FTIR measurements

The plasma samples were stored at -80°C until analysis in this work. The samples were handled in a randomized and blinded fashion where the person performing the measurement had no access to either sample identifiers or associated medical records. In advance of the FTIR measurements, one aliquot per plasma sample was thawed and again centrifuged for 10 min at 2000 g. The supernatant was distributed into 50 μL measurement tubes and refrozen at -80°C . Thus, all FTIR measurements were performed upon two freeze-thaw cycles. Measurements were performed in the liquid phase using a commercially available automated spectrometer (MIRA Analyzer, Clade.io) with a flow-through transmission cuvette made of calcium fluoride (CaF_2) with a path length of approximately 8 μm . The spectra were acquired with a resolution of 4 cm^{-1} in a spectral range between 950 cm^{-1} and 3050 cm^{-1} . A water reference spectrum was recorded after each sample measurement to reconstruct the IR absorption spectra. To track experimental errors over extended time,⁹⁶ a measurement of quality control serum (QC, pooled human serum, BioWest, Nuaille, France) was performed after every five plasma measurements. Each measurement sequence contained up to 40 samples, including the plasma and QC samples. After each measurement sequence, the spectrometer was carefully cleaned and re-qualified according to the manufacturer's recommendations. Measurements of sample set #1 were performed over the span of 3 months. Measurements of sample set #2 were performed approximately 2.7 years prior to the measurements of sample set #1 and also spanned over 3 months. As CaF_2 is slightly soluble in water, the volume and path length of the measurement cuvette increases over time. Due to wearing off, different transmission cuvettes were used between measurements of the first and second sample sets. The spectra of the QCs were used to evaluate the measurement error. In our previous study, involving the same experimental procedures, we found that the measurement error was small when compared to the inter-person biological variability of IR spectra of serum.⁷⁸

Measurement preprocessing

Preprocessing of the IR spectra was performed as in our previous study.²¹ If the liquid sample contains less water than the reference (pure water), negative absorption occurs. This was corrected by a previously described approach where a scaled water absorption spectrum was added to each sample spectrum using a coefficient optimized such that the first derivative of the signal between 2000 and 2300 cm^{-1} is minimal.^{21,78,97} Subsequently, all spectra were truncated to 1000–3000 cm^{-1} and the region between 1800 and 2800 cm^{-1} was removed since, in plasma, it observes no significant and biologically relevant absorbance. Each measured spectrum was then treated as a vector and normalized using the L2 norm, unless otherwise mentioned in the figure legend. The normalization helps to correct for measurement noise and spectral differences that can arise from changes in the total molecular concentration in

the samples (e.g., owing to variations in sample collection, storage/handling, and processing) as described in previous work.^{78,98} Figure S2 depicts the effect normalization had on the comparability between different groups of samples.

QUANTIFICATION AND STATISTICAL ANALYSIS

Classification analysis

An L2-regularized logistic regression algorithm was used for the classification analysis. To enable multilabel predictions, where an input can be assigned to multiple labels (i.e., phenotypes), a chaining method was employed.^{42,99} The strategy consisted of fitting one logistic regression classifier per modeled label. The architecture and label order of the classification chain was depicted in Figure 2B. When training, each classifier was fit on a training set of IR spectra, in addition to the true labels of the investigated phenotypes. The first classifier in the chain had no additional information other than the input IR spectra and the true labels relevant to its prediction (dyslipidemia). Subsequent classifiers had additional access to the true labels from all preceding phenotypes when training. When making predictions, only the input IR spectra were used and the predictions of each preceding classifier got passed on to all other classifiers coming within the chain since all the true labels are hidden from the classifier chain. Naturally, the order of the classification had an effect on the predictive performance of the multilabel classification. To investigate its effects and select a model with the optimal classification order, an exhaustive grid search was carried out on all possible permutations of ordering the 5 labels to examine the predictive performance of each multilabel classification variant. The grid search was only carried out on measurements from sample set #1 where each label order was tested in a 10-fold cross-validation. A large number of label arrangements were found to lead to similar predictive efficacies (Table S4). It was reassuring to observe that the top performing models followed analytically comparable label arrangements – e.g., the classifier detecting whether an individual was healthy remained toward the end of the chain. From the total of 120 possible label arrangements, the order which maximized the exact match ratio was selected since this investigation on sample set #1 showed that consistent results would be achievable using different arrangements. The commonly used metrics of exact match ratio, Hamming score, and Hamming loss were used as measures of overall multilabel performance (further explained below).^{43–45,100} Based on the probabilities given by the classifiers, a threshold of 0.5 was selected to identify each positive class prediction from a negative class prediction. Furthermore, the ROC curve was examined for each logistic regression within the chain to observe how each binary classification performed across all possible prediction thresholds. Classifier validation is described in the following section.

Validation of multilabel and binary classifications

All reported metrics of classification performance were validated on unseen test samples (Figure S1). Relevant to the results depicted in Figures 2C, 2D, 3C and 3D, three types of validations were performed: (1) a 10-fold cross-validation on sample set #1; (2) another 10-fold cross-validation on sample set #2; and (3) a sample set-independent validation. For the sample set-independent validation, our aim was to investigate how the classifications performed when tested on samples measured years apart and on entirely different individuals from the training samples. Since a large portion of individuals overlapped between the two sample sets utilized in our study, we could not directly train on sample set #1 and test on sample set #2. Instead, we held-out approximately 10% of the individuals that overlapped between the two sample sets to be used for classifier testing. We then trained the classifier on sample set #1, using individuals not included in the held-out set. The trained classifier was then tested on the held-out portion of individuals from sample set #2. This procedure was then repeated 10 times, holding-out a different 10% of test individuals in each iteration. This allowed us to retain a large number of samples for both classifier training and testing while not violating the goal of training and testing on fully different individuals from the two samples sets. Classification performance metrics (the exact match ratio, Hamming score, Hamming loss, and ROC curves) were averaged across all test splits of the data and reported along with their standard deviations.

Multilabel classification metrics

In multilabel classifications, unlike binary or multiclass classifications, one input may be simultaneously associated with multiple outputs/labels. Interpreting the efficacy of such predictions is more complex as predictions may be entirely correct, partially correct to varying degrees, or entirely incorrect. Therefore, several metrics that judge the prediction efficacy differently should be used to judge the multilabel prediction. In this study, three established metrics were used: Exact match ratio,^{42,45} Hamming score,^{43–45} and Hamming loss.^{42,43,45} Descriptions of each metric are provided below. An illustrative example of interpreting each metric is provided in Table S5.

The exact match ratio is a strict metric that treats partially correct predictions as entirely incorrect. Given a sample $\mathbf{x}_i$, the prediction is scored with a 1 only if the prediction $\hat{\mathbf{y}}_i$ exactly matches the true label vector $\mathbf{y}_i$ associated with the sample. Otherwise, the prediction is scored with a 0. This evaluation was performed for every sample and averaged to provide the proportion of exactly matched predictions. Its values range from 0 to 1, where higher values indicate better prediction efficacy. The metric is expressed as the following where I is the indicator function:

$$\text{Exact Match Ratio} = \frac{1}{n_{\text{samples}}} \sum_{i=1}^{n_{\text{samples}}} I(\mathbf{y}_i = \hat{\mathbf{y}}_i)$$

The Hamming score, also known as the Jaccard similarity coefficient, considers partially correct predictions to different degrees. Given a sample, it is defined as the proportion of predicted positive labels matching the true labels to the total number of possible positive (predicted and actual) labels for that sample. A predicted positive label here implies the existence of a certain label/condition (e.g., a phenotype is observed for a given sample). The metric does not reward a model correctly predicting that a label does not apply to a given sample (true negative), while it penalizes false positive predictions. The metric is therefore more strict than simply calculating the fraction of predicted labels that match the true labels. Its values range from 0 to 1, where higher values indicate better prediction efficacy, and is expressed as the following:

$$\text{Hamming Score} = \frac{1}{n_{\text{samples}}} \sum_{i=1}^{n_{\text{samples}}} \frac{|\mathbf{y}_i \cap \hat{\mathbf{y}}_i|}{|\mathbf{y}_i \cup \hat{\mathbf{y}}_i|}$$

The Hamming loss is defined as the fraction of incorrectly predicted labels across all instances and labels. It considers each class individually and counts the number of times the model's predictions differ from the actual labels. Its values range from 0 to 1, where lower values indicate better prediction efficacy (i.e., lower label misclassification rate). To keep a consistent interpretation between the three metrics, where higher values indicate better predictive efficacy, we calculated 1 - Hamming loss as follows:

$$1 - \text{Hamming loss} = 1 - \frac{1}{n_{\text{samples}} * n_{\text{labels}}} \sum_{i=1}^{n_{\text{samples}}} \sum_{j=1}^{n_{\text{labels}}} \mathbb{I}(y_{ij} \neq \hat{y}_{ij})$$

In the multilabel analysis applied in our study, we considered five labels. These labels can occur in any of 12 different clinically feasible combinations, each with varying prevalence (Figure 2A). We performed a stochastic procedure to establish a benchmark for comparing the prediction value in relation to random chance. For each of the 12 unique label combinations, we randomly sampled 1000000 5-dimensional vectors that describe the labels of the samples observed in our dataset. These sampled label vectors thus preserve the clinically feasible label combinations that may occur and the prevalence of each label combination in the population (i.e., the sampling preserves the distribution of the labels). The same sampling procedure was then repeated but with a shuffling procedure, thus breaking the correspondence between the two sets of 100000 5-dimensional vectors. We then computed the three evaluation metrics for a classifier making clinically feasible predictions that follow the true label distributions but was unable to learn when to make which prediction. These estimates served as a benchmark to assess the multilabel model's predictive efficacy and were depicted by dashed lines in Figure 2C.

Regression analysis

An L2-regularized linear regression algorithm (ridge regression) was used for the regression analysis. IR spectra from both sample sets were combined together and used as the model inputs. Predictive performance was estimated by the root mean squared error (RMSE) and coefficient of determination (R^2) in 10-fold cross-validations. The metrics of evaluations were calculated on the test sets, averaged across the 10 cross-validation splits, and reported along with their standard deviation. The predicted values from the test splits were further plotted and compared to the clinically determined ones (Figure 4B).

Feature selection analysis

No explicit feature selection algorithm was applied prior to the classification and regression analysis. Instead, our approach relied on the relatively large sample sizes and the established ability of multivariate predictive analysis to discern optimal weights through L2 (ridge) regularization,¹⁰¹ alongside comprehensive testing of predictive efficacy on unseen test samples.

Analysis software

Data analysis was performed using Python (version 3.8.8). The open-source Scikit-learn package (version 0.24.1) was used for its implementations of the logistic regression and linear regression algorithms.⁸³

Cell Reports Medicine, Volume 5

Supplemental information

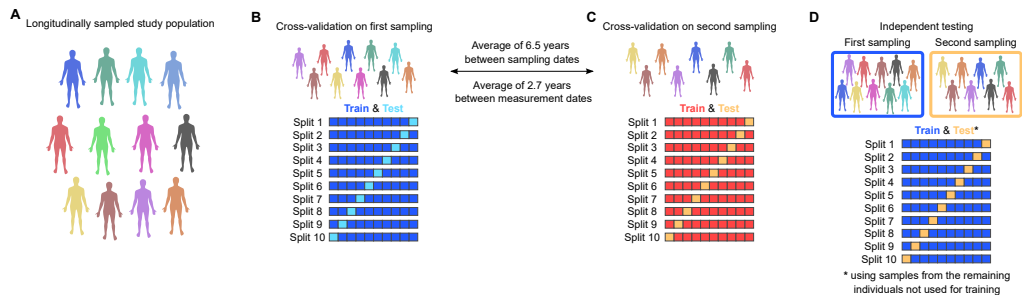
**Plasma infrared fingerprinting with
machine learning enables single-measurement
multi-phenotype health screening**

Tarek Eissa, Cristina Leonardo, Kosmas V. Kepesidis, Frank Fleischmann, Birgit Linkohr, Daniel Meyer, Viola Zoka, Marinus Huber, Liudmila Voronina, Lothar Richter, Annette Peters, and Mihaela Zigman

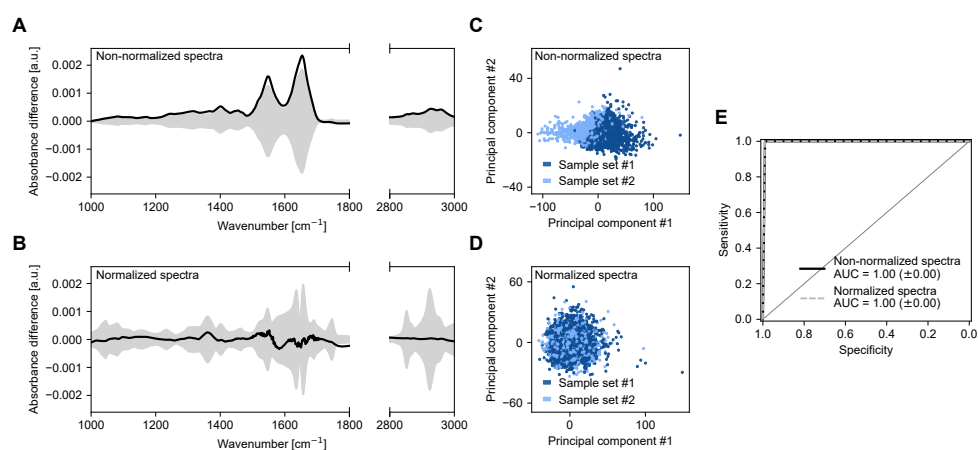
Appendix A

	Sample set #1 (3044 samples)	Sample set #2 (2140 samples)	Combined (5184 samples)
Male proportion, %	48.23	48.79	48.46
Age, mean ($\pm$ SD), years	56.19 ($\pm$ 13.24)	60.19 ($\pm$ 12.30)	57.84 ($\pm$ 13.01)
Body mass index, mean ($\pm$ SD), kg/m ²	27.62 ($\pm$ 4.81)	27.77 ($\pm$ 4.92)	27.68 ($\pm$ 4.85)
Total cholesterol, mean ($\pm$ SD), mmol/l	5.58 ($\pm$ 1.02)	5.58 ($\pm$ 1.01)	5.58 ($\pm$ 1.02)
LDL cholesterol, mean ($\pm$ SD), mmol/l	3.51 ($\pm$ 0.90)	3.48 ($\pm$ 0.90)	3.50 ($\pm$ 0.90)
HDL cholesterol, mean ($\pm$ SD), mmol/l	1.45 ($\pm$ 0.37)	1.69 ($\pm$ 0.48)	1.55 ($\pm$ 0.44)
Triglycerides, mean ($\pm$ SD), mmol/l	1.42 ($\pm$ 0.97)	1.38 ($\pm$ 0.81)	1.40 ($\pm$ 0.90)
Hematocrit, mean ($\pm$ SD), l/l	0.41 ($\pm$ 0.03)	0.41 ($\pm$ 0.03)	0.41 ($\pm$ 0.03)
Erythrocytes, mean ($\pm$ SD), /pl	4.53 ($\pm$ 0.42)	4.60 ($\pm$ 0.43)	4.56 ($\pm$ 0.42)
Leukocytes, mean ($\pm$ SD), /nl	5.98 ($\pm$ 1.88)	5.98 ($\pm$ 1.66)	5.98 ($\pm$ 1.79)
Platelets, mean ($\pm$ SD), /nl	253.99 ($\pm$ 62.29)	229.52 ($\pm$ 55.84)	243.89 ($\pm$ 60.91)
HbA1c, mean ($\pm$ SD), mmol/mol	37.21 ($\pm$ 6.80)	37.05 ($\pm$ 6.99)	37.14 ($\pm$ 6.88)
Fasting glucose, mean ($\pm$ SD), mmol/l	5.46 ($\pm$ 1.08)	5.69 ($\pm$ 1.19)	5.55 ($\pm$ 1.13)
Albumin, mean ($\pm$ SD), g/l	44.57 ($\pm$ 3.38)	42.54 ($\pm$ 3.17)	43.73 ($\pm$ 3.44)
Creatinine, mean ($\pm$ SD), μ mol/l	80.45 ($\pm$ 21.58)	80.57 ($\pm$ 20.67)	80.50 ($\pm$ 21.21)

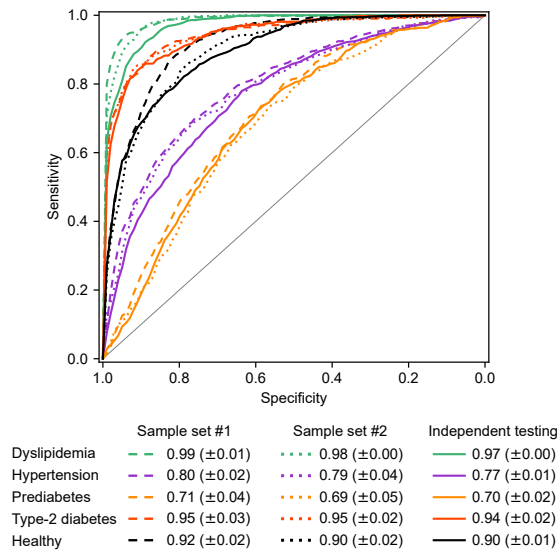
Supplementary Table S1. Description of study cohort, related to STAR Methods. Two time-separated samplings of a longitudinal population-based cohort were considered.



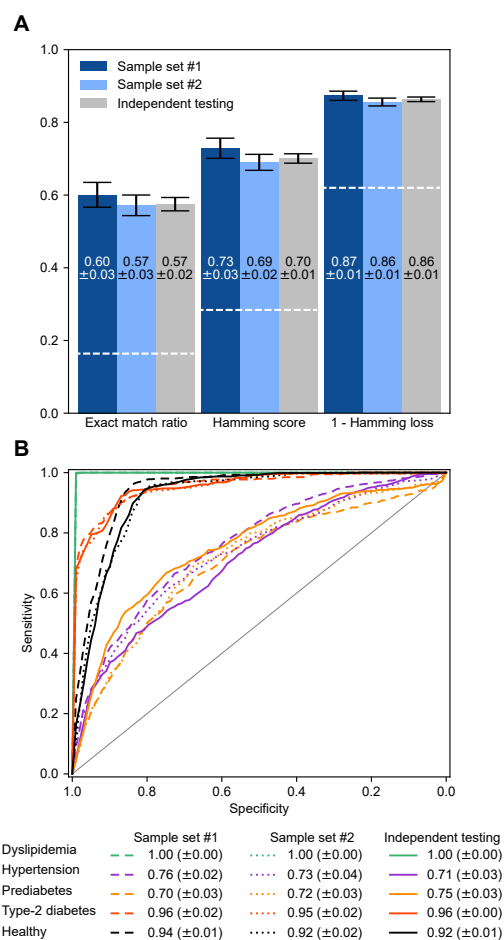
Supplementary Figure S1. Model validation procedure, related to STAR Methods. **(A)** Studied population consisting of naturally occurring heterogeneity between individuals that was sampled at two time points, thus longitudinal. **(B)** On the first sample set of the two, 10-fold cross-validation was performed to obtain metrics of model performance in 10 train and test splits of the samples. Each split was trained on approximately 90% of the samples and tested on approximately 10% of the remaining samples. **(C)** On the second sample set, the same cross-validation procedure was performed – independently of the first. **(D)** In 10 splits, model training was performed on 90% of the first sample set, leaving out a different set of individuals in each iteration. Model testing was then performed on the individuals from the second sample set who were not included in the training data from the first sample set for each data split. The model was thus tested on a dataset of independent samples and individuals.



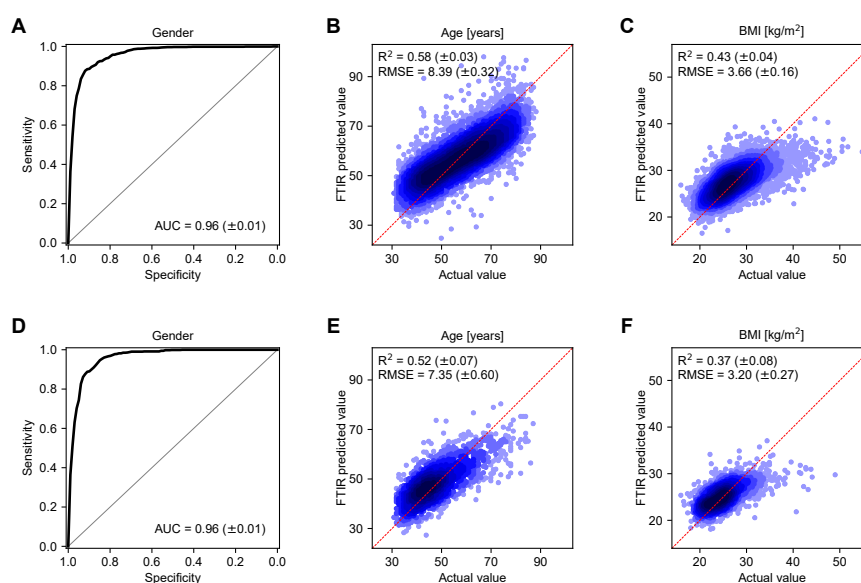
Supplementary Figure S2. Effects of variations in sample and measurement conditions on IR molecular fingerprints, related to STAR Methods. (A–B) Difference between the mean measurement of the first sample set and the mean measurement of the second on (A) non-normalized spectra and (B) normalized spectra. Gray-shaded areas depict the standard deviation of the IR fingerprints. (C–D) Principal component analysis was applied on the combined set of spectra from the two sample sets on (C) non-normalized spectra and (D) normalized spectra, color-coding the measurements from each sample set. (E) Binary classifications were carried out to determine to which sample set a measurement belonged to. Both sample sets were combined into one dataset, and the classifier was trained to assign each measurement into one sample set to determine whether there exists differences between the IR spectra of the sample sets. Two ROC curves of the classifications are depicted – one when applied to normalized spectra and one when applied to non-normalized spectra. The ROC curves are depicted on test samples as determined via cross-validation. Both classifiers were capable of perfectly determining to which sample set each measurement belonged to – revealing that measurement variations existed between the sample sets (even on normalized spectra), despite the overlapping cloud appearance depicted in (D).



Supplementary Figure S3. Receiver operating characteristic (ROC) curves for binary classifications modeling each listed outcome against all other outcomes (one-versus-rest) without the classifier chaining, related to Figure 2. For each classification, the mean cross-validated ROC curve of the test sets is depicted and the area under the curve is listed below, along with its standard deviation. Model validation was carried out as described in the main text. Compared to the ROC curves depicted in Figure 2D in the main text, the binary classification analysis here was performed with independent binary classifiers that do not utilize the chain mechanism.



Supplementary Figure S4. Capacity of clinical analytes to predict common phenotypes, related to Figure 2. Twelve clinical analytes (total cholesterol, LDL cholesterol, HDL cholesterol, triglycerides, hematocrit, erythrocytes, leukocytes, platelets, HbA1c, fasting glucose, albumin, and creatinine) were used to compare the performance of simultaneously detecting the phenotypes of interest with the performance of using the IR molecular fingerprints as exclusive predictors (shown in Figure 2 in the main text). (A–B) Test predictive performance of a model using the clinical analytes as the exclusive predictors. Model validation was carried out as described in the main text.

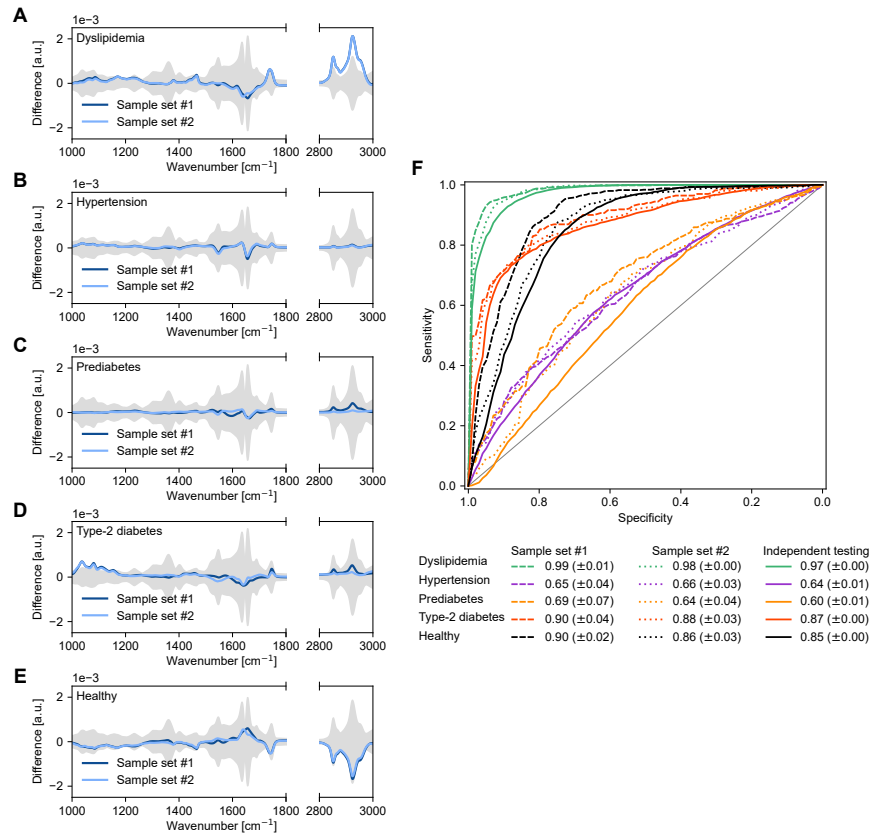


Supplementary Figure S5. Prediction of anthropometric parameters using IR molecular fingerprints, related to Figure 2. **(A)** Test ROC curves for distinguishing males from females. Classification performance was estimated by repeatedly training on 90% of the first sample set and testing on the second sample set in 10 iterations, ensuring no overlap between the longitudinally sampled individuals. **(B–C)** Prediction error of regression models trained to estimate **(B)** age and **(C)** body mass index (BMI) of the sampled individuals, combining samples from both sets. The predicted values follow from the test sets of models fit on training sets in a 10-fold cross-validation where each point represents a measurement. **(D–F)** Same procedure, but applied on the healthy subset of the population to reduce the contribution of the health phenotypes on the anthropometric parameter predictions.

Appendix A

		Sample set #1		Sample set #2	
		Controls	Cases	Controls	Cases
Dyslipidemia	<i>n</i>	1298	1298	856	856
	Male proportion, %	50.8	50.4	51.6	51.8
	Age, mean ($\pm$ SD), years	56.3 ($\pm$ 13.3)	57.8 ($\pm$ 12.1)	60.7 ($\pm$ 11.5)	60.6 ($\pm$ 11.3)
	Body mass index, mean ($\pm$ SD), kg/m ²	27.5 ($\pm$ 4.8)	28.2 ($\pm$ 4.5)	28.4 ($\pm$ 4.8)	28.6 ($\pm$ 4.7)
Hypertension	<i>n</i>	757	757	551	551
	Male proportion, %	50.5	50.5	49.7	49.7
	Age, mean ($\pm$ SD), years	59.4 ($\pm$ 11.3)	61.1 ($\pm$ 11.4)	63.1 ($\pm$ 10.6)	64.6 ($\pm$ 10.4)
	Body mass index, mean ($\pm$ SD), kg/m ²	28.0 ($\pm$ 4.1)	28.5 ($\pm$ 4.2)	27.9 ($\pm$ 4.3)	28.5 ($\pm$ 4.4)
Prediabetes	<i>n</i>	492	492	375	375
	Male proportion, %	52.8	52.8	57.1	57.1
	Age, mean ($\pm$ SD), years	63.1 ($\pm$ 11.0)	63.2 ($\pm$ 11.0)	65.1 ($\pm$ 11.1)	65.2 ($\pm$ 11.0)
	Body mass index, mean ($\pm$ SD), kg/m ²	29.7 ($\pm$ 4.7)	29.7 ($\pm$ 4.8)	29.6 ($\pm$ 4.8)	29.6 ($\pm$ 4.8)
Type-2 diabetes	<i>n</i>	335	335	295	295
	Male proportion, %	59.4	59.4	60.0	60.0
	Age, mean ($\pm$ SD), years	66.9 ($\pm$ 9.3)	67.1 ($\pm$ 9.3)	69.2 ($\pm$ 9.9)	69.4 ($\pm$ 10.1)
	Body mass index, mean ($\pm$ SD), kg/m ²	31.1 ($\pm$ 5.0)	31.2 ($\pm$ 5.1)	30.9 ($\pm$ 5.2)	31.1 ($\pm$ 5.3)
Healthy	<i>n</i>	715	715	526	526
	Male proportion, %	45.5	45.5	43.9	43.9
	Age, mean ($\pm$ SD), years	51.0 ($\pm$ 11.2)	49.3 ($\pm$ 11.4)	56.1 ($\pm$ 10.5)	54.5 ($\pm$ 10.8)
	Body mass index, mean ($\pm$ SD), kg/m ²	26.2 ($\pm$ 3.9)	25.8 ($\pm$ 4.0)	26.5 ($\pm$ 3.9)	26.0 ($\pm$ 4.1)

Supplementary Table S2. Description of cohorts used in case-control matched classifications, related to Figure 2 and Figure S6. Matching was performed based on gender, age, and body mass index of the sampled individuals in a pair-wise fashion.

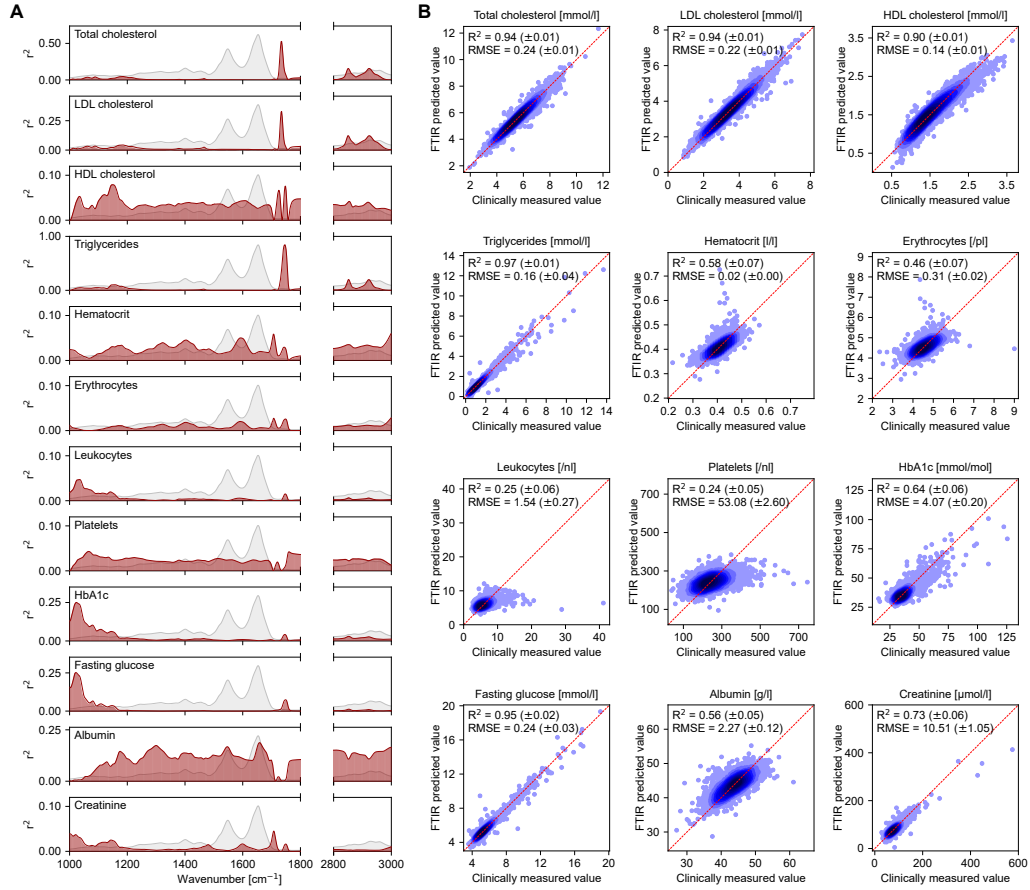


Supplementary Figure S6. Phenotype detection in gender, age, and body mass index matched cohorts using IR molecular fingerprints, related to Figure 2 and Table S2. (A–E) Difference between the mean spectral measurement of cases and controls for each condition in the first sample set (dark blue) and second sample set (light blue). The “healthy” category includes individuals having none of the other listed conditions. (F) ROC curves for the binary classifications detecting each phenotype. The mean cross-validated ROC curve of the test sets is depicted and the area under the curve is listed, along with its standard deviation, for each binary classification.

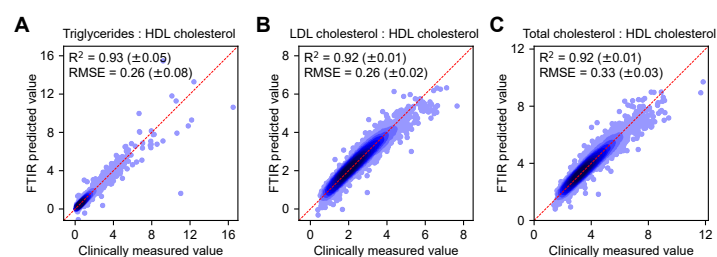
Appendix A

Sample set #1 (baseline)						
No. of risk factors	Sample count	Visceral obesity	Hypertriglyceridemia	Low HDL cholesterol	Hypertension	Hyperglycemia
0	547	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
1	709	473 (66.7%)	30 (4.2%)	40 (5.6%)	105 (14.8%)	61 (8.6%)
2	659	555 (84.2%)	146 (22.2%)	169 (25.6%)	264 (40.1%)	184 (27.9%)
3	548	504 (92.0%)	318 (58.0%)	238 (43.4%)	296 (54.0%)	288 (52.6%)
4	365	347 (95.1%)	333 (91.2%)	278 (76.2%)	214 (58.6%)	288 (78.9%)
5	163	163 (100.0%)	163 (100.0%)	163 (100.0%)	163 (100.0%)	163 (100.0%)
Sample set #2 (follow-up)						
No. of risk factors	Sample count	Visceral obesity	Hypertriglyceridemia	Low HDL cholesterol	Hypertension	Hyperglycemia
0	291	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
1	562	434 (77.2%)	14 (2.5%)	7 (1.2%)	39 (6.9%)	68 (12.1%)
2	488	432 (88.5%)	107 (21.9%)	50 (10.2%)	130 (26.6%)	257 (52.7%)
3	375	361 (96.3%)	215 (57.3%)	135 (36.0%)	172 (45.9%)	242 (64.5%)
4	290	278 (95.9%)	281 (96.9%)	216 (74.5%)	121 (41.7%)	264 (91.0%)
5	97	97 (100.0%)	97 (100.0%)	97 (100.0%)	97 (100.0%)	97 (100.0%)

Supplementary Table S3. Breakdown of metabolic risk factors for pre-MetS and MetS investigations, related to Figure 3. Each row represents a group of individuals with varying numbers of concurrent risk factors. The sample counts and percentages are listed according to the distribution of risk factors observed in each group. Distributions of the first and second sample sets are listed separately. Samples with unknown values for any of the risk factors were excluded.



Supplementary Figure S7. Classical clinical analytes reflected in non-normalized IR molecular fingerprints, related to Figure 4. **(A)** Pearson correlation coefficient (red curves) between the concentrations of each clinical analyte and the absorbance at each wavenumber between 1000 and 3000 cm^{-1} for non-normalized spectra. The mean absorbance spectrum of all measured IR spectra ($n = 5184$) is overlaid in gray on each panel as a visual reference for the shape of the spectrum. **(B)** Performance of quantitatively predicting clinical analytes using the non-normalized IR spectra of the measured population. Regression algorithms were trained for each parameter to capture the relations between the clinically-measured values and the multivariate spectral features. The predicted values follow from the test sets of 10-fold cross-validations. Each point represents a measurement. The mean coefficient of determination (R^2) and the root mean squared error (RMSE) are listed for each parameter, along with their standard deviations across the test splits. The diagonal (dashed red line) is a visual reference for a perfect fit.



Supplementary Figure S8. Performance of quantitatively predicting cholesterol ratios using IR molecular fingerprints, related to Figure 4. Ratios were calculated by dividing the clinically measured mmol/l concentrations of triglycerides by HDL cholesterol (A), LDL cholesterol by HDL cholesterol (B), and total cholesterol by HDL cholesterol (C). Regression algorithms were trained for each ratio to capture the relations between the calculated values and the multivariate spectral features. Predicted values follow from test sets of models fit on training sets in a 10-fold cross-validation. Each point represents a measurement. Model performance was measured by the mean coefficient of determination (R^2) and the root mean squared error (RMSE) and is listed for each parameter, along the standard deviation, across the test sets of the cross-validation. The diagonal (dashed red line) represents a visual reference for a perfect fit.

Appendix A

Supplementary Table S4. Predictive performance of multilabel classification using different label arrangements for optimal model selection, related to STAR Methods. To select an optimal classification order, several permutations were tested to observe their influence on the overall predictive efficacy. For each listed order, model performance was estimated by cross-validating on only sample set #1 to avoid introducing model selection bias by leaking information from sample set #2 (which was used for the independent testing). The mean and standard deviation of the calculated exact match ratio, Hamming score, and Hamming loss is listed for each of the possible 120 ordering of the 5 binary classifiers in the chain. The last entry in the table lists the performance of a non-chained model where each of the 5 phenotypes was detected with an independent binary classifier. Abbreviations: T2D – type-2 diabetes; PD – prediabetes.

#	Label order	Exact match ratio	Hamming score	1 - Hamming loss
1	Dyslipidemia → Hypertension → T2D → PD → Healthy	0.58 (±0.03)	0.71 (±0.03)	0.86 (±0.01)
2	Hypertension → T2D → Dyslipidemia → PD → Healthy	0.58 (±0.03)	0.71 (±0.03)	0.86 (±0.01)
3	Hypertension → Dyslipidemia → T2D → PD → Healthy	0.57 (±0.03)	0.71 (±0.03)	0.86 (±0.01)
4	PD → Hypertension → T2D → Dyslipidemia → Healthy	0.57 (±0.03)	0.70 (±0.03)	0.86 (±0.01)
5	Dyslipidemia → Hypertension → PD → T2D → Healthy	0.57 (±0.03)	0.70 (±0.03)	0.86 (±0.01)
6	T2D → Hypertension → Dyslipidemia → PD → Healthy	0.57 (±0.03)	0.71 (±0.03)	0.86 (±0.01)
7	PD → Hypertension → Dyslipidemia → T2D → Healthy	0.57 (±0.04)	0.70 (±0.03)	0.86 (±0.01)
8	Dyslipidemia → PD → Hypertension → T2D → Healthy	0.57 (±0.04)	0.70 (±0.03)	0.86 (±0.01)
9	Hypertension → PD → T2D → Dyslipidemia → Healthy	0.57 (±0.03)	0.70 (±0.03)	0.86 (±0.01)
10	Hypertension → PD → Dyslipidemia → T2D → Healthy	0.57 (±0.03)	0.70 (±0.03)	0.86 (±0.01)
11	Hypertension → Dyslipidemia → PD → T2D → Healthy	0.57 (±0.03)	0.70 (±0.03)	0.86 (±0.01)
12	PD → Dyslipidemia → Hypertension → T2D → Healthy	0.57 (±0.04)	0.70 (±0.03)	0.86 (±0.01)
13	Dyslipidemia → Hypertension → Healthy → PD → T2D	0.57 (±0.03)	0.70 (±0.03)	0.86 (±0.01)
14	PD → T2D → Hypertension → Dyslipidemia → Healthy	0.57 (±0.04)	0.70 (±0.03)	0.86 (±0.01)
15	Dyslipidemia → T2D → Hypertension → PD → Healthy	0.57 (±0.03)	0.70 (±0.03)	0.86 (±0.01)
16	T2D → Dyslipidemia → Hypertension → PD → Healthy	0.57 (±0.03)	0.70 (±0.03)	0.86 (±0.01)
17	Hypertension → T2D → PD → Dyslipidemia → Healthy	0.57 (±0.03)	0.70 (±0.03)	0.86 (±0.01)
18	Dyslipidemia → Hypertension → PD → Healthy → T2D	0.57 (±0.03)	0.70 (±0.03)	0.86 (±0.01)
19	Hypertension → Dyslipidemia → Healthy → PD → T2D	0.57 (±0.03)	0.70 (±0.03)	0.86 (±0.01)
20	PD → Hypertension → Dyslipidemia → Healthy → T2D	0.57 (±0.03)	0.70 (±0.03)	0.86 (±0.01)
21	T2D → Hypertension → PD → Dyslipidemia → Healthy	0.57 (±0.03)	0.70 (±0.03)	0.86 (±0.01)
22	Dyslipidemia → PD → Hypertension → Healthy → T2D	0.57 (±0.04)	0.70 (±0.03)	0.86 (±0.01)
23	Dyslipidemia → PD → T2D → Hypertension → Healthy	0.57 (±0.04)	0.70 (±0.03)	0.86 (±0.01)
24	PD → T2D → Dyslipidemia → Hypertension → Healthy	0.57 (±0.04)	0.70 (±0.03)	0.86 (±0.01)
25	Hypertension → PD → Dyslipidemia → Healthy → T2D	0.57 (±0.03)	0.70 (±0.03)	0.86 (±0.01)
26	Hypertension → Dyslipidemia → PD → Healthy → T2D	0.57 (±0.03)	0.70 (±0.03)	0.86 (±0.01)
27	PD → Dyslipidemia → Hypertension → Healthy → T2D	0.57 (±0.04)	0.70 (±0.03)	0.86 (±0.01)
28	PD → Dyslipidemia → T2D → Hypertension → Healthy	0.57 (±0.04)	0.70 (±0.03)	0.86 (±0.01)
29	Dyslipidemia → Hypertension → T2D → Healthy → PD	0.57 (±0.03)	0.70 (±0.03)	0.86 (±0.01)
30	Dyslipidemia → Hypertension → Healthy → T2D → PD	0.57 (±0.03)	0.70 (±0.03)	0.86 (±0.01)
31	PD → Hypertension → T2D → Healthy → Dyslipidemia	0.57 (±0.03)	0.70 (±0.03)	0.86 (±0.01)
32	T2D → PD → Hypertension → Dyslipidemia → Healthy	0.57 (±0.04)	0.70 (±0.03)	0.86 (±0.01)
33	Hypertension → T2D → Dyslipidemia → Healthy → PD	0.57 (±0.03)	0.70 (±0.03)	0.86 (±0.01)
34	Hypertension → PD → T2D → Healthy → Dyslipidemia	0.57 (±0.03)	0.70 (±0.03)	0.86 (±0.01)
35	Hypertension → Dyslipidemia → T2D → Healthy → PD	0.57 (±0.03)	0.70 (±0.03)	0.86 (±0.01)
36	PD → T2D → Hypertension → Healthy → Dyslipidemia	0.57 (±0.04)	0.70 (±0.03)	0.86 (±0.01)
37	Dyslipidemia → T2D → PD → Hypertension → Healthy	0.57 (±0.04)	0.70 (±0.03)	0.86 (±0.01)
38	T2D → Dyslipidemia → PD → Hypertension → Healthy	0.57 (±0.04)	0.70 (±0.03)	0.86 (±0.01)
39	Hypertension → Dyslipidemia → Healthy → T2D → PD	0.57 (±0.03)	0.70 (±0.03)	0.86 (±0.01)
40	T2D → Hypertension → Dyslipidemia → Healthy → PD	0.57 (±0.03)	0.70 (±0.03)	0.86 (±0.01)
41	Hypertension → T2D → PD → Healthy → Dyslipidemia	0.57 (±0.03)	0.70 (±0.03)	0.86 (±0.01)
42	T2D → PD → Dyslipidemia → Hypertension → Healthy	0.57 (±0.04)	0.70 (±0.03)	0.86 (±0.01)
43	T2D → Hypertension → PD → Healthy → Dyslipidemia	0.57 (±0.03)	0.70 (±0.03)	0.86 (±0.01)
44	Dyslipidemia → T2D → Hypertension → Healthy → PD	0.57 (±0.04)	0.70 (±0.03)	0.86 (±0.01)
45	T2D → Dyslipidemia → Hypertension → Healthy → PD	0.57 (±0.03)	0.70 (±0.03)	0.86 (±0.01)
46	T2D → PD → Hypertension → Healthy → Dyslipidemia	0.57 (±0.04)	0.70 (±0.03)	0.86 (±0.01)
47	PD → Hypertension → Healthy → Dyslipidemia → T2D	0.56 (±0.04)	0.69 (±0.03)	0.85 (±0.01)
48	Dyslipidemia → PD → T2D → Healthy → Hypertension	0.56 (±0.04)	0.70 (±0.03)	0.86 (±0.01)
49	PD → Hypertension → Healthy → T2D → Dyslipidemia	0.56 (±0.03)	0.69 (±0.03)	0.85 (±0.01)
50	PD → T2D → Dyslipidemia → Healthy → Hypertension	0.56 (±0.04)	0.70 (±0.03)	0.86 (±0.01)
51	Hypertension → PD → Healthy → Dyslipidemia → T2D	0.56 (±0.03)	0.69 (±0.03)	0.85 (±0.01)
52	PD → Dyslipidemia → T2D → Healthy → Hypertension	0.56 (±0.04)	0.70 (±0.03)	0.86 (±0.01)
53	Hypertension → PD → Healthy → T2D → Dyslipidemia	0.56 (±0.03)	0.69 (±0.03)	0.85 (±0.01)
54	Dyslipidemia → PD → Healthy → Hypertension → T2D	0.56 (±0.03)	0.70 (±0.02)	0.86 (±0.01)
55	PD → Dyslipidemia → Healthy → Hypertension → T2D	0.56 (±0.03)	0.70 (±0.02)	0.86 (±0.01)
56	Dyslipidemia → PD → Healthy → T2D → Hypertension	0.56 (±0.03)	0.70 (±0.02)	0.86 (±0.01)
57	PD → Dyslipidemia → Healthy → T2D → Hypertension	0.56 (±0.03)	0.70 (±0.02)	0.86 (±0.01)
58	Dyslipidemia → T2D → PD → Healthy → Hypertension	0.56 (±0.04)	0.70 (±0.03)	0.86 (±0.01)
59	T2D → Dyslipidemia → PD → Healthy → Hypertension	0.56 (±0.04)	0.70 (±0.03)	0.86 (±0.01)
60	T2D → PD → Dyslipidemia → Healthy → Hypertension	0.56 (±0.04)	0.70 (±0.03)	0.86 (±0.01)

Continued on next page.

Appendix A

Supplementary Table S4. *Continued from previous page.*

#	Label order	Exact match ratio	Hamming score	1 - Hamming loss
61	Hypertension → T2D → Healthy → Dyslipidemia → PD	0.56 (±0.03)	0.69 (±0.03)	0.86 (±0.01)
62	Hypertension → Healthy → PD → Dyslipidemia → T2D	0.56 (±0.03)	0.69 (±0.03)	0.85 (±0.01)
63	Hypertension → T2D → Healthy → PD → Dyslipidemia	0.56 (±0.03)	0.69 (±0.03)	0.86 (±0.01)
64	T2D → Hypertension → Healthy → Dyslipidemia → PD	0.56 (±0.03)	0.69 (±0.03)	0.86 (±0.01)
65	Hypertension → Healthy → PD → T2D → Dyslipidemia	0.56 (±0.03)	0.69 (±0.03)	0.85 (±0.01)
66	Dyslipidemia → Healthy → Hypertension → PD → T2D	0.56 (±0.03)	0.69 (±0.02)	0.85 (±0.01)
67	T2D → Hypertension → Healthy → PD → Dyslipidemia	0.56 (±0.03)	0.69 (±0.03)	0.86 (±0.01)
68	Hypertension → Healthy → Dyslipidemia → PD → T2D	0.56 (±0.03)	0.69 (±0.03)	0.85 (±0.01)
69	Dyslipidemia → Healthy → PD → Hypertension → T2D	0.55 (±0.03)	0.69 (±0.02)	0.85 (±0.01)
70	Dyslipidemia → Healthy → PD → T2D → Hypertension	0.55 (±0.03)	0.69 (±0.02)	0.85 (±0.01)
71	Hypertension → Healthy → Dyslipidemia → T2D → PD	0.55 (±0.03)	0.69 (±0.03)	0.86 (±0.01)
72	Dyslipidemia → Healthy → Hypertension → T2D → PD	0.55 (±0.03)	0.69 (±0.03)	0.86 (±0.01)
73	Dyslipidemia → Healthy → T2D → Hypertension → PD	0.55 (±0.03)	0.69 (±0.02)	0.86 (±0.01)
74	Hypertension → Healthy → T2D → Dyslipidemia → PD	0.55 (±0.03)	0.68 (±0.03)	0.86 (±0.01)
75	Hypertension → Healthy → T2D → PD → Dyslipidemia	0.55 (±0.03)	0.69 (±0.03)	0.86 (±0.01)
76	Dyslipidemia → Healthy → T2D → PD → Hypertension	0.55 (±0.03)	0.69 (±0.03)	0.86 (±0.01)
77	Dyslipidemia → T2D → Healthy → Hypertension → PD	0.55 (±0.03)	0.69 (±0.03)	0.86 (±0.01)
78	T2D → Dyslipidemia → Healthy → Hypertension → PD	0.55 (±0.03)	0.69 (±0.03)	0.86 (±0.01)
79	Dyslipidemia → T2D → Healthy → PD → Hypertension	0.55 (±0.04)	0.69 (±0.03)	0.86 (±0.01)
80	T2D → Dyslipidemia → Healthy → PD → Hypertension	0.55 (±0.03)	0.69 (±0.03)	0.86 (±0.01)
81	PD → T2D → Healthy → Hypertension → Dyslipidemia	0.54 (±0.03)	0.67 (±0.02)	0.85 (±0.01)
82	PD → T2D → Healthy → Dyslipidemia → Hypertension	0.54 (±0.03)	0.67 (±0.02)	0.85 (±0.01)
83	T2D → PD → Healthy → Hypertension → Dyslipidemia	0.53 (±0.03)	0.67 (±0.02)	0.85 (±0.01)
84	T2D → PD → Healthy → Dyslipidemia → Hypertension	0.53 (±0.03)	0.67 (±0.02)	0.85 (±0.01)
85	PD → Healthy → Dyslipidemia → Hypertension → T2D	0.53 (±0.03)	0.67 (±0.02)	0.84 (±0.01)
86	PD → Healthy → Dyslipidemia → T2D → Hypertension	0.53 (±0.03)	0.67 (±0.02)	0.84 (±0.01)
87	PD → Healthy → Hypertension → Dyslipidemia → T2D	0.53 (±0.03)	0.67 (±0.02)	0.84 (±0.01)
88	PD → Healthy → T2D → Dyslipidemia → Hypertension	0.53 (±0.03)	0.67 (±0.02)	0.84 (±0.01)
89	PD → Healthy → T2D → Hypertension → Dyslipidemia	0.53 (±0.02)	0.66 (±0.02)	0.84 (±0.01)
90	PD → Healthy → Hypertension → T2D → Dyslipidemia	0.53 (±0.03)	0.66 (±0.02)	0.84 (±0.01)
91	Healthy → Dyslipidemia → Hypertension → PD → T2D	0.53 (±0.03)	0.66 (±0.02)	0.84 (±0.01)
92	Healthy → PD → Dyslipidemia → Hypertension → T2D	0.52 (±0.02)	0.66 (±0.02)	0.84 (±0.01)
93	Healthy → Hypertension → PD → Dyslipidemia → T2D	0.52 (±0.03)	0.66 (±0.02)	0.84 (±0.01)
94	Healthy → PD → Dyslipidemia → T2D → Hypertension	0.52 (±0.02)	0.66 (±0.02)	0.84 (±0.01)
95	Healthy → Dyslipidemia → PD → T2D → Hypertension	0.52 (±0.03)	0.66 (±0.02)	0.84 (±0.01)
96	Healthy → Dyslipidemia → PD → Hypertension → T2D	0.52 (±0.02)	0.66 (±0.02)	0.84 (±0.01)
97	Healthy → PD → Hypertension → Dyslipidemia → T2D	0.52 (±0.02)	0.66 (±0.02)	0.84 (±0.01)
98	Healthy → Hypertension → PD → T2D → Dyslipidemia	0.52 (±0.02)	0.66 (±0.02)	0.84 (±0.01)
99	Healthy → PD → T2D → Dyslipidemia → Hypertension	0.52 (±0.02)	0.66 (±0.02)	0.84 (±0.01)
100	Healthy → Hypertension → Dyslipidemia → PD → T2D	0.52 (±0.02)	0.66 (±0.02)	0.84 (±0.01)
101	Healthy → PD → T2D → Hypertension → Dyslipidemia	0.52 (±0.02)	0.66 (±0.02)	0.84 (±0.01)
102	T2D → Healthy → Dyslipidemia → Hypertension → PD	0.52 (±0.03)	0.66 (±0.02)	0.85 (±0.01)
103	Healthy → PD → Hypertension → T2D → Dyslipidemia	0.52 (±0.02)	0.66 (±0.02)	0.84 (±0.01)
104	Healthy → Dyslipidemia → Hypertension → T2D → PD	0.52 (±0.03)	0.66 (±0.02)	0.85 (±0.01)
105	Healthy → Dyslipidemia → T2D → Hypertension → PD	0.52 (±0.03)	0.66 (±0.02)	0.85 (±0.01)
106	T2D → Healthy → PD → Dyslipidemia → Hypertension	0.52 (±0.03)	0.66 (±0.02)	0.85 (±0.01)
107	T2D → Healthy → PD → Hypertension → Dyslipidemia	0.52 (±0.03)	0.66 (±0.03)	0.85 (±0.01)
108	T2D → Healthy → Dyslipidemia → PD → Hypertension	0.52 (±0.03)	0.66 (±0.02)	0.85 (±0.01)
109	T2D → Healthy → Hypertension → Dyslipidemia → PD	0.52 (±0.03)	0.66 (±0.03)	0.85 (±0.01)
110	Healthy → T2D → Dyslipidemia → Hypertension → PD	0.52 (±0.02)	0.66 (±0.02)	0.85 (±0.01)
111	Healthy → Dyslipidemia → T2D → PD → Hypertension	0.52 (±0.03)	0.66 (±0.02)	0.85 (±0.01)
112	Healthy → Hypertension → Dyslipidemia → T2D → PD	0.52 (±0.02)	0.66 (±0.02)	0.85 (±0.01)
113	T2D → Healthy → Hypertension → PD → Dyslipidemia	0.52 (±0.03)	0.66 (±0.03)	0.85 (±0.01)
114	Healthy → T2D → PD → Dyslipidemia → Hypertension	0.52 (±0.03)	0.66 (±0.02)	0.85 (±0.01)
115	Healthy → T2D → Dyslipidemia → PD → Hypertension	0.52 (±0.03)	0.66 (±0.02)	0.85 (±0.01)
116	Healthy → Hypertension → T2D → Dyslipidemia → PD	0.52 (±0.02)	0.66 (±0.02)	0.85 (±0.01)
117	Healthy → Hypertension → T2D → PD → Dyslipidemia	0.52 (±0.03)	0.66 (±0.02)	0.85 (±0.01)
118	Healthy → T2D → PD → Hypertension → Dyslipidemia	0.52 (±0.02)	0.66 (±0.02)	0.85 (±0.01)
119	Healthy → T2D → Hypertension → Dyslipidemia → PD	0.52 (±0.03)	0.66 (±0.02)	0.85 (±0.01)
120	Healthy → T2D → Hypertension → PD → Dyslipidemia	0.52 (±0.03)	0.66 (±0.02)	0.85 (±0.01)
121	Non-chained model (label order not applicable)	0.50 (±0.02)	0.64 (±0.02)	0.86 (±0.01)

Appendix A

Input	True label ($\mathbf{y}_i$)	Model output ($\hat{\mathbf{y}}_i$)	Exact match ratio	Hamming score	1 - Hamming loss
$\mathbf{x}_1$	[1 0 0 0 0]	[0 1 0 0 0]	0.00	0/2 = 0.00	3/5 = 0.60
$\mathbf{x}_2$	[1 0 0 0 0]	[1 1 0 0 0]	0.00	1/2 = 0.50	4/5 = 0.80
$\mathbf{x}_3$	[1 1 0 0 0]	[1 0 1 0 0]	0.00	1/3 = 0.33	3/5 = 0.60
$\mathbf{x}_4$	[1 1 0 0 0]	[1 1 1 0 0]	0.00	2/3 = 0.67	4/5 = 0.80
$\mathbf{x}_5$	[1 1 1 0 0]	[1 1 1 1 0]	0.00	3/4 = 0.75	4/5 = 0.80
$\mathbf{x}_6$	[1 1 1 1 0]	[1 1 1 1 0]	1.00	4/4 = 1.00	5/5 = 1.00
Average			0.17	0.54	0.77

Supplementary Table S5. Example of evaluating multilabel classification metrics, related to STAR Methods. The exact match ratio, Hamming score, and Hamming loss metrics were computed and provided as examples to demonstrate the accuracy assessment between predictions and true labels. $\mathbf{x}_i$ represents an input sample, $\mathbf{y}_i$ is a five-dimensional vector of the true labels for the sample, and $\hat{\mathbf{y}}_i$ is the vector of predicted labels. Erroneous label predictions are highlighted in red.

Appendix B

The perils of molecular interpretations from vibrational spectra of complex samples

Copyright notice: © 2024 The Author(s). This is an open access article distributed under the terms of the Creative Commons CC-BY license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The CC-BY 4.0 license is printed in Appendix E.

The Perils of Molecular Interpretations from Vibrational Spectra of Complex Samples

Tarek Eissa, Liudmila Voronina, Marinus Huber, Frank Fleischmann, and Mihaela Žigman*

Abstract: Vibrational spectroscopy is a widely used technique for chemical characterizations across various analytical sciences. Its applications are increasingly extending to the analysis of complex samples such as biofluids, providing high-throughput molecular profiling. While powerful, the technique suffers from an inherent limitation: The overlap of absorption information across different spectral domains hinders the capacity to identify individual molecular substances contributing to measured signals. Despite the awareness of this challenge, the difficulty of analyzing multi-molecular spectra is often underestimated, leading to unsubstantiated molecular interpretations. Here, we examine the prevalent overreliance on spectral band assignment and illuminate the pitfalls of correlating spectral signals to discrete molecular entities or physiological states without rigorous validation. Focusing on blood-based infrared spectroscopy, we provide examples illustrating how peak overlap among different substances, relative substance concentrations, and preprocessing steps can lead to erroneous interpretations. We advocate for a viewpoint shift towards a more careful understanding of complex spectra, which shall lead to either accepting their fingerprinting nature and leveraging machine learning analysis – or involving additional measurement modalities for robust molecular interpretations. Aiming to help translate and improve analytical practices within the field, we highlight the limitations of molecular interpretations and feature their viable applications.

Introduction

Infrared (IR) spectroscopy has traditionally been taught as a powerful modality for analyzing the structure of individual organic compounds, identifying simple chemicals, or for use in industrial quality control and process monitoring procedures.^[1–5] More recently, its applications have expanded to the quantitative analysis of complex multi-molecular mixtures, including innovative applications in biomedical spectroscopy and photonics.^[1,2,6–14] This expansion is partly attributed to advancements in instrumentation – such as improved spectral resolution, a broader availability of Fourier transform IR (FTIR) spectrometers, and, lately, laser-based spectroscopic methods.^[1,2,15–19] Additionally, progress in computational methods and molecular biology have further driven IR spectroscopy to become a prominent tool for high-throughput probing of biological specimens.^[1,2,8,20]

IR spectroscopy probes the composition of a given mixture by simultaneously measuring the resonant vibrational response of present molecular structures when excited by IR radiation. The principal advantage of the method is its ability to deliver information on a wide variety of molecular constituents within a sample in a label-free one-shot measurement, requiring no prior knowledge and minimal sample preparation. When analyzing single substances or mixtures of few organic molecules, it is generally possible to distinguish between spectroscopic signatures of different molecular functional groups, rely on well-defined spectral band patterns, and thus identify specific substances with high confidence.^[1,3,4] The former cannot be performed when analyzing spectra of more complex multi-molecular samples, such as biofluids, cellular components, or other biological media. The overlap in absorption peaks or interference between different molecular components hinders the capacity to link spectral signals of different functional groups to specific substances or even molecular classes (e.g., carbohydrates, lipids, or proteins). Thus, while IR spectroscopy is highly specific to identifying functional groups, performing peak interpretations can be challenging, and the method generally suffers from low molecular specificity when analyzing complex mixtures to profile biological systems.^[21–25]

Nevertheless, the desire to link spectral signals to specific substances in a clinical scenario is omnipresent, and misinterpretations of complex spectra remain widespread across peer-reviewed literature. In our experience, the question about the underlying molecular origin of spectral signals often arises in the scientific discourse. This is usually

[*] T. Eissa, L. Voronina, M. Huber, F. Fleischmann, M. Žigman
 Ludwig-Maximilians-Universität München (LMU), Chair of Experimental Physics - Laser Physics, Garching, Germany
 E-mail: mihaela.zigman@mpq.mpg.de

T. Eissa, M. Huber, F. Fleischmann, M. Žigman
 Max Planck Institute of Quantum Optics (MPQ), Laboratory for Attosecond Physics, Garching, Germany

T. Eissa
 Technical University of Munich (TUM), School of Computation, Information and Technology, Garching, Germany

© 2024 The Authors. Angewandte Chemie International Edition published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

not the right question or viewpoint, as its answers typically depend on oversimplified assumptions. This situation underscores a broader issue: The classical teaching of IR spectroscopy, which serves well for simple matrices, does not directly extend to the analysis of molecularly complex biological media. While many experts in the field are aware and share this same perspective, there is still a pressing need for a broader viewpoint shift in how spectral data is interpreted and presented, moving towards a framework that accommodates the complexity of biological systems.

In this Scientific Perspective, we critically examine the limitations inherent to analyzing complex spectra – a topic that, in our opinion, needs to be more directly and comprehensively addressed in scientific literature. We focus on IR spectroscopy of human blood derivatives – native serum and plasma. We highlight examples from previous

studies where molecular interpretations range from reaching reasonable and most likely correct conclusions to ones that are highly speculative and questionable. Using data from several case-control and health diagnostics scenarios we have considered previously, we demonstrate (i) how molecular interpretations can be misled by the overlap of spectrally similar yet biologically very different molecular substances; (ii) how measurement preprocessing can affect data interpretations; and (iii) how linking low-abundant substances, e.g., DNA/RNA, or protein biomarkers such as cancer antigen 125 (CA125) or prostate-specific antigen (PSA), to spectral signals is compromised by measurement sensitivity and background variability. Our analyses and conclusions are based on different case examples of molecules found in cell-free human blood – proteins, lipid particles, nucleic acids, and water-soluble metabolites (Sup-



Tarek Eissa began his studies in Computer Science at the American University in Cairo before moving to the Technical University of Munich, where he earned his master's in Data Engineering and Analytics. He is currently completing his PhD with the Broadband Infrared Diagnostics (BIRD) team, jointly between the Chairs of Laser Physics, Ludwig Maximilians University in Munich, and Bioinformatics, Technical University of Munich. Driven by his interest in data science, Tarek's research centers on analyzing blood-based infrared spectra for clinical diagnostics.



Liudmila Voronina studied at the Moscow Institute of Physics and Technology, before moving to the Swiss Federal Institute of Technology Lausanne, where she obtained her PhD in Physical Chemistry. In 2017, she joined the Broadband Infrared Diagnostics group at the Chair of Laser Physics, Ludwig Maximilians University in Munich as a Postdoc. Mobility fellow of the Swiss National Science Foundation. Liudmila has since become a senior scientist, expanding her research interests from spectroscopy of individual biomolecules to applications of liquid biopsies.



Marinus Huber studied Physics at the Friedrich Schiller University of Jena and the Ludwig Maximilians University in Munich, completing his master's thesis at Harvard University. For his PhD at Ludwig Maximilians University, he developed new laser-based infrared spectroscopy methods for medical diagnostics. He continued this work as a postdoc at the Max Planck Institute of Quantum Optics, Friedrich Schiller University, and the University of Kaiserslautern-Landau, focusing on innovative applications of infrared spectroscopy in analyzing liquids and cells.



spectroscopy to fingerprint complex human samples.

Frank Fleischmann studied Biology at the Technical University of Munich, and earned his PhD in Phytopathology from the Technical University of Munich. Following his postdoctoral research in forest pathology at the same university, he worked with the genetics service provider IMGM Laboratories in Martinsried. In 2018, Frank joined the Broadband Infrared Diagnostics (BIRD) team at the Chair of Laser Physics, Ludwig Maximilians University in Munich, where his research focuses on establishing and employing infrared molecular spectro-



molecular spectroscopy to analyze and map human health.

Mihaela Žigman studied Molecular Biology at the University of Ljubljana and earned her PhD from the University of Vienna. During postdoctoral research at the Institute of Molecular Biotechnology of the Austrian Academy of Sciences in Vienna, and at the Fred Hutchinson Cancer Research Center in Seattle, she investigated molecular mechanisms controlling cell division. Following her research at the Ruprecht Karls University in Heidelberg, she joined the Ludwig Maximilians University in Munich in 2015. Mihaela established the Broadband Infrared Diagnostics (BIRD) research, leveraging infrared molecular spectroscopy to analyze and map human health.

porting Information Table S1) – for which we measured IR spectra. We further utilize spectra from several thousand blood serum and plasma samples experimentally acquired in our laboratory over the last decade.

Ultimately, we urge users and professionals in the field of biomedical IR spectroscopy to accept the approach's fingerprinting nature, which inherently lacks sufficient molecular specificity. Importantly, this property does not prevent IR spectra from being specific to an individual or a physiological state. To fully capture the informational content of IR molecular fingerprints, appropriate machine learning approaches can be used.^[1,7,24,26–28] In contexts where molecular interpretation of experimental IR spectroscopic signatures is deemed necessary, orthogonal molecular-specific approaches should be involved, or the sample should be prepared in a way that reduces its molecular complexity.^[10,29–33] While we focus on IR spectroscopy of liquid blood derivatives, the principles we discuss generally apply to vibrational spectroscopy of diverse biological samples (e.g., cerebrospinal fluid or interstitial fluid).^[34] We aim to help and guide the establishment of vibrational

spectroscopy as a reliable biomedical tool while acknowledging its limitations. We believe that the proposed approach shall foster broader acceptance of the technological strengths inherent to the technique and support bridging the gap to bring the technology closer to clinical applications.

Misleading Assignments of Spectral Regions to Molecular Classes

The desire to systematically interpret the information encoded in complex spectra has led to categorizing the mid-IR into distinct regions associated with specific molecular classes. This resulted in the emergence of depictions within scientific literature that typically define discrete spectral regions responsible for protein, lipid, carbohydrate, and nucleic acid absorption.^[35–40] We provide one example of such a representation in Figure 1A. While the majority of absorption signals of these molecular classes are indeed present within the defined spectral regions, several impor-

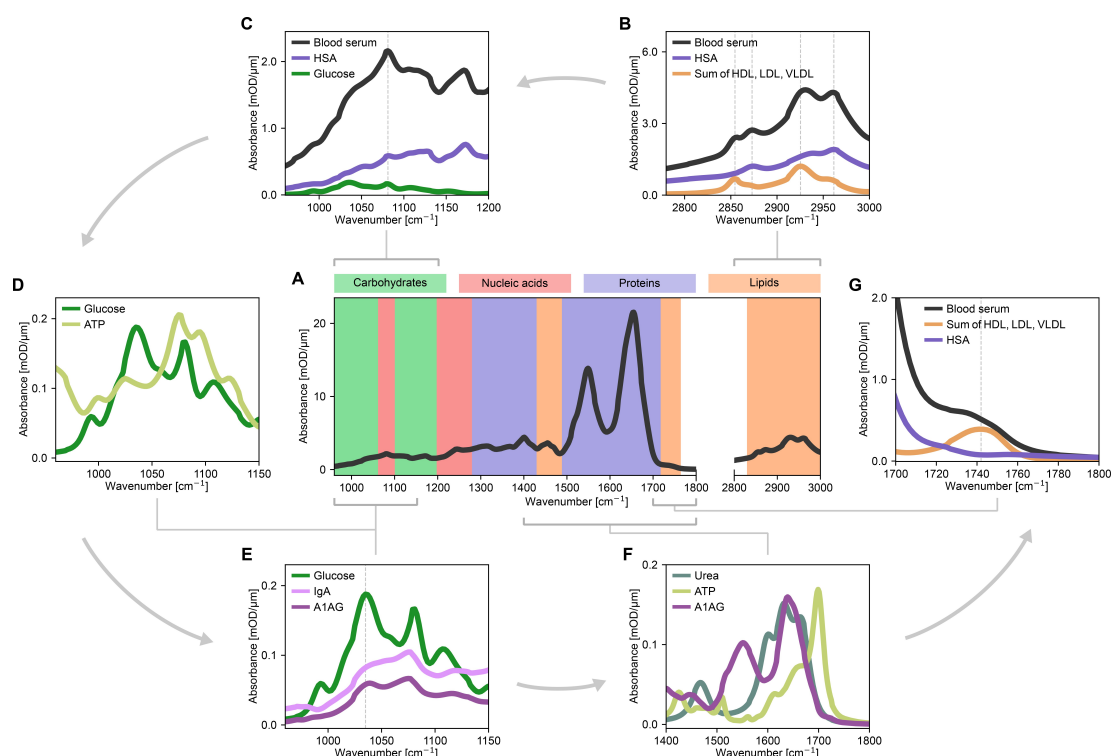


Figure 1. Misleading molecular assignments of spectral regions. (A) Descriptive labeling of the mid-IR commonly assumed in scientific literature to link different molecular classes to discrete spectral regions. A typical spectrum of human blood serum is shown in black. (B–G) Comparison between the absorption of various blood serum components recorded separately and scaled by their typical concentrations in healthy individuals (Supporting Information Table S1). Gray arrows indicate the counterclockwise sequence of the panels. HSA - Human serum albumin; HDL - High-density lipoprotein; LDL - Low-density lipoprotein; VLDL - Very-low-density lipoprotein; ATP - Adenosine triphosphate; IgA - Immunoglobulin A; A1AG - Alpha-1-acid glycoprotein.

tant issues render this approach oversimplified and can encourage a perspective that overlooks complexities in peak interpretations.

First, there is the straightforward yet important consideration of the relative concentrations of different molecular classes in cell-free human venous blood. The most prominent absorption bands of any single protein are the amide-I and -II bands, located between 1500–1700 cm^{-1} .^[2,41] Proteins, however, have other weaker absorption bands that cover all the “fingerprint” region that spans 960–1500 cm^{-1} , as well as the 2800–3000 cm^{-1} range.^[2] On the other hand, the IR absorption of glucose – the most abundant dissolved carbohydrate in blood plasma – is indeed strongest in the region 960–1200 cm^{-1} , and the absorption bands of lipids do appear in the regions labeled in Figure 1A. The total absorption of any molecular class, however, is a sum of the products of concentrations of each constituent and their absorption spectrum. The high concentrations of many proteins make their absorption the strongest in any of the regions mentioned above. To demonstrate this, we compare the absorption of the single most abundant protein in cell-free blood – human serum albumin (HSA) – to the absorption of lipid particles (Figure 1B) and glucose (Figure 1C). Evidently, the absorption of HSA alone is at least as significant in these regions as any of the other molecules in the so-called “lipid” and “carbohydrate” regions, or even more dominant.

The considerations mentioned above are not limited to the abundance of proteins. For instance, some water-soluble metabolites contribute to the absorption between 960–1150 cm^{-1} , making the “carbohydrate” label of this spectral region also questionable. As an example, we show that the IR absorption of glucose in its typical concentration is comparable to that of adenosine-triphosphate (ATP), a molecule that does not belong to the carbohydrate class (Figure 1D). As for the “nucleic acid” region highlighted in Figure 1A, it should be emphasized that the IR absorption of DNA and RNA – the primary nucleic acids – cannot be attributed to any vibrational response of cell-free blood. These substances are present in concentrations far below the detection limits of the current technology – a point that we further address and discuss later in the text.

Importantly, dividing the spectral regions into the four categories (proteins, carbohydrates, lipids, and nucleotides) does not reflect the actual composition of cell-free blood. Lipids, for instance, are not water-soluble and are present in blood plasma as part of lipoprotein particles – mostly HDL, LDL, and VLDL – in which triacylglycerols, cholesterol, and phospholipids are found in various proportions.^[42,43] Furthermore, lipid particles contain a high percentage of proteins, up to 33 % in the case of HDL.^[43] As a result, the experimental IR absorption spectra of HDL, LDL, and VLDL contain a clear protein spectral signature, in addition to the absorption bands in the lipid region (Supplementary Figure S1). Thus, while strong lipid absorption signals can appear within the highlighted regions shown in Figure 1A, their absorbance should not be confined to such a restricted range, as an ensemble of diverse molecules compose each lipid particle.

Carbohydrates are primarily present in cell-free blood not as separate molecules but as glycans – bound to proteins in the process of their post-translational modification.^[44] Due to the chemical similarity of free carbohydrates (e.g., glucose) and glycans, attributing a spectral change to an alteration in the former or the latter is challenging without additional analysis. To illustrate this, we compare the absorption of two glycosylated proteins in their typical concentrations to that of glucose (Figure 1E). It is evident that the most prominent absorption peak of glucose overlaps with the spectral band present in the glycoprotein vibrational spectra.

The examples discussed so far may suggest that one can confidently attribute spectral contributions of proteins, given their high concentrations. We challenge this assumption in Figure 1F, which shows the amide-I spectral region that is typically used to describe changes in the protein secondary structure.^[2,22,41,45–49] We depict the IR absorption of alpha-1-acid glycoprotein (A1AG), one of the twelve most abundant proteins in blood plasma, compared to the absorption of urea and ATP. It is evident that both metabolites exhibit absorption peaks that could be mistaken for a change in the protein's secondary structure. This purely spectroscopic consideration leads to more general biomedical concerns about using protein secondary structure change as a biomarker, as discussed previously.^[21,22]

It is also important to acknowledge that there exists a spectral region that can be quite confidently correlated with a specific molecular class – the peak centered around 1742 cm^{-1} (Figure 1G). The total absorbance in this region is predominantly due to the presence of lipid particles. Therefore, it would be reasonable to infer a lipid-centric contribution when a strong signal is observed there, while considering the overall molecular composition of lipid particles as discussed above.

Given these considerations, we advocate for caution when using descriptive labels for discrete regions of the IR spectrum, even when only intended as a foundational guide. Such labels may lead researchers to misinterpret the spectral signals they observe, which often requires a more critical perspective and experimentation with additional molecular techniques.

Spectral Similarity of Different Substances

Having demonstrated the limitations of attributing spectral regions to molecular classes, we now take a more quantitative approach to illustrate the potential for overlap among the spectra of individual organic substances (Figure 2). We calculated the hit quality index (HQI)^[50] between pairs of various substances to assess the IR absorption similarity of different components found in cell-free blood. These include the spectra of lipid particles, proteins, and water-soluble metabolites. Given two spectra, represented as one-dimensional vectors $\mathbf{u}$ and $\mathbf{v}$ of absorption values, the HQI measures the squared cosine of the angle between the two vectors [Eq. 1].

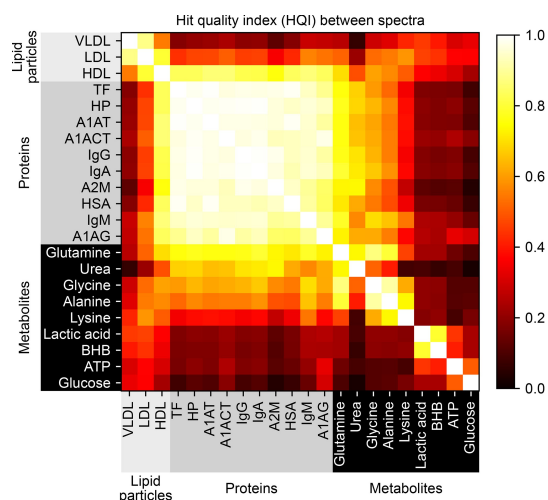


Figure 2. Similarity among IR spectra of organic substances present in blood. Color intensity represents the level of similarity between pairs of IR absorption spectra (960–3000 cm⁻¹) of single substances. Brighter colors indicate higher spectral overlap. VLDL - Very low-density lipoprotein; LDL - Low-density lipoprotein; HDL - High-density lipoprotein; TF - Transferrin; HP - Haptoglobin; A1AT - Alpha-1-antitrypsin; A1ACT - Alpha-1-antichymotrypsin; IgG - Immunoglobulin G; IgA - Immunoglobulin A; A2M - Alpha-2-macroglobulin; HSA - Human serum albumin; IgM - Immunoglobulin M; A1AG - Alpha-1-acid glycoprotein; BHB - Beta-hydroxybutyric acid; ATP - Adenosine triphosphate.

$$HQI = \frac{(\mathbf{u} \cdot \mathbf{v})^2}{(\mathbf{u} \cdot \mathbf{u})(\mathbf{v} \cdot \mathbf{v})} \quad (1)$$

HQI values closer to one indicate higher collinearity and, therefore, more similar spectral profiles.

The spectra of all investigated individual proteins were found to be very similar to one another (Figure 2). This stands to reason, as proteins have a large molecular structure made of amino acids that share similar functional groups (namely, the amide bonds). Yet, this analysis clearly demonstrates that linking any spectral signal to a single protein “biomarker” is likely to be flawed, as variations in the concentration of nearly any protein could produce comparable spectral signals.

Lipid particles are very complex, comprising a mix of cholesterol, triglycerides, phospholipids, and proteins. This leads to a broad range of vibrational modes. Notably, the spectra of LDL and VLDL showed considerable similarity to one another (Figure 2, see also Figure S1). However, HDL, being enriched with apolipoprotein A1, reflected a spectral profile more closely aligned with that of single proteins than either LDL or VLDL did.

Comparatively, metabolites are smaller and more diverse in terms of their chemical structures and functional groups. This diversity leads to more unique vibrational modes for each metabolite and narrow peaks, resulting in more distinct IR spectra (Figure 2). Nevertheless, some metabolite spectra

share similarities to protein spectra. For example, urea – as we previously mentioned, and was highlighted by Ollesch et al.^[51] – and, naturally, amino acids, such as glutamine, share these similarities. Furthermore, some metabolite pairs, such as glycine and alanine or lactic acid and beta-hydroxybutyric acid (BHB), can share similar spectral profiles with one another. Therefore, caution should be taken even when linking spectral signals to smaller molecules. Out of all the substances investigated, glucose was found to generally have the most unique spectral profile, which likely explains the high efficacy of quantifying its concentrations from IR spectra of blood plasma, serum, and whole blood.^[14,52–55]

Ambiguity of Signals in Complex Mixtures

While the spectra of individual organic substances can be very similar, the problem of overlapping signals becomes much more prominent when considering the plethora of substances in biological samples. Take a health phenotyping scenario as an example. The typical study reveals how some spectral features significantly differ between two (or more) phenotypes of interest. In many instances, a discussion ensues that relies on peak positions to decipher the bonds and chemical nature of the underlying molecules responsible for such spectral differences.^[45,47,48,56–64] Here, we illustrate how the signal of one substance can be almost perfectly replicated by a combination of only a few other substances (Figure 3) – deeming such speculative assignments unproven.

We modeled the spectra of different proteins and lipid particles as a linear combination of other substances [Eq. 2].

$$\mathbf{y} \approx \sum_{i=1}^m \mathbf{x}_i \cdot c_i \quad (2)$$

$\mathbf{y}$ represents the target spectrum to be reproduced at a given concentration, $\mathbf{x}_i$ are spectra of other substances, and c_i represents their respective concentrations. The concentrations c_i were determined by a computational optimization that minimizes the reconstruction error. To keep it simple, we limited the number of substances used to reconstruct the target to four ($m = 4$).

We found that the IR spectra of all the proteins and lipid particles we considered could indeed be reproduced to a high degree, while only using relatively few substances (Figure 3, Supplementary Figure S2). Interestingly, the metabolite spectra were heavily utilized in reconstructing the target to capture the smaller peaks apparent within the larger molecular structures. This analysis demonstrates, albeit with an extreme example, that observing an IR spectral signal that resembles an individual substance may indeed be caused by a mix of other possibly fully independent molecules that may not even contribute to the same signaling pathways nor be mechanistically interconnected.

To provide a parallel demonstration of a more realistic application, we reconsidered one of our previous works.^[29]

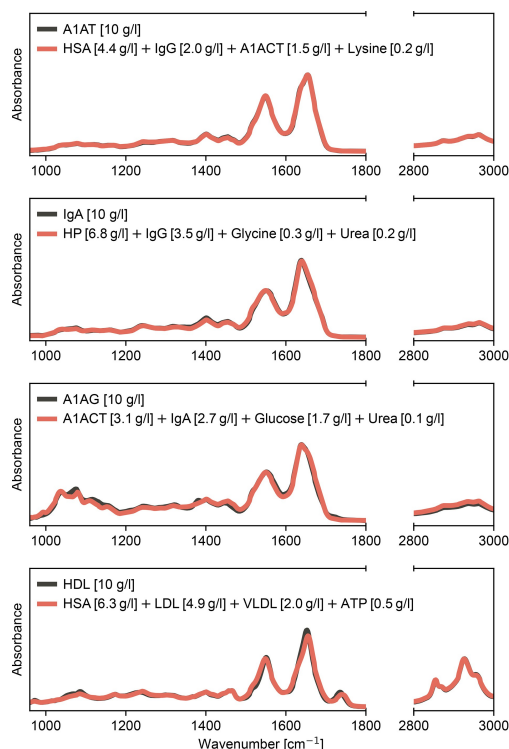


Figure 3. Ambiguity of individual substances in complex mixtures. Spectra of individual substances (black) of larger molecular structures were reconstructed by combining the spectra of other substances (red). The linear combinations were determined by scaling each component with the listed concentrations to yield a reproduced spectrum of the target at 10 g/l. Supplementary Figure S2 depicts the same analysis on an expanded set of proteins and lipid particles. A1AT - Alpha-1-antitrypsin; HSA - Human serum albumin; IgG - Immunoglobulin G; A1ACT - Alpha-1-antichymotrypsin; IgA - Immunoglobulin A; HP - Haptoglobin; A1AG - Alpha-1-acid glycoprotein; HDL - High-density lipoprotein; LDL - Low-density lipoprotein; ATP - Adenosine triphosphate; VLDL - Very-low-density lipoprotein.

This study investigated a case-control biomedical diagnostic application to detect spectral differences due to lung cancer with FTIR spectroscopy of blood serum. Mass spectrometry-based proteomics profiling of the same sample set helped identify the differences in protein concentrations between cases and controls (Figure 4A). In the prior work, we found that the differences between the spectra of case and control individuals observed in bulk serum spectra could be largely reproduced from a reduced molecular model that relied on three proteins (Figure 4B, upper panel). These proteins (HSA, HP, and A1AG) were identified as the ones that changed the most in absolute concentration between the case and control individuals, as measured by proteomics.

Now, we demonstrate that this very previously described lung cancer IR signal can be derived when other combinations of three individual proteins are considered (Figure 4B,

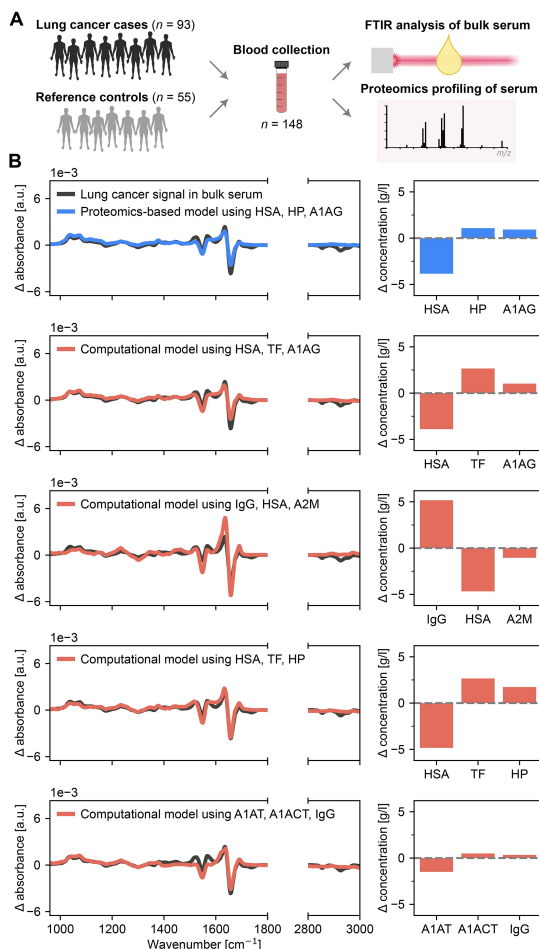


Figure 4. Replicating lung cancer-induced spectral deviations with different proteins. (A) Blood serum was collected from case and reference control individuals to investigate spectral changes due to presence of lung cancer. Two measurements were performed on the same sample sets: Bulk serum IR absorption spectroscopy and proteomics-based profiling, as described in detail previously.^[29] (B) Black curves depict the difference in the mean IR measurements between case and control individuals. Blue curve (upper panel) depicts a reproduced signal based on the three proteins that changed the most between case and control individuals, as determined through proteomics. Remaining curves (red) reproduce the same signal with other combinations of three proteins that were determined computationally (i.e., without direct evidence linking this protein combination to the true signal). Bar graphs depict the concentration changes in the modeled proteins that yielded the blue and red curves. HSA - Human serum albumin; HP - Haptoglobin; A1AG - Alpha-1-acid glycoprotein; TF - Transferrin; IgG - Immunoglobulin G; A2M - Alpha-2-macroglobulin; A1AT - Alpha-1-antitrypsin; A1ACT - Alpha-1-antichymotrypsin; IgG - Immunoglobulin G.

remaining panels). Interestingly, in one of the four examples we present, the lung cancer signal could be largely replicated

with changes in concentrations of A1AT, A1ACT, and IgG – three proteins that were not even relevant in the proteomics-based model. This further showcases that without supplementing the IR approach with yet further analytical methods capable of molecular identification (e.g., various -omics approaches), conclusions on which molecules contribute to discriminative signals can very likely lead to inaccurate conclusions.

Signals of Low Abundant Molecules

In some cases apparent in scientific literature, molecules of a very low concentration are linked to specific wavenumbers and suggested to be responsible for observed signals in the analyzed specimens.^[45,46,61,62,64–70] Typical examples are circulating cell-free RNA/DNA and known biomarkers such as CA125, mannose, or PSA. Using cell-free DNA and RNA as case examples, here we demonstrate that such substances are so far below the detection limit that the measurement variability certainly conceals their expected contribution to the possibly detectable IR spectral signals (Figure 5).

The median concentrations of circulating DNA and RNA have been reported to range between 6.8–10 ng/ml in cell-free blood.^[71–73] Highest individual concentrations can

reach near 100 ng/ml. At their median concentrations, these substances are around three orders of magnitude below the detection limit of a typical IR spectrometer (Figure 5).^[19,74] Even if their concentrations were 1000 times higher than the median, they would remain well below the threshold necessary to reliably link their contributions to spectral signals, especially given the between-person biological variability of cell-free blood.

Accordingly, when analyzing cell-free blood, any observed changes in IR absorption cannot be attributed to variations in circulating DNA/RNA concentrations. It is important to note that DNA/RNA in tissues such as the liver, spleen, or heart can reach rather high concentrations that may shape IR spectra. Therefore, the conclusions drawn from this analysis should not be generalized to other biological specimens.

Chemical Modifications of Molecules

Some attempts to interpret changes between two groups of vibrational signatures rely on the spectra of cell-free blood components recorded in isolation from their environment or on tables of characteristic vibrational modes.^[56,57,59,60,67,75] This approach overlooks the interactions between various blood components, such as extracellular molecular transport by effective diffusion or assembly into higher-order complexes that modulate and regulate their dispersal (e.g., via proteins or encapsulation of molecules in extracellular vesicles).^[42,76] This very study also suffers from this limitation to some degree: The formation of non-covalent macro-molecular complexes may lead to shifts in vibrational frequencies that are overlooked when merely aggregating the contributions of individual components. However, a much larger change in the IR spectroscopic signatures of molecules, especially small ones, can be caused by the formation of additional covalent bonds. Here, we consider two examples of the many modifications that molecules undergo in their life cycle (Figure 6).

Protein glycosylation is a complex post-translational modification that is sensitive to changes in the health status of individuals.^[77] For instance, altered sialylation has been reported in multiple cancers and liver diseases.^[77] The IR absorption spectrum of N-Acetylneuraminic acid – the most common sialic acid – is depicted in Figure 6A. The sharp peak at 1730 cm⁻¹ has been previously assigned to the –COOH vibration of sialic acid and used to monitor the degree of protein sialylation.^[78–80] We then compared the IR absorption spectrum of bovine fetuin with its desialylated form, asialofetuin. The difference between their spectra clearly lacks the sharp peak at 1730 cm⁻¹ that is apparent in free sialic acid (Figure 6A). Instead, the peak at 1035 cm⁻¹ may be used to characterize the presence of sialic acid. In a complex mixture, however, this peak would at least partially be masked by the variability of other glycans and thus not detected.

The region of the IR spectrum that is often referred to as “carbohydrate” is also sometimes labeled as “fingerprint” – because it is sensitive to the overall changes in the

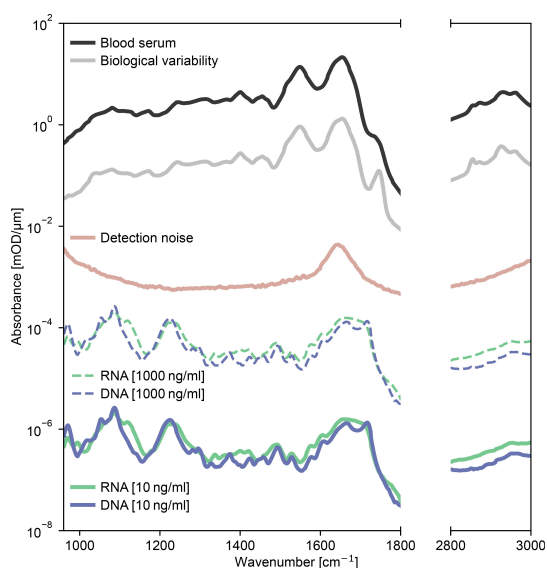


Figure 5. Contributions of cell-free nucleic acids relative to IR detection limits. Black curve depicts the IR absorbance of a typical human serum sample. Gray curve depicts the standard deviation of serum spectra across different individuals. Brown curve depicts the detection noise of the IR measurement, determined by calculating the standard deviation of repeated measurements of a technical replicate (pure water). Dashed curves depict the IR absorbance spectra of RNA (green) and DNA (blue) at 100x their median concentrations in cell-free blood, while solid lines represent their absorbance at their typical median concentrations. Y-axis is plotted on a logarithmic scale.

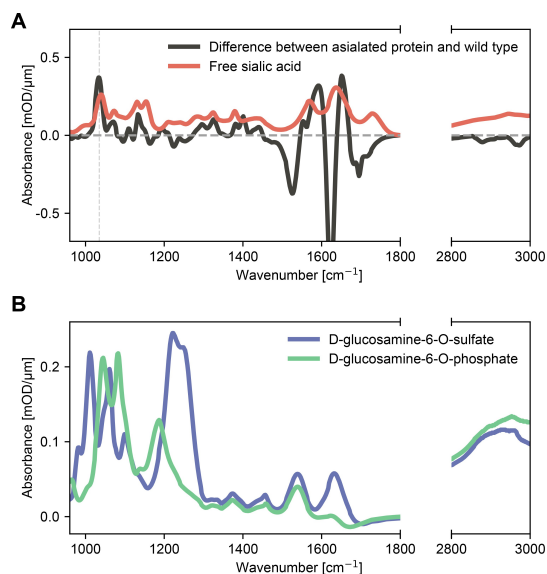


Figure 6. Impact of chemical modifications on IR spectra. (A) Absorption spectrum of free sialic acid (N-acetylneuraminic acid, red) compared to the absorption differences obtained by subtracting the IR spectrum of asialofetuin (the desialylated form of fetuin) from that of fetuin (black). (B) Absorption spectra of D-glucosamine-6-O-sulphate (blue) and D-glucosamine-6-O-phosphate (green).

molecular structure. Although the absorption pattern there can be used to confirm the presence of a specific molecule,^[1] the change in this pattern due to a chemical modification of the molecule is challenging to predict without quantum chemical calculations. To illustrate this, we compared the IR absorption spectra of substituted glycans: D-glucosamine-6O-sulphate and D-glucosamine-6O-phosphate (Figure 6B). The difference in their functional groups causes a shift in the characteristic asymmetric stretching band from 1186 cm⁻¹ to 1222 cm⁻¹, as expected. However, in addition to that, there was a significant shift of all the bands below 1200 cm⁻¹. It is crucial to note that all these changes were caused by one chemical modification, with no changes in the carbohydrate content.

The presented examples serve as a reminder that vibrational spectra of molecular structures, especially small ones, are highly sensitive to their surroundings and chemical modifications. This is especially true for the “fingerprint” region of the IR spectrum, but not limited to it. It is also important to acknowledge that post-translational modifications, such as sulphation, phosphorylation, methylation, glycosylation, etc., contribute not only their specific IR signatures directly, but might also affect other spectral regions by changing the 3D structure of the whole macromolecule.

Effects of Measurement Preprocessing

To further compound the issue, molecular interpretation in spectral datasets is often performed without considering the preprocessing steps performed prior to data analysis. Common preprocessing steps include vector/peak normalization, baseline correction, offset correction, and smoothing – as comprehensively described in prior work.^[28,81,82] While such data manipulations can help reduce noise and make datasets more comparable, some preprocessing steps can lead to artifacts in the analysis, including correlations in the noise structures of the data.^[28] These artifacts can result in false peaks that range from being very apparent in the data to being subtle and hard to detect. Here, we considered vector normalization (L2) as an example of a widely applied method to improve measurement reproducibility^[13,28,81,83] and demonstrate how it can mislead peak interpretations.

Consider a case-control scenario in which only the concentration of one molecule changes between two groups of samples. To model such a situation computationally, we spiked (added) the spectrum of HSA to a typical spectrum of blood serum. We then calculated the difference between the unspiked and spiked spectra (Figure 7A). We performed this analysis both with and without the normalization step,

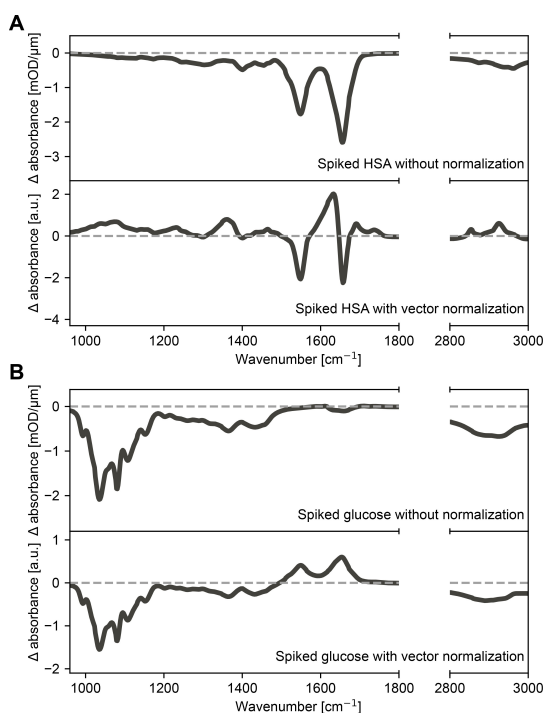


Figure 7. Effect of vector normalization on spectral interpretations. (A) Human serum albumin (HSA) was spiked to a typical blood serum spectrum. The curves depict the difference between the unspiked and spiked spectra, with and without vector normalization. (B) Same investigation, but spiking the spectrum of glucose instead of HSA.

applied prior to subtraction. Very evidently, the resulting signal becomes hard to interpret when normalization was used, and associating this signal to the change in the HSA concentration would be nearly impossible. Although the negative absorbance near the amide-I and -II regions was somewhat preserved in the normalized depiction, artifacts of positive “absorbance” were introduced across the remaining regions of the spectrum.

We then repeated the above analysis, but computationally spiking in the glucose spectrum instead (Figure 7B). The shape of the signal in the region between 960–1200 cm^{-1} was mostly preserved between the normalized and non-normalized depictions. However, artifacts of positive absorbance now appeared near the amide-I and -II regions – despite the absence of any significant absorbance in the spectrum of glucose in this region. Such an artifact could easily mislead one to think that there was a difference in protein content between the two groups of spectra. In reality, only the glucose concentration was altered.

Generally, applying measurement preprocessing steps may be necessary, particularly also to Raman spectroscopy.^[84] Therefore, we do not suggest to eliminate these steps. Rather, in conjunction with Morais et al.,^[28] we advocate for their cautious use, considering their potential impact on the data analysis. In some instances, this impact could make peak interpretations simply unfeasible.

Univariate and Multivariate Spectral Analysis

The task of analyzing complex vibrational spectra necessitates robust statistical methods that can effectively uncover subtle yet critical spectral information often masked by overlapping signals. For instance, the power of multivariate statistical methods has long been recognized in spectral data analysis, including in foundational work of FTIR spectroscopy.^[85] Several studies offer constructive tutorials and guidelines on selecting appropriate data analysis methods for spectral data.^[1,7,24,26–28,86] Here, we do not go through an extensive description of the technicalities of data analysis methods. Instead, we focus on their capacity to capture information on molecular substances from spectral datasets. First, we demonstrate how simple univariate analysis can be insightful, and then we point out its limitations, which necessitate the application of more advanced multivariate methods (Figure 8).

As an example, we revisited one of our previous studies.^[14] There, blood plasma from a large population-based cohort was profiled with FTIR spectroscopy (Figure 8A). Laboratory clinical chemistry analyses were performed on the same blood samples to quantify the concentrations of triglycerides (TG), LDL cholesterol (LDL-C), and HDL cholesterol (HDL-C) (Figure 8A,B). This allowed us to correlate the information content of the plasma IR spectra with the clinically determined concentrations of these lipid-centric analytes.

First, we calculated the Pearson correlation coefficient (r) between the concentrations of the three analytes and the IR absorbance observed in bulk plasma at each univariate

spectral feature (Figure 8C). As a reference, we compared this correlation spectrum with that of the individual substance spectrum relevant to each analyte. Given that TG constitutes 50 % of VLDL dry mass,^[43] we compared the correlation of TG with that of the VLDL spectrum.

For TG and LDL-C, we found a high level of agreement between the univariate correlation spectrum and that of the individual lipid particles (Figure 8C). In particular, the narrow region between 1745–1750 cm^{-1} revealed a strong correlation between bulk plasma absorbance and the concentrations of the two substances. Specifically for TG, the peak at 1748 cm^{-1} aligned with that of the VLDL reference spectrum and at a very high correlation ($r = 0.92$). This suggested that spectral signals relating to TG at this peak were largely not masked by the presence of any other substances. Surprisingly, however, the region of 2800–3000 cm^{-1} revealed a lower correlation level compared to the aforementioned region of 1745–1750 cm^{-1} for both TG and LDL-C. Given that the spectra of the pure substances have peaks with the strongest intensity between 2800–3000 cm^{-1} , this suggested that the contributions of VLDL and LDL were overshadowed by other substances that also contribute spectral signals within that region, e.g., proteins, in line with Figure 1B,G.

In contrast, HDL-C revealed relatively weak correlations across the whole spectral range (Figure 8C). At the amide bands, for instance, no significant correlation strength was observed. This suggested that the plethora of other proteins in cell-free blood masked the signals resulting from HDL. Similarly, in the regions between 1745–1750 cm^{-1} and 2800–3000 cm^{-1} , no strong correlation was observed. This suggested that the signal of HDL-C was, again, overshadowed by the presence of other substances – namely, TG and LDL-C – which together have a much higher concentration than HDL-C (Figure 8B). Altogether, this analysis demonstrated that the HDL-C concentration had no single distinctive spectral feature that could be reliably linked to its concentration, unlike TG and LDL-C.

Compared to the prior analyses that relied on a relatively simple univariate approach, multivariate analysis is capable of considering the interactions among features. Such an aspect provides a more granular view of the information contained within the complex spectra as subtle dependencies between the absorption peaks can be captured – ones that would be otherwise obscured in a univariate context. To put this into practice, we trained a ridge regression model to quantify the clinically determined concentrations of the three lipid analytes given an input plasma spectrum (Figure 8D). When all the features in the spectral range were considered, a very high quantification efficacy was observed for all three analytes, with R^2 values ranging from 0.90 to 0.97. Notably, even HDL-C was well quantified despite its lack of single distinctive features. This analysis highlights the power of multivariate algorithms to extract comprehensive insight from the information-rich IR spectra.

It is also crucial to recognize that the information across several spectral features is highly co-dependent. As previously discussed, individual substances can have many

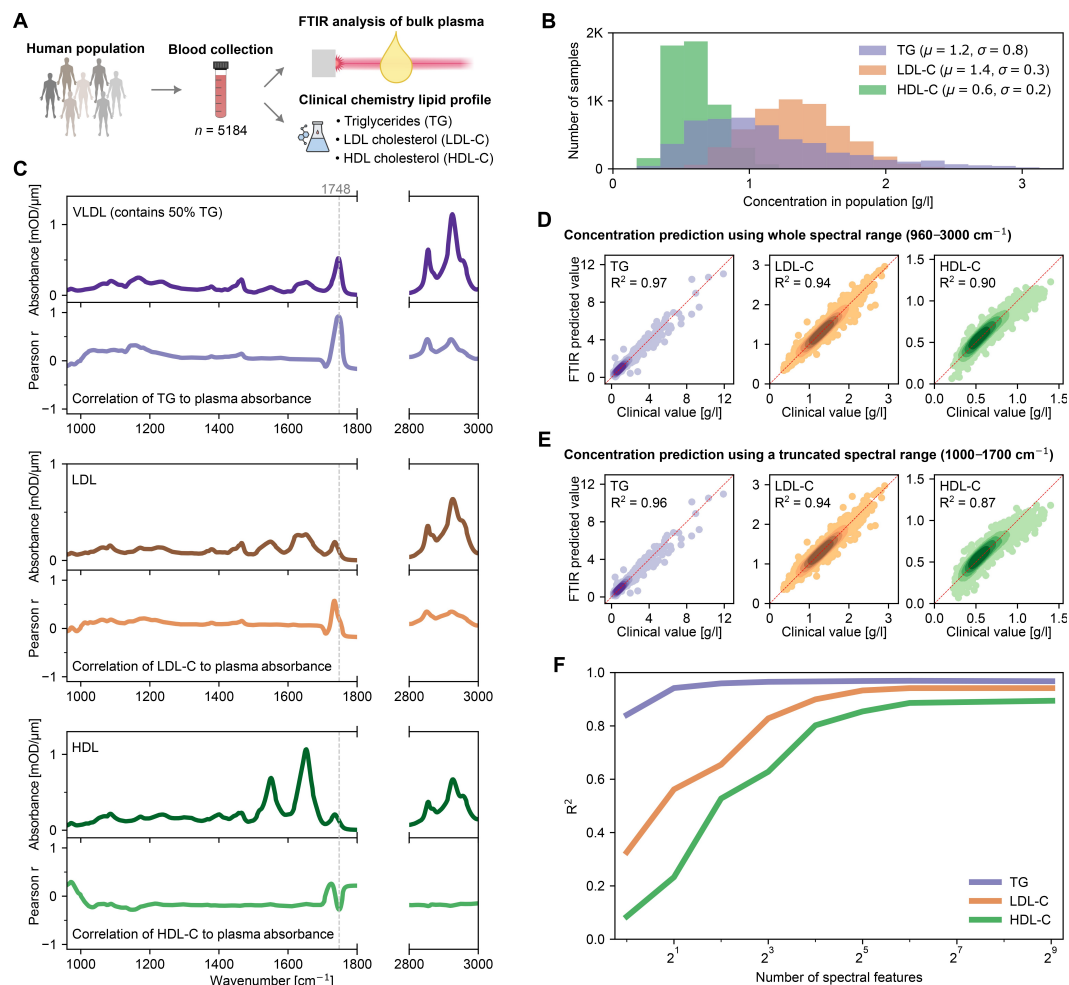


Figure 8. Unraveling the informational content of complex spectra with univariate and multivariate statistical analysis. (A) Blood samples were collected from a large, population-based cohort for FTIR spectroscopy and clinical laboratory analysis to determine lipid analyte concentrations. (B) Distributions of lipid analyte concentrations in the population. (C) Absorbance spectra of single lipid particles (darker colors, upper panels) and the correlation between plasma absorption to the concentrations of the three lipid analytes (brighter colors, lower panels). (D) Multivariate regression models were trained to quantify the concentrations of each lipid analyte given an input spectrum of plasma. Each point represents the model's prediction for a plasma sample against the clinical measurement. A 10-fold cross-validation was applied. Only the predictions of the test sets of the cross-validation splits are depicted. (E) Same multivariate regression analysis, but using a truncated spectral range as input to the model here. (F) Dependency between the number of spectral features given to the regression model and the quantification efficiency of each lipid analyte. A forward feature selection algorithm was applied to determine the optimal spectral features at each point. All correlation and regression analyses were applied on non-normalized spectra.

peaks across a wide range of frequencies. To illustrate the magnitude of this point, we performed the above regression analysis again, but paradoxically, using only the spectral range between 1000–1700 cm^{-1} – i.e., the range that had minimal univariate correlation to the lipid analytes. Surprisingly, the quantification efficacy for TG, LDL-C, and HDL-C remained nearly unchanged (Figure 8E). This further demonstrates the magnitude of the multi-collinearity

present within the spectral features and highlights that linking biological substances to specific spectral bands is highly questionable – especially when multivariate analysis methods are employed.

As demonstrated above, considering a large number of spectral features provides a more accurate overview of the breadth of molecular information in the IR spectra. To further examine this, we quantitatively determined the

number of spectral features that were required to effectively quantify the concentrations of the lipid analytes. We examined the quantification efficacy for each studied analyte, considering a varying amount of spectral features given to the regression model (Figure 8F). These features were iteratively selected by a forward selection algorithm to determine the most optimal input features at each point. For TG, utilizing only one distinctive feature (1748 cm^{-1}) had already resulted in a high quantification efficacy. Including just one more feature resulted in a nearly saturated quantification efficacy, revealing that the signal of TG can be adequately isolated using only two spectral features. Comparatively, for LDL-C and HDL-C, utilizing one or two spectral features proved to be insufficient. Specifically for HDL-C, nearly 64 different spectral features were required to reach the full potential of correlating its signal to the plasma spectra. Given that most substances in blood likely do not have such a distinctive feature as TG does, utilizing a low number of features or, e.g., considering spectral ratios, is therefore unlikely to provide a comprehensive view of the spectral informational content.

To summarize, when relating an IR absorption spectrum of a bulk mixture to certain components of interest:

1. univariate analysis (or use of a limited selection of peaks) is only appropriate when the signal of the substance is very strong and not masked by other compounds;
2. multivariate analysis may be capable of capturing information about a substance that is heavily overshadowed by the contributions of other compounds;
3. spectral features are highly co-dependent, which allows for analyzing the signature of substances in spectral regions where they do not have prominent peaks; and
4. to harness the power of multivariate analysis, a substantial amount of spectral features are needed to discern complex feature inter-dependencies (beyond spectral ratios).

Navigating the Pitfalls of Spectral Interpretations

The question of how to most effectively analyze vibrational spectra of complex biological samples has no straightforward answer. Each analytical task is unique, targets different biological phenomena, is specific to different medical scenarios, and faces different requirements and challenges. Generally, we as a community would like to have a realistic overview of the molecules and their combinations we are profiling when analyzing complex biological matrices. Equally important is raising awareness about what we cannot possibly observe as it is beyond the technical limitations of the current technology and analytical setting. Furthermore, we would like to develop robust analytical workflows that accurately model alterations in a studied system (e.g., characteristics of the human phenotype). Below, we discuss how to approach these tasks.

First, to produce a more realistic view of the composition of the IR spectrum from cell-free blood, we decomposed it into the contributions of its primary molecular classes: metabolites, proteins, and lipid particles (Figure 9A). It is

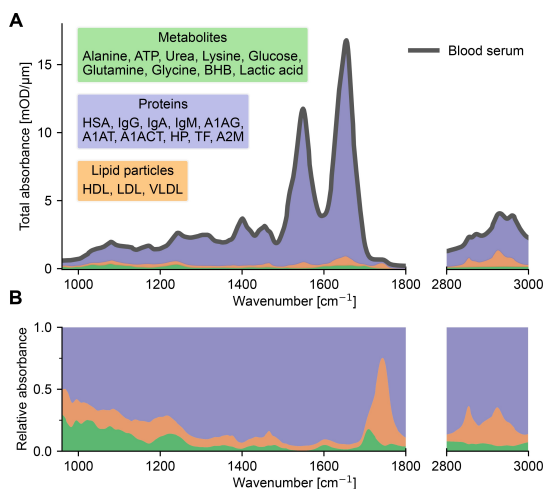


Figure 9. Composite makeup of a typical infrared spectrum of venous blood serum. (A) Stacked representation of the primary molecular classes found in cell-free blood. Individual component spectra of each substance were scaled according to their typical concentrations in healthy adults (Supporting Information Table S1) and combined to estimate the cumulative spectral contributions. Molecular classes were stacked in sequence, with metabolites at the base, followed by lipid particles, and proteins on top. Black curve represents the contribution of their total sum, approximating the spectrum of blood serum. Supplementary Figure S3 compares the approximated spectrum with a true serum spectrum. (B) Relative absorption contribution of each molecular class, illustrating the proportions of total absorption across the spectral range. ATP - Adenosine triphosphate; BHB - Beta-hydroxybutyric acid; HSA - Human serum albumin; IgG - Immunoglobulin G; IgA - Immunoglobulin A; IgM - Immunoglobulin M; A1AG - Alpha-1-acid glycoprotein; A1AT - Alpha-1-antitrypsin; A1ACT - Alpha-1-antichymotrypsin; HP - Haptoglobin; TF - Transferrin; A2M - Alpha-2-macroglobulin.

evident that proteins dominate the absolute absorption signals across the spectral range, considering that lipid particles also contain a proportion of proteins. When examining the relative contributions of each molecular class, their influence on absorption signals significantly varies across the spectral range (Figure 9B). These representations should not imply that one molecular class provides more information over another. For example, glucose, a metabolite, can be reliably determined in blood plasma, serum, and whole blood using IR spectroscopy as previously mentioned.^[14,52–55] Rather, these representations further underscore the notion that the IR spectrum should not be overly simplified into discrete regions exclusively linked to specific molecular classes. Each substance and each molecular class contributes to the composite nature of the overall IR spectrum, which offers a holistic view of the molecular landscape within the sample.

When confronted with changes in the IR absorption spectrum of a biological sample, researchers may be tempted to seek simple, intuitive explanations. However, as demonstrated in our examples, such approaches should be avoided unless robustly validated. Machine learning and multivariate

statistical methods are instrumental in scenarios where the molecular components are numerous and the relationships between them are complex. Nevertheless, it is crucial that these statistical models are trained on reproducible, high-quality, well-characterized datasets and that the results are reasonable within the biological context of the study. This typically requires large datasets that accurately reflect demographic distributions, include appropriate control reference samples, and involve diverse pathophysiological strata (e.g., comorbidities or stages of a disease). Furthermore, biological systems are inherently dynamic, and thus, statistical models should be regularly updated with new data. Adapting and learning from new inputs will help improve the statistical models' accuracy and relevance. This is especially important in clinical, medical settings, where training data must mirror realistic data variations caused by inherent biological variabilities (e.g., lifestyle changes and aging), sample collection variations, differences in sample storage regimes, measurement device maintenance, and other technical variations typically encountered in practice.^[87] Furthermore, independent external validation of developed models on samples from different clinical studies is essential to more robustly determine whether results are reproducible.

If molecular interpretations of changes in the IR absorption spectrum are necessary, we suggest involving complementary analytical techniques. Techniques such as mass spectrometry, nuclear magnetic resonance (NMR)-based omics, and clinical chemistry provide orthogonal information that aids in this process.^[14,29,88,89] Another fruitful avenue is to reduce the molecular complexity via chemical fractionation – e.g., liquid chromatography,^[30] protein precipitation^[29] or ultrafiltration.^[10,31–33] IR measurements can then be performed on the individual fractions, which are less complex in terms of their molecular makeup and might be easier to interpret. These strategies offer independent confirmation of molecular identities and concentrations, enabling further verification of the findings. Furthermore, the results should be compared to other biomarker research efforts, and the molecular-specific mechanistic and pathway information that the medical community sometimes requires should be provided. Once the molecular mechanisms are provided and validated, the IR approach can be used in its high-throughput mode – without the involvement of orthogonal measurement approaches.

Generally, the conceptional suggestions presented here can be broadly applied to IR spectroscopic examinations of other complex biological systems beyond blood plasma. Extensions to model paradigms such as yeast, bacterial, algal, and mammalian cells, as well as animal, human, and plant tissues can be envisioned. However, it is important to consider that the IR spectral signatures of different model paradigms can significantly vary due to their distinct molecular makeup (e.g., the content of proteins vs. carbohydrates is different between animal and plant models^[90]) and there may be instances where the assignment of spectral regions to specific molecular entities is feasible. For instance, certain wavenumbers in cellular spectra can be qualitatively linked to nucleic acid content,^[91] or sialic acid

content may be identified at specific wavenumbers in spectra of extracellular polymeric substances.^[79] Nevertheless, spectral molecular assignments must be performed cautiously and, ideally, supplemented with additional experimentation, as previously discussed.

Transparency and the critical evaluation of the reported results are also essential aspects. Documentation of all analytical steps and assumptions for interpretations should be clearly communicated in peer-reviewed works. For instance, it is crucial to detail the data preprocessing steps and their impact on the observed spectral signals. The spectroscopic technique's dynamic range and sensitivity limits must always be considered and recognized, especially when dealing with low-abundant substances. Another rarely discussed aspect revolves around the sample collection methodology (e.g., if a variety of sample collection tubes were used within the same study). For instance, blood plasma collection tubes, as opposed to serum, use ethylenediaminetetraacetic acid (EDTA) or other anticoagulants within the blood draw tubes. EDTA typically has a high enough concentration that it contributes to absorption signals of bulk plasma spectra to a significant extent (Supplementary Figure S4). Changes in the blood draw and sample processing workflows or the use of samples from different clinical studies and clinical centers could therefore lead to noticeable variations in the signals when studying certain biological phenomena, potentially confounding the results.

Finally, the value of multidisciplinary research must be recognized. A purely chemical, spectroscopic perspective is no longer sufficient on its own. It must be complemented by expertise in biological/medical sciences, computational analyses, and physical interpretations. Encouraging collaboration across these disciplines integrates knowledge from different perspectives. Collaborative efforts can drive innovative solutions to the challenges of interpreting complex spectra and lead to the development of new analysis techniques.

Conclusion

In this perspective, we highlighted the critical yet often overlooked intricacies of analyzing vibrational spectral data, particularly IR spectroscopy of cell-free blood. The method's capacity to provide molecular fingerprints is unmatched in its simplicity, speed, and molecular breadth – making it a prime candidate for inclusion in high-throughput diagnostic workflows. To bring the technology further, and to mitigate it to practical use in possible medical diagnostics, its limitations, however, must always be recognized.

Our illustrative case studies utilized thousands of blood-based spectra in addition to spectra of predominant proteins, major lipid particles, and key metabolites. These single-component spectra together make up the primary components of cell-free blood. Yet, we acknowledge that this is not an exhaustive list of all blood constituents and that our approach did not consider inter-molecular interactions.^[92] Nonetheless, by expanding the scope of substances analyzed

and considering the interactions between these molecular components, isolating the signal of specific molecules in complex mixtures becomes even more challenging, as our analysis demonstrates.

Moving forward, we advocate caution when performing IR spectral interpretations to avoid speculative conclusions that could undermine the applicability as well as acceptability of research, particularly in clinical settings. Instead, the high-throughput fingerprinting nature of the approach should be accepted. By embracing both its capabilities and limitations, we can more effectively establish robust spectroscopic analytical workflows.

Acknowledgements

This work was supported by LMU Munich, Centre for Advanced Laser Applications (CALA). We are grateful to Ferenc Krausz for the research environment and support, Maya Hintz for contributing to spectroscopic measurements, Michael Trubetskov for reviewing and feedback on the manuscript, and Annette Peters and Birgit Linkohr for their contributions to clinical study management. Open Access funding enabled and organized by Projekt DEAL.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Keywords: Vibrational Spectroscopy • Infrared Spectroscopy • Molecular Interpretation • Spectral Assignment • Blood Analysis

- [1] P. R. Griffiths, J. A. De Haseth, J. D. Winefordner, *Fourier Transform Infrared Spectrometry*, Chemical Analysis: A Series of Monographs on Analytical Chemistry and Its Applications, Wiley-Blackwell, Chichester, England, 2 edition **2007**.
- [2] A. Barth, P. Haris, *Biological and Biomedical Infrared Spectroscopy*, Advances in biomedical spectroscopy, IOS Press **2009**.
- [3] I. Fleming, D. Williams, *Spectroscopic Methods in Organic Chemistry*, Springer International Publishing **2019**.
- [4] B. H. Stuart, *Infrared Spectroscopy: Fundamentals and Applications*, Wiley **2004**.
- [5] P. J. Larkin, *Infrared and Raman Spectroscopy*, Elsevier **2018**.
- [6] A. L. Mitchell, K. B. Gajjar, G. Theophilou, F. L. Martin, P. L. Martin-Hirsch, *Journal of Biophotonics* **2014**, 7, 153–165.
- [7] M. J. Baker, J. Trevisan, P. Bassan, R. Bhargava, H. J. Butler, K. M. Dorling, P. R. Fielden, S. W. Fogarty, N. J. Fullwood, K. a Heys, C. Hughes, P. Lasch, P. L. Martin-Hirsch, B. Obinaju, G. D. Sockalingum, J. Sulé-Suso, R. J. Strong, M. J. Walsh, B. R. Wood, P. Gardner, F. L. Martin, *Nature Protocols* **2014**, 9, 1771.
- [8] S. De Bruyne, M. M. Speckaert, J. R. Delanghe, *Critical Reviews in Clinical Laboratory Sciences* **2017**, 55, 1–20.
- [9] F. L. Martin, J. G. Kelly, V. Llabjani, P. L. Martin-Hirsch, I. I. Patel, J. Trevisan, N. J. Fullwood, M. J. Walsh, *Nature Protocols* **2010**, 5, 1748.
- [10] W. Petrich, K. B. Lewandrowski, J. B. Muhlestein, M. E. H. Hammond, J. L. Januzzi, E. L. Lewandrowski, R. R. Pearson, B. Dolenko, J. Früh, M. Haass, M. M. Hirsch, W. Köhler, R. Mischler, J. Möcks, J. Ordóñez-Llanos, O. Quarder, R. Somorjai, A. Staib, C. Sylvén, G. Werner, R. Zerback, *The Analyst* **2009**, 134, 1092.
- [11] T. Lubinski, B. Plotka, S. Janik, L. Canini, W. Mäntele, *Journal of Diabetes Science and Technology* **2020**, 15, 6–10.
- [12] M. Huber, K. V. Kepesidis, L. Voronina, F. Fleischmann, E. Fill, J. Hermann, I. Koch, K. Milger-Kneidinger, T. Kolben, G. B. Schulz, F. Jokisch, J. Behr, N. Harbeck, M. Reiser, C. Stief, F. Krausz, M. Žigman, *eLife* **2021**, 10.
- [13] M. Huber, K. V. Kepesidis, L. Voronina, M. Božić, M. Trubetskov, N. Harbeck, F. Krausz, M. Žigman, *Nature Communications* **2021**, 12.
- [14] T. Eissa, C. Leonardo, K. V. Kepesidis, F. Fleischmann, B. Linkohr, D. Meyer, V. Zoka, M. Huber, L. Voronina, L. Richter, A. Peters, M. Žigman, *Cell Reports Medicine* **2024**, 5, 101625.
- [15] A. Schwaighofer, M. Montemurro, S. Freitag, C. Kristament, M. J. Culzoni, B. Lendl, *Analytical Chemistry* **2018**, 90, 7072–7079.
- [16] A. S. Kowligy, H. Timmers, A. J. Lind, U. Elu, F. C. Cruz, P. G. Schunemann, J. Biegert, S. A. Diddams, *Science Advances* **2019**, 5.
- [17] C. Krafft, J. Popp, *Analytical and Bioanalytical Chemistry* **2014**, 407, 699–717.
- [18] I. Pence, A. Mahadevan-Jansen, *Chemical Society Reviews* **2016**, 45, 1958–1979.
- [19] I. Pupeza, M. Huber, M. Trubetskov, W. Schweinberger, S. A. Hussain, C. Hofer, K. Fritsch, M. Poetzlberger, L. Vamos, E. Fill, T. Amotchkina, K. V. Kepesidis, A. Apolonski, N. Karpowicz, V. Pervak, O. Pronin, F. Fleischmann, A. Azzeer, M. Žigman, F. Krausz, *Nature* **2020**, 577, 52–59.
- [20] D. Finlayson, C. Rinaldi, M. J. Baker, *Analytical Chemistry* **2019**, 91, 12117–12128.
- [21] M. Diem, *Journal of Biophotonics* **2018**, 11.
- [22] M. Jackson, H. H. Mantsch, *Critical Reviews in Biochemistry and Molecular Biology* **1995**, 30, 95–120.
- [23] J. Coates, *Interpretation of Infrared Spectra, A Practical Approach* **2000**.
- [24] L. Wang, B. Mizaikoff, *Analytical and Bioanalytical Chemistry* **2008**, 391, 1641–1654.
- [25] D. I. Ellis, R. Goodacre, *The Analyst* **2006**, 131, 875.
- [26] R. Gautam, S. Vanga, F. Arie, S. Umapathy, *EPJ Techniques and Instrumentation* **2015**, 2.
- [27] R. Bhargava, *Annual Review of Analytical Chemistry* **2023**, 16, 205–230.
- [28] C. L. M. Morais, K. M. G. Lima, M. Singh, F. L. Martin, *Nature Protocols* **2020**, 15, 2143–2162.
- [29] L. Voronina, C. Leonardo, J. B. Mueller-Reif, P. E. Geyer, M. Huber, M. Trubetskov, K. V. Kepesidis, J. Behr, M. Mann, F. Krausz, M. Žigman, *Angewandte Chemie International Edition* **2021**, 60, 17060.
- [30] L. Voronina, F. Fleischmann, J. Šimunović, C. Ludwig, M. Novokmet, M. Žigman, *Analytical Chemistry* **2024**.
- [31] C. Hughes, M. Brown, G. Clemens, A. Henderson, G. Monjardez, N. W. Clarke, P. Gardner, *Journal of Biophotonics* **2014**, 7, 180–188.
- [32] H. J. Byrne, F. Bonnier, J. McIntyre, D. R. Parachalil, *Clinical Spectroscopy* **2020**, 2, 100004.

- [33] I. Elsohaby, J. T. McClure, C. B. Riley, J. Bryanton, K. Bigsby, R. A. Shaw, *Journal of Pharmaceutical and Biomedical Analysis* **2018**, *150*, 413–419.
- [34] J. M. Bader, V. Albrecht, M. Mann, *Molecular & Cellular Proteomics* **2023**, *22*, 100577.
- [35] E. Gray, H. J. Butler, R. Board, P. M. Brennan, A. J. Chalmers, T. Dawson, J. Goodden, W. Hamilton, M. G. Hegarty, A. James, M. D. Jenkinson, D. Kernick, E. Lekka, L. J. Livermore, S. J. Mills, K. O'Neill, D. S. Palmer, B. Vaqas, M. J. Baker, *BMJ Open* **2018**, *8*, e017593.
- [36] O. Richards, C. Jenkins, H. Griffiths, E. Paczkowska, P. R. Dunstan, S. Jones, M. Morgan, T. Thomas, J. Bowden, A. Nakimuli, M. Nair, C. A. Thornton, *Frontiers in Global Women's Health* **2021**, *1*.
- [37] X. Yang, Q. Ou, K. Qian, J. Yang, Z. Bai, W. Yang, Y. Shi, G. Liu, *Frontiers in Oncology* **2021**, *11*.
- [38] A. Sala, D. J. Anderson, P. M. Brennan, H. J. Butler, J. M. Cameron, M. D. Jenkinson, C. Rinaldi, A. G. Theakstone, M. J. Baker, *Cancer Letters* **2020**, *477*, 122–130.
- [39] X. Yang, T. Fang, Y. Li, L. Guo, F. Li, F. Huang, L. Li, *Optik* **2019**, *180*, 189–198.
- [40] D. A. Scott, D. E. Renaud, S. Krishnasamy, P. Meriç, N. Buduneli, S. Undefinedetinkalp, K.-Z. Liu, *Diabetology & Metabolic Syndrome* **2010**, *2*.
- [41] H. Fabian, W. Mäntele, *Infrared Spectroscopy of Proteins* **2001**.
- [42] D. Voet, J. G. Voet, *Biochemistry, international adaptation*, John Wiley & Sons, Nashville, TN, 4 edition **2021**.
- [43] R. A. Davis, J. E. Vance, *Structure, assembly and secretion of lipoproteins*, page 473–493, Elsevier **1996**.
- [44] B. Lin, X. Qing, J. Liao, K. Zhuo, *Cells* **2020**, *9*, 1022.
- [45] S. Kino, M. Kanamori, Y. Shimoda, K. Niizuma, H. Endo, Y. Matsuura, *BMC Cancer* **2024**, *24*.
- [46] J. Titus, E. Viennois, D. Merlin, A. G. Unil Perera, *Journal of Biophotonics* **2016**, *10*, 465–472.
- [47] J. T. Borden, A. Man, D. A. Scott, K.-Z. Liu, *Journal of Molecular Medicine* **2003**, *81*, 788–794.
- [48] H. Ghimire, C. Garlapati, E. A. M. Janssen, U. Krishnamurti, G. Qin, R. Aneja, A. G. U. Perera, *Cancers* **2020**, *12*, 1708.
- [49] K. Thumanu, S. Sangrajang, T. Khuaprema, A. Kalalak, W. Tanthanuch, S. Pongpiachan, P. Heraud, *Journal of Biophotonics* **2014**, *7*, 222–231.
- [50] J. D. Rodriguez, B. J. Westenberger, L. F. Buhse, J. F. Kauffman, *Analytical Chemistry* **2011**, *83*, 4061–4067.
- [51] J. Ollesch, D. Theegarten, M. Altmayer, K. Darwiche, T. Hager, G. Stamatis, K. Gerwert, *Biomedical Spectroscopy and Imaging* **2016**, *5*, 129–144.
- [52] Y. Shen, A. Davies, E. Linfield, P. Taday, D. Arnone, T. Elsey, *Journal of Biological Physics* **2003**, *29*, 129–133.
- [53] G. Hoşafçı, O. Klein, G. Oremek, W. Mäntele, *Analytical and Bioanalytical Chemistry* **2006**, *387*, 1815–1822.
- [54] C. Petibois, G. Cazorla, A. Cassaigne, G. Déleris, *Clin. Chem.* **2001**, *47*, 730.
- [55] R. A. Shaw, S. Kotowich, M. Leroux, H. H. Mantsch, *Annals of Clinical Biochemistry: International Journal of Laboratory Medicine* **1998**, *35*, 624–632.
- [56] S. Yao, H. Bao, S. M. Nuguri, L. Yu, Z. Mikulik, M. M. Osuna-Diaz, K. R. Sebastian, K. V. Hackshaw, L. Rodriguez-Saona, *Biomedicine* **2023**, *11*, 712.
- [57] U. Zelig, E. Barlev, O. Bar, I. Gross, F. Flomen, S. Mordechai, J. Kapelushnik, I. Nathan, H. Kashtan, N. Wasserberg, O. Madhala-Givon, *BMC Cancer* **2015**, *15*.
- [58] P. Carmona, M. Molina, M. Calero, F. Bermejo-Pareja, P. Martínez-Martín, A. Toledano, *Journal of Alzheimer's Disease* **2013**, *34*, 911–920.
- [59] X. Wang, X. Shen, D. Sheng, X. Chen, X. Liu, *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **2014**, *122*, 193–197.
- [60] S. Guo, G. Wei, W. Chen, C. Lei, C. Xu, Y. Guan, T. Ji, F. Wang, H. Liu, *Biomolecules* **2022**, *12*, 1815.
- [61] L. Lechowicz, M. Chrapek, J. Gaweda, M. Urbaniak, I. Konieczna, *Molecular Biology Reports* **2016**, *43*, 1321–1326.
- [62] J. R. Hands, G. Clemens, R. Stables, K. Ashton, A. Brodbelt, C. Davis, T. P. Dawson, M. D. Jenkinson, R. W. Lea, C. Walker, M. J. Baker, *Journal of Neuro-Oncology* **2016**, *127*, 463–472.
- [63] C. S. Carvalho, A. C. A. Silva, T. J. P. S. Santos, A. A. Martin, A. C. dos Santos Fernandes, L. E. Andrade, L. Raniero, A rheumatoid arthritis study by Fourier transform infrared spectroscopy, in A. MahadevanJansen, W. Petrich (Editors), *Biomedical Vibrational Spectroscopy V: Advances in Research and Industry*, volume 8219, International Society for Optics and Photonics, SPIE **2012** page 821910.
- [64] M. Paraskevaidi, C. L. M. Morais, K. M. G. Lima, J. S. Snowden, J. A. Saxon, A. M. T. Richardson, M. Jones, D. M. A. Mann, D. Allsop, P. L. Martin-Hirsch, F. L. Martin, *Proceedings of the National Academy of Sciences* **2017**, *114*.
- [65] P. Giamougiannis, C. L. M. Morais, B. Rodriguez, N. J. Wood, P. L. Martin-Hirsch, F. L. Martin, *Analytical and Bioanalytical Chemistry* **2021**, *413*, 5095–5107.
- [66] D. Sheng, Y. Wu, X. Wang, D. Huang, X. Chen, X. Liu, *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **2013**, *116*, 365–369.
- [67] H. Ghimire, X. Hu, G. Qin, A. G. Unil Perera, *Biomedical Optics Express* **2020**, *11*, 4679.
- [68] E. Barlev, U. Zelig, O. Bar, C. Segev, S. Mordechai, J. Kapelushnik, I. Nathan, F. Flomen, H. Kashtan, R. Dickman, O. Madhala-Givon, N. Wasserberg, *Journal of Gastroenterology* **2015**, *51*, 214–221.
- [69] V. E. Sitnikova, M. A. Kotkova, T. N. Nosenko, T. N. Kotkova, D. M. Martynova, M. V. Uspenskaya, *Talanta* **2020**, *214*, 120857.
- [70] N. Mateus Pereira de Souza, B. Hunter Machado, A. Koche, L. Beatriz Fernandes da Silva Furtado, D. Becker, V. Antonio Corbellini, A. Rieger, *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **2023**, *288*, 122135.
- [71] R. Meddeb, Z. A. A. Dache, S. Thezenas, A. Otandault, R. Tanos, B. Pastor, C. Sanchez, J. Azzi, G. Tusch, S. Azan, C. Mollevi, A. Adenis, S. El Messaoudi, P. Blache, A. R. Thierry, *Scientific Reports* **2019**, *9*.
- [72] M. H. Larson, W. Pan, H. J. Kim, R. E. Mauntz, S. M. Stuart, M. Pimentel, Y. Zhou, P. Knudsgaard, V. Demas, A. M. Aravanis, A. Jamshidi, *Nature Communications* **2021**, *12*.
- [73] V. Swarup, M. Rajeswari, *FEBS Letters* **2007**, *581*, 795–799.
- [74] T. Eissa, K. V. Kepesidis, M. Zigman, M. Huber, *Analytical Chemistry* **2023**, *95*, 6523–6532.
- [75] I. Maitra, C. L. M. Morais, K. M. G. Lima, K. M. Ashton, R. S. Date, F. L. Martin, *The Analyst* **2019**, *144*, 7447–7456.
- [76] P. Müller, A. F. Schier, *Developmental Cell* **2011**, *21*, 145–158.
- [77] B. P. Chatterjee, G. Mondal, U. Chatterjee, *Proceedings of the National Academy of Sciences, India Section B: Biological Sciences* **2014**, *84*, 865–874.
- [78] D. R. de Graaff, S. Felz, T. R. Neu, M. Pronk, M. C. van Loosdrecht, Y. Lin, *Water Research* **2019**, *155*, 343–351.
- [79] I. S. Pinel, H. B. Kleikamp, M. Pabst, J. S. Vrouwenfelder, M. C. van Loosdrecht, Y. Lin, *Applied Sciences* **2020**, *10*, 7694.
- [80] M. Khajehpour, J. L. Dashnau, J. M. Vanderkooi, *Analytical Biochemistry* **2006**, *348*, 40–48.
- [81] P. Lasch, *Chemometrics and Intelligent Laboratory Systems* **2012**, *117*, 100–114.
- [82] H. J. Butler, B. R. Smith, R. Fritsch, P. Radhakrishnan, D. S. Palmer, M. J. Baker, *The Analyst* **2018**, *143*, 6121–6134.

- [83] T. W. Randolph, *Cancer Biomarkers* **2006**, 2, 135.
- [84] J. Thyr, T. Edvinsson, *Angewandte Chemie* **2023**, 135.
- [85] P. R. Griffiths, *Science* **1983**, 222, 297–302.
- [86] S. Guo, J. Popp, T. Bocklitz, *Nature Protocols* **2021**, 16, 5426–5459.
- [87] T. Eissa, M. Huber, B. Obermayer-Pietsch, B. Linkohr, A. Peters, F. Fleischmann, M. Žigman, *PNAS Nexus* **2024**, 3, 449.
- [88] J. Bujok, M. Gąsior-Głogowska, M. Marszałek, N. Trochanowska-Pauk, F. Zigo, A. Pavlak, M. Komorowska, T. Walski, *BioMed Research International* **2019**, 2019, 1–9.
- [89] P. Lasch, J. Schmitt, M. Beekes, T. Udelhoven, M. Eiden, H. Fabian, W. Petrich, D. Naumann, *Analytical Chemistry* **2003**, 75, 6673–6678.
- [90] R. Milo, R. Phillips, *Cell Biology by the Numbers*, Garland Science **2015**.
- [91] P. Zucchini, E. Mitri, S. Kenig, F. Billè, G. Kourousias, D. E. Bedolla, L. Vaccari, *Analytical Chemistry* **2016**, 88, 12090–12098.
- [92] P. I. Haris, G. T. Robillard, A. A. Van Dijk, D. Chapman, *Biochemistry* **1992**, 31, 6279–6284.

Manuscript received: June 20, 2024

Version of record online: November 7, 2024



Supporting Information

The Perils of Molecular Interpretations from Vibrational Spectra of Complex Samples

*T. Eissa, L. Voronina, M. Huber, F. Fleischmann, M. Žigman**

Material and Methods

Blood sample collection and processing

All blood samples utilized in this work were considered in our previous work^[1–3], where details on the clinical studies, sample collection, and sample processing were provided.

Details relating to lung cancer and proteomic analysis (Fig. 4) were provided in Voronina et al.^[2] Lung cancer detection was a part of a multi-institutional study on lung, breast, bladder, and prostate cancers with corresponding benign pathologies in the same organs as well as non-symptomatic subjects^[1]. Participants provided written informed consent for the study under research study protocol #17-141 and broad consent under research study protocol #17-182, both of which were approved by the Ethics Committee of the Ludwig-Maximillan-University (LMU) of Munich. The study complies with all relevant ethical regulations and was conducted according to Good Clinical Practice (ICH-GCP) and the principles of the Declaration of Helsinki. The clinical trial is registered (ID DRKS00013217) at the German Clinical Trials Register (DRKS).

Details relating to the lipid parameter analysis (Fig. 8) were provided in Eissa et al.^[3] Blood samples were collected as a part of the Cooperative Health Research in the Region of Augsburg (KORA) study^[4] – which comprised of an age- and sex-stratified sample of participants randomly drawn from the resident registration offices within Southern Germany. The KORA study methods were approved by the ethics committee of the Bavarian Chamber of Physicians, Munich (EC No. 06068).

Chemicals, reagents, and measurements of infrared absorption spectra

Proteins and lipid particles of the highest available purity were purchased from Sigma Aldrich GmbH (Taufkirchen, Germany). They were diluted in deionized water to concentrations of 1–5 mg/mL, and their IR absorption spectra were measured. If the presence of additional salts was detected, the buffer was exchanged to deionized water using Amicon Ultra centrifugal filters with 10 kDa cutoff, purchased from Sigma Aldrich GmbH (Taufkirchen, Germany). The concentration of the proteins in the resulting solution was then measured using a Nanodrop One UV-Vis spectrophotometer (Thermo Fischer Scientific GmbH, Dreieich, Germany), and the UV absorption of the solution before buffer exchange was used for device calibration.

To assess how the pH of the solution and the resulting protein protonation state affected the IR spectra of the proteins, we recorded the absorption spectra of several model proteins in PBS buffer (pH 7.4). We then subtracted the IR spectrum of the buffer from that of the protein in a buffer in order to obtain the IR signature of a pure protein in a controlled protonation state. When comparing these spectra with the ones recorded in pure water, we did not observe any significant differences in the ratios of the primary absorption bands, their shapes, or intensities. Based on experiences across a variety of our experimentation, we consider that the details of the IR spectra of the proteins related to the pH would not affect any of our conclusions.

The water-soluble metabolites were purchased from Sigma Aldrich GmbH (Taufkirchen, Germany), diluted in deionized water, and their IR absorption was measured. The same procedure was followed for glycans, purchased from Dextra Laboratories (Reading, UK).

Pure DNA from calf thyme and RNA from bakers yeast were purchased from Sigma Aldrich GmbH (Taufkirchen, Germany) and dissolved in deionized water. Their concentration and purity (A260/280 ratio >1.9) were assessed using a Nanodrop One UV-Vis spectrophotometer with the built-in methods for double-stranded DNA and single-stranded RNA. Based on NanoDrop results, the concentrations were adjusted to 1 mg/mL for the FTIR measurement. The obtained spectra were scaled down to the concentrations shown in Fig. 5, as those are below the detection limit.

Measurements of liquid biofluids and single components were all performed in a hydrated, fluid state using an automated FTIR device MIRA-Analyzer (Clade GmbH, Esslingen, Germany) with a flow-through transmission cuvette (CaF₂ with 8 μ m path length). The spectra were acquired with a resolution of 4 cm^{−1} and an averaging time of 45 s. Prior to every sample, a water reference spectrum was measured to reconstruct the IR absorption spectra. The spectra were obtained in the range 930–3050 cm^{−1} and truncated to 960–3000 cm^{−1}.

Preprocessing of all absorption spectra was performed as follows. To account for the water substituted by the molecular constituents in the sample compared to the pure water reference, a water absorption spectrum was added to the sample spectrum with a coefficient optimized such that the first derivative of the signal in the region of 1850–2150 cm^{−1} was minimized, similar to our prior work^[1–3,5]. Subsequently, the minimum of the absorption in this region was subtracted from the spectrum (“offset correction”) ^[2]. Vector normalization was applied to the spectra in the analysis relevant to lung cancer (Fig. 4), as performed in the prior work^[2]. Vector normalization (L2) was also applied in part of the analysis related to the normalization analysis (Fig. 7), as described in the figure legend.

Mathematical optimizations and statistical analyses

All mathematical and statistical analysis was performed in a Python environment (version 3.8.8) using custom scripts.

To reproduce a spectral signal given a set of substances (Figs. 3, 4, S2), the “minimize” function from the SciPy Python package (version 1.6.2) was used^[6]. The objective function calculated the mean squared error between the target signal to be reproduced and the input set of spectra [Eq. 2].

For the regression analysis to quantify the concentrations of lipid analytes (Fig. 8), a linear regression algorithm with a ridge penalty was used. A cross-validation was performed on the available spectra to quantify the concentrations of the analytes which were determined from clinical chemistry analysis^[3]. Mean R^2 values of the test predictions were reported, and all test set prediction errors were depicted (Fig. 8D,E). The Scikit-learn Python package (version 0.24.1) was used for its implementations of the ridge regression algorithm^[7].

Table S1. Individual molecular substances considered in this work, consisting of ten proteins, three lipid particles, and nine metabolites, along with their typical concentrations in healthy adults.

Substance	Abbreviation	Concentration [g/l] ^[a]	Molecular class
Human serum albumin	HSA	42.65	Protein
Immunoglobulin G	IgG	9.55	Protein
Immunoglobulin A	IgA	3.36	Protein
Alpha-2-macroglobulin	A2M	3.20	Protein
Transferrin	TF	2.66	Protein
Haptoglobin	HP	1.74	Protein
Immunoglobulin M	IgM	1.71	Protein
Alpha-1-antitrypsin	A1AT	1.49	Protein
Alpha-1-acid glycoprotein	A1AG	1.03	Protein
Alpha-1-antichymotrypsin	A1ACT	0.55	Protein
Low-density lipoprotein	LDL	3.11	Lipid particle
Very low-density lipoprotein	VLDL	2.40	Lipid particle
High-density lipoprotein	HDL	2.00	Lipid particle
Adenosine triphosphate	ATP	1.52	Metabolite
D-Glucose	Glucose	0.90	Metabolite
Urea	Urea	0.24	Metabolite
Lactic acid	Lactic acid	0.09	Metabolite
Glutamine	Glutamine	0.07	Metabolite
Lysine	Lysine	0.05	Metabolite
Alanine	Alanine	0.04	Metabolite
Beta-hydroxybutyric acid	BHB	0.04	Metabolite
Glycine	Glycine	0.03	Metabolite

[a] Typical concentrations were estimated from existing literature for proteins^[8–13] and metabolites^[14]. Concentrations of lipid particles were estimated from the studied KORA population-based cohort (Fig. 8B)^[3,4]. Specifically, HDL and LDL particle concentrations were estimated from HDL and LDL cholesterol levels – assuming that HDL and LDL particles are respectively made of 30% and 45% cholesterol and cholesteryl ester^[15]. VLDL particle concentrations were estimated from triglyceride concentrations – assuming that triglycerides make up 50% of VLDL particles^[15].

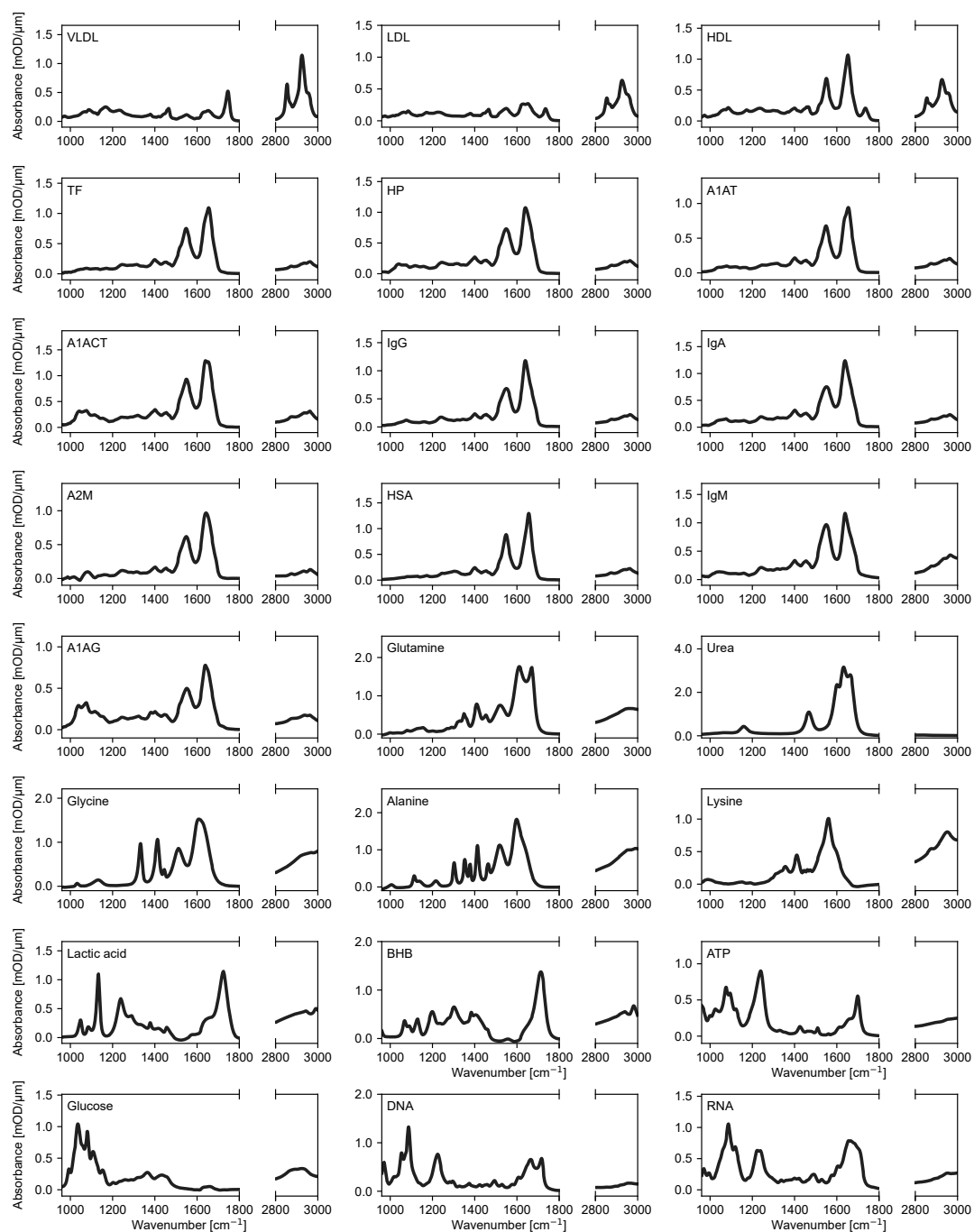


Figure S1. FTIR absorption spectra of the individual molecular substances. All substances are depicted at a concentration of 5 g/l.

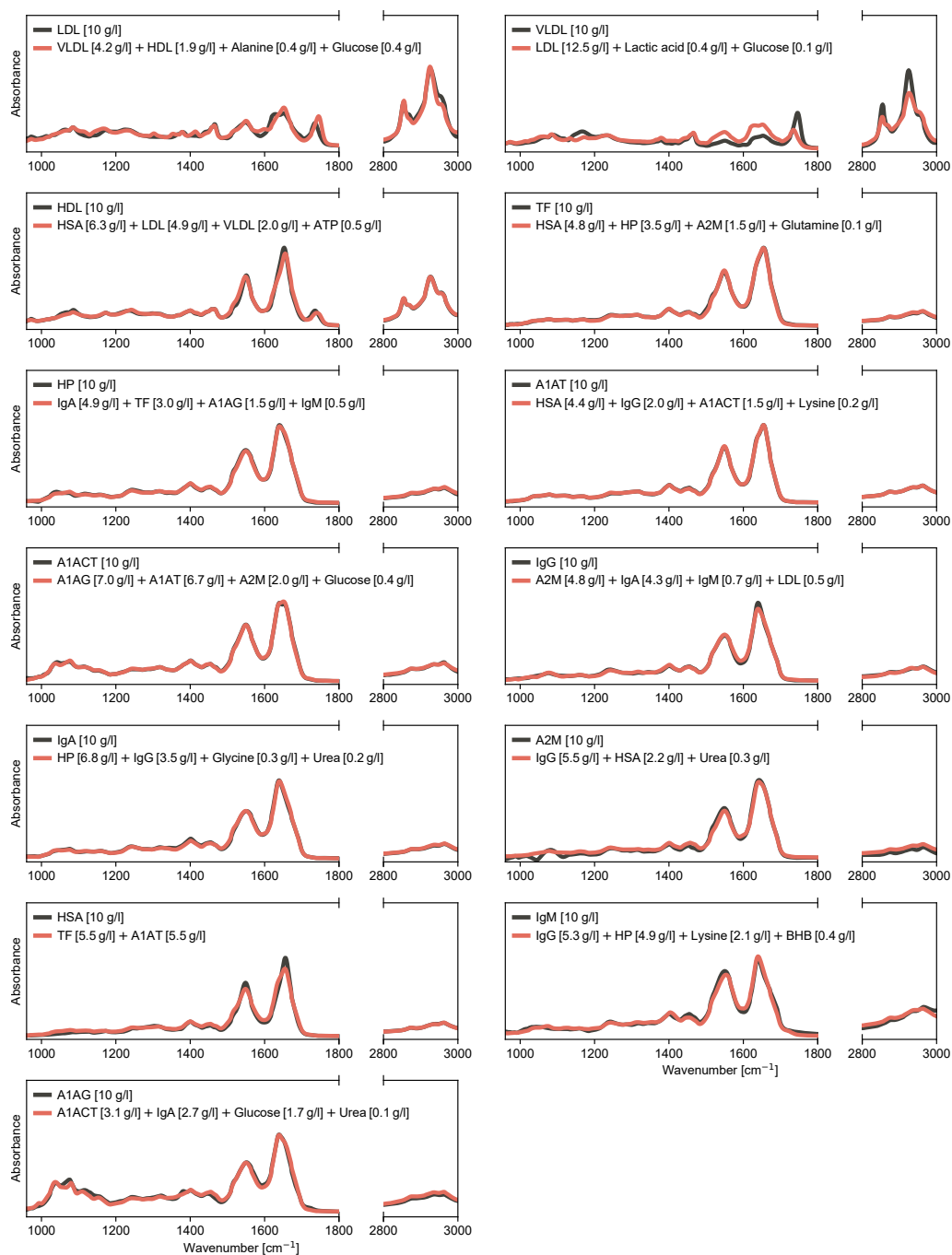


Figure S2. Reconstructing the spectrum of a single substance (black) from a linear combination of other substances (red) scaled to different concentrations. Related to Fig. 3.

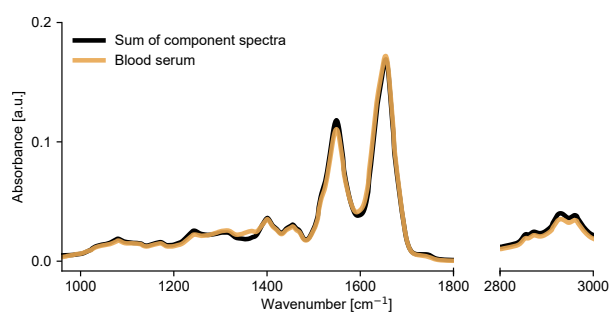


Figure S3. Comparing an experimental blood serum spectrum with a model constructed from individual component spectra. The modeled spectrum (black) was generated through the addition of individual component spectra of different metabolites, proteins, and lipid particles, each scaled to their typical concentrations in healthy adults (Table S1). The experimental spectrum (yellow) represents the mean spectrum of bulk serum from a nominally healthy adult cohort.

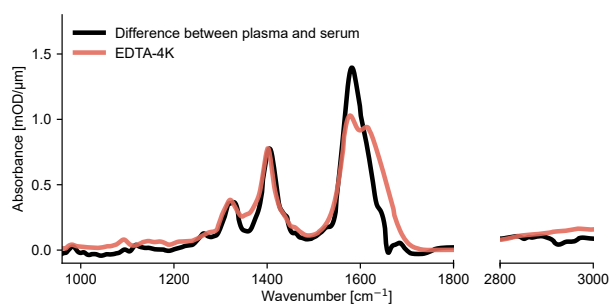


Figure S4. Difference between absorption spectra of blood plasma and serum. Both plasma and serum were processed from the same whole blood samples for different individuals. The black curve depicts the difference between plasma and serum spectra, processed from the same whole blood sample, and subsequently averaged across the different individuals. The red curve depicts the spectrum of EDTA-4K (Ethylenediaminetetraacetic acid tetrapotassium salt), the anticoagulant used for plasma collection. EDTA-4K was measured in a NaCl background and its spectrum was scaled to align the peak near 1400 cm^{-1} .

Supplementary References

- [1] M. Huber, K. V. Kepesidis, L. Voronina, F. Fleischmann, E. Fill, J. Hermann, I. Koch, K. Milger-Kneidinger, T. Kolben, G. B. Schulz, F. Jokisch, J. Behr, N. Harbeck, M. Reiser, C. Stief, F. Krausz, M. Žigman, *eLife* **2021**, *10*.
- [2] L. Voronina, C. Leonardo, J. B. Mueller-Reif, P. E. Geyer, M. Huber, M. Trubetskov, K. V. Kepesidis, J. Behr, M. Mann, F. Krausz, M. Žigman, *Angewandte Chemie International Edition* **2021**, *60*, 17060.
- [3] T. Eissa, C. Leonardo, K. V. Kepesidis, F. Fleischmann, B. Linkohr, D. Meyer, V. Zoka, M. Huber, L. Voronina, L. Richter, A. Peters, M. Žigman, *Cell Reports Medicine* **2024**, *5*, 101625.
- [4] R. Holle, M. Happich, H. Löwel, H. Wichmann, *Das Gesundheitswesen* **2005**, *67*, 19–25.
- [5] M. Huber, K. V. Kepesidis, L. Voronina, M. Božić, M. Trubetskov, N. Harbeck, F. Krausz, M. Žigman, *Nature Communications* **2021**, *12*.
- [6] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, Í. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, SciPy 1.0 Contributors, *Nature Methods* **2020**, *17*, 261.
- [7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, *Journal of Machine Learning Research* **2011**, *12*, 2825.
- [8] R. F. Ritchie, G. E. Palomaki, L. M. Neveux, O. Navolotskaia, *Journal of Clinical Laboratory Analysis* **1998**, *12*, 371–377.
- [9] R. F. Ritchie, G. E. Palomaki, L. M. Neveux, O. Navolotskaia, T. B. Ledue, W. Y. Craig, *Journal of Clinical Laboratory Analysis* **1999**, *13*, 273–279.
- [10] R. F. Ritchie, G. E. Palomaki, L. M. Neveux, O. Navolotskaia, T. B. Ledue, W. Y. Craig, *Journal of Clinical Laboratory Analysis* **2000**, *14*, 284–292.
- [11] A. Kratz, M. Ferraro, P. M. Sluss, K. B. Lewandrowski, *New England Journal of Medicine* **2004**, *351*, 1548–1563.
- [12] J. Housley, *Journal of Clinical Pathology* **1968**, *21*, 27–31.
- [13] F. Licastrò, E. Masliah, S. Pedrini, L. J. Thal, *Dementia and Geriatric Cognitive Disorders* **2000**, *11*, 25–28.
- [14] N. Psychogios, D. D. Hau, J. Peng, A. C. Guo, R. Mandal, S. Bouatra, I. Sinelnikov, R. Krishnamurthy, R. Eisner, B. Gautam, N. Young, J. Xia, C. Knox, E. Dong, P. Huang, Z. Hollander, T. L. Pedersen, S. R. Smith, F. Bamforth, R. Greiner, B. McManus, J. W. Newman, T. Goodfriend, D. S. Wishart, *PLoS ONE* **2011**, *6*, e16957.
- [15] R. A. Davis, J. E. Vance, *Structure, assembly and secretion of lipoproteins*, page 473–493, Elsevier **1996**.

Appendix C

Limits and prospects of molecular fingerprinting for phenotyping biological systems revealed through *in silico* modeling

Copyright notice: © 2023 The Author(s). This is an open access article distributed under the terms of the Creative Commons CC-BY license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The CC-BY 4.0 license is printed in Appendix E.

Limits and Prospects of Molecular Fingerprinting for Phenotyping Biological Systems Revealed through *In Silico* Modeling

Tarek Eissa, Kosmas V. Kepesidis, Mihaela Zigman,* and Marinus Huber*

Cite This: *Anal. Chem.* 2023, 95, 6523–6532

Read Online

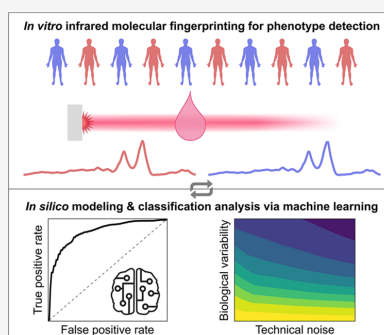
ACCESS |

Metrics & More

Article Recommendations

Supporting Information

ABSTRACT: Molecular fingerprinting via vibrational spectroscopy characterizes the chemical composition of molecularly complex media which enables the classification of phenotypes associated with biological systems. However, the interplay between factors such as biological variability, measurement noise, chemical complexity, and cohort size makes it challenging to investigate their impact on how the classification performs. Considering these factors, we developed an *in silico* model which generates realistic, but configurable, molecular fingerprints. Using experimental blood-based infrared spectra from two cancer-detection applications, we validated the model and subsequently adjusted model parameters to simulate diverse experimental settings, thereby yielding insights into the framework of molecular fingerprinting. Intriguingly, the model revealed substantial improvements in classifying clinically relevant phenotypes when the biological variability was reduced from a between-person to a within-person level and when the chemical complexity of the spectra was reduced. These findings quantitatively demonstrate the potential benefits of personalized molecular fingerprinting and biochemical fractionation for applications in health diagnostics.



Vibrational spectroscopy by means of Raman or infrared techniques is a powerful analytical platform capable of characterizing molecular samples at any state of matter in a label-free manner.^{1,2} Since essentially every molecule exhibits a unique vibrational spectrum, spectroscopic approaches are able to quantify individual molecular contributions in complex matrices.³ Although determining changes in the nature and quantity of individual molecular species is challenged by overlapping spectral bands, the vibrational spectrum still reflects the overall molecular composition of a given sample and is therefore referred to as its “molecular fingerprint”. Statistical or machine learning methods can identify spectral patterns specific to molecular phenotypes and consequently classify samples.

Molecular fingerprinting by vibrational spectroscopy has been increasingly applied to biomedical problems.⁴ The approach has been used to classify bacteria and cell (sub)types,^{5,6} distinguish between benign and malignant tissues,^{1,2} and identify diseases based on fingerprint spectra of biofluids.^{2,7}

Although successfully applied, the prospects and fundamental limitations of vibrational fingerprinting for certain applications, such as for clinically relevant questions, remain largely unexplored.^{2,8} The challenge is that a variety of technical and data acquisition aspects impact the measured spectra and thus can affect the classification accuracy.^{9–11} At the same time, every living system exhibits an inherent level of biological variability,^{10,12,13} further challenging the unambiguous identification of different molecular states. While previous work has examined the variations caused by technical, data acquisition,

and biological aspects on recorded spectra,^{10,11,14} disentangling their individual contribution on the classification accuracy remains experimentally challenging. This often requires large-scale studies involving different measurement instruments, different laboratories, and different protocols for sample handling, which may be difficult to realize due to resource limitations.¹⁵

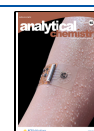
In silico investigations performed via computer simulations are particularly powerful in this respect as they can address problems that are practically or experimentally challenging.^{16,17} By creating computational models capable of mimicking the behavior of biological systems, considering various sources of noise or heterogeneity, one can rapidly gain insights into the underlying mechanisms that define the behavior of a system under differing simulated conditions.¹⁸

To investigate the fundamental capacity of vibrational fingerprinting, we propose an *in silico* approach that generates artificial infrared spectra (Figure 1). Our approach is based on modeling the molecular composition of a given biological system in a defined molecular state and transforming this

Received: October 25, 2022

Accepted: March 15, 2023

Published: April 12, 2023



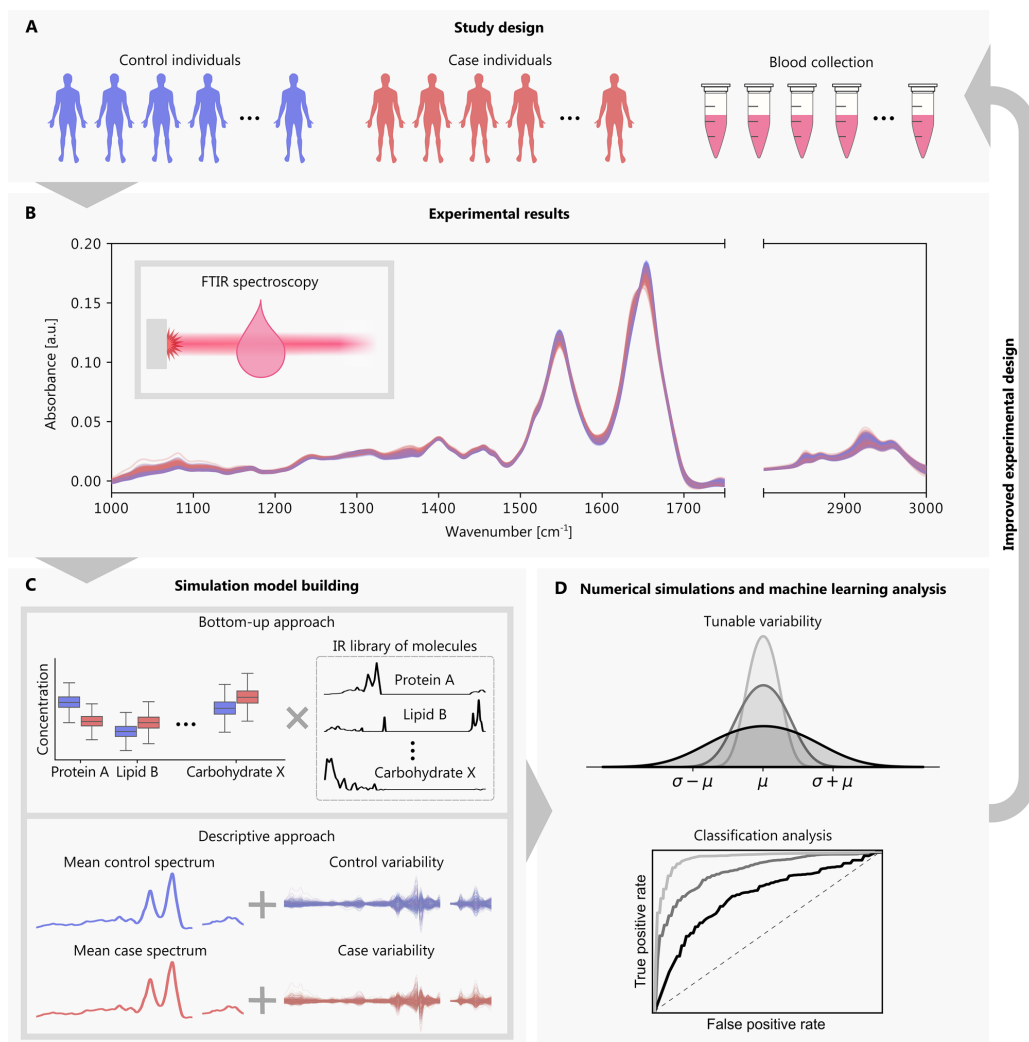


Figure 1. Overview of the *in silico* model and its application to infrared fingerprinting for disease detection. (A) Blood samples are collected from case and control individuals, and (B) blood-based samples are measured with an infrared spectrometer. Machine learning algorithms are applied to the spectral data set, and a value of classification accuracy is retrieved. (C) Same problem can be investigated with an *in silico* model using either a bottom-up approach based on single component spectra or a descriptive approach based on measured infrared fingerprints of blood-based samples. (D) Within the frame of the model, the influence of simulated experimental conditions is investigated to gain insight into the effects different simulated conditions have on the predictive capacity of machine learning models. Based on these findings, (A) initial study design or (B) measurement mode can be adapted to improve the performance of the envisioned application.

composition into a measurable quantity—the resulting infrared spectrum. With this model, analytical sources of variation can be considered to obtain realistic spectra that have comparable properties to experimental observations. Not only does this allow for the generation of simulated “measurement events” in any number, but also enables the precise control of crucial parameters that may impact the recorded spectra. Our model provides foundational insights into the factors affecting the underlying measurement approach—molecular fingerprinting of complex biological systems—and carries the capacity to guide

future experiments without the need for exhaustive sample collections and measurements.

The model is described and tested on a previously published application—infrared fingerprinting of blood sera to detect lung and prostate cancer.⁷ Our descriptions are kept as general as possible to facilitate their transferability to other fingerprinting techniques and applications. Furthermore, we provide a toolbox (written in Python),¹⁹ including the generated data, to allow for convenient applications of the proposed model. Although applied here to infrared spectra, the model can be applied to

other molecular fingerprinting approaches, such as Raman²⁰ or nuclear magnetic resonance²¹ spectroscopy.

Altogether, we obtain excellent agreement between the results of experimental and simulated data. By systematically adjusting all model parameters, we explored the potentials and limitations of molecular fingerprinting and thus contribute to accelerating its real-world applications.

METHODS

In Silico Model. In essence, the problems tackled by molecular pattern recognition using machine learning are often very related. Depending on the application, by detecting changes in qualitative or quantitative aspects, a distinction between two or more molecular states is desired. In practice, this distinction is only possible to a certain extent due to the inherent biological or sample variability as well as measurement-specific errors. Additionally, often a limited amount of data is available to robustly validate conclusions drawn by analyzing differences between the studied states. Therefore, it can be difficult to make conclusive statements about the potential and limitations of the method in question or to isolate the molecular signal of interest. Even in a scenario where the signal is already known (e.g., the spectrum of a specific molecule), the expected accuracy of distinguishing between different states often requires numerous measurements due to the multiple sources of variation. To facilitate the investigation of these points, the *in silico* model should be able to generate spectra which meet the following requirements:

- Reflect the biological or sample variability of a given sample pool.
- Incorporate differences between different molecular states (e.g., cases and controls).
- Consider characteristics of the physical measurement, namely the noise introduced by the measurement device.

We describe two approaches to generate the artificial spectra. In a so-called bottom-up approach, the contribution of individual molecular species in a given sample is considered. Although intuitive, this approach cannot be implemented in the case of infrared spectroscopy of highly complex samples as we discuss below. Instead, it provides the motivation and basis for implementing another so-called descriptive approach which follows a similar mathematical formulation. Although the model can be applied to study different samples types with varying chemical complexities and to several molecular fingerprinting techniques, infrared molecular fingerprints of blood serum will serve as the guiding principle for assembling the model.

Bottom-Up Approach. It has been shown that by breaking down a molecularly complex sample into its individual components, the corresponding infrared spectrum can be reasonably well approximated with a linear combination of individual component spectra of the most abundant molecular species.²² Thereby, each molecular spectrum $\mathbf{x}_i$ is scaled to its respective concentration c_i with $\sum_{i=1}^M c_i \cdot \mathbf{x}_i$.

This idea can be also utilized to create spectra of a given population. The normal concentration ranges for the most abundant blood components in healthy individuals—which in total make up more than 99% of the molecular mass—can be obtained from the literature.^{23,24} Using these facts, the concentration parameter c_i could be replaced with a random variable $c(\mu_i, \sigma_i)$ that models the distribution of each component for a healthy cohort. Thereby, a set of spectra that model the

molecularly complex samples could be simulated where each spectrum is modeled as a statistical outcome Y with:

$$Y = \sum_{i=1}^M c(\mu_i, \sigma_i) \cdot \mathbf{x}_i \quad (1)$$

We term this model formulation as the bottom-up approach.

It is important to note that the bottom-up approach contains several simplifications and assumptions. Specifically, it is assumed that the concentrations of individual molecules or molecular classes are independent, which is not the case. In reality, biological networks are interconnected, and molecular changes within living organisms are occurring in a correlated fashion.²⁵ In cases where these correlations are well known, introducing dependencies between the corresponding random variables would, in principle, account for this behavior.

The assumption that an infrared spectrum of a complex sample can be described by a linear combination of individual component spectra scaled to their concentrations can be regarded as sufficiently fulfilled. Although molecular spectra are influenced by inter-molecular interactions^{26,27} and environmental conditions (e.g., temperature and pH-value), several studies suggest that complex blood spectra can be described by linear superposition of individual single component spectra. For infrared spectroscopy, it is known that the strength of a molecular spectrum scales linearly with concentration over many orders of magnitude.^{3,28} Furthermore, it has been shown that the concentration of several different molecules in blood can be determined by linear regression.^{29–32} Moreover, it was shown that a linear combination of the spectra of the most abundant molecular species can be used to describe experimentally observed differences in the spectra of blood serum from a healthy cohort and from patients with lung cancer.²²

While the bottom-up approach is straightforward to envision, access to the associated molecular spectra is essential for creating the model. This is a crucial limitation since, to our knowledge, there exists no consistent database with infrared spectra of the majority of molecules contained in complex biological systems. While previous work suggests that 12 selected protein spectra can model the shape of blood serum spectra,²² our results show that significantly more are required to represent the molecular complexity and biological variability adequately. Taken together, from the current perspective, it is unlikely that the bottom-up can be implemented when studying a complex blood-based matrix. In principle, however, it can be used to simulate the spectra of simpler systems, such as for pharmaceutical samples that can be described as a mixture of a few well-known substances.³³

Descriptive Approach. To mitigate the limited access to individual component infrared spectra, we introduce an alternative descriptive approach. Here, we use a set of m experimentally derived spectra $\mathbf{s}_i$ that describe the samples of a particular biological matrix (e.g., blood serum of healthy individuals). By utilizing a random variable $\beta(\mu_i, \sigma_i)$ assuming a Gaussian distribution, the spectrum of a measured biological matrix can now be expressed as another statistical outcome Y with:

$$Y = \sum_{i=1}^m \beta(\mu_i, \sigma_i) \cdot \mathbf{s}_i \quad (2)$$

Thereby, the mathematical structure of the formulation is similar to the bottom-up approach, but utilizing a different set of calibration vectors to realize the model.

One difference between the two model formulations is that we cannot rely on the literature to obtain the concentration ranges for individual molecules and construct a realistic behavior for the biological variability. Therefore, a calibration procedure is required to adjust the statistical variables and select the experimentally derived spectra such that the resulting spectra reflect the biological variability of the measured samples. In addition, we must validate that assuming a Gaussian distribution creates spectral cohorts that adequately match experimentally measured cohorts.

To calibrate the biological variability, we use a set of experimentally measured blood serum spectra b_i and calculate the expected value and variance of the descriptive approach:

$$Y = \sum_{i=1}^m \beta(\mu_i, \sigma_i) \cdot s_i = \sum_{i=1}^m \mu_i \cdot s_i + \sum_{i=1}^m \beta(0, \sigma_i) \cdot s_i \quad (3)$$

$$E(Y) = \sum_{i=1}^m \mu_i \cdot s_i \quad (4)$$

$$\text{Var}(Y) = \sum_{i=1}^m \sigma_i^2 \cdot s_i^2 \quad (5)$$

Under the condition that the expected value $E(Y)$ should correspond to the mean of the experimentally measured spectra $\bar{b}$, we can replace the sum of the scaled vectors $\sum_{i=1}^m \mu_i \cdot s_i$ in eqs 3 and 4 by $\bar{b}$. Additionally, by comparing the variance of eq 5 to the sample variance $\frac{1}{m} \sum_{i=1}^m (b_i - \bar{b})^2$, we set $\sigma_i = \frac{1}{\sqrt{m}}$ and $s_i = b_i - \bar{b}$. With this, the expected value and variance of the generated spectra become equal to the experimentally determined values.

Since the considered number of measured spectra m may be smaller than the number of different molecular species M within the sample, this alternative representation may not be able to reproduce the complete molecular complexity. However, considering that we can often use hundreds of measurements for modeling, this approach should be able to capture variations related to a similar number of molecular species and thus asymptotically approximate most of the biological variability.

Incorporating Measurement Errors. In addition to modeling the biological properties, it is important to factor in the contributions of noise introduced by the measurement device to assess its influence on the application of interest.

When using a commercial Fourier-transform infrared (FTIR) spectrometer, the dominant noise type inherent to the spectral measurement itself is additive white noise. Such noise can be added to a modeled spectrum as a vector ϵ :

$$Y = \bar{b} + \sum_{i=1}^m \left(\beta\left(0, \frac{1}{\sqrt{m}}\right) \cdot (b_i - \bar{b}) \right) + \epsilon \quad (6)$$

To determine a realistic estimate for ϵ , we repeatedly measured water samples with the FTIR device used for the blood sera measurements and calculated the standard deviation observed across the spectral features (Figure S1 in the Supporting Information). The measurement noise coefficient ϵ was thus calibrated to be a random Gaussian vector with a mean 0 and a spectrally dependent standard deviation.

Other sources of error in infrared spectroscopy which result from baseline drift³ or sample delivery and preparation^{10,14} can

also be considered by introducing multiplicative noise and drift vectors in the model.²⁸ When transferring this model to Raman-based fingerprinting applications, it may be necessary to consider further, partly nonlinear noise sources³⁴ to obtain realistic results. As we show below, for the application of the model presented in this work, it is sufficient to consider only additive white noise to obtain realistic results.

Incorporating Differences between Molecular States.

Changes in the physiology of an organism, e.g., due to the onset of a disease, may change the molecular composition of the analyzed sample. Thus, we can choose a new set of experimentally measured spectra b^* that characterize the new molecular state Y^* and use that as a basis for calibration. As later shown, for certain applications, the discriminant features can be largely explained by a single vector d . By utilizing an additional Gaussian variable $\delta(\mu_d, \sigma_d)$ which scales the introduced vector d and accounts for its variability, a modeled spectrum of an alternative class can be reduced to:

$$Y^* = Y + \delta(\mu_d, \sigma_d) \cdot d \quad (7)$$

With this modeling platform, an arbitrary number of samples can be generated for different molecular states, resulting in a generated cohort that reflects the statistical properties of the measured sample pool $b_1, \dots, b_m$ and modeled measurement noise $\epsilon(0, \sigma_{H_2O})$. The statistical quantities β and ϵ can then be scaled by additional factors to simulate different levels of variability for these coefficients in a created cohort of n measurements. In our application, the molecular outcome Y represents an artificial measurement of a control sample and Y^* represents an artificial measurement of a cancer case sample. Unless explicitly stated otherwise, the descriptive approach is used where independent experimental measurements of cases and controls form the basis of biological variability calibrations.

Experimental FTIR Spectra and Machine Learning Analysis. The experimental FTIR spectra used in this study largely overlap with spectra obtained from a previous study that involved the detection of lung and prostate cancer from blood serum.⁷ Newly measured blood sera samples from different individuals were included in this study to increase the sample size and thus improve the statistical robustness of the results. All participants provided written informed consent for the study under research study protocol #17-141 and under research study protocol #17-182, both of which were approved by the Ethics Committee of the Ludwig-Maximilian-University (LMU) of Munich. Our study complies with all relevant ethical regulations and was conducted according to Good Clinical Practice (ICH-GCP) and the principles of the Declaration of Helsinki. The clinical trial is registered (ID DRKS00013217) at the German Clinical Trials Register (DRKS).

In total, we considered spectra from 523 lung cancer patients, 411 prostate cancer patients, and nonsymptomatic references pair-matched to each individual from each cancer entity. A detailed breakdown of the cohorts is provided in Table S1 in the Supporting Information. Information regarding the study design, statistical matching, sample collection, sample handling, FTIR measurements, and spectral preprocessing is detailed in the previous study.⁷ Machine learning analysis performed on the experimental and simulated data is detailed in Section 4 in the Supporting Information. In essence, an L2-regularized logistic regression was used as a classification algorithm. Classifications were assessed using the receiver operating characteristic (ROC)

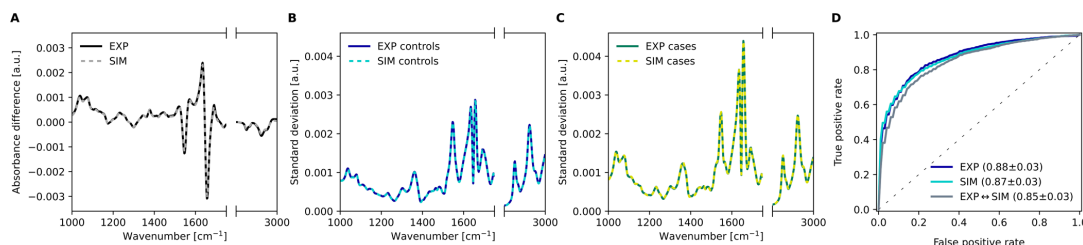


Figure 2. Validation of the model calibration procedure for the lung cancer detection case study. Spectral cohorts of individuals were simulated to model the properties of measured experimental cohorts using the same sample sizes ($n = 523$ cases vs 523 controls). (A) Differential fingerprint, defined as the difference between the mean of the case and control spectral features, for the experimental cohort (black) and the simulated cohorts (gray) as averaged across 10 simulation repetitions. (B, C) Standard deviation of the control spectral features (B) and case spectral features (C) for the experimental cohort (blue) and simulated cohorts (cyan) as averaged across 10 simulation repetitions. (D) ROC curve for binary case–control classifications when training and testing on experimental samples (blue), training and testing on simulated samples (cyan), and training on simulated samples and testing on experimental samples (gray). The AUCs are listed in the figure legend along with their standard deviations across the cross-validation splits. Figure S2 in the Supporting Information shows consistent results with the prostate cancer application.

curve, and the area under the curve (AUC) was used as a summary metric of predictive performance.

RESULTS

By applying our simulation model to a clinically relevant scenario, we studied the potential and limitations of detecting lung and prostate cancer in binary case–control settings. Using experimentally measured FTIR spectra, we validated that comparable results were obtained between artificially generated cohorts and experimental cohorts. Next, we investigated how classification performance was influenced by varying model parameters that control the cohort size, measurement noise, biological variability, and molecular complexity. Our findings from the lung cancer application are provided within the main text and figures. Figures relating to the prostate cancer conveyed consistent results with the lung cancer application and are thus provided in Sections 6–9 in the Supporting Information.

Validation of Simulated Data. After performing the calibration procedure described in the methods, we validated that the simulation model was able to create cohorts that effectively captured the intrinsic properties of experimental cohorts for our lung cancer case study (Figure 2A–D). We simulated 10 data sets of case and control samples with the same sample counts as our experimental cohorts. Each simulated cohort was created with a different set of generated random numbers (i.e., a different random seed), thereby, minimizing the effects of random perturbations. Across the 10 created data sets, we calculated the difference between the mean spectrum of case and control samples and their standard deviations across the spectrum. The resulting statistical properties were then averaged for the 10 simulated data sets and compared to the same properties of the experimental cohort.

In this comparison, we observed a high degree of agreement in the statistical features of the spectra between simulated and experimental data (Figure 2A–C). The differential fingerprint retained its structure in simulation (Figure 2A), and the standard deviations around the mean spectrum were well modeled for case and control samples (Figure 2B,C). Since our simulated cohorts were calibrated to follow the experimental cohort, it was unsurprising to find such a high degree of similarity in these properties.

Relating to our aim of studying the influence of noise factors on predictive modeling, we compared the classification performance of detecting cancer cases from controls for both

experimental and simulated data (Figure 2D). An L2-regularized logistic regression classifier was used for the predictive modeling to help ensure classifier robustness against noise (Sections 4 and 11 in the Supporting Information for more details). For the experimental data, model performance was estimated in a 10-times repeated 10-fold cross-validation—totaling 100 folds. For the simulated data sets, a 10-fold cross-validation was carried out on each of the 10 simulated data sets—thereby, also totaling 100 folds. With this pipeline, we estimated mean ROC-AUCs of 0.88 and 0.87 for the lung cancer classification with the experimental data and the simulated data, respectively (Figure 2D, blue and cyan curves). The ROC-AUC values that we observed fall within the standard deviation across the cross-validation folds, validating that the degree of class separation was well modeled for our simulated data following the descriptive calibration procedure.

As an additional validation, we compared the predictive performance of a model trained exclusively on simulated data and tested on experimental data. To ensure that no experimental data used for testing contributed to the model training, we split the experimental data into 10 folds. For each fold, we held out an experimental test set, fit a predictive model on simulated samples calibrated on the remaining experimental samples, and tested the predictive model on the held-out set. As previously described, we created simulated data sets of the same cohort sizes as our experimental data and repeated the simulation 10 times, resulting in a total of 100 ROC-AUC estimates. With this approach, we still achieved a similar performance to what we expected from the experimental data estimates, with a mean ROC-AUC of 0.85 (Figure 2D, gray curve). When the machine learning model was trained on even larger sets of simulated data with $n = 100,000$ of balanced cases and controls, the experimentally obtained ROC-AUC values can be fully recovered (Figure S6 in the Supporting Information). Taken together, these results suggest that the proposed descriptive modeling approach is able to capture all relevant properties and features for the classification of the studied system.

We repeated the above numerical experiment but with a more simplistic model for generating the cancer cohort. Instead of modeling cases based on experimental measurements, we introduced a single discriminant feature vector $\mathbf{d}$ to describe them (eq 7 based on the experimentally obtained differential fingerprints depicted in Figure 2A). Surprisingly, the model was still able to reproduce the experimental results to a large extent

(Figure S8 in the Supporting Information). Although the approach based on modeling the case measurements using experimental case spectra (Figure 2A–D) better captured the properties of the spectra, the more simplistic approach indicated that most of the relevant changes between the case and control state could be explained by a single vector.

Influence of Cohort Size. The simulation approach allows for generating artificial cohorts in arbitrary counts, which enabled us to carry out an investigation on the influence of cohort size on predictive performance. We used the previously calibrated parameters of the simulation model and tested out different cohort sizes (Figure 3). For each cohort size tested, 100

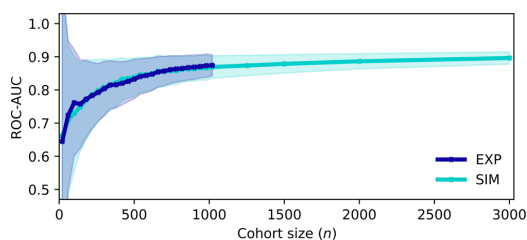


Figure 3. Effect of cohort size on lung cancer detection. Spectral cohorts consisting of balanced cases and controls were simulated at changing sample counts. The cross-validated ROC-AUC is plotted against the cohort size for the simulated samples (cyan). For comparison, sets of experimental samples were randomly selected and cross-validated upon to model the effects of changing cohort sizes with experimental observations (blue). The solid curves depict the average scores along with their standard deviations in the shaded region. Figure S3 in the Supporting Information shows consistent results with the prostate cancer application.

artificial cohorts were generated and the ROC-AUC was again determined in 10-fold cross-validation on each cohort size. In

addition, a similar investigation was carried out using the experimental observations, where samples were randomly selected in 100 different iterations, selecting a different set of measurements in each iteration for each cohort size, and cross-validated upon (10-fold).

We observed a similar dependence between the cohort size and the performance of detecting lung cancer for simulated and experimental data (Figure 3). This investigation revealed that increasing the size of the simulated data sets beyond the number of experimentally available data ($n = 1046$) only marginally improved the classification performance and reduced the relative standard deviation as the plateau of asymptotic performance was reached. It was apparent, however, that the classification performance significantly suffered when smaller cohorts were used (e.g., $n \leq 200$) compared to using larger cohorts (e.g., with $n \geq 1000$). Moreover, the standard deviation became intolerably large with smaller sample sizes, making it difficult to construct conclusive estimates when only limited cohort sizes are available.

This result showed that the cohort sizes used were sufficiently large, which is reflected by the well-known fact that the predictive performance of a machine learning classifier reaches a plateau above a certain cohort size.^{35,36} Furthermore, it served as an additional validation that the simulation model was able to reproduce the results obtained from experimental observations.

Influence of Biological Variability and Measurement Noise. After demonstrating that the calibration procedure generated artificial cohorts that captured the properties of our experimental results, we examined the potential effects of measurement noise and biological variability on the performance of detecting lung cancer (Figure 4). When describing the simulation model, we considered the random coefficients β and ϵ , modeling the biological variability and measurement noise in levels similar to that of the experimental observations. The standard deviations of these two coefficients are tunable, thus enabling the scaling of these levels noise factors seen within the

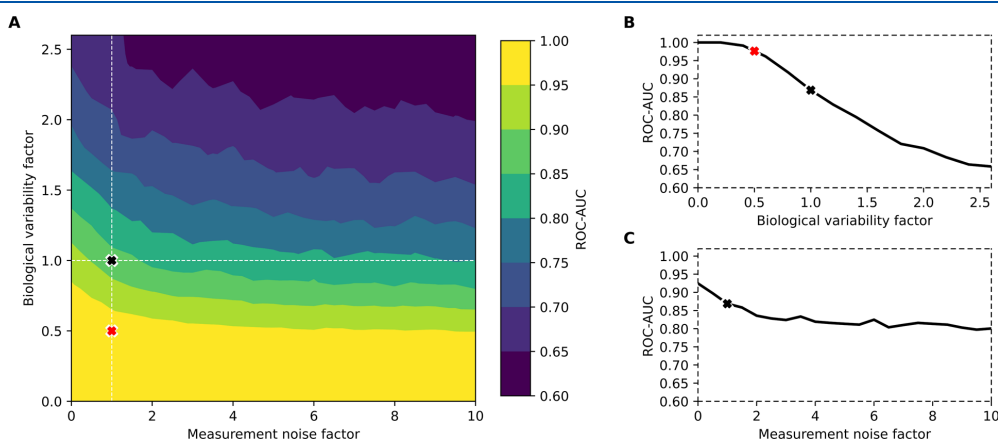


Figure 4. Effect of biological variability and measurement noise on lung cancer detection. (A) Classification performance was measured by the mean cross-validated ROC-AUC on multiple spectral cohorts simulated at changing levels of biological variability and measurement noise. The black “x” marks the point of calibration where the simulated cohorts model the experimental levels of between-person biological variability and measurement noise. The red “x” marks an estimate for reduced biological variability in a longitudinal, self-referencing, scenario as estimated from previous work.¹⁴ (B) Slice of (A) when the measurement noise factor is held constant at the calibrated level and the biological variability factor is changing. (C) Slice of (A) when the biological variability factor is held constant at the calibrated level and the measurement noise factor is changing. Figure S4 in the Supporting Information shows consistent results with the prostate cancer application.

spectra and thereby simulating different experimental conditions. We generated artificial cohorts considering varying levels of these noise factors and examined their effects on classification performance. For each level of measurement noise and biological variability considered, we simulated 10 cohorts and cross-validated (10-fold) on each simulated cohort—averaging the ROC-AUC scores across the 100 total test data splits.

Overall, we found that the dominating source of noise, which masked away the signal separating the classes, was the level of biological variability (Figure 4A). With no added biological variability, the binary classifiers were able to distinguish the clear signal of the diseases, achieving near-perfect to perfect ROC-AUCs within the explored levels of measurement noise. With increasing levels of biological noise, the disease signals were increasingly masked away, leading to sharp declines in ROC-AUCs down to the plateau of observing random chance models. Our point of calibration fell within a region where minor changes to the biological variability have significant effects on the performance of distinguishing cases from controls (Figure 4B).

Previously, it was found that the level of within-person biological variability in molecular fingerprints over a period of 6 months was near a factor of 2 less, on average, than the level of between-person variability of individuals in groups.¹⁴ This allowed us to estimate the potential classification performance of detecting cancer in a personalized, longitudinal health-monitoring scenario. By tuning our random coefficient β to model the level of within-person variability (i.e., $\beta \times 0.5$), we estimated that such a classification would yield a ROC-AUC of 0.98 for lung cancer detection (Figure 4A,B).

Compared to the effect of biological variability, varying the level of measurement noise was found to have a substantially smaller effect on the classification performance (Figure 4A,C). Eliminating the modeled measurement noise (i.e., $\epsilon \times 0$) yielded an estimated ROC-AUC of 0.92. Nevertheless, varying both the biological variability and measurement noise simultaneously revealed that the effect of the measurement noise on the retrieved ROC-AUC also depended on the level of biological variability. For instance, when the level of biological variability was highest, reducing the measurement noise resulted in substantial classification performance gains (in terms of Δ ROC-AUC). However, when the biological variability was lowest, measurement noise had little effect on the retrieved ROC-AUC—with the classification performing with perfect efficiency. Thus, advances in spectroscopic methods which result in further reductions of measurement noise may be particularly promising for problems with low classification efficiencies (e.g., due to weak disease signatures).

Influence of Molecular Complexity. Within the proposed simulation model, the biological variability of the molecularly complex blood-based spectra was modeled according to two aspects: the total strength of biological variability, determined by the statistical variable β , and the structure of the underlying biological variability as determined by the set of m unique and independent calibration vectors b_i of a given sample pool. Each vector represents the absorbance spectrum combining all molecular constituents in a given molecular sample, combining all their respective concentrations. As previously described, changing the number of independent vectors considered when creating an artificial sample would mathematically correspond to changing the number of unique molecular species creating the full blood-serum spectrum.

To model this, we kept the mean measurement $\bar{b}$ for each sample pool constant and randomly selected m different experimental calibration vectors b_i to create artificial cohorts at varying values of m (Figure 5A–C). Since $\bar{b}$ is held constant,

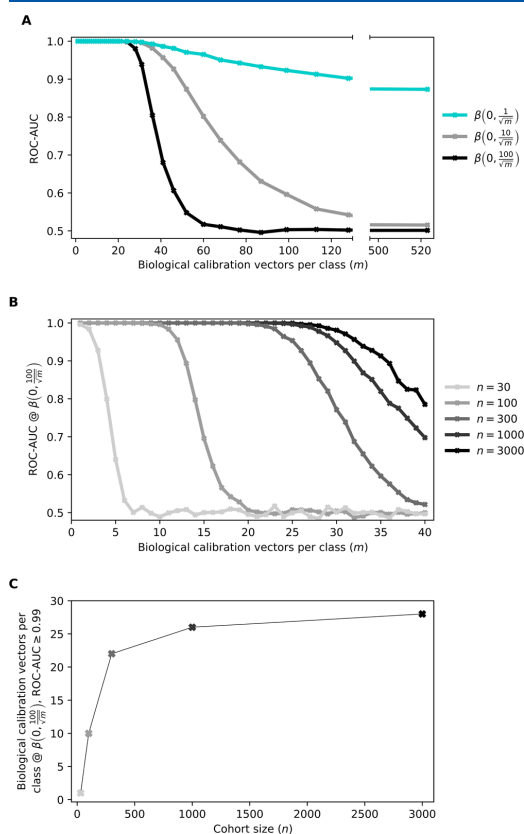


Figure 5. Influence of molecular complexity on lung cancer detection. (A) Multiple spectral cohorts were simulated at changing levels of m (i.e., the number of calibration vectors defining each molecular state) and changing levels of biological variability (β). Cohorts were created with the same cohort size as the experimental measurements and cross-validated upon to estimate the ROC-AUC as a measure of prediction performance. (B) Multiple data sets were simulated at changing levels of m and changing cohort sizes (n) to investigate the effects of the sample size on the trends observed in (A) at the highest level of biological variability investigated. (C) Number of calibration vectors m required to achieve a ROC-AUC ≥ 0.99 at the cohort sizes investigated in (B). Figure S5 in the Supporting Information shows consistent results with the prostate cancer application.

we assume that altering the molecular complexity has no direct effect on the disease signal, but in fact, increasing complexity serves to mask it away. To note as well, the case and control cohort spectra are each based on a separate set of vectors, and thus, the entire set of case and control spectra is based on $2 \times m$ independent vectors.

In the first numerical experiment, we kept the relative strength of biological variability at the same level as previously calibrated, utilizing the random coefficient $\beta(0, \frac{1}{\sqrt{m}})$. The measurement

noise coefficient ϵ was also kept at the calibrated level when creating the simulated samples. Cohorts for lung cancer were created with the same sample sizes as our experimental measurements. This process was repeated 100 times, each with a different set of m randomly selected experimental measurements, and cross-validated (10-fold) on each simulated data set. We observed that when the molecular complexity was modeled with $m \leq 20$ independent biological calibration vectors for each sample pool, perfect class separation can be achieved with ROC-AUCs of 1.0 (Figure 5A, cyan curve). Increasing the molecular complexity of the spectra past this, using calibration vectors that did not contribute to the cancer disease signal, added a level of molecular complexity that reduced the prediction confidence of the classification models.

This observation prompted the question of whether there might be a more general relationship between molecular complexity, biological variability, and the ability of infrared fingerprinting to distinguish between two classes. Therefore, we repeated the previous experiment, but with larger levels of variability for β , representing the strength of the biological variability. We found that perfect separation was possible for $m \leq 20$, irrespective of the level of biological variability in our explored domain (Figure 5A, gray and black curves). From a mathematical point of view, this suggests that at such a threshold, the underlying system of equations to be solved is over-determined and thus has a unique solution that a predictive model is able to find.

Further investigations revealed that the threshold value for m , for which perfect separation is still possible, can be increased with increasing cohort size but only at an exponentially decreasing rate (Figure 5B,C). In other words, more individual components could be detected at a higher level of sensitivity by measuring exponentially more samples. However, this threshold certainly depends on the correlation of the calibration vectors, the spectral bandwidth, and resolution, and can vary depending on the application.

We would like to note that the threshold values obtained here can only be transferable under idealistic conditions. In particular, the application of interest must not be tainted by any measurement noise, and the signal of interest, as well as all other molecules, must be of similar concentration. Nevertheless, these results clearly reflect that the capacity of infrared molecular fingerprinting of complex samples can be drastically increased—if the number of components of the analyzed matrix in which the spectral marker is embedded is significantly reduced below 100.

DISCUSSION

Capturing biological phenotypes through molecular fingerprinting forms the basis for many innovative applications, such as disease detection. Nevertheless, the full potential and limitations of these approaches are not fully understood due to the complex interplay of measurement noise, biological variability, and the actual distinctive molecular pattern.

To investigate these effects, we systematically adjusted the parameters of our validated *in silico* model. We found that improved classification was achievable by decreasing the levels of biological variability and measurement noise. Reducing measurement noise promises to improve classification efficiency by a few percentage points (in terms of ROC-AUC scores) and can potentially be exploited with next-generation infrared spectrometers.³⁷ A greater advantage came from reducing biological variability, which is more challenging to realize as it is inherent to any biological setting—from a single gene³⁸ to an

organismal level.^{12,13} A possible strategy here is populational stratification, or grouping individuals into defined strata (e.g., by age, body mass index, and lifestyle factors), to reduce the biological variability within each stratum. Although widely established in genomics,³⁹ it needs to be carefully controlled to not lead to spurious associations. Furthermore, the concept of longitudinal self-referencing could be incorporated in fingerprinting for disease diagnostics and health monitoring.⁴⁰ The within-person biological variability of blood-based infrared fingerprints is near a factor of 2 lower than the between-person variability when monitored over a span of 6 months.¹⁴ Using our model, we estimated that this would improve the ROC-AUC from 0.88 for group comparison to 0.98 for longitudinal self-referencing when detecting lung cancer.

Further investigations showed that identifying spectral patterns using infrared molecular fingerprinting may be fundamentally limited by the molecular complexity of the analyzed samples. By reducing the number of different molecular species considered for generating the cohorts, perfect class separation was achievable when less than 20 unique spectra were used for each sample pool—regardless of the level of biological variability. By increasing the cohort size, near-perfect separation could be achieved even with increasing molecular complexity—albeit, this requires exponentially larger cohort sizes. This indicated that infrared fingerprinting of complex biological samples (e.g., blood-based media) is only able to unambiguously determine a certain number of independent variables, or molecular species, within complex samples. The actual number of detectable independent variables certainly depends on measurement properties such as the spectral bandwidth, spectral range, and the type of sample under investigation. Such questions can be investigated with the presented model but would exceed the scope of this study. Considering our applications using blood-serum, we found that reducing the chemical sample complexity (e.g., via chromatography) could be a viable option to improve the classification performance.

Critically seen, we initially developed our approach to calculate realistic fingerprint spectra of blood serum by considering the individual spectral contributions of the majority of molecular species contained therein. Since this requires information on individual molecular concentrations in blood for different groups of the human population and also the use of a large number of individual component spectra, with the majority unavailable, we could not realize this approach. Instead, we chose a descriptive approach that relies on measured spectra of the molecularly complex samples from observational studies. We have, however, shown that both approaches follow a similar mathematical formulation. To assess whether they have similar explanatory power, a direct comparison is necessary and which is already ongoing. At the same time, we believe that the bottom-up approach can facilitate the analysis of simpler systems, such as pharmaceutical samples,³³ to investigate the effects of factors like molecular mixing and variance of molecular concentrations introduced through manufacturing.

A disadvantage of the descriptive approach is that the best agreement with experimental results is only achieved when healthy and diseased sample pools were calibrated separately. Choosing a simplified approach that describes the occurrence of a disease by a single discriminant vector, but otherwise only considers the biological variability of the healthy cohort, reduces the agreement with the experimental results (Figure S8 in the Supporting Information). Nevertheless, this approach has its

advantages. It is well suited for exploratory analysis of any fingerprinting settings and could help estimate to what extent a change in a certain molecule (e.g., a studied biomarker) would be detectable by fingerprinting. For such investigations, one would only need to record the spectra of the target molecule independently and then use the calibrated data from this work to generate simulated spectra with and without contributions from the target spectra.

In general, the core idea of the proposed *in silico* model can be transferred to other molecular fingerprinting techniques, such as Raman spectroscopy²⁰ or nuclear magnetic resonance spectroscopy,²¹ and to other applications involving multiclass classifications. Furthermore, the model can be applied to study the proper applications of machine learning algorithms—from investigating the effects of noise on hyperparameter tuning (Section 11 in the Supporting Information) to how different classification algorithms perform under different simulated conditions. To facilitate the use of the model, we provide a Python toolbox alongside executable scripts to reproduce and extend the results of this study (Sections 2 and 3 in the Supporting Information).

CONCLUSIONS

In this study, we describe an *in silico* approach capable of modeling molecular fingerprints of complex biological systems. We calibrated our model using experimentally measured infrared spectra of blood sera and applied it to the detection of lung and prostate cancer. Excellent agreement between the statistical properties and classification results of simulated and experimental data was achieved for both applications. The validity of the approach was further supported by the finding that machine learning models trained on simulated data and tested on experimentally measured data delivered comparable results. We demonstrated that the newly developed model enables investigations of the potential and limitations of the molecular fingerprinting framework and that it can serve as a platform for exposing opportunities to advance molecular fingerprinting applications.

ASSOCIATED CONTENT

Data Availability Statement

Datasets and Python scripts: <https://github.com/tarek-eissa/in-silico-spectral-generator>.

Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.analchem.2c04711>.

Experimental cohort description, data and code availability, additional methodological details, and supplementary figures (PDF)

Datasets and Python scripts (ZIP)

AUTHOR INFORMATION

Corresponding Authors

Mihaela Zigman — Department of Laser Physics, Ludwig Maximilian University of Munich (LMU), 85748 Garching, Germany; Laboratory for Attosecond Physics, Max Planck Institute of Quantum Optics (MPQ), 85748 Garching, Germany; orcid.org/0000-0001-8306-1922; Email: mihaela.zigman@mpq.mpg.de

Marinus Huber — Department of Laser Physics, Ludwig Maximilian University of Munich (LMU), 85748 Garching, Germany; Laboratory for Attosecond Physics, Max Planck

Institute of Quantum Optics (MPQ), 85748 Garching, Germany; orcid.org/0000-0001-5309-4475; Email: marinus.huber@mpq.mpg.de

Authors

Tarek Eissa — Department of Laser Physics, Ludwig Maximilian University of Munich (LMU), 85748 Garching, Germany; Department of Informatics, Technical University of Munich (TUM), 85748 Garching, Germany; orcid.org/0000-0002-8932-2553

Kosmas V. Kepesidis — Department of Laser Physics, Ludwig Maximilian University of Munich (LMU), 85748 Garching, Germany; Laboratory for Attosecond Physics, Max Planck Institute of Quantum Optics (MPQ), 85748 Garching, Germany; orcid.org/0000-0002-6391-7743

Complete contact information is available at: <https://pubs.acs.org/10.1021/acs.analchem.2c04711>

Author Contributions

M.H. proposed the *in silico* model; T.E. performed the research and analyzed the data with assistance from K.V.K., M.H., and M.Z.; T.E. and M.H. wrote the manuscript with assistance from K.V.K. and M.Z.

Funding

Open access funded by Max Planck Society.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We would like to thank Ferenc Krausz for providing the research environment that made this work possible. We also would like to thank Michael Trubetskov, Jacqueline Hermann, and Ferenc Krausz for the fruitful discussions and their constructive feedback on this work.

REFERENCES

- (1) Butler, H. J.; Ashton, L.; Bird, B.; Cinque, G.; Curtis, K.; Dorney, J.; Esmonde-White, K.; Fullwood, N. J.; Gardner, B.; Martin-Hirsch, P. L.; Walsh, M. J.; McAinsh, M. R.; Stone, N.; Martin, F. L. *Nat. Protoc.* **2016**, *11*, 664–687.
- (2) Baker, M. J.; Trevisan, J.; Bassan, P.; Bhargava, R.; Butler, H. J.; Dorling, K. M.; Fielden, P. R.; Fogarty, S. W.; Fullwood, N. J.; Heys, K. A.; Hughes, C.; Lasch, P.; Martin-Hirsch, P. L.; Obinaju, B.; Sockalingum, G. D.; Sulé-Suso, J.; Strong, R. J.; Walsh, M. J.; Wood, B. R.; Gardner, P.; Martin, F. L. *Nat. Protoc.* **2014**, *9*, 1771–1791.
- (3) Griffiths, P. R.; de Haseth, J. A. *Fourier Transform Infrared Spectrometry*; John Wiley & Sons, 2007; 171.
- (4) Paraskevaidi, M.; Matthew, B. J.; Holly, B. J.; Hugh, B. J.; Thulya, C. P. V.; Loren, C.; StJohn, C.; Peter, G.; Callum, G.; Sergei, K. G.; Kamila, K.; Maria, K.; Kássio, L. M. G.; Pierre, M. H. L.; Evangelos, P.; Savithri, P.; John, A. A.; Alexandra, S.; Marfran, S.; Josep, S. S.; Gunjan, T.; Michael, W.; Bayden, W. *Appl. Spectrosc. Rev.* **2021**, *56*, 804–868.
- (5) Martin, F. L.; Kelly, J. G.; Llabjani, V.; Martin-Hirsch, P. L.; Patel, I. I.; Trevisan, J.; Fullwood, N. J.; Walsh, M. J. *Nat. Protoc.* **2010**, *5*, 1748–1760.
- (6) Lee, K. S.; Landry, Z.; Pereira, F. C.; Wagner, M.; Berry, D.; Huang, W. E.; Taylor, G. T.; Kneipp, J.; Popp, J.; Zhang, M.; Cheng, J.-X.; Stocker, R. *Nat. Rev. Methods Primers* **2021**, *1*, 80.
- (7) Huber, M.; Kepesidis, K.; Voronina, L.; Fleischmann, F.; Fill, E.; Hermann, J.; Koch, I.; Milger-Kneidinger, K.; Kolben, T.; Schulz, G. B.; Jokisch, F.; Behr, J.; Harbeck, N.; Reiser, M.; Stief, C.; Krausz, F.; Zigman, M. *Elife* **2021**, *10*, No. e68758.
- (8) Diem, M. J. *Biophotonics* **2018**, *11*, No. e201800064.
- (9) Guo, S.; Popp, J.; Bocklitz, T. *Nat. Protoc.* **2021**, *16*, 5426–5459.

- (10) Kwak, J. T.; Reddy, R.; Sinha, S.; Bhargava, R. *Anal. Chem.* **2012**, *84*, 1063–1069.
- (11) Bhargava, R. *Anal. Bioanal. Chem.* **2007**, *389*, 1155–1169.
- (12) Eling, N.; Morgan, M. D.; Marioni, J. C. *Nat. Rev. Genet.* **2019**, *20*, 536–548.
- (13) Eling, N.; Morgan, M. D.; Marioni, J. C. *Nat. Rev. Genet.* **2019**, *20*, 562.
- (14) Huber, M.; Kepesidis, K.; Voronina, L.; Božić, M.; Trubetskoy, M.; Harbeck, N.; Krausz, F.; Žigman, M. *Nat. Commun.* **2021**, *12*, 1511.
- (15) Guo, S.; Beleites, C.; Neugebauer, U.; Abalde-Cela, S.; Afseth, N. K.; Alsamad, F.; Anand, S.; Araujo-Andrade, C.; Aškrić, S.; Avci, E.; Baia, M.; Baranska, M.; Baria, E.; de Carvalho, L. A. E. B.; de Bettignies, P.; Bonifacio, A.; Bonnier, F.; Brauchle, E. M.; Byrne, H. J.; Chourpa, I.; Cicchi, R.; Cuisinier, F.; Culha, M.; Dahms, M.; David, C.; Duponchel, L.; Duraipandian, S.; El-Mashtoly, S. F.; Ellis, D. I.; Eppe, G.; Falgayrac, G.; Gamulin, O.; Gardner, B.; Gardner, P.; Gerwert, K.; Giamarellos-Bourboulis, E. J.; Gizurarson, S.; Gnyba, M.; Goodacre, R.; Grysan, P.; Guntinas-Lichius, O.; Helgadottir, H.; Grošev, V. M.; Kendall, C.; Kiselev, R.; Kolbach, M.; Krafft, C.; Krishnamoorthy, S.; Kubryck, P.; Lendl, B.; Loza-Alvarez, P.; Lyng, F. M.; Machill, S.; Malherbe, C.; Marro, M.; Marques, M. P. M.; Matuszyk, E.; Morasso, C. F.; Moreau, M.; Muhamadali, H.; Mussi, V.; Nottingher, I.; Pacia, M. Z.; Pavone, F. S.; Penel, G.; Petersen, D.; Piot, O.; Rau, J. V.; Richter, M.; Rybarczyk, M. K.; Salehi, H.; Schenke-Layland, K.; Schlücker, S.; Schosserer, M.; Schütze, K.; Sergio, V.; Sinjab, F.; Smulko, J.; Sockalingum, G. D.; Stiebing, C.; Stone, N.; Untereiner, V.; Vanna, R.; Wieland, K.; Popp, J.; Bocklitz, T. *Anal. Chem.* **2020**, *92*, 15745–15756.
- (16) Netz, R. R.; Eaton, W. A. *Proc. Natl. Acad. Sci. U. S. A.* **2021**, *118*, No. e2022753118.
- (17) Segatta, F.; Russo, M.; Nascimento, D. R.; Presti, D.; Rigodanza, F.; Nenov, A.; Bonvicini, A.; Arcioni, A.; Mukamel, S.; Maiuri, M.; Muccioli, L.; Govind, N.; Cerullo, G.; Garavelli, M. *J. Chem. Theory Comput.* **2021**, *17*, 7134–7145.
- (18) Székely, T., Jr.; Burrage, K. *Comput. Struct. Biotechnol. J.* **2014**, *12*, 14–25.
- (19) Eissa, T. In Silico Spectral Generator. *GitHub repository*; GitHub, 2022, <https://github.com/tarek-eissa/in-silico-spectral-generator>.
- (20) Traynor, D.; Behl, I.; O'Dea, D.; Bonnier, F.; Nicholson, S.; O'Connell, F.; Maguire, A.; Flint, S.; Galvin, S.; Healy, C. M.; Martin, C. M.; O'Leary, J. J.; Malkin, A.; Byrne, H. J.; Lyng, F. M. *Nat. Protoc.* **2021**, *16*, 3716–3735.
- (21) Wallner-Liebmann, S.; Tenori, L.; Mazzoleni, A.; Dieber-Rotheneder, M.; Konrad, M.; Hofmann, P.; Luchinat, C.; Turano, P.; Zatloukal, K. *J. Proteome Res.* **2016**, *15*, 1787–1793.
- (22) Voronina, L.; Leonardo, C.; Mueller-Reif, J. B.; Geyer, P. E.; Huber, M.; Trubetskoy, M.; Kepesidis, K. V.; Behr, J.; Mann, M.; Krausz, F.; Žigman, M. *Angew. Chem., Int. Ed.* **2021**, *60*, 17060–17069.
- (23) Anderson, N. L.; Anderson, N. G. *Mol. Cell. Proteomics* **2002**, *1*, 845–867.
- (24) Marshall, W. J.; Lapsley, M.; Day, A.; Ayling, R. *Clinical Biochemistry E-Book: Metabolic and Clinical Aspects*; Elsevier Health Sciences, 2014.
- (25) Wu, S.; Chen, D.; Snyder, M. P. *Curr. Opin. Chem. Biol.* **2022**, *66*, No. 102101.
- (26) Popova, A. V.; Hinch, D. K. *Biophys. J.* **2003**, *85*, 1682–1690.
- (27) Mayerhöfer, T. G.; Ilchenko, O.; Kutsyk, A.; Popp, J. *Appl. Spectrosc.* **2022**, *76*, 92–104.
- (28) Solheim, J. H.; Zimmermann, B.; Tafintseva, V.; Dzurendová, S.; Shapaval, V.; Kohler, A. *Molecules* **2022**, *27*, 1900.
- (29) Isensee, K.; Kröger-Lui, N.; Petrich, W. *Analyst* **2018**, *143*, 5888–5911.
- (30) Schwaighofer, A.; Brandstetter, M.; Lendl, B. *Chem. Soc. Rev.* **2017**, *46*, 5903–5924.
- (31) Perez-Guaita, D.; Sanchez-Illana, A.; Ventura-Gayete, J.; Garrigues, S.; de la Guardia, M. *Analyst* **2014**, *139*, 170–178.
- (32) Kobayashi-Kirschvink, K. J.; Nakaoka, H.; Oda, A.; Kamei, K. F.; Noshio, K.; Fukushima, H.; Kanesaki, Y.; Yajima, S.; Masaki, H.; Ohta, K.; Wakamoto, Y. *Cell Syst.* **2018**, *7*, 104–117.e4.
- (33) Eerdenbrugh, B.; van Taylor, L. S. *Int. J. Pharm.* **2011**, *417*, 3–16.
- (34) Bocklitz, T.; Walter, A.; Hartmann, K.; Rösch, P.; Popp, J. *Anal. Chim. Acta* **2011**, *704*, 47–56.
- (35) Beleites, C.; Neugebauer, U.; Bocklitz, T.; Krafft, C.; Popp, J. *Anal. Chim. Acta* **2013**, *760*, 25–33.
- (36) Cox, J. *Nat. Biotechnol.* **2023**, *41*, 33–43.
- (37) Pupeza, I.; Huber, M.; Trubetskoy, M.; Schweinberger, W.; Hussain, S. A.; Hofer, C.; Fritsch, K.; Poetzlberger, M.; Vámos, L.; Fill, E.; Amotchkina, T.; Kepesidis, K. V.; Apolonski, A.; Karpowicz, N.; Pervak, V.; Pronin, O.; Fleischmann, F.; Azzeer, A.; Žigman, M.; Krausz, F. *Nature* **2020**, *577*, 52–59.
- (38) Ozbudak, E. M.; Thattai, M.; Kurtser, I.; Grossman, A. D.; van Oudenaarden, A. *Nat. Genet.* **2002**, *31*, 69–73.
- (39) Price, A. L.; Zaitlen, N. A.; Reich, D.; Patterson, N. *Nat. Rev. Genet.* **2010**, *11*, 459–463.
- (40) Rose, S. M. S.-F.; Contrepois, K.; Moneghetti, K. J.; Zhou, W.; Mishra, T.; Mataraso, S.; Dagan-Rosenfeld, O.; Ganz, A. B.; Dunn, J.; Hornburg, D.; Rego, S.; Perelman, D.; Ahadi, S.; Sailani, M. R.; Zhou, Y.; Leopold, S. R.; Chen, J.; Ashland, M.; Christle, J. W.; Avina, M.; Limcaoco, P.; Ruiz, C.; Tan, M.; Butte, A. J.; Weinstock, G. M.; Slavich, G. M.; Sodergren, E.; McLaughlin, T. L.; Haddad, F.; Snyder, M. P. *Nat. Med.* **2019**, *25*, 792–804.

Supporting information for: “Limits and prospects of molecular fingerprinting for phenotyping biological systems revealed through *in silico* modeling”

Tarek Eissa^{1,3}, Kosmas V. Kepesidis^{1,2}, Mihaela Zigman^{1,2,*}, and Marinus Huber^{1,2,*}

¹Ludwig Maximilian University of Munich (LMU), Department of Laser Physics, 85748 Garching, Germany

²Max Planck Institute of Quantum Optics (MPQ), Laboratory for Attosecond Physics, 85748 Garching, Germany

³Technical University of Munich (TUM), Department of Informatics, 85748 Garching, Germany

* Correspondence: mihaela.zigman@mpq.mpg.de, marinus.huber@mpq.mpg.de

Contents

1	Description of experimental cohort	S-2
2	Analysis code.....	S-2
3	Data access.....	S-2
4	Machine learning analysis	S-2
5	Modeled FTIR measurement noise.....	S-3
6	Validation of simulated data for the prostate cancer application	S-3
7	Influence of cohort size for the prostate cancer application	S-4
8	Influence of biological variability and measurement noise for the prostate cancer application ...	S-4
9	Influence of molecular complexity for the prostate cancer application	S-5
10	Simulating larger cohorts for model validation	S-6
11	Influence of noise on predictive model tuning	S-7
12	Simplified <i>in silico</i> model using a single vector to model case measurements.....	S-8
	References.....	S-8

1 Description of experimental cohort

Table S1 provides sample size, age, sex, and body mass index (BMI) descriptions of the experimental lung and prostate cancer cohorts. These cohorts were used for both - calibrating the biological variability of the simulated samples and for validating the statistical properties of the simulated cohorts. Details about the study design, sample collection and handling, and measurements of the samples are described in a previously published work.¹

	Lung cancer cohort		Prostate cancer cohort	
	Cases	Controls	Cases	Controls
n	523	523	411	411
% male	54%	54%	100%	100%
Age (mean $\pm$ SD)	67.2 $\pm$ 9.7 years	64.2 $\pm$ 10.4 years	65.2 $\pm$ 9.1 years	59.2 $\pm$ 15.9 years
BMI (mean $\pm$ SD)	25.7 $\pm$ 4.9 kg/m ²	26.4 $\pm$ 5.0 kg/m ²	27.1 $\pm$ 4.0 kg/m ²	26.7 $\pm$ 4.9 kg/m ²

Table S1: Description of the experimental lung cancer and prostate cancer cohorts. For each entity, cancer cases were pair-matched to non-symptomatic controls based on sex, age, and BMI.

2 Analysis code

To facilitate the use of the model, the Python code used for generating the artificial spectra is deposited in a GitHub repository.² There, we further provide Python scripts (on the Jupyter Notebook framework) to perform and reproduce all simulations shown in the results of the main text of the article.

3 Data access

Making use of the proposed model, we provide a dataset of generated lung and prostate cancer cohorts of FTIR spectra which model the data used within this work ($n = 100000$ for each cancer entity) in a GitHub repository.² The dataset was created using our experimental blood serum measurements for each cancer entity and the matched non-symptomatic controls. The artificial cohorts were created by following the described procedure for calibrating the biological variability using experimental measurements and the measurement noise using repeated water measurements. The water measurements ($n = 400$) used for calibrating the measurement noise are also deposited, along with the generated artificial cohorts. This data can serve as an input for generating additional artificial cohorts for *in silico* investigations, as well as be employed as a basis for gaining insights when limited numbers of serum-based measurements are at hand.

4 Machine learning analysis

Classification performance was evaluated by calculating the area under the receiver operating characteristic curve (ROC-AUC) of trained L2-penalized logistic regressions. To estimate the ROC-AUC score on unseen test samples, the data was first split according to a 10-fold stratified cross-validation. On the training splits of each fold, an exhaustive grid search was carried out using an inner 3-fold stratified cross-validation to determine an optimal value for the cost parameter C of the logistic regression³ – the value which maximizes the ROC-AUC estimated on the inner validation sets. Once an optimal value was determined, the logistic regression was re-fit on all samples of a given training split and tested on the corresponding outer set of testing samples to provide an unbiased estimate of performance on unseen data. The outer 10-fold cross-validation was repeated for some problems and was mentioned in the corresponding section of the results. The ROC-AUC scores across all folds were averaged and reported along with their standard deviation. When investigating the effect of the cost parameter C (Section 11 in the Supporting Information), the inner cross-validation and grid search steps were skipped and the results with all C values were shown. Data standardization – done by subtracting the mean and dividing by the standard deviation of each spectral feature – was applied as a preprocessing step before training and testing the logistic regression models. All mean and standard deviation values were calculated from the training splits and applied to their corresponding testing splits. The logistic regression, cross-validation, and grid search algorithms were used as implemented in the Scikit-Learn open-source Python package (version 0.24.1).⁴

5 Modeled FTIR measurement noise

To determine a realistic estimate for the measurement noise of the FTIR measurement device, a total of 400 water samples were repeatedly measured with the FTIR spectrometer. The standard deviation between the performed measurements across the spectral features was subsequently calculated. Fig. S1A depicts water measurements performed on the FTIR spectrometer. These measurements were used to calibrate the additive measurement noise introduced into the model (illustrated in Fig. S1B).

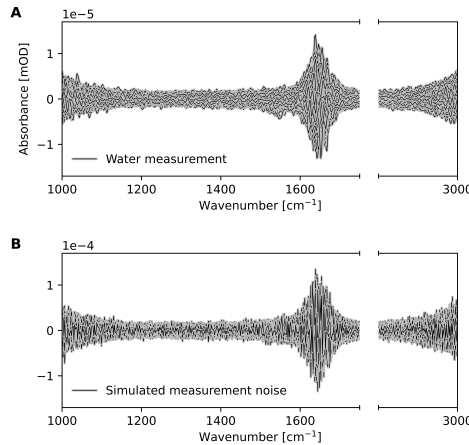


Figure S1: Calibration of the measurement noise using repeated water FTIR measurements.

(A) Water reference absorbance spectra ($n = 400$) with the mean of all water measurements subtracted. (B) A set of random vectors generated based on the standard deviation of the repeated water measurements across the wavenumbers on an arbitrary y-axis. The vectors in (B) illustrate the calibrated measurement noise coefficient ϵ in the simulation model and were scaled by a factor of 9.15 to account for the L2 vector normalization in the preprocessing of the sera measurements.

6 Validation of simulated data for the prostate cancer application

Fig. S2 supplements Fig. 2 in the main text which validated the model calibration procedure. Here, the validation was applied to the prostate cancer case study.

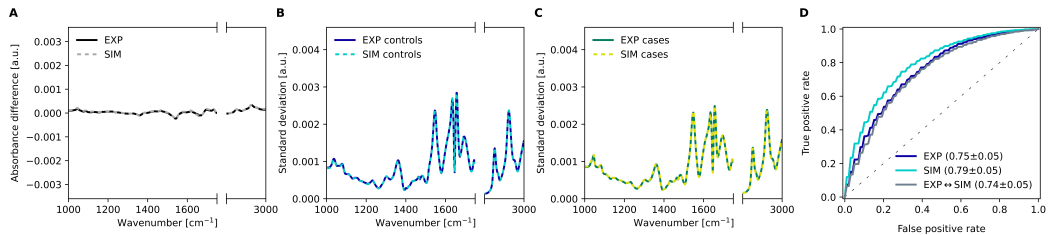


Figure S2: Validation of the model calibration procedure for the prostate cancer case study.

Spectral cohorts of individuals were simulated to model the properties of measured experimental cohorts using the same sample sizes ($n = 411$ cases vs. 411 controls). (A) The differential fingerprint, defined as the difference between the mean of case and control spectral features, for the experimental cohort (black) and the simulated cohorts (gray) as averaged across 10 simulation repetitions. (B-C) The standard deviation of the control spectral features (B) and case spectral features (C) for the experimental cohort (blue) and simulated cohorts (cyan) as averaged across 10 simulation repetitions. (D) The ROC curve for binary case-control classifications when training and testing on experimental samples (blue), training and testing on simulated samples (cyan), and training on simulated samples and testing on experimental samples (gray). The AUCs are listed in the figure legend along with their standard deviations across the cross-validation splits.

7 Influence of cohort size for the prostate cancer application

Fig. S3 supplements Fig. 3 in the main text which investigated the effect of the cohort size on classifier performance for experimental and simulated cohorts. Here, the validation was applied to the prostate cancer case study.

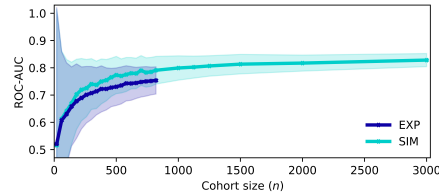


Figure S3: Effect of cohort size on prostate cancer detection.

Spectral cohorts consisting of balanced case and control samples were simulated at changing sample counts. The cross-validated ROC-AUC is plotted against the cohort size for the simulated samples (cyan). For comparison, sets of experimental samples were randomly selected and cross-validated upon to model the effects of changing cohort sizes with the experimental observations (blue). The solid curves depict the averaged scores along with their standard deviations in the shaded region.

8 Influence of biological variability and measurement noise for the prostate cancer application

Fig. S4 supplements Fig. 4 in the main text which investigated the effects of measurement noise and biological variability on classifier performance. Here, the validation was applied to the prostate cancer case study.

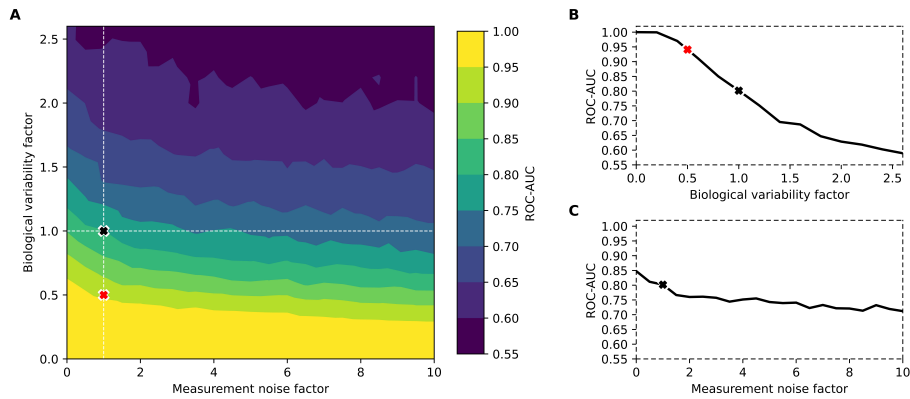


Figure S4: Effect of biological variability and measurement noise on prostate cancer detection.

(A) Classification performance was measured by the mean cross-validated ROC-AUC on multiple spectral cohorts simulated at changing levels of biological variability and measurement noise. The black “x” marks the point of calibration where the simulated cohorts model the experimental levels of between-person biological variability and measurement noise. The red “x” marks an estimate for reduced biological variability in a longitudinal, self-referencing, scenario as estimated from previous work.⁵ (B) Depicts a slice of (A) when the measurement noise factor is held constant at the calibrated level and the biological variability factor is changing. (C) Depicts a slice of (A) when the biological variability factor is held constant at the calibrated level and the measurement noise factor is changing.

9 Influence of molecular complexity for the prostate cancer application

Fig. S5 supplements Fig. 5 in the main text which investigated the effect of chemical molecular complexity on classifier performance. Here, the validation was applied to the prostate cancer case study.

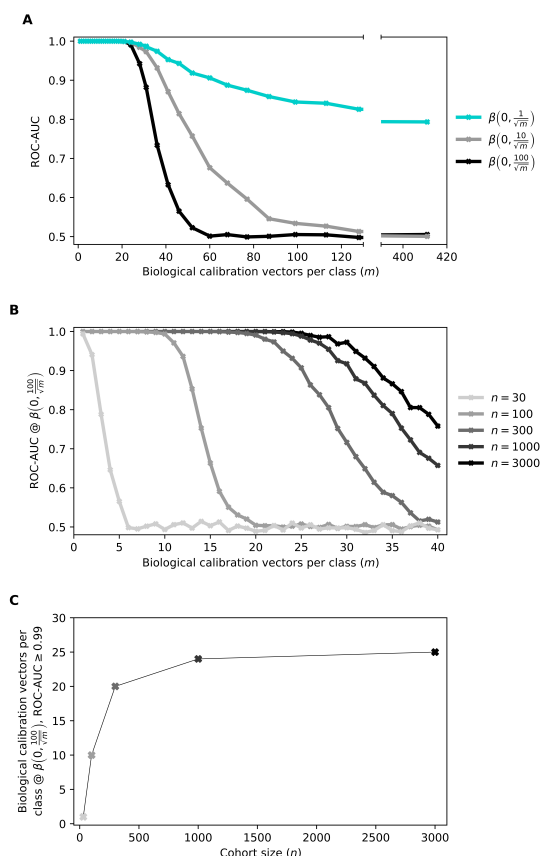


Figure S5: Influence of molecular complexity on prostate cancer detection.

(A) Multiple spectral cohorts were simulated at changing levels of m (i.e., the number of calibration vectors defining each molecular state) and changing levels of biological variability (β). Cohorts were created with the same cohort size as the experimental measurements and cross-validated upon to estimate the ROC-AUC as a measure of prediction performance. (B) Multiple datasets were simulated at changing levels of m and changing cohort sizes (n) to investigate the effects of the sample size on the trends observed in (A) at the highest level of biological variability investigated. (C) The number of calibration vectors m required to achieve a ROC-AUC ≥ 0.99 at the cohort sizes investigated in (B).

10 Simulating larger cohorts for model validation

Fig. S6 supplements Fig. 2D in the main text and S2D. Here, larger simulated cohorts were created to validate a model trained on simulated data and tested on experimental data. In comparison to Fig. 2D and S2D, the ROC-AUC values could be fully recovered using the larger sample sizes for both the lung (Fig. S6A) and prostate cancer (Fig. S6B) applications.

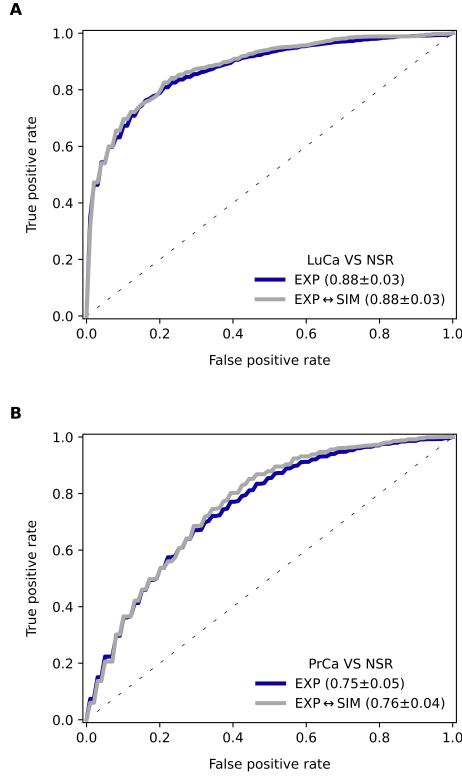


Figure S6: Receiver operating characteristic curves for binary classifications when trained on $n = 100000$ simulated samples of balanced classes and tested on unseen sets of experimental samples in a cross-validation (in gray). The blue curve depicts the result when cross-validating on the experimental samples. The areas under the curves and their standard deviations are listed in the figure legends for the lung cancer (A) and prostate cancer (B) applications.

11 Influence of noise on predictive model tuning

Regularization plays a pivotal role in the predictive modeling of ill-posed problems such as the classifications explored within this study. Techniques of regularization help predictive models mitigate the effects of multicollinearity and noise present in the training data.^{6,7} By introducing an L2 penalty, a supervised machine learning model works to solve the following generalized optimization problem, given a set of instance-label pairs $(\mathbf{x}_i, y_i)$:

$$\min_{\mathbf{w}} \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^N \xi(\mathbf{w}; \mathbf{x}_i, y_i)$$

where $\xi(\mathbf{w}; \mathbf{x}_i, y_i)$ is a defined error, or loss, function used in a predictive model. In other words, the optimization works to find the weight vector $\mathbf{w}$ that minimizes the error between predicted outcomes and the known outcomes. To avoid a solution where $\mathbf{w}$ becomes unduly large in magnitude, overfitting to peculiarities in the training data, the L2 penalty term is introduced and the extent of its involvements is controlled by the free parameter $C > 0$. The closer the value of C is to 0, the sparser the solution – where the entries of $\mathbf{w}$ are closer to 0. The optimal value for C can be determined by cross-validating to find the value which strikes the balance between allowing the model to learn the best possible value for $\mathbf{w}$ and not overfit.

We used our simulation model to investigate the effects of regularization on the performance of detecting lung and prostate cancer using L2-regularized logistic regression models at varying measurement noise and biological variability levels (Fig. S7A-D). We simulated artificial cohorts at increments of increasing measurement noise, holding the biological variability at the calibrated level, and subsequently cross-validated on each simulated dataset with different values for the parameter C . A similar evaluation was repeated to examine the effect of increasing biological variability while keeping the measurement noise at the calibrated level.

We found that with heavy classifier regularization (e.g., $C = 2^{-12}$), the classification performance on unseen data samples was relatively robust to increasing levels of measurement noise (Fig. S7A and C). Such heavily regularized models, however, did not provide optimal classification performance at the explored levels of measurement noise and led to “underfit” models that failed to capture the complexity of the problem. With little-to-no measurement noise, classifiers with weaker regularization strengths (e.g., $C = 2^{12}$) consistently outperformed in testing performance. Since such weakly regularized models are very susceptible to overfitting, increasing measurement noise quickly led to the loss of generalization and the classifier under-performed on the testing sets while perfectly separating the training samples.

Comparatively, the optimal regularization parameter was less sensitive to varying levels of biological variability (Fig. S7B and D). The optimal regularization strength ($C = 2^6$), remained optimal in the explored biological variability domain. Moreover, for this type of noise, all explored values for C suffered in a similar way from increasing noise. This revealed that the optimal penalty does not depend on the biological variability and thus the effects of increased biological variability cannot be mitigated by tuning a better classifier.

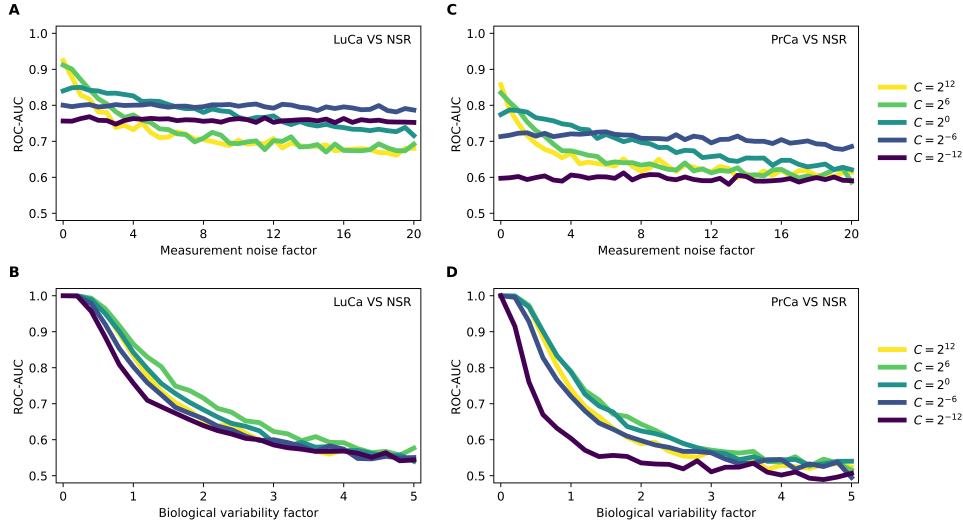


Figure S7: Influence of measurement noise and biological variability on classifier regularization.

Binary classifiers, with changing regularization strengths (parameter C), were fit to detect cancer on spectral cohorts simulated at different levels of measurement noise (first row) and biological variability (second row). (A-B) Depict the effect on detecting lung cancer while (C-D) depict the effect on detecting prostate cancer. Performance, as measured by the ROC-AUC, was estimated by cross-validating and averaging the score on each cohort created.

12 Simplified *in silico* model using a single vector to model case measurements

Fig. S8 repeats the validation of the numerical simulation shown in Fig. 2 in the main text and supporting Fig. S2, but by introducing a single discriminant feature vector $\mathbf{d}$ to describe the case measurements and only considering the biological variability of the non-symptomatic references to calibrate the model (equation 7 in the main text).

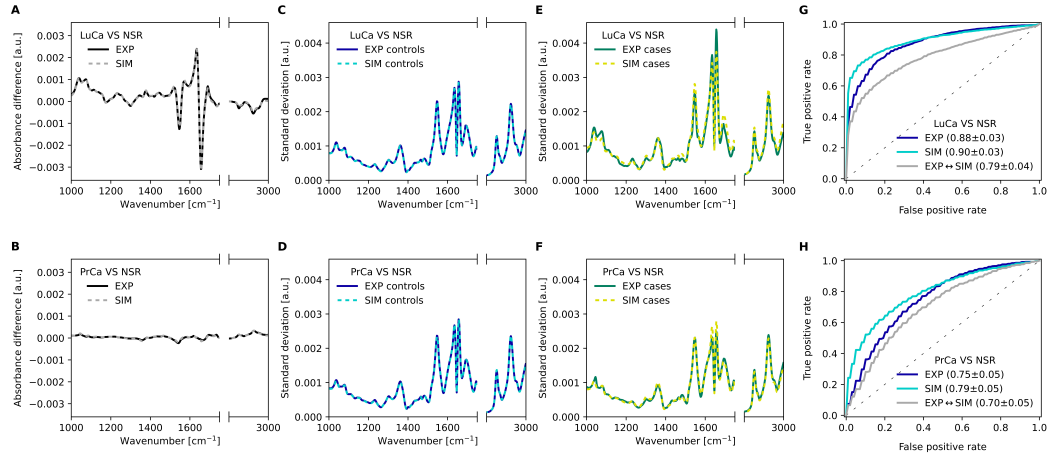


Figure S8: Validation of the calibration procedure for the model variant that uses a single discriminant vector to distinguish between cases and controls. In this investigation, the biological variability was calibrated based on the control measurements for each cancer cohort. (A–B) The differential fingerprint, defined as the difference between the mean of the case and control spectral features, for the experimental cohort and the simulated cohorts as averaged across 10 simulation repetitions. This differential fingerprint was introduced to each simulated case measurement to model the difference between the classes using a single vector for the lung (A) and prostate (B) cancer applications. (C–D) The standard deviation of the control spectral features for the experimental cohort and simulated cohorts as averaged across 10 simulation repetitions for both lung cancer (C) and prostate cancer (D) applications. (E–F) The standard deviation of the case spectral features for the experimental cohort and simulated cohorts was averaged across 10 simulation repetitions for both lung cancer (E) and prostate cancer (F) applications. For the simulated case measurements (E–F), the biological variability was scaled to minimize the squared error between the simulated standard deviation spectrum and the experimentally derived one. (G–H) The receiver operating characteristic curves (ROCs) for binary case-control classifications of lung cancer (G) and prostate cancer (H) when training and testing on experimental samples (blue), training and testing on simulated samples (cyan), and training on simulated samples and testing on experimental samples (gray). The areas under the curves are listed in the figure legends along with their standard deviations across the cross-validation splits.

References

- (1) Huber, M.; Kepesidis, K. v.; Voronina, L.; Fleischmann, F.; Fill, E.; Hermann, J.; Koch, I.; Milger-Kneidinger, K.; Kolben, T.; Schulz, G. B.; Jokisch, F.; Behr, J.; Harbeck, N.; Reiser, M.; Stief, C.; Krausz, F.; Zigman, M. Infrared Molecular Fingerprinting of Blood-Based Liquid Biopsies for the Detection of Cancer. *Elife* **2021**, *10*. <https://doi.org/10.7554/eLife.68758>.
- (2) Eissa, T. In Silico Spectral Generator. *GitHub repository*. GitHub 2022. <https://github.com/tarek-eissa/in-silico-spectral-generator>.
- (3) Fan, R. E.; Chang, K. W.; Hsieh, C. J.; Wang, X. R.; Lin, C. J. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research* **2008**, *9*.
- (4) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*.
- (5) Huber, M.; Kepesidis, K. v.; Voronina, L.; Božić, M.; Trubetskov, M.; Harbeck, N.; Krausz, F.; Žigman, M. Stability of Person-Specific Blood-Based Infrared Molecular Fingerprints Opens up Prospects for Health Monitoring. *Nat Commun* **2021**, *12* (1) <https://doi.org/10.1038/s41467-021-21668-5>.
- (6) Bauer, F.; Lukas, M. A. Comparing Parameter Choice Methods for Regularization of Ill-Posed Problems. *Math Comput Simul* **2011**, *81* (9), 1795–1841. <https://doi.org/10.1016/j.matcom.2011.01.016>.
- (7) Zou, H.; Hastie, T. Regularization and Variable Selection via the Elastic Net. *J R Stat Soc Series B Stat Methodol* **2005**, *67* (2), 301–320.

Appendix D

CODI: Enhancing machine learning-based molecular profiling through contextual out-of-distribution integration

Copyright notice: © 2024 The Author(s). This is an open access article distributed under the terms of the Creative Commons CC-BY license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The CC-BY 4.0 license is printed in Appendix E.

CODI: Enhancing machine learning-based molecular profiling through contextual out-of-distribution integration

Tarek Eissa ^{a,b,c,*}, Marinus Huber ^{a,b}, Barbara Obermayer-Pietsch ^d, Birgit Linkohr ^e, Annette Peters ^{e,f}, Frank Fleischmann ^{a,b} and Mihaela Žigman ^{a,b,*}

^aChair of Experimental Physics - Laser Physics, Ludwig-Maximilians-Universität München, Bavaria 85748, Germany

^bLaboratory for Attosecond Physics, Max Planck Institute of Quantum Optics, Bavaria 85748, Germany

^cSchool of Computation, Information and Technology, Technical University of Munich, Bavaria 85748, Germany

^dDepartment of Internal Medicine, Division of Endocrinology and Diabetology, Medical University, Styria 8010, Austria

^eInstitute of Epidemiology, Helmholtz Zentrum München, Bavaria 85764, Germany

^fChair of Epidemiology, Institute for Medical Information Processing, Biometry and Epidemiology, Medical Faculty, Ludwig-Maximilians-Universität München, Bavaria 81377, Germany

*To whom correspondence should be addressed: Email: tarek.eissa@mpq.mpg.de (T.E.); mihaela.zigman@mpq.mpg.de (M.Ž.)

Edited By: Lydia Kavradi

Abstract

Molecular analytics increasingly utilize machine learning (ML) for predictive modeling based on data acquired through molecular profiling technologies. However, developing robust models that accurately capture physiological phenotypes is challenged by the dynamics inherent to biological systems, variability stemming from analytical procedures, and the resource-intensive nature of obtaining sufficiently representative datasets. Here, we propose and evaluate a new method: Contextual Out-of-Distribution Integration (CODI). Based on experimental observations, CODI generates synthetic data that integrate unrepresented sources of variation encountered in real-world applications into a given molecular fingerprint dataset. By augmenting a dataset with out-of-distribution variance, CODI enables an ML model to better generalize to samples beyond the seed training data, reducing the need for extensive experimental data collection. Using three independent longitudinal clinical studies and a case-control study, we demonstrate CODI's application to several classification tasks involving vibrational spectroscopy of human blood. We showcase our approach's ability to enable personalized fingerprinting for multiyear longitudinal molecular monitoring and enhance the robustness of trained ML models for improved disease detection. Our comparative analyses reveal that incorporating CODI into the classification workflow consistently leads to increased robustness against data variability and improved predictive accuracy.

Keywords: data augmentation, molecular analytics, machine learning, variability modeling, out-of-distribution

Significance Statement

Analyzing molecular fingerprint data is challenging due to multiple sources of biological and analytical variability. This variability hinders the capacity to collect sufficiently large and representative datasets that encompass realistic data distributions. Consequently, the development of machine learning models that generalize to unseen, independently collected samples is often compromised. Here, we introduce Contextual Out-of-Distribution Integration (CODI), a versatile framework that enhances traditional classifier training methodologies. The concept of CODI is to incorporate information about possible out-of-distribution variations into a given training dataset, augmenting it with simulated samples that better capture the true data distribution. This allows the classification to achieve improved predictive performance on samples beyond the original training distribution.

Introduction

Technological advances in molecular analytics increasingly enable the probing of biological systems. Distinguishing between physiologically relevant states from quantitative molecular fingerprints presents a new opportunity for *in vitro* phenotyping. Extensive efforts are thus dedicated to developing standardized procedures involving streamlined biological sampling, post-

collection handling, and sensitive quantitative measurements. Nevertheless, empirical datasets are susceptible to diverse sources of variability, both analytical and inherently biological (1–9). Obtaining a dataset that reflects a realistic data distribution is often resource-intensive, costly, and, in some cases, impossible. This applies especially in the context of clinical studies, covering all pathophysiological strata, studying rare disease, or



Competing Interest: The authors declare no competing interests.

Received: April 9, 2024. **Accepted:** September 7, 2024

© The Author(s) 2024. Published by Oxford University Press on behalf of National Academy of Sciences. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

longitudinally probing the same system over time. Exploratory studies are thus often limited in size and scope, making it challenging for a given “training” set to be representative of the true unseen “test” domain. Consequently, when applying a developed machine learning (ML) model to independently collected and experimentally measured samples, the model may fail to achieve the expected efficacy (8, 10–15).

While traditional approaches often rely on standardizing experimental workflows and creating computer-aided processing techniques to reduce unwanted empirical noise (7, 16–20), complete noise removal is likely unattainable. Failure to account for noise and distributional shifts may obscure the true biological patterns of interest, misleading an ML algorithm into utilizing information that is unlikely to be reproduced. This failure is due to violating the assumption underlying (supervised) ML algorithms that the training and testing data are independent and identically distributed (i.i.d.) (11, 13, 21–23). To decode the information contained within a dataset, accounting for analytical and biological variability is critical to handle data domain shifts and ensure successful model generalization.

The concept of out-of-distribution (OOD) generalization has very recently garnered attention in ML research to address the

shortcomings of i.i.d. assumptions (22, 24, 25). This paradigm shift acknowledges the unpredictability of unseen data, prompting exploration into methods that better accommodate distributional shifts to generalize beyond the training set. OOD generalization has been extensively explored in computer vision and natural language processing tasks (24–27). However, there is a critical lack in the development of OOD generalization techniques in molecular analytics involving vibrational spectroscopy, NMR spectroscopy, and mass spectrometry, as well as in clinical chemistry analytics.

To address these challenges, here we develop and empirically test a hybrid experimental and computational modeling strategy. We explore OOD generalization in the context of molecular analytics and propose to recognize the variations arising from analytical paradigms as integral components of real-world observations (Fig. 1). We introduce Contextual Out-of-Distribution Integration (CODI), a strategy that paradoxically embraces measurement variability and the inherent complexities of biological systems, transforming them into valuable properties that can be utilized. CODI first involves experimental data to evaluate their distributional characteristics. Following the characterization, we deliberately introduce these distributional characteristics into a studied, independent, dataset through the *in silico* generation of synthetic

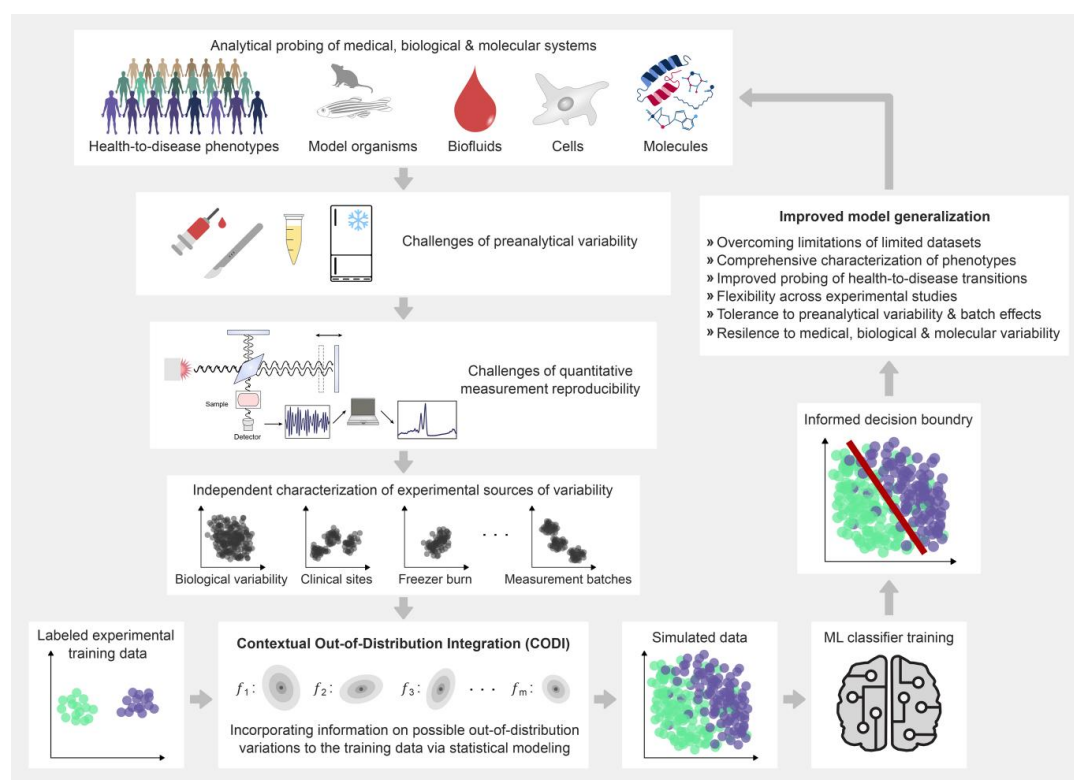


Fig. 1. Overview of problem context and CODI's methodology. Given a biological, medical, or molecular system of interest, we are presented with a task of classifying distinct groups of samples. However, variations stemming from several biological and (pre)analytical aspects in the empirical workflow may impact captured measurements in different ways. CODI leverages independently characterized sources of variability and incorporates them into a labeled training set of experimental observations. This process generates simulated samples with a more representative data distribution. Training an ML classifier on simulated samples enables it to learn a decision boundary that separates classes of samples in a more informed manner, increasing the likelihood of generalizing to unseen test samples. This approach enables sample characterizations that are more robust to variations in the empirical workflow.

data. These synthetic data mimic the system(s) of interest while expanding the distribution of the original training set to incorporate information about sources of variance that were crucially OOD and missing.

To establish the concept and evaluate it in a realistic setting, we apply our method on experimental infrared (IR) spectroscopic data to aid in vitro blood-based diagnostics. The advantage here lies in cross-molecular fingerprinting, where quantitative analytical measurements capture the breadth of changes in the molecular landscape of complex samples as indicators of systemic health and disease. We test our method in the framework of three independent longitudinal clinical studies spanning up to an 8-year follow-up period (28–30), as well as a case–control study to detect four common cancers (31). Our results demonstrate that integrating CODI into the classification pipeline enables the creation of more representative datasets, arbitrarily large in size, that empower ML algorithms to more effectively capture reproducible signals in biological datasets. Ultimately, we showcase how the proposed framework leads to significantly improved classification output on unseen, independently measured test samples, ensuring robust predictions despite shifts in data distribution.

Results

Characterizing empirically observed variability

We previously introduced an *in silico* model that generates 1D spectra of complex biological samples, focusing on IR absorption spectra (32). Our initial work explored the impact of varying levels of between-person biological variability on classification efficiency in simulated case–control conditions. Building on this foundation, we extend the model beyond the theoretical framework. We generate data simulating longitudinal and/or case–control settings, accounting for diverse sources of possible variability that we experimentally characterize. To assess its practical applications, we explore the capacity of the modeling framework to computationally generate larger training sets that are more robust to biological and analytical variabilities.

CODI is a relatively simple statistical procedure that relies on characterizing data distribution patterns (Fig. 1). In a generalized form, we capture the differences between sets of experimental observations ($\mathbf{u}_i - \mathbf{v}_j$), henceforth called calibration measurements. These calibration measurements can be adjusted based on different pools of available measurements that reflect diverse sources of data variability. For instance, quality control samples can be repeatedly measured under varying laboratory conditions, resulting in a measurement set $\{\mathbf{u}_i \mid i = 1, \dots, m\}$. By setting $\mathbf{v}_j = \bar{\mathbf{u}}$, the calibration set would consist of mean-centered observations that reflect the deviations of each control measurement. Alternatively, rather than mean-centering the observations, $\mathbf{u}_i$ and $\mathbf{v}_j$ can represent measurements of paired samples collected, processed, or measured under different scenarios. For instance, $\mathbf{u}_i$ and $\mathbf{v}_j$ can be measurements of two samples processed by different operators from the same mother tube or measurements of the same sample on different experimental instruments. These sources of variability can be adjusted to different analytical settings, depending on the expected sources of deviation.

After the characterization of measurement deviations, these differences ($\mathbf{u}_i - \mathbf{v}_j$) are scaled by a random variable that assumes a Gaussian distribution and then combined over the entire variability calibration set. This aggregation may then be added onto an independent experimental training seed measurement $\mathbf{x}_k$ to create a new simulated measurement that is now modeled as a statistical outcome. Such a simulation approach can be

repeatedly applied to generate a cohort of simulated measurements in arbitrary size. The generated cohort, as a whole, would reflect the variability properties observed between $\mathbf{u}_i$ and $\mathbf{v}_j$ onto $\mathbf{x}_k$. If the set of calibration measurements would reflect a source of variability that was unobserved in a given training set of measurements $\{\mathbf{x}_k \mid k = 1, \dots, n\}$, a new level of variability would be introduced onto the training set of measurements. This strategy allows for the creation of realistic synthetic data without needing to fine tune free parameters controlling the data generation. Further detailed descriptions are in the Materials and methods and [Supplementary Information](#).

In our example applications of CODI, we introduced several distinct sets of calibration measurements to model different sources of variability that may be observed in IR spectral measurements of blood-based media (Fig. 2a). Within these calibration measurements are characteristics of empirical variability stemming from inherent biological factors, variations in sample collection and handling, as well as instrument-specific measurement noise and drifts ([Supplementary Information](#)).

When addressing biological variability, the calibration measurements $\mathbf{u}_i, \mathbf{v}_j$ are selected to be experimental measurements of the same individual over time, capturing a level of within-person biological variability (Fig. 2a, upper left), as defined previously (30). Alternatively, opting to set the calibration measurements to be of different individuals would yield a level of between-person biological variability, as demonstrated previously (32).

Further variations that stem from different clinical sample collection sites, clinical study protocols, and sample handling procedures may be effectively represented by selecting calibration measurements characteristic of samples derived from different clinical studies (Fig. 2a, upper right). The same concept can be extended to model realistic variations that arise from experimental procedures like sample storage temperature and duration, aliquoting procedures, and measurement device drifts. For example, quality control (QC) samples, may be subjected to diverse handling and storage conditions. Performing measurements of QCs under different operating conditions for the measurement device, including instances of recalibration, routine maintenance, or changes in the surrounding environment would enable the QC measurement dataset to mimic potential variations in both laboratory procedures and instrumental drifts (Fig. 2a, lower left). Further, independent, measurements of technical replicates (e.g. pure water) performed over extended periods can facilitate a clearer distinction between instrumental device noise and laboratory variations (Fig. 2a, lower right).

The overarching goal of characterizing diverse sources of variability is to realistically simulate the data distribution that may be encountered in the empirical workflow. It is crucial to recognize that this can be achieved through the utilization of measurements that are independent of the original training set and unrelated to the specific questions posed by it (i.e. class-invariant). Therefore, the characterized source of variability can be repeatedly used across a diversity of classification tasks. When extending the concept to other molecular systems or measurement modalities, similar calibration sets may be adapted to characterize the variations relevant to the studied conditions.

Introducing experimental variability in silico

As an illustrative example, we applied the CODI framework to five experimental spectra of blood plasma to generate a larger set of simulated measurements that reflect an increased level of variance (Fig. 2b). The five original spectra may be considered to be a training set, with each measurement representing a labeled

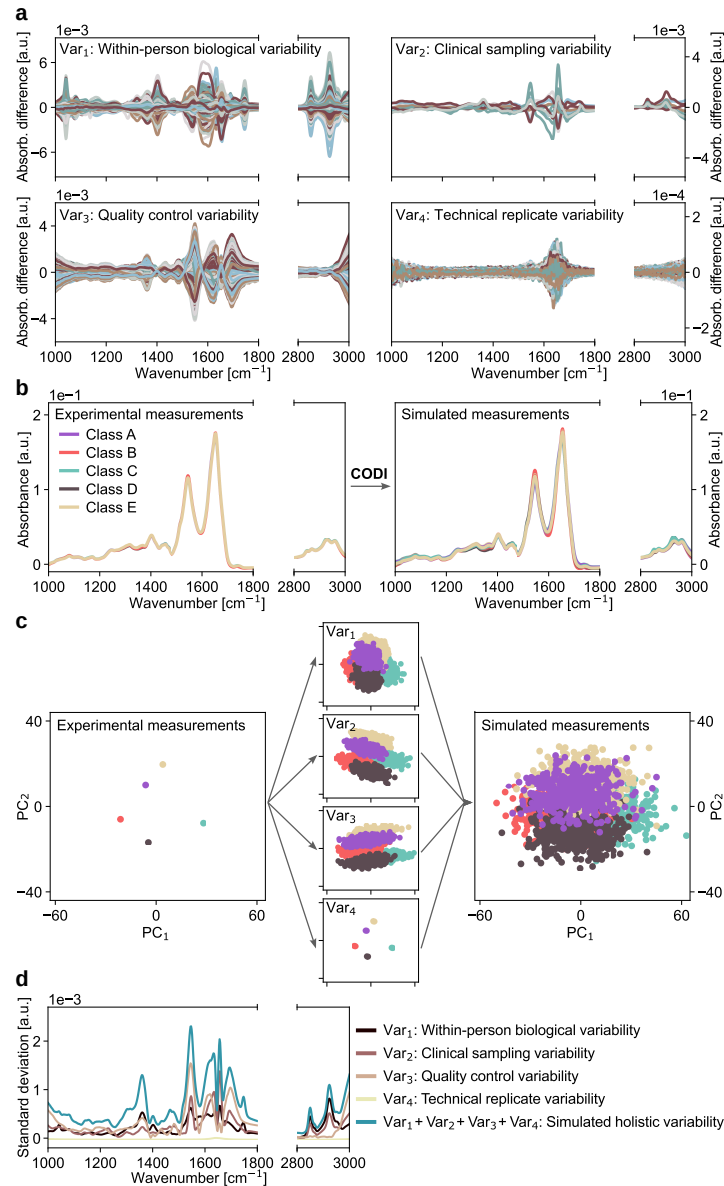


Fig. 2. Applying CODI to introduce measurement variability onto illustrative experimental IR spectra. a) Four distinct sources of possible variability were characterized from calibration sets of experimental observations. The curves depict how each measurement differs from its expected mean. b) By applying CODI, the four sources of variability were introduced to five illustrative experimental blood-based spectra (left panel) to generate a larger set of simulated spectra (right panel). c) Principal component analysis (PCA) on the five illustrative experimental spectra (left panel), on a larger simulated set of spectra that introduced only one out of the four characterized sources of variability (middle panel), and on a simulated set of spectra that introduced all four sources of variability to the five illustrative experimental spectra (right panel). d) Comparison of the standard deviation across the spectral range for each characterized source of variability and the standard deviation of simulated measurements, resulting in an overall increased standard deviation.

class (Fig. 2b, left). Using the five measurements as a seed input, CODI enabled the generation of a larger and more representative training set of measurements (Fig. 2b, right).

Principal component analysis (PCA) applied to the original and simulated measurements reveals that each source of measurement

variability affected the spatial distribution of the seed data differently across the first two components (Fig. 2c). In other words, each set of calibration measurements—modeling different variability properties—affected linearly independent data features. Once all four sources of measurement variability were incorporated into

the original seed data (Fig. 2c, right), the simulated measurements occupied a larger cloud of data points, while still maintaining their distinct cluster centroids.

Similarly, examining the measurement standard deviation shows that the simulated measurements had a higher standard deviation than each individual source of empirical variability (Fig. 2d). While it may seem counter-intuitive that a simulated dataset with increased variance could offer added value compared to the existing experimental observations, this variance contains valuable, usable information. The principle relies on the assumption that the simulated measurements include OOD measurement events that are likely to occur when presented with additional experimental observations.

So far, we have shown how CODI can introduce additional sources of variability into an existing dataset to enrich its information content. The value of the method for real-world applications is examined in the following sections.

Application to longitudinal study settings

In longitudinal clinical studies that involve the collection of biological specimens tainted by attrition and loss-to-follow-up over time, great efforts are required to gather sufficiently large datasets. Typically, individuals participate in an initial baseline sample collection, followed by extended waiting periods for subsequent collections from the same individuals. In situations where only few samples are initially available per individual, the

challenge arises in extrapolating meaningful insights to later collected and measured follow-ups—owing to the dynamic nature of the empirical procedure as previously described. To examine whether our proposed approach offers added value when severely limited samples are available for analysis, we first employ CODI in the context of longitudinal analyses (Fig. 3).

We utilized samples from three independent clinical studies that followed individuals over multiyear periods (Fig. 3a). The Lasers4Life-LG study cohort (30) comprised of 31 individuals that repeatedly donated blood samples at irregular follow-up intervals. The study commenced with a 7-week baseline monitoring period, during which 288 samples were collected through repeated donations. The initial baseline donation period was followed by three additional donations, spanning up to 4.5 years, during which one sampling point was considered per individual. In the BioPersMed study (28), a subcohort of 44 individuals repeatedly participated over an 8-year follow-up period, with a 2-year interval between each donation. In the KORA study (29, 33), a subcohort of 2015 individuals participated in two donations, separated by a 6.5-year follow-up interval. Blood plasma was processed from all samples and measured via absorption IR spectroscopy (Materials and methods).

Empowering long-term molecular profiling

The concept of identifying individuals from a given population based on different biofluids has been demonstrated with IR

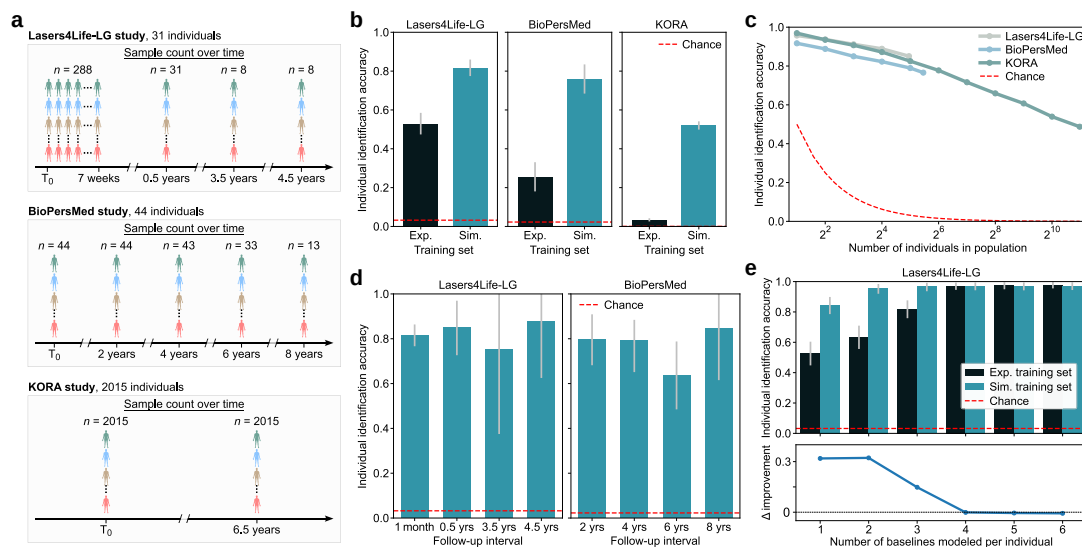


Fig. 3. CODI enhances personalized fingerprinting through more accurate long-term molecular profiling. a) Setup of three independent longitudinal clinical studies in which same individuals repeatedly participated in venous blood sampling over time. Experimental IR spectroscopic measurements were performed on blood plasma and utilized in this analysis. b) Individual identification efficacy utilizing only a single baseline IR measurement per individual across the three study cohorts. Bars depict the classification accuracy using experimental ("Exp.") baseline measurements for training and using simulated ("Sim.") measurements generated by CODI derived through a baseline IR measurement per individual as a seed input. Classifier testing was performed on the same experimental follow-up measurements of each individual for both training approaches. c) Dependence of identification accuracy on number of individuals in the populations (i.e. number of classes). Individuals were randomly selected, at varying population sizes, and CODI was applied using a single baseline IR measurement per individual as a seed input. Classifier testing was performed on experimental follow-up measurements of each individual included in the training set. d) Dependence between the identification accuracy and the follow-up time axis on applications involving the Lasers4Life-LG and BioPersMed cohorts. e) Modeling an increasing number of experimental baseline measurements per individual as a training seed. Bars depict the identification accuracy using simulated training sets and experimental training sets. Model testing was conducted on the same experimental follow-ups beyond the first 6 baseline measurements per individual. The lower panel illustrates the difference between accuracy derived from the experimental and simulated training sets.

spectroscopy, NMR, and mass spectrometry as fingerprinting modalities (30, 34–36). We previously demonstrated that plasma- and serum-based IR fingerprints can identify individuals over a 6-month follow-up period (30). This application inherently relies on the stability of measurements over a study period, often requiring the comparison of measurements acquired at different times despite inevitable experimental drifts (37). In previous works (30, 34–36), several measurements from the same individual were required to adequately train a multiclass classifier to distinguish between different individuals over a given follow-up period (typically 8–42 samples).

We set out to test the possible value of CODI for enhancing longitudinal studies, using the individual identification task as a readout metric. To test the limits of the framework, here we considered a scenario with severely limited training data—relying only upon a single experimental observation per class for training, i.e. one measurement per individual (Fig. 3b). We utilized the first baseline measurement of each individual to train a multiclass classifier to identify individuals from their follow-up measurements. We then compared this prediction efficiency to a classifier trained on a simulated set of measurements generated through CODI that used the experimental baselines as initial seed data. Within the CODI framework, we modeled the four previously described sources of variability (Fig. 2a), which were characterized from data sources independent of the experimental seed data and follow-up test data (Supplementary Information). This step was crucial to ensure that no leakage from the test data occurred when introducing the variance to the training seed. Through CODI, we generated 1,000 simulated measurements per individual that were then used to train the classifier.

This investigation revealed that the classifier trained on simulated measurements had a remarkably improved prediction capacity over the classifier trained directly on experimental measurements (Fig. 3b). Across the three cohorts, the individual identification accuracy improved from 0.53 to 0.82, from 0.26 to 0.76, and from 0.03 to 0.52. This demonstrated that the informed incorporation of data variance was indeed capable of enabling better generalization to unseen test samples.

To examine which sources of data variability aided the most in boosting the classification efficiency, we re-performed the above analysis, but systematically eliminated one of the four sources of variability we incorporated in the CODI framework (Fig. S1). We found that the within-person biological variability over time and measurement variability of the same quality control samples over the course of measurements were the most critical contributors to the success of the classification. Including the variability stemming from technical replicates and clinical sampling had minimal impact on the classification, compared to the two aforementioned sources of variability. This highlighted the importance of incorporating information on the data distribution that was truly missing from the original experimental training set—not just the incorporation of (random) added variance.

For the above classification task, it is crucial to recognize that the population size varied between the cohorts. The decreased accuracy observed in the KORA cohort (involving 2015 individuals) is thus not directly comparable to that of the Lasers4Life-LG cohort (involving only 31 individuals). This is due to the fact that the more individuals exist in a dataset, the more likely it is that their fingerprints will overlap with one another—making the task of identifying individuals more challenging. Despite this, it was very surprising and encouraging to observe that nearly half of 2015 individuals can be identified from IR molecular fingerprints when combined with the proposed modeling approach—and requiring

only the venous blood sampling of a single baseline sample per individual.

Observing that the identification accuracy decreased with an increasing population size prompted us to further investigate this dependency (Fig. 3c). We first trained a classifier on simulated measurements utilizing the first experimental baseline measurements of only 2 individuals and, as previously, tested on their follow-ups. This procedure was repeated several times, using 2 other randomly selected individuals. We then performed the same procedure, but on 4, 8, 16, and so on individuals. This analysis revealed that the identification accuracy, depending on the population size, follows a nearly perfect logarithmic trend. This intriguing finding draws from information theory and may be explored further to quantify the informational content of diverse molecular fingerprints. Remarkably, these results were reproducible on three independent cohorts, revealing that a similar identification accuracy can be achieved when the datasets involve similar population sizes.

In the above application, follow-up measurements of all individuals were pooled together and the accuracy of identification was averaged, independent of the follow-up time axis. This prompted the question of whether the individual identification accuracy was dependent on the time interval between follow-up measurements and the baseline (Fig. 3d). In other words, is it more difficult to identify an individual 8 years after their baseline sample was assessed than from a 2-year follow-up? To investigate this, we grouped the follow-up measurements by their time differences to the baseline and examined whether any temporal trend was observed in the identification accuracy (Fig. 3d). This analysis was only possible on the Lasers4Life-LG and BioPersMed cohorts, since the available KORA cohort only involved one follow-up. Here, we revealed that the identification accuracy did not depend on how far off the follow-up was from the baseline. Very surprisingly, the accuracy remained relatively stable even over an 8-year follow-up period (Fig. 3d). Although it is crucial to recognize that the number of test samples in the later follow-up years was limited (Fig. 3a), this is the very first experimental result over such long-lived fingerprint stability.

Altogether, the above investigations were made possible by CODI, which enabled applications that were previously unfeasible with limited experimental observations.

Personalized multibaseline modeling

In the above quest of examining the value of the CODI framework, we relied on a single baseline measurement per individual. We then questioned to what extent can the classification be made more robust when more training instances per class are available. Specifically, considering that a single baseline measurement may be an outlier, we examined the dependence between the number of training instances per individual and the identification accuracy (Fig. 3e). For this analysis to be properly investigated, individuals would have to be repeatedly sampled in a given baseline monitoring period. Among the three clinical studies, only the Lasers4Life-LG study facilitated such a setup (30).

We utilized data from up to the first 6 baseline measurements per individual from the Lasers4Life-LG cohort to be used as an experimental seed for training. Then, we simulated a training set that consisted of 1,000 measurements per baseline and investigated how the identification accuracy depended on how many experimental baselines were modeled per individual. Classifier testing was performed on the remaining follow-ups that were beyond the first 6 baselines of each individual. This investigation revealed that the identification accuracy following the

simulation-based approach could indeed be improved when more than one baseline measurement was modeled per individual (Fig. 3e, blue bars). When using only one baseline measurement, an accuracy near 0.85 was achieved. Surprisingly, including only one additional baseline already led to an improvement of a nearly perfect prediction efficiency, achieving an accuracy of 0.96.

As a comparable benchmark, we again examined the dependence between the identification accuracy and the number of baselines modeled per individual, but now training the classifier directly on experimental measurements (Fig. 3e, black bars). The classifier was first trained on one experimental baseline per individual, then again on two, and up to 6 baselines of each individual. Testing was performed as previously—on the remaining follow-ups beyond the first 6 training baselines per individual. Here, it was again revealed that the simulation-based approach had a significant advantage over the experimental approach—but only when few observations per class were available (≤ 3 baselines per individual). Once ≥ 4 experimental baselines per individual were available for training, the experimental approach had also achieved a near perfect prediction efficiency and thus no advantage was seen by applying CODI. This underscored the impact of our proposed modeling paradigm in contexts with only limited experimental datasets. Once sufficiently large experimental datasets are available, the simulation-based training approach may not provide an advantage over training directly on experimental data.

Altogether, these findings show that the CODI framework can enable the establishment of a more reliable “baseline” per individual—one that is more resilient to analytical and biological variations and can more robustly enable ML generalization. We further demonstrate that IR molecular fingerprints are highly stable and individual-specific. Previously, this was only demonstrated on the time frame of 6 months (30). In the current study, we extend these findings to a medically relevant time frame of 8 years. These results form the foundation for future applications of blood-based IR fingerprinting as a modality of personalized monitoring of human health over time, potentially requiring a small number of samples to establish a reliable baseline per individual.

Comparison to domain-agnostic augmentation schemes

The CODI framework inherently relies on *a priori* information on potential sources of measurement variability. In contrast, domain-agnostic augmentation methods employ generic transformations on input seed data to simulate new observations (e.g. introducing random additive or multiplicative noise). By eliminating the need for *a priori* information, domain-agnostic augmentation strategies are practically easier to implement than CODI. To examine whether our approach yielded an advantage over other augmentation methods, we re-performed the above analysis by applying several methods of augmenting the spectral measurements (Fig. S2). We found that the CODI strategy of generating training sets significantly outperformed all other methods of domain-agnostic augmentation that randomly manipulated the spectra. This underscored the value of incorporating contextual *a priori* information into the data augmentation process to enable the classification to generalize beyond the original training set.

Impact of increased variance on classification efficacy

As the basis of CODI is to introduce variance to a given dataset, this approach also carries the concern that the simulated variance may

significantly exceed what is typically expected from the domain of possible empirical observations. An excessive amount of variance may obscure the underlying discriminative signals of interest, leading the classification to perform poorly on unseen test data. To investigate the extent of this concern, we systematically increased the variance introduced by CODI and examined its effects on the classification task of longitudinally identifying individuals (Fig. S3). We found that excessive variance resulted in a worsened classification efficacy, particularly when few samples were generated. Generating larger datasets, however, mitigated these risks and made the classification more robust to the excessive variance. This emphasized the need to generate a large number of simulated samples with our proposed approach. This finding is consistent with the known principle that a higher ratio of features to sample size increases the likelihood that an ML model will fit to noise rather than the targeted underlying patterns (38).

Cross-specimen generalization

Molecular profiling applications involve the use of diverse sample specimens—e.g. serum or plasma as cell-free products of systemic blood (Fig. 4a). Selecting an appropriate specimen is typically made in a study design phase, considering factors like ease of collection and biological relevance (39, 40). However, limitations may arise from a preemptive selection as insights gained from one specimen may not generalize when transferred to another. For instance, assume a dataset of plasma spectra is available. Later, the need arises to classify and compare unlabeled spectra that originate from serum samples. This prompted an intriguing question: how well would a classifier trained on plasma spectra perform when tested on serum spectra? The straightforward answer is that the classification is likely to fail, due to underlying molecular differences between the specimens (41). Effective classification necessitates the inclusion of training instances from different specimens, each with sufficient representation to capture class-specific distributions—a highly resource-intensive process. As proof-of-principle, here we demonstrate the potential versatility of the CODI framework to enable such a domain adaptation application, while minimizing the need for extensive biological dataset collection.

The IR spectra of plasma and serum share many characteristics, due to their relatively similar molecular profiles (Fig. 4b). The main spectral variations stem from the plasma preparation process, which involves the use of ethylenediaminetetraacetic acid (EDTA), as demonstrated in previous work (42). To achieve effective classification flexibility between specimens, their differences must be well-characterized. This can be achieved by calculating differences between experimental plasma and serum measurements of the same collected blood sample (Fig. 4c). With CODI, we incorporated such characterized differences into an independent experimental dataset of plasma spectra to generate simulated spectra that resemble a mixture between the specimens (Fig. 4d).

Next, we revisited the task of identifying individuals from a given population as a metric to estimate the capacity of cross-specimen generalization. We performed this investigation in the Lasers4Life-LG cohort (Fig. 4e), where both serum and plasma were available from the same individuals at all blood donations. The dataset was split into a training set, consisting of 4–12 donations per individual, and a test set, consisting of the remaining follow-up donations. As a benchmark, we first examined the efficacy of training two classifiers—one trained on experimental plasma spectra and one trained on simulated plasma spectra, testing

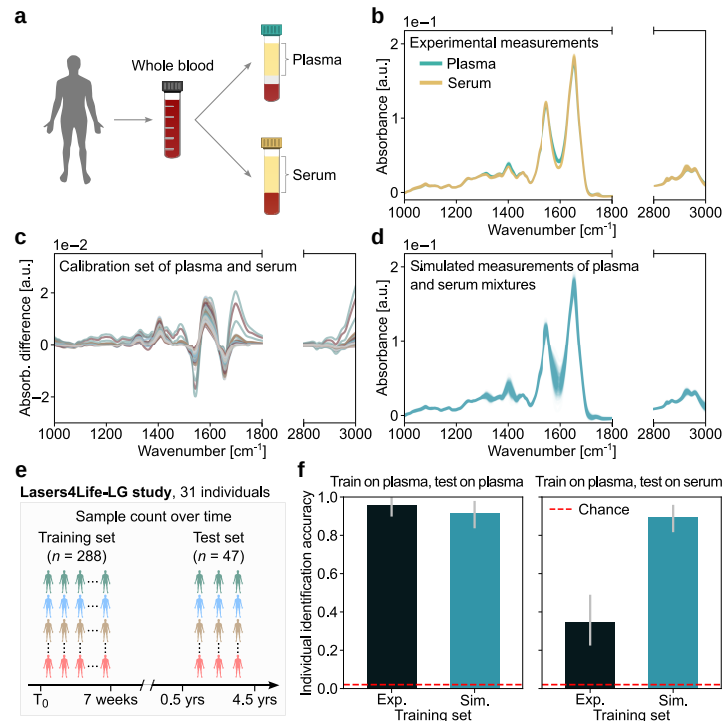


Fig. 4. CODI enables classification flexibility across biological specimen variants. a) Plasma and serum were collected as cell-free products of whole venous blood. b) Experimental spectra were measured from several plasma and serum samples of the same individuals. c) Differences between spectra of plasma and serum, processed from the same whole blood sample, were calculated to reveal the characteristic variations between the specimens. d) CODI enabled the creation of simulated spectra of plasma/serum mixtures by utilizing the characteristic variations between the specimens as a calibration set. e) Setup of Lasers4Life-LG cohort in which the same individuals repeatedly participated in venous blood sampling over time. Donations were split into a training set and a test set, with blood plasma and serum processed from all donations. f) Individual identification accuracy utilizing the Lasers4Life-LG cohort as a basis for training and testing. Left panel depicts the accuracy of classifiers trained on experimental plasma fingerprints and simulated plasma fingerprints—testing both classifiers on experimental plasma fingerprints. Right panel depicts the accuracy of a classifier trained on experimental plasma fingerprints and simulated fingerprints of plasma/serum mixtures—testing both classifiers on experimental serum fingerprints.

both on plasma spectra (Fig. 4f, left panel). This investigation confirmed the earlier discovery—in that, with sufficiently large experimental training sets, both experimental- and simulation-based classifiers perform similarly.

We then applied the same classification procedure, but testing on the serum measurements (Fig. 4f, right panel). For this analysis, we employed CODI to generate a training set of plasma/serum mixtures. In order to eliminate the risk of test data leakage, the differences between plasma and serum samples were characterized from blood samples of an independent cohort of individuals (Supplementary Information). A substantial drop in prediction efficiency was observed for the classifier trained on experimental plasma spectra, achieving an accuracy of 0.34. Remarkably, the classifier trained on simulated mixture spectra nearly fully recovered the initial classification efficiency—achieving an accuracy of 0.89. This unexpected finding demonstrated that CODI enabled the creation of a dataset that can even be robust to variations in biological specimen characteristics.

Altogether, this proof-of-principle analysis further demonstrated the potential of CODI to overcome analytical limitations, enabling a classification transfer despite significant measurement deviations. For one, there is no need to re-collect a large number of specimens when deviations occur in the sample collection

procedure. One can leverage a limited set of measurements that characterize differences between specimens, collected in a class-independent fashion. CODI may then extend ML applications to different specimen variations. Nevertheless, here we only demonstrated such potential on serum and plasma. If the specimens widely vary in their molecular composition and reflection of physiology (e.g. blood-based vs. urine- or saliva-based media), this approach may not perform as effectively as demonstrated here. A promising avenue for future exploration may involve adapting a classifier trained on EDTA plasma for use with citrate samples (43).

Generalization to independently acquired datasets

A crucial aspect in determining how well a medical diagnostic assay is likely to perform is to test it on unseen samples. In biodiagnostic applications, a cross-validation procedure is commonly applied to get an estimate of true (external) classification performance. However, if a bias exists in the collected dataset, e.g. confounding information caused by measurement “batch effects,” the estimated performance may not be reproduced when the classifier is truly externally validated (44). To test this in a relevant medical application, we considered our previous work (31)—in which four cancer entities were classified against nonsymptomatic cancer-free controls. In contrast to our prior work which employed

cross-validations (31), here the samples were initially split into a training and a test set, then measured independently (Fig. 5a, left panel).

As a benchmark, we first investigated the performance of classifying each cancer entity, relying on a cross-validation procedure (Fig. 5b, gray curves/bars). The cross-validation was performed

exclusively on the training set of experimental samples and the receiver operating characteristics (ROC) curve of the validation splits was examined for each cancer entity. Next, we trained a classifier on the training set of experimental samples, testing it on the test set of experimental samples (Fig. 5b, black curves/bars). This investigation revealed that the classification efficiency

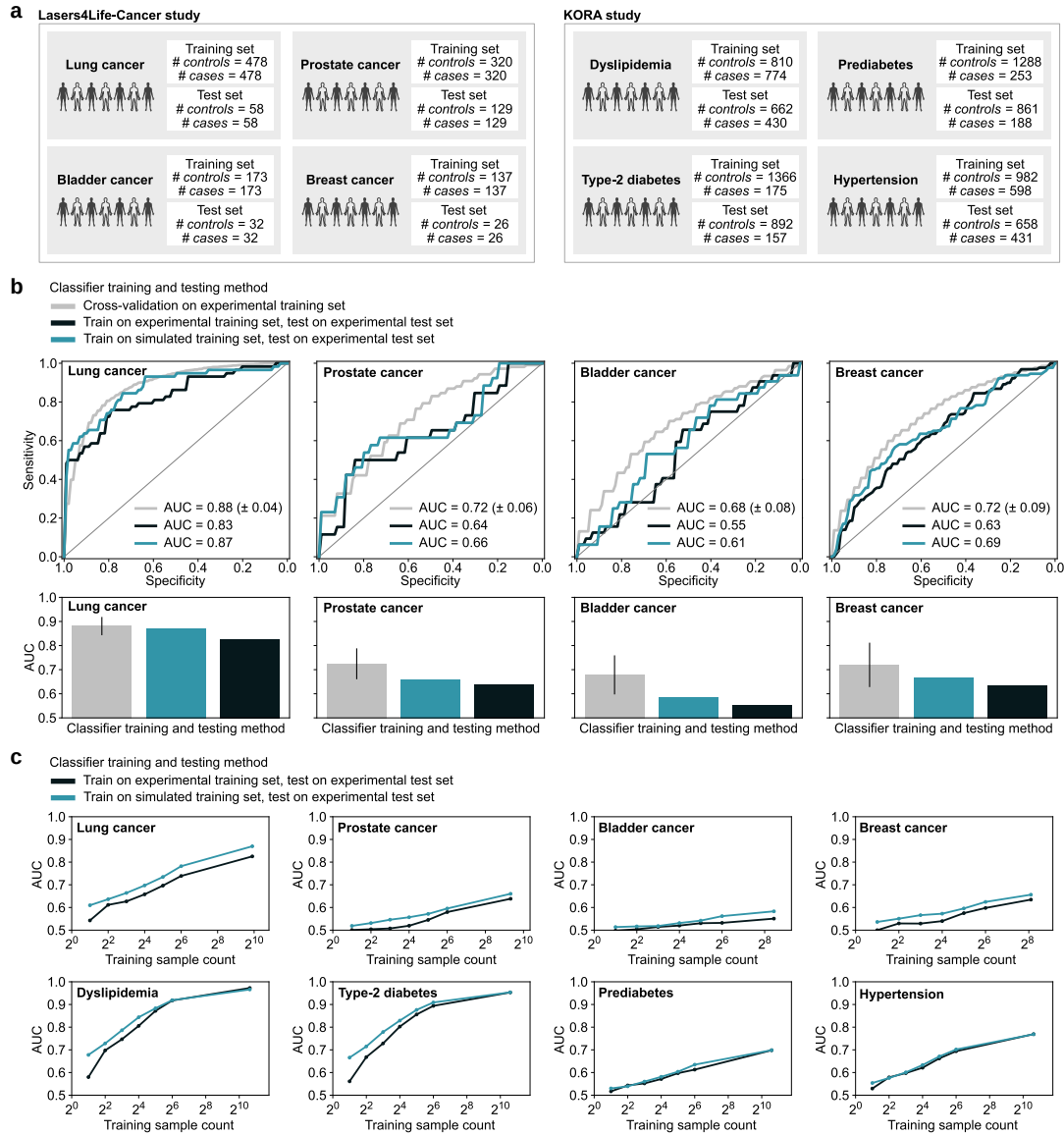


Fig. 5. CODI recovers lost classification efficacy on independently measured case-control test sets. a) Setup of eight binary classifications, spanning diverse health conditions, in two independent clinical studies. IR spectroscopy of blood plasma was performed on all samples. Training and test sample sets were measured independently under different measurement device conditions—including a gap in measurement time, measurement device maintenance, and component replacements. b) Cancer detection was investigated under three different setups of estimating classification efficiency. For the simulation-based training, CODI was employed to introduce measurement variability into the training set measurements. All ROC curves are depicted for the validation/test splits of the data (upper panel), along with the estimated AUCs (lower panel). c) Classification efficiency when training a classifier on experimental observations and applying CODI to train a classifier on simulated data utilizing varying training sample counts as a basis for training. Classifier testing was performed exclusively on held-out experimental test sets.

across the four cancer entities decreased when tested on the later-measured samples. This validated the prior notion that cross-validation estimates may not be entirely reproduced with a train-test split classification setup.

Next, we questioned whether CODI can aid in such a scenario. In principle, by introducing class-invariant empirical variability into a training set of measurements, we can practically make the learning task more difficult for the classifier. Potentially, this would enable the classifier to appropriately weigh features that are more robust to measurement artifacts, making it rely on information that is likely to be reproduced in unseen data.

To test this, we employed CODI to introduce added variance to the training set of measurements (Supplementary Information). Among these were added levels of between-person biological variability, calculated from independent cohorts of individuals, as well as variability from quality control samples—none of which involved the clinical samples from the test sets. Across the four cancer entities, we revealed that an improvement in prediction efficiency was indeed observed when testing the classifiers on the held-out test sets (Fig. 5b, blue curves/bars). The most impressive improvement was for the lung cancer application, where the area under the ROC curve (AUC) was nearly fully recovered and was comparable to the prior cross-validation estimate. For the remaining cancer entities, the CODI framework still provided an advantage, though not to the same extent as lung cancer. This may be partly attributed to the occurrence of measurement artifacts in the training set that happened to correlate with the outcome of interest, leading to an overly optimistic AUC estimate during cross-validation. It may also be partially due to the generally smaller sample sizes used for testing the classifier, and the randomly selected test samples included cases and controls that were more difficult to distinguish than those in the training set (e.g. due to inherent physiological variations that interfere with the cancer signals).

Nevertheless, compared to directly training on experimental observations, including CODI consistently led to improved classification output on independently measured test samples.

Influence of experimental training cohort size

To further examine under which conditions CODI facilitates a more robust classification, we repeated the above case-control investigations but varied the number of experimental observations utilized for training (Fig. 5c). In addition to the previous cancer applications, here we also examined case-control applications involving IR fingerprints from the KORA cohort (33)—focusing on detecting common health physiologies (Fig. 5a, right). Given the longitudinal design of the KORA cohort, we randomly selected 50% of the measurements from the first sample donation for the training set. The second sample donation, measured independently 2.7 years after the first (Supplementary Information), served as the test set, including only samples from individuals not included in the training set.

We then randomly selected samples from the training sets at varying cohort sizes to train several classifiers on each subset of selected samples (Fig. 5c). First, we trained directly on the experimental observations, always testing on the held-out test sets (Fig. 5c, black curves). Unsurprisingly, the smaller the training set was, the worse the classifier performed on the experimental test sets. We then employed CODI to generate simulated datasets that utilized the experimental observations at each sample count as seed input (Fig. 5c, blue curves). Here, it was revealed the classification with CODI almost consistently outperformed the experimental modeling approach across the varying sample counts available as a basis for training. Notably, for the detection of dyslipidemia and

type-2 diabetes, two conditions with strong molecular deviations reflected in IR fingerprints, CODI provided the largest advance when smaller training sample counts were available.

For the detection of prediabetes and hypertension, no clear advantage was observed by incorporating CODI into the classification workflow. This could either stem from the experimental training data already closely resembling the test data distribution, or because the variability introduced by CODI failed to effectively capture the distribution shifts present in the test data. While no advantage was observed for these two conditions, including CODI did not have an adverse impact on the classification. This observation suggested that integrating CODI into the classification pipeline may be an effective standard practice as it either enhances prediction performance or, at minimum, does not impair it.

Altogether, our findings show exciting promise for the proposed CODI framework. The value of the method has been demonstrated in the context of several biomedically relevant applications, where the method achieved improved ML classification output for several practical applications.

Computational efficiency

The computations behind CODI follow a straightforward procedure. Essentially, its practical implementation involves generating random numbers, matrix multiplication, and matrix addition. This allows the strategy to efficiently scale when generating large datasets, depending on the size of the seed input and variability calibration datasets (Fig. S4). The typical runtime to generate a dataset ranges from milliseconds to a few seconds on standard hardware—see Supplementary section “Computational Efficiency.”

Discussion

Multimolecular profiling and computational modeling offer promising avenues to advance our understanding of biological systems. In this study, we introduced CODI, a framework designed to enrich collected datasets to facilitate robust analytics for probing molecular systems. Across several experimental settings, we rigorously tried and tested the framework to demonstrate its validity. We examined how different analytical and biological variations influenced IR molecular fingerprints and revealed the framework's advantage in overcoming the limitations of unrepresentative observational datasets.

Effectively, the datasets generated through CODI enable an ML algorithm to better capture latent information present in a studied dataset, guiding it to distinguish which features are most relevant and reproducible. Such a strategy is particularly valuable when collecting large, representative datasets presents a limitation. This is exemplified in the context of studying pathophysiological phenomena through molecular profiling (e.g. omics studies). In such instances, biological experimentation or medical studies demand substantial involvement, including the probing of a significant number of subjects, considerable time for phenotypic evolution, and the added constraints of intricate sample collection and handling (1, 2, 4, 9, 45–47). Another layer of complexity comes from the fact that biological variations at an organismal level are inherent—due to the dynamics of biological systems and human physiologies (e.g. recycling, turnover, rhythmic oscillations, aging) (1, 48–52). These challenges are further compounded by the often involved quantitative measurement procedures. Factors like the wear and tear of measurement device components, routine maintenance, and sensitivity to environmental conditions all may lead to “batch effects” that are often specific to analytical approaches (3, 5, 7, 20, 53, 54). Ultimately,

these challenges hinder the generalization of insights to unseen, later collected and measured samples.

The OOD generalization problem is well-known in ML research, and the development of methods to address it is likely to receive increasing attention (22, 24, 25). Previous works have proposed several strategies to address domain shifts between source and target data. Some strategies include developing regularizers for learning domain-invariant data representations (55–58), training a collection of models that infer different patterns of the data (59), engineering proxy features that are robust to distributional shifts (60, 61), and employing augmentation techniques through synthetic sample creation (26, 62). Typically, the developed strategies are focused on applications related to image analysis and natural language processing tasks. For instance, data augmentation in image analysis often involves manipulating training instances through geometric modifications (e.g. rotation, skewing, cropping), color adjustments, and noise introductions, which can facilitate generalization to practical variances (26). Data augmentation has also been applied to measurements of biological signals from electroencephalography (EEG), electromyography (EMG), Raman, and near-IR spectra (63–70). While existing augmentation methods often involve random noise introductions, signal warping, and decomposition of available datasets, CODI extends the concept by taking advantage of additional calibration measurements to augment a dataset with actual empirical variance. CODI thus offers a tangential augmentation strategy to model domain shifts, which is especially valuable for molecular fingerprinting applications—where inferring potential measurement variances without empirical observations is challenging. Further research that builds upon OOD concepts can greatly benefit molecular fingerprinting methods, given the difficulties associated with obtaining sufficiently representative datasets.

In our investigations, we addressed the topic of barely supervised learning (71)—where the set of labeled training samples is limited to very few observations per class. Given the experimental constraints of populational sampling over year-long time-frames, we examined whether the number of sequential samplings of the same individual over time could be minimized with CODI. With the example of IR fingerprinting, we surprisingly identified that only a single baseline measurement is sufficient to follow-up and identify an individual in a population at a later time point. Although the identification of the same individual in a heterogeneous population is only a distant approximation to identifying physiologically relevant deviations, it presents a foundation for the concept of longitudinal probing. Our generic framework can be quickly adopted to possibly spare unnecessary samplings and inform future prospective studies.

Further applications of CODI to IR spectroscopic fingerprinting showcased its versatility and potential impact in aiding model generalization. We observed a remarkable level of comparability across experimental data collected over almost a decade, underlining the method's ability to improve classification efficacy on independently measured test sets. The adaptability of the framework extended to a proof-of-principle application that involved training a classifier on one sample medium (plasma) and applying it to another (serum). This application demonstrated the potential of CODI to streamline cross-specimen dataset analyses in various biological and biomedical applications. Such a strategy may be particularly valuable when gathering data sets from retrospective studies or online repositories to help ensure specimen comparability to another envisioned application—i.e. domain adaptation applications (72, 73). Another promising use of CODI would be to harmonize data obtained from different

measurement devices (e.g. several spectrometers made by the same or different manufacturers), potentially improving ML model transferability between them (5).

It is imperative to emphasize that our proposed method shall not be positioned as a replacement for improving study designs, better standardization of classical analytical procedures, and computational data preprocessing steps. Such aspects remain essential when establishing a molecular profiling platform to help satisfy i.i.d. assumptions between train-test datasets. In addition to efforts aimed at ensuring train-test dataset comparability, the motivation of our method is to work around the instances when the assumption is violated due to inevitable sources of error and variability that cannot be eliminated.

An inherent limitation of the proposed method is its reliance on *a priori* knowledge about the sources of possible empirical variations. This presents a challenge as gathering such information may require extensive experimental evaluations of biological and analytical variations. Furthermore, the characterized sources of variability must adequately represent the true possible domain of empirical variability for any successful application. Therefore, continuous refinement through more controlled experiments to include several, independent, sources of variance holds potential to further enhance the utility of the framework. Nevertheless, once the domain of possible variance is successfully characterized for a given molecular system and analytical procedure, the same characterizations may be repeatedly utilized in diverse applications.

While our practical investigations focused on blood-based IR spectroscopy to aid *in vitro* diagnostics, applications of CODI are not limited to this context. The principle and mathematical foundation of CODI are sufficiently generic to be translated to examining diverse biological systems, medical problems, measurement modalities, and ML tasks. Applications involving NMR spectroscopy, mass spectrometry, and Raman spectroscopy serve as direct extensions that can be explored with the same approach and code implementation. Additionally, CODI holds potential for applications related to cell typing and the integration of single-cell multi-modal omics data—given the inherent challenges associated with obtaining accurate measurements of cell type populations at scale, where out-of-distribution measurement events are prevalent (74, 75). Altogether, the presented framework establishes a foundation for future explorations to enhance the robustness of molecular analytics while conserving the resources required to gather representative datasets.

Materials and methods

Additional details on implementing CODI, our applications, datasets, experimental procedures, and ML analyses are provided in the [Supplementary Information](#).

In brief, the simulation model behind CODI is designed for extensibility, such that it can be applied in various applications and for different measurement modalities. The modeling framework involves utilizing seed observations ($\mathbf{s}_i \mid i = 1, \dots, m$) that are representative of properties intrinsic to a phenomenon of interest. For instance, a seed observation might represent the mean observation of a class of samples (e.g. healthy control), while another corresponds to a different class (e.g. disease sample). Variability is introduced to the seed observations $\mathbf{s}_i$ through the addition of random functions $f_1, f_2, \dots, f_m$, simulating a measurement as a statistical variable $\mathbf{Y}$ in the following generalized form:

$$\mathbf{Y} = \mathbf{s}_i + f_1 + f_2 + \dots + f_m.$$

Repeatedly applying the above model would generate a cohort of simulated measurements, in arbitrary size, centered around $\mathbf{s}$, and incorporating the variations introduced by $f_1, f_2, \dots, f_m$.

The functions $f_1, f_2, \dots, f_m$ can be characterized by *ab initio* calculations or bottom-up models that each represent a source of expected data variance. However, the former is often specialized and problem-specific. An alternative descriptive approach that relies on collecting datasets of calibration measurements which incorporate the levels of expected variance can be easily applied to a variety of problems. For instance, quality control samples can be subjected to different freezer-storage durations, number of freeze/thaw cycles, and aliquoting of samples by different operators. The quality control samples can then be repeatedly measured under different measurement device conditions. The variance observed in this calibration dataset would be reflective of potential sources of variance in handling and measuring samples from the original dataset modeling a certain phenomenon (e.g. biofluids of cases and controls). This variance can then be introduced by defining the function f_1 . Other potential sources of data variance, such as biological variability, can be modeled by using additional calibration datasets reflective of the variability sources.

When employing CODI, it is crucial to ensure that the introduced variability is derived from samples independent of the test samples to prevent data leakage. Across our applications, several sources of variability were modeled. For the variability calculated from QC (commercially purchased pooled human serum) and technical replicate (water) samples, there is no risk of data leakage as these samples are independent of all clinical study samples. The remaining sources of variability were derived from measurements of clinical study samples. To eliminate the risk of data leakage when investigating a classification task in one of the clinical studies, any additional variance missing from the experimental training set was modeled from the other, entirely independent clinical studies. Thus, the test sets were completely held-out when simulating data and training classifiers.

Four independent study cohorts were utilized in this study: Lasers4Life-LG (30), BioPersMed (28), KORA (29, 33), and Lasers4Life-Cancer (31). Lasers4Life-LG involved the collection of blood serum and plasma from 31 healthy individuals initially sampled up to 13 times over a 7-week period, with an additional follow-up after 6 months as detailed in a previous publication (30). Since this initial publication, the same individuals were invited to participate in two additional sample donations at 3.5 and 4.5 years post their initial involvement. The Lasers4Life-LG study was approved by the Ethics Committee of the Ludwig-Maximilian-University (LMU) of Munich and all participants provided written informed consent (research study protocol #17-532). BioPersMed is an ongoing population-based cohort performed at the Medical University Graz, Austria (28). Repetitive examinations of participants were conducted in 2-year intervals. In the current study, we utilized blood plasma samples and medical data from a subset of 44 healthy individuals (out of 1,022 participants). The BioPersMed study was approved by the Ethics Committee of the Medical University of Graz, Austria (EC Nr. 24-224 ex 11/12; project application number 4008_22). KORA is a population-based cohort in Southern Germany (29). The cohort comprised of an age- and gender-stratified sample of participants randomly drawn from the resident registration offices within the study area. In the current study, we utilized blood plasma samples and medical data from the second and third visits (named KORA-F4 and KORA-FF4, respectively). The available KORA-F4 data consisted of 3,044 samples, while the KORA-FF4 data consisted of 2,140

samples. A subset of 2015 individuals participated in both samplings, while 1,154 individuals participated in only one of the samplings. The KORA-F4 and KORA-FF4 study methods were approved by the ethics committee of the Bavarian Chamber of Physicians, Munich (EC No. 06068). Lasers4Life-Cancer is a case-control study cohort involving several cancer entities where both serum and plasma are collected (31). The samples utilized in this study largely overlapped with samples from our previous study (31)—although the measurement procedures differed (Supplementary Information). Since the original publication, blood plasma and serum samples from different individuals were collected and included in this study. Case samples were collected prior to cancer-related treatment. Nonsymptomatic controls were pair-matched to cancer cases by age, gender, and body mass index. All participants in the Lasers4Life-Cancer study provided written informed consent for the study under research study protocol #17-141 and under research study protocol #17-182, both of which were approved by the Ethics Committee of the LMU of Munich. The clinical trial is registered at the German Clinical Trials Register (ID DRKS00013217).

Experimental measurements were performed on a Fourier transform IR (FTIR) spectrometer, with liquid samples injected into a flow cell and spectra recorded in transmission mode as in previous studies (30, 31, 33). The spectrometer underwent routine maintenance, with components replaced as needed throughout all measurements (Supplementary Information).

Multiclass classifications were performed using a nearest neighbor algorithm for applications involving a single training instance per class. For multiclass applications involving more than more training instance per class, a linear discriminant analysis (LDA) algorithm was applied. PCA was applied prior to multiclass classifications (Fig. S5 for details on the effects of PCA). Binary classifications were performed using a logistic regression with a ridge penalty. All classification metrics were reported on held-out test experimental samples. Additional details are provided in the Supplementary Information.

Acknowledgments

We thank Jacqueline Hermann for clinical study management and manuscript feedback, Viola Zoka, and Daniel Meyer for measurement contributions, and Michael Trubetskov for manuscript feedback.

Supplementary Material

Supplementary material is available at PNAS Nexus online.

Funding

This work was supported by LMU Munich, Centre for Advanced Laser Applications (CALA). The KORA study was initiated and financed by Helmholtz Zentrum München–German Research Center for Environmental Health, which is funded by the German Federal Ministry of Education and Research (BMBF) and by the State of Bavaria. Data collection in the KORA study is done in cooperation with the University Hospital of Augsburg. The BioPersMed cohort study was funded by the Austrian Research Fund, as a COMET K-project (No. 825329), supported by the Austrian Federal Ministry of Transport, Innovation and Technology (BMVIT) and the Austrian Federal Ministry of Economics and Labour/the Federal Ministry of Economy, Family and Youth (BMWA/BMWFJ) as well as the Styrian Business Promotion Agency (SFG).

Author Contributions

T.E. conceived the project; T.E., M.H., and M.Ž. contributed to study design; T.E. performed the research and analyzed the data with assistance from M.H. and M.Ž.; M.Ž., B.O.P., B.L., and A.P. oversaw, organized, and coordinated the clinical studies; F.F. supervised and contributed to the experimental measurements and coordinated data transfers; T.E. and M.Ž. wrote the manuscript with assistance from M.H. and F.F.; All authors revised and approved the manuscript.

Preprint

This manuscript was posted on a preprint: <https://doi.org/10.1101/2024.06.15.598503>.

Data Availability

All data supporting the findings of this study are included within the article and the [Supplementary Information](#). Requests for data relating to the KORA study should be sent to kora.passt@helmholtz-muenchen.de and are subject to approval by the KORA Board. BioPersMed cohort data relevant to this study are available upon reasonable request (barbara.obermayer@medunigraz.at). Data from Lasers4Life-LG and Lasers4Life-Cancer studies can not be publicly available due to privacy regulations under GDPR (EU) 2016/679, reasonable requests can be addressed to the responsible scientist (mihaela.zigman@mpq.mpg.de). An implementation of CODI is available as a Python package (pycodi). Please refer to the following GitHub repository: <https://github.com/tarek-eissa/codi>.

References

- Batool SM, et al. 2023. Extrinsic and intrinsic preanalytical variables affecting liquid biopsy in cancer. *Cell Rep Med*. 4(10):101196.
- Bowen RA, Remaley AT. 2014. Interferences from blood collection tube components on clinical chemistry assays. *Biochem Med (Zagreb)*. 24(1):31–44.
- Čuklina J, et al. 2021. Diagnostics and correction of batch effects in large-scale proteomic studies: a tutorial. *Mol Syst Biol*. 17(8):e10240.
- Dvinge H, et al. 2014. Sample processing obscures cancer-specific alterations in leukemic transcriptomes. *Proc Natl Acad Sci U S A*. 111(47):16802–16807.
- Guo S, et al. 2020. Comparability of Raman spectroscopic configurations: a large scale cross-laboratory study. *Anal Chem*. 92(24):15745–15756.
- Kwak JT, Reddy R, Sinha S, Bhargava R. 2011. Analysis of variance in spectroscopic imaging data from human tissues. *Anal Chem*. 84(2):1063–1069.
- Morais CLM, Lima KMG, Singh M, Martin FL. 2020. Tutorial: multivariate classification for vibrational spectroscopy in biological samples. *Nat Protoc*. 15(7):2143–2162.
- Perez-Guaita D, Ventura-Gayete J, Pérez-Rambla C, Sancho-Andreu M, Garrigues S. 2013. Evaluation of infrared spectroscopy as a screening tool for serum analysis. *Microchem J*. 106:202–211.
- Yin P, Lehmann R, Xu G. 2015. Effects of pre-analytical processes on blood samples used in metabolomics studies. *Anal Bioanal Chem*. 407(17):4879–4892.
- Check E. 2004. Proteomics and cancer: running before we can walk? *Nature*. 429(6991):496–497.
- Cohen JP, et al. 2021. Problems in the deployment of machine-learned models in health care. *Can Med Assoc J*. 193(35):E1391–E1394.
- Goetz L, Seedat N, Vandersluijs R, van der Schaar M. 2024. Generalization—a key challenge for responsible AI in patient-facing clinical applications. *NPJ Digit Med*. 7(1):126.
- Moreno-Torres JG, Raeder T, Alaiz-Rodríguez R, Chawla NV, Herrera F. 2012. A unifying view on dataset shift in classification. *Pattern Recognit*. 45(1):521–530.
- Obermeyer Z, Emanuel EJ. 2016. Predicting the future—big data, machine learning, and clinical medicine. *N Engl J Med*. 375(13):1216–1219.
- Zech JR, et al. 2018. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med*. 15(11):e1002683.
- Gagnon-Bartsch JA, Speed TP. 2011. Using control genes to correct for unwanted variation in microarray data. *Biostatistics*. 13(3):539–552.
- Livera AMD, et al. 2015. Statistical methods for handling unwanted variation in metabolomics data. *Anal Chem*. 87(7):3606–3615.
- Molania R, et al. 2022. Removing unwanted variation from large-scale RNA sequencing data with PRPS. *Nat Biotechnol*. 41(1):82–95.
- Molania R, Gagnon-Bartsch JA, Dobrovic A, Speed TP. 2019. A new normalization for nanostring nCounter gene expression data. *Nucleic Acids Res*. 47(12):6073–6083.
- Peng M, Li Y, Wamsley B, Wei Y, Roeder K. 2021. Integration and transfer learning of single-cell transcriptomes via cFIT. *Proc Natl Acad Sci U S A*. 118(10):e2024383118.
- Chong Y, et al. 2023. Machine learning of spectra-property relationship for imperfect and small chemistry data. *Proc Natl Acad Sci U S A*. 120(20):e2220789120.
- Liu J, et al. 2023. Towards out-of-distribution generalization: a survey, arXiv, arXiv:2108.13624, preprint: not peer reviewed. <https://doi.org/10.48550/arXiv.2108.13624>
- Mirkes EM, et al. 2022. Domain adaptation principal component analysis: base linear method for learning with out-of-distribution data. *Entropy*. 25(1):33.
- Li X, et al. 2022. Uncertainty modeling for out-of-distribution generalization. In: International Conference on Learning Representations (ICLR).
- Zhang X, et al. 2022. Towards unsupervised domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). p. 4910–4920.
- Mikolajczyk A, Grochowski M. 2018. Data augmentation for improving deep learning in image classification problem. In: 2018 International Interdisciplinary PhD Workshop (IIPhDW). IEEE.
- Shorten C, Khoshgoftaar TM, Furht B. 2021. Text data augmentation for deep learning. *J Big Data*. 8(1):101.
- Haudum CW, et al. 2022. Cohort profile: 'biomarkers of personalised medicine' (BioPersMed): a single-centre prospective observational cohort study in Graz/Austria to evaluate novel biomarkers in cardiovascular and metabolic diseases. *BMJ Open*. 12(4):e058890.
- Holle R, Happich M, Löwel H, Wichmann H. 2005. KORA - a research platform for population based health research. *Das Gesundheitswesen*. 67(S 01):19–25.
- Huber M, et al. 2021. Stability of person-specific blood-based infrared molecular fingerprints opens up prospects for health monitoring. *Nat Commun*. 12(1):1511.
- Huber M, et al. 2021. Infrared molecular fingerprinting of blood-based liquid biopsies for the detection of cancer. *eLife*. 10:e68758.

- 32 Eissa T, Kepesidis KV, Zigman M, Huber M. 2023. Limits and prospects of molecular fingerprinting for phenotyping biological systems revealed through in silico modeling. *Anal Chem.* 95(16): 6523–6532.
- 33 Eissa T, et al. 2024. Plasma infrared fingerprinting with machine learning enables single-measurement multi-phenotype health screening. *Cell Rep Med.* 5(7):101625.
- 34 Assfalg M, et al. 2008. Evidence of different metabolic phenotypes in humans. *Proc Natl Acad Sci U S A.* 105(5):1420–1424.
- 35 Wallner-Liebmann S, et al. 2016. Individual human metabolic phenotype analyzed by ¹H-NMR of saliva samples. *J Proteome Res.* 15(6):1787–1793.
- 36 Yousri NA, et al. 2014. Long term conservation of human metabolic phenotypes and link to heritability. *Metabolomics.* 10(5): 1005–1017.
- 37 Moqri M, et al. 2023. Biomarkers of aging for the identification and evaluation of longevity interventions. *Cell.* 186(18): 3758–3775.
- 38 Raudys S, Jain A. 1991. Small sample size effects in statistical pattern recognition: recommendations for practitioners. *IEEE Trans Pattern Anal Mach Intell.* 13(3):252–264.
- 39 Baker MJ, et al. 2014. Using fourier transform IR spectroscopy to analyze biological materials. *Nat Protoc.* 9(8):1771–1791.
- 40 Chetwynd AJ, Dunn WB, Rodriguez-Blanco G. 2017. Collection and preparation of clinical samples for metabolomics. In: *Advances in experimental medicine and biology.* Springer International Publishing. p. 19–44.
- 41 Yu Z, et al. 2011. Differences between human plasma and serum metabolite profiles. *PLoS One.* 6(7):e21230.
- 42 Eissa T, Voronina L, Huber M, Fleischmann F, Žigman M. 2024. The perils of molecular interpretations from vibrational spectra of complex samples. *Angewandte Chemie International Edition*, pages Accepted Manuscript, In Press.
- 43 Stanisewska E, Bartosz AK, Malek K, Baranska M. 2013. An effect of anticoagulants on the FTIR spectral profile of mice plasma. *Biomed Spectrosc Imaging.* 2(4):317–330.
- 44 Soneson C, Gerster S, Delorenzi M. 2014. Batch effect confounding leads to strong bias in performance estimates obtained by cross-validation. *PLoS One.* 9(6):e100335.
- 45 Cameron JM, et al. 2020. Exploring pre-analytical factors for the optimisation of serum diagnostics: progressing the clinical utility of ATR-FTIR spectroscopy. *Vib Spectrosc.* 109: 103092.
- 46 González-Domínguez R, González-Domínguez Á, Sayago A, Fernández-Recamales Á. 2020. Recommendations and best practices for standardizing the pre-analytical processing of blood and urine samples in metabolomics. *Metabolites.* 10(6):229.
- 47 Pérez-Guaita D, Quintás G, Farhane Z, Tauler R, Byrne HJ. 2022. Combining pharmacokinetics and vibrational spectroscopy: MCR-ALS hard-and-soft modelling of drug uptake in vitro using tailored kinetic constraints. *Cells.* 11(9):1555.
- 48 Eling N, Morgan MD, Marioni JC. 2019. Challenges in measuring and understanding biological noise. *Nat Rev Genet.* 20(9): 536–548.
- 49 Hawkrigde AM, Muddiman DC. 2009. Mass spectrometry-based biomarker discovery: toward a global proteome index of individuality. *Annu Rev Anal Chem.* 2(1):265–277.
- 50 López-Otín C, Kroemer G. 2021. Hallmarks of health. *Cell.* 184(1): 33–63.
- 51 López-Otín C, Blasco MA, Partridge L, Serrano M, Kroemer G. 2013. The hallmarks of aging. *Cell.* 153(6):1194–1217.
- 52 Rose SMS-F, et al. 2019. A longitudinal big data approach for precision health. *Nat Med.* 25(5):792–804.
- 53 Guo S, Popp J, Bocklitz T. 2021. Chemometric analysis in Raman spectroscopy from experimental design to machine learning-based modeling. *Nat Protoc.* 16(12):5426–5459.
- 54 Haghverdi L, Lun ATL, Morgan MD, Marioni JC. 2018. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol.* 36(5): 421–427.
- 55 Arjovsky M, Bottou L, Gulrajani I, Lopez-Paz D. 2019. Invariant risk minimization, arXiv, arXiv:1907.02893, preprint: not peer reviewed. <https://doi.org/10.48550/arXiv.1907.02893>
- 56 Bellot A, van der Schaar M. 2022. Accounting for unobserved confounding in domain generalization, arXiv, arXiv:2007.10653, preprint: not peer reviewed. <https://doi.org/10.48550/arXiv.2007.10653>
- 57 Krueger D, et al. 2021. Out-of-distribution generalization via risk extrapolation (rex). In: *Proceedings of the 38th International Conference on Machine Learning of Proceedings of Machine Learning Research.* vol. 139. PMLR. p. 5815–5826.
- 58 Sun B, Saenko K. 2016. Deep coral: correlation alignment for deep domain adaptation. In: *Hua G, Jégou H, editors. Computer Vision – ECCV 2016 Workshops.* Cham: Springer International Publishing. p. 443–450.
- 59 Teney D, Abbasnejad E, Lucey S, Van den Hengel A. 2022. Evading the simplicity bias: training a diverse set of models discovers solutions with superior OOD generalization. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* p. 16761–16772.
- 60 Li C, Zhang D, Huang W, Zhang J. 2023. Cross contrasting feature perturbation for domain generalization. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision.* p. 1327–1337.
- 61 Mazaheri B, Mastakouri A, Janzing D, Hardt M. 2023. Causal information splitting: engineering proxy features for robustness to distribution shifts. In: *Shpitser I ERJ, editor. Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence of Proceedings of Machine Learning Research.* vol. 216. PMLR. p. 1401–1411.
- 62 Yao H, et al. 2022. Improving out-of-distribution robustness via selective augmentation. In: *International Conference on Machine Learning.* PMLR. p. 25407–25437.
- 63 Bjerrum EJ, Glahder M, Skov T. 2017. Data augmentation of spectral data for convolutional neural network (CNN) based deep chemometrics, arXiv, arXiv:1710.01927, preprint: not peer reviewed. <https://doi.org/10.48550/arXiv.1710.01927>
- 64 Freer D, Yang G-Z. 2020. Data augmentation for self-paced motor imagery classification with c-LSTM. *J Neural Eng.* 17(1): 016041.
- 65 Guo S, et al. 2018. Model transfer for Raman-spectroscopy-based bacterial classification. *J Raman Spectrosc.* 49(4):627–637.
- 66 Lebrun A, et al. 2022. Pushing the limits of surface-enhanced Raman spectroscopy (SERS) with deep learning: identification of multiple species with closely related molecular structures. *Appl Spectrosc.* 76(5):609–619.
- 67 Lotte F. 2015. Signal processing approaches to minimize or suppress calibration time in oscillatory activity-based brain-computer interfaces. *Proc IEEE.* 103(6):871–890.
- 68 Tsinganos P, Cornelis B, Cornelis J, Jansen B, Skodras A. 2020. Data augmentation of surface electromyography for hand gesture recognition. *Sensors.* 20(17):4892.
- 69 Wang F, Zhong S, Peng J, Jiang J, Liu Y. 2018. Data augmentation for EEG-based emotion recognition with deep convolutional neural networks. In: *MultiMedia modeling.* Springer International Publishing. p. 82–93.

-
- 70 Zanini RA, Colombini EL. 2020. Parkinson's disease EMG data augmentation and simulation with DCGANs and style transfer. *Sensors*. 20(9):2605.
- 71 Sohn K, et al. 2020. Fixmatch: simplifying semi-supervised learning with consistency and confidence. *Adv Neural Inf Process Syst*. 33:596–608.
- 72 Bareinboim E, Pearl J. 2016. Causal inference and the data-fusion problem. *Proc Natl Acad Sci U S A*. 113(27): 7345–7352.
- 73 Kyono T, van der Schaar M. 2021. Exploiting causal structure for robust model selection in unsupervised domain adaptation. *IEEE Trans Artif Intell*. 2(6):494–507.
- 74 Argelaguet R, Cuomo ASE, Stegle O, Marioni JC. 2021. Computational principles and challenges in single-cell data integration. *Nat Biotechnol*. 39(10):1202–1215.
- 75 Dorkenwald S, et al. 2023. Multi-layered maps of neuropil with segmentation-guided contrastive learning. *Nat Methods*. 20(12): 2011–2020.

SUPPLEMENTARY INFORMATION

CODI: Enhancing machine learning-based molecular profiling through contextual out-of-distribution integration

Tarek Eissa^{a,b,c,*}, Marinus Huber^{a,b}, Barbara Obermayer-Pietsch^d, Birgit Linkohr^e, Annette Peters^{e,f}, Frank Fleischmann^{a,b}, Mihaela Žigman^{a,b,*}

^a Ludwig-Maximilians-Universität München, Chair of Experimental Physics - Laser Physics, Garching, Germany

^b Max Planck Institute of Quantum Optics, Laboratory for Attosecond Physics, Garching, Germany

^c Technical University of Munich, School of Computation, Information and Technology, Garching, Germany

^d Department of Internal Medicine, Division of Endocrinology and Diabetology, Medical University, Graz, Austria

^e Helmholtz Zentrum München, Institute of Epidemiology, Neuherberg, Germany

^f Chair of Epidemiology, Institute for Medical Information Processing, Biometry and Epidemiology, Medical Faculty, Ludwig-Maximilians-Universität München, Munich, Germany

* Correspondence: tarek.eissa@mpq.mpg.de; mihaela.zigman@mpq.mpg.de

CONTENTS

1	Extended Methods	S2
2	Choosing the variability functions	S7
3	Comparative analysis of CODI to domain-agnostic augmentation techniques	S8
4	Effects of introducing an excessive amount of variance	S9
5	Computational efficiency	S10
6	Effects of principal component analysis on classification efficiency	S11
	References	S12

1 EXTENDED METHODS

SIMULATION MODEL

In the following subsections, we provide a general framework of the simulation model behind CODI to facilitate its extensibility to other applications and molecular fingerprinting modalities. Technical details on our specific applications of the model are provided later in the text.

Generalized representation

CODI relies on introducing information about possible sources of variations (biological, technical, etc.) that may arise in an analytical setting to an existing set of experimental observations $\{\mathbf{x}_i \mid i = 1, \dots, n\}$ where each $\mathbf{x}_i$ is a numerical vector. In a generalized form, the sources of variability can be denoted as functions $f_1, f_2, \dots, f_m$, each representing a model for distinct aspects of variation. These functions are assumed to be random vectors in the same space as the experimental observations, and taking a probability distribution centered around 0.

From the set of experimental observations $\mathbf{x}_i$, we can extract another set of experimentally-derived seed observations $\{\mathbf{s}_i \mid i = 1, \dots, m\}$ that model intrinsic properties of a sub-group in the original set of data – e.g., different classes. With these, a resulting simulated measurement can be modeled as a statistical variable $\mathbf{Y}$, using the following generalized form:

$$\mathbf{Y} = \mathbf{s}_i + f_1 + f_2 + \dots + f_m \quad (1)$$

Repeatedly applying the above model would therefore generate a cohort of simulated measurements, in arbitrary size, centered around $\mathbf{s}_i$ and incorporating variations introduced by $f_1, f_2, \dots, f_m$.

Setting the seed measurements

Defining the experimentally-derived seed measurements $\mathbf{s}_i$ depends on the available input dataset of experimental measurements $\mathbf{x}_i$ and the type of measurement event we wish to simulate.

In a straightforward implementation, we can set $\mathbf{s}_i = \mathbf{x}_i$. This formulation would allow us to introduce a level of variability to each measurement-specific observation. Repeatedly applying the model for each $\mathbf{x}_i$ would create several sets of measurements, where each set is centered around each experimental observation. In an alternative definition, we can set $\mathbf{s}_i = \bar{\mathbf{x}}$. This would allow us to introduce a level of variability to the mean measurement of the experimental set $\mathbf{x}_i$, thereby, creating a cohort of measurements centered around their expected value.

If all measurements in $\mathbf{x}_i$ reflect a certain group or class of samples (e.g., healthy individuals), the simulated cohort as a whole would reflect a measurement distribution of that class of samples. Trivially, when the dataset $\mathbf{x}_i$ is switched to involve measurements representative of a different class (e.g., disease cases), the simulated cohort would reflect a distribution centered around the alternative outcome.

Introduction of variability modeling functions

The variations introduced by the functions $f_1, f_2, \dots, f_m$ may be characterized by *ab initio* calculations or bottom-up models that each represent a source of expected data variance. However, the former is highly specialized and problem-specific. An alternative descriptive approach that relies on collecting datasets of calibration measurements which incorporate the levels of expected variance can be easily applied to a variety of problems.

Suppose an independent calibration dataset $\{\mathbf{u}_i \mid i = 1, \dots, l\}$ is available that reflects a given source of measurement variability, where each $\mathbf{u}_i$ is a numerical vector in the same dimensions as the experimental seeds $\mathbf{x}_i$. Here, f_1 can take the following form:

$$f_1 = \sum_i \beta \left(0, \frac{1}{\sqrt{l}}\right) \cdot (\mathbf{u}_i - \bar{\mathbf{u}}) \quad (2)$$

where $\beta \left(0, \frac{1}{\sqrt{l}}\right)$ represents one random number drawn from a Gaussian distribution with mean 0 and standard deviation $\frac{1}{\sqrt{l}}$. By scaling each mean-centered calibration measurement with β and aggregating over all l measurements, the random function f_1 would have the same variance as the dataset $\{\mathbf{u}_i \mid i = 1, \dots, l\}$. The formal and empirical proof for this was described in our previous study where f_1 generated a level of measurement variance that is observed between different individuals [1]. If the variance source in question does not follow a Gaussian distribution, the random variable β may assume a more appropriate probability distribution, or methods of transforming features to normality may be employed.

Alternative to the representation in Eq. 2 that utilizes mean-centered observations, the function f_1 can calculate the differences between pairs of observations:

$$f_1 = \sum_{i,j} \beta \left(0, \frac{1}{\sqrt{l}}\right) \cdot (\mathbf{u}_i - \mathbf{u}_j) \quad (3)$$

For instance, in our applications of CODI, we calculated the differences between pairs of serum and plasma spectra that were processed from the same whole blood sample. This representation is thus more appropriate when the calibration measurements are derived from sources that share a common origin.

Additional functions $f_2, f_3, \dots, f_m$ may be modeled similar to Eqs. 2 and 3, by utilizing different calibration datasets that reflect other sources of measurement variability. In our example applications of the model in this study, the utilized calibration datasets reflected levels of within- or between-person biological variability, along with several sources of analytical variability.

Descriptions on the model variants applied in our applications of the model, along with of how the calibration datasets were defined, is detailed in following sections of the text.

Simulating case-control measurements

For analysis that involved simulating measurements of cases and controls in our applications (e.g., detection of cancer), the input dataset took the following form, where the superscript denotes a case or control:

$$\mathbf{D} = \{\mathbf{x}_1^{(+1)}, \mathbf{x}_2^{(-1)}, \mathbf{x}_3^{(-1)}, \dots, \mathbf{x}_n^{(+1)}\} \quad (4)$$

Two model definitions were applied based on Eq. 1, once by setting $\mathbf{s}_i = \bar{\mathbf{x}}^{(+1)}$ and once by setting $\mathbf{s}_i = \bar{\mathbf{x}}^{(-1)}$. This results in the following model formulations:

$$\begin{aligned} \mathbf{Y}^{(+1)} &= \bar{\mathbf{x}}^{(+1)} + f_1 + f_2 + \dots + f_m \\ \mathbf{Y}^{(-1)} &= \bar{\mathbf{x}}^{(-1)} + f_1 + f_2 + \dots + f_m \end{aligned} \quad (5)$$

Both model variants were then repeatedly applied to generate a dataset containing case and control measurements for each disease detection application separately.

Simulating longitudinally-captured measurements

For an application that involved simulating longitudinal measurements of the same individuals, the input dataset took the following form:

$$\mathbf{D} = \{\mathbf{x}_0^{(a)}, \mathbf{x}_1^{(a)}, \mathbf{x}_2^{(a)}, \mathbf{x}_0^{(b)}, \dots, \mathbf{x}_0^{(z)}, \mathbf{x}_1^{(z)}\} \quad (6)$$

where the superscript denotes different individuals and the subscript denotes their visit number.

For each individual, we may then utilize their baseline measurement (visit 0) to simulate a set of measurements centered around that baseline in the following form:

$$\begin{aligned} \mathbf{Y}^{(a)} &= \mathbf{x}_0^{(a)} + f_1 + f_2 + \dots + f_m \\ \mathbf{Y}^{(b)} &= \mathbf{x}_0^{(b)} + f_1 + f_2 + \dots + f_m \\ &\vdots \\ \mathbf{Y}^{(z)} &= \mathbf{x}_0^{(z)} + f_1 + f_2 + \dots + f_m \end{aligned} \quad (7)$$

Rather than only using the baseline measurements, the same concept can be extended to incorporate additional measurements obtained during subsequent visits for each individual. The process can therefore generate an arbitrary number of measurements for each individual – whether based on a single baseline measurement or more.

Simulating between-person biological variability

To model variations in measurements between different individuals, the calibration dataset $\mathbf{u}_i$ should include experimentally obtained measurements of different individuals. Ideally, all measurements captured for different individuals should be performed in a similar experimental workflow, following the same analytical protocol. This helps ensure that the variance of the calibration dataset is driven primarily by the biological differences between different individuals, excluding the contributions of other sources of data variance. Here, f_1 may be set to model the between-person biological variability and would be defined in the same way described in Eq. 2.

Simulating within-person biological variability

To model variations in measurements of the same individual over time, the calibration dataset $\mathbf{u}_i$ should include several experimentally obtained measurements per individual. Similar to the longitudinal dataset definition provided in Eq. 6, we can formulate another calibration dataset of l entries that model within-person deviations in the following form:

$$\mathbf{D} = \{\mathbf{u}_0^a - \bar{\mathbf{u}}^a, \mathbf{u}_1^a - \bar{\mathbf{u}}^a, \mathbf{u}_2^a - \bar{\mathbf{u}}^a, \dots, \mathbf{u}_0^z - \bar{\mathbf{u}}^z, \mathbf{u}_1^z - \bar{\mathbf{u}}^z\} \quad (8)$$

In other terms, the calibration vectors are mean-centered per individual across their longitudinal visits and the whole set of individual-specific deviations are pooled into one dataset.

Utilizing the dataset defined in Eq. 8, the function f_1 can be set to model within-person biological variability and would be defined similarly to Eq. 2 with the following form:

$$f_1 = \sum_{d \in \mathbf{D}} \beta \left(0, \frac{1}{\sqrt{l}}\right) \cdot \mathbf{d} \quad (9)$$

With this, the function f_1 thus models the within-person variability as observed across multiple individuals.

Simulating analytical variability

To model variations that may arise from a given analytical procedure, additional experimental measurements that are subjected to experimental variations introduced during the analytical procedure can be utilized. This includes variations that may arise from sample preparation, environmental factors, and instrumental errors. In general, describing the reproducibility of collected data can be achieved by conducting multiple measurements of replicates. Specifically defining how to handle and measure the replicates is dependent on the analytical workflow applied and the measurement technique.

Measurements of replicates can be performed on samples with a high level of technical reproducibility (e.g., water samples) or on quality control samples that exhibit stability in chemical composition similar to that of the intended application (e.g., blood-based samples pooled from thousands of individuals). With blood-based quality control samples, for instance, variations could further include freezer-storage durations, number of freeze/thaw cycles, aliquoting of samples by different operators, varied experimental laboratory room temperature, and sample storage in tubes from different manufacturers.

Technicalities specific to the measurement device may also be intentionally varied when performing measurements – e.g., filter changes, software versions, and cuvette lifetimes. The replicate samples may also be performed across multiple measurement devices, and environmental conditions, including temperature and humidity, to further enrich a calibration dataset.

The variations that arise from the repeated replicate measurements would then be modeled in the same fashion described in Eq. 2. f_2 , for instance, may encompass all the sources of measurement variability that are reflected in a blood-based quality control sample, while, f_3 may encompass the sources of measurement variability that are reflected in water measurements.

Modeling such factors, in a way that best mimics what may be encountered in application, makes a given training dataset more resilient to molecular changes or measurement device operating conditions that may arise from an analytical workflow.

Defining which variability functions to include

Successful applications of the CODI framework relies on the careful selection of appropriate sources of variability $f_1, f_2, \dots, f_m$. This selection depends on the application of interest and what sources of measurement variability are expected to be observed. For instance, if the envisioned application is to model case-control measurements, it is more critical to include a level of between-person biological variability than to include a level of within-person biological variability. Conversely, if the goal is to simulate how an individual-specific measurement varies over time, including a level of within-person biological variability is necessary. Sources of analytical errors may be independent of the envisioned application of an analytical procedure (e.g., measurement device drifts, or sample storage variations). Therefore, such variability functions may be included, regardless of the application setting.

In a model validation phase, where experimental test samples are available, it must be ensured that the calibration datasets utilized do not include any of the test samples. This avoids leaking any information or statistical properties of the test samples into the training samples.

In our applications, the variations between quality control samples, between different studies, and between technical replicate measurements were always included in the model. The within-person biological variability was only included for longitudinal applications, while the between-person biological variability was included for case-control applications. The variations between plasma and serum were only included when transferring a model trained on plasma measurements to its application on serum measurements. Descriptions of what calibration measurements were utilized for each application are provided in the following sections of the text.

CLINICAL STUDIES

Lasers4Life-LG study

The Lasers4Life-LG study cohort comprised of 31 nominally healthy individuals, initially sampled up to 13 times over a 7-week period, with an additional follow-up after 6 months as detailed in a previous publication [2]. Since this initial publication, the same individuals were invited to participate in two additional sampling sessions at 3.5 and 4.5 years post their initial involvement. In both the latter two samplings, 8 individuals participated. Blood plasma and serum were collected from all participants throughout the study.

BioPersMed study

The BioPersMed study is an ongoing population-based cohort at the Medical University Graz, Austria [3]. Repetitive examinations of participants were conducted in 2-year intervals. In the current study, we utilized blood plasma samples and medical data from a subset of 44 nominally healthy individuals out of 1022 participants. All 44 individuals participated in the baseline sampling, and follow-up visits ranged up to 8 years. Further detailed description of the study design was previously published [3].

KORA study

The KORA study is a population-based cohort in Southern Germany [4]. The study comprised of an age- and gender-stratified sample of participants randomly drawn from the resident registration offices within the study area. In the current study, we utilized blood plasma samples and medical data from the second and third participant visits (named KORA-F4 and KORA-FF4, respectively). The available KORA-F4 data consisted of 3044 samples, while the KORA-FF4 data consisted of 2140 samples. A subset of 2015 individuals participated in both samplings, while 1154 individuals participated in only one of the samplings. The samplings were separated by an average of 6.5 follow-up years. Data collection methods and standardized sample collections have been described in detail elsewhere [4–7].

Lasers4Life-Cancer study

Samples utilized for the cancer analysis were derived from the case-control Lasers4Life-Cancer study. The samples utilized in this study largely overlap with samples utilized in our previous study that involved the detection of four common cancer entities (lung, prostate, bladder, and breast) [8]. Newly collected blood plasma and serum samples from different individuals were included to increase the sample size and thus improve the statistical robustness of the results. Case samples were collected prior to any cancer-related treatment (i.e., therapy-naïve). Non-symptomatic individuals were used as controls. For each cancer entity, controls were pair-matched to the cancer cases by age, gender, and body mass index (BMI).

EXPERIMENTAL PROCEDURES

Infrared spectroscopy was performed on liquid samples using a Fourier transform IR (FTIR) spectrometer (MIRA Analyzer, CLADE GmbH, Esslingen, Germany). After a sample was injected into a flow cell (window material calcium fluoride) with approx. 8 μm of optical pathlength, the IR spectrum of the sample was recorded in transmission mode and subsequently a reference spectrum of the transport medium (i.e., calcium fluoride saturated water) was recorded. The spectra were acquired with a resolution of 4 cm^{-1} in a spectral range between 950 cm^{-1} and 3050 cm^{-1} . The raw data provided by the MIRA Analyzer is the absorption spectrum of the sample subtracted by the reference spectrum. Preprocessing of the resulting IR spectra was performed as described previously [8].

The instrument was maintained by the manufacturer on a yearly basis. The components that were replaced during the yearly or routine user maintenance with potential effect on the IR spectra are the following: light source (after three years); desiccant cartridges (yearly, or when necessary); flow cell when the optical path length exceeded the limit (typically after several thousands of samples), 2 μm flow cell pre-filter (typically after 200-300 samples).

Sample handling

All clinical samples involved were stored at -80°C after being processed to serum or plasma. The transport to the measurement laboratory was performed on dry ice, and further storage was at -80°C . Original samples (0.3 to 1.0 mL) were thawed at 4°C , and centrifuged for 10 min at 2000 g. The supernatant was aliquoted into the measurement tubes (50 - 100 μL per tube) and refrozen until measured.

Purchased pooled human serum was used as quality control (QC) samples. 3 liters (100 mL flasks, BioWest, Nuaillé, France) were ordered and stored in original flasks at -80°C . Prior to use, these were thawed at 4°C , filtered through a 0.45 μm filter, and aliquoted to 50-100 μL aliquots that were kept at -80°C until use. Each sample measurement batch started and ended with a QC serum measurement. In addition, a QC sample was measured after every 5 clinical samples. A measurement batch consisted of 25 to 40 samples. Although the same lot of the QC serum was used, slight variability between the flasks were measured, possibly due to the difference in storage duration and technical handling.

To evaluate the technical variability of the FTIR spectrometer, pure water samples were measured. Although not measured daily, several hundreds of water measurements were performed over the years. Due to the pre-processing of measurement data by the FTIR device (i.e., subtraction of the reference spectrum), the resulting deviations reflect the technical noise of the measurement procedure.

Measurements of clinical samples

Samples of the Lasers4Life-LG study were measured in a fully randomized order over all visits but separately for serum and plasma (single flow cell). BioPersMed study samples were measured in a fully randomized order (single flow cell). KORA-F4 and KORA-FF4 study samples were measured in two campaigns separated by an average of 2.7 years between the measurement dates of each sampling. Three flow cells were used throughout both measurement campaigns (KORA-F4: single flow cell; KORA-FF4: two flow cells).

For the Lasers4Life-Cancer study, the whole study was split into a training and a test sample set prior to any measurements. Within these sets, the measurement order was fully randomized. The samples of the training set were measured with three flow cells. A gap of four weeks occurred prior to measuring the test set samples. Within the gap, fully independent measurements were routinely performed. Measurements of the test set were performed on another, fourth, flow cell to simulate potential measurement drifts.

All samples were measured in small batches of 25-40 samples plus QC serum as described above. After each measurement batch, the device was cleaned according to the manufacturer's recommendations. A performance qualification test of the instrument was performed daily before the first sample batch or after the clearance of technical issues.

SIMULATION AND CLASSIFICATION ANALYSIS

Individual identification in longitudinal monitoring

Simulated datasets of longitudinal measurements were generated using the model formulation presented in Eq. 7. For the analyses illustrated in Fig. 3B–D, the baseline measurement of each individual served as the initial seed input. In the analysis depicted in Fig. 3E, the model formulation in Eq. 7 was iteratively applied, utilizing up to the first 6 baselines per individual as seed inputs.

Four variability functions were employed to introduce variability to each seed input: (1) within-person biological variability; (2) clinical sampling variability; (3) quality control variability; and (4) technical replicate variability. The within-person variability was always modeled by only using data of two of the three utilized longitudinal cohorts (depicted in Fig. 3A). When the individual identification classification was investigated in one of the three cohorts, the within-person variability was characterized from the

two other cohorts. Since the within-person biological variability relied on utilizing follow-up measurements (Eq. 9), this step ensured that no information was leaked to the training sets from the test sets (the follow-ups). For each experimental baseline, 1000 simulated measurements were generated to ensure a sufficiently large sample size, reaching a plateau of classification performance (Supplementary Fig. S3).

Before classification, a data standardization step (mean is 0, standard deviation is 1) and principal component analysis (PCA) were applied. For the data standardization, the mean and standard deviation were calculated only from the training sets to standardize both the training and test sets. PCA was applied to the training sets – keeping components that explain 99.99% of the total variance. The loading vectors from PCA were then applied to both the training sets and test sets before training – see Supplementary Fig. S5 for details on the effects of PCA. After the preprocessing prior to classification, a linear discriminant analysis (LDA) algorithm was used for multi-class classifications when more than one observation per class was available for training. When only one instance per class was available in the training set, a K-nearest neighbor (KNN) algorithm was for classifications (with $k = 1$). Classifier testing was always performed on held-out samples from follow-ups that were not used to create training datasets. Confidence intervals (CIs) were constructed by bootstrap resampling of the test data, following the approach applied in previous work [9]. The [2.5, 97.5] percentile boundaries were chosen to construct a 95% CI interval of the classification accuracy.

Generalization between plasma and serum specimens

For the application that involved creating simulated spectra of plasma and serum mixtures (Fig. 4), a similar simulation and classification workflow was applied as described in the prior section. However, in addition to the four previously described sources of variability, an additional function was introduced to model deviations between serum and plasma measurements (Eq. 3). The calibration datasets used to characterize the differences between IR spectral measurements of blood plasma and serum (depicted in Fig. 4C) were calculated from the Lasers4Life-Cancer study, where plasma and serum were collected at the same sampling occasion. The individual identification application was then applied to the Lasers4Life-LG cohort (Fig. 4F) – an independent cohort that involved different individuals than the Lasers4Life-Cancer cohort.

Classifier testing on independent test sets

An L2-regularized logistic regression algorithm was used for the binary (case-control) classification analysis. Data was initially split into a training and test set (Fig. 5A). Three different setups of evaluating the classification efficiency were performed. First, a 10-times repeated 10-fold stratified cross-validation was carried out on the training set of samples. Classification efficiency was evaluated on the validation splits of the cross-validation by calculating the area under the receiver operating characteristic curve (AUC). The AUC was averaged across the validation splits and reported along with its standard deviation. In the second setup, the classifier was trained directly on the experimental data. The AUC was then calculated from the unseen, initially held out, test set. In the third setup, classifier training was performed on a simulated set of samples – based on the training set as seed data. The AUC was again calculated from the unseen, initially held out, experimental test set.

When creating the simulated datasets, four variability functions were employed to introduce variability to each seed input: (1) between-person biological variability; (2) clinical sampling variability; (3) quality control variability; and (4) technical replicate variability. The data sources for calculating the level of between-person variability involved the training set of samples, as well as samples from the other independent studies – i.e., samples from all utilized clinical studies, outside of the test set. The model formulation in Eq. 5 was applied, where the mean measurement of cases and controls served as seed measurements for each binary classification task. For each task, 100000 samples were generated per class.

ANALYSIS SOFTWARE AND CODE AVAILABILITY

Analyses in this study were performed using custom scripts written in Python (v.3.8.8). The open-source packages NumPy (v.1.21.2), scikit-learn (v.0.24.1), and matplotlib (v.3.5.1) were used. An implementation of CODI is available as a Python package. Please refer to the following GitHub repository for details on its usage: <https://github.com/tarek-eissa/codi>.

2 CHOOSING THE VARIABILITY FUNCTIONS

The sources of variability that we utilized in our applications of CODI were characterized by sets of calibration measurements that reflect a level of data variance. As previously described, within these calibration measurements were characteristics of empirical variability stemming from inherent biological factors, variations in sample collection and handling, as well as instrument-specific measurement noise and drifts. Specifically, in our longitudinal analyses, we characterized four main sources of variability: (1) within-person biological variability, (2) clinical sampling variability, (3) quality control variability, and (4) technical replicate variability. Here, we examined how each of these sources of variability contributed to the success of the classification.

Fig. S1 depicts an analysis where the classification task was to identify an individual given one baseline measurement per individual/class – testing the classification on follow-up measurements. Similar to the analysis depicted in Fig. 3A–B in the main text, this analysis was also carried out on the three independent longitudinal cohorts. The CODI framework was employed to generate simulated training data based on all four sources of variability (Fig. S1, blue bars). We then systematically removed one of the variability sources considered in the simulation model (keeping the other three) to examine whether the classification accuracy was affected (Fig. S1, remaining bars). We found that with the absence of either the within-person biological variability or the quality control variability, the classification accuracy significantly suffered. Specifically for the application on the samples from KORA clinical study, a very significant drop in the prediction accuracy was observed when the simulation model did not include the variability stemming from the quality control samples. Interestingly, the KORA samples — the baseline measurements and follow-ups — were measured at very different times, years apart (see section “Measurements of clinical samples” above). Given that the quality control measurement include information about potential measurement drifts that may be observed when performing measurements at very different times, this highlighted the importance of the informed inclusion of variance that is characteristic of what may be observed in realistic applications. On the other hand, the variability introduced by the clinical sampling and technical replicate calibration measurements had minimal impact on the classification accuracy – revealing that, for our specific use-cases, these variability sources are not crucial to account for in the modeling process. Nevertheless, no significant loss in the classification accuracy was observed when all four sources of variability were included in the simulation model. Thus, we included all four sources in our analyses.

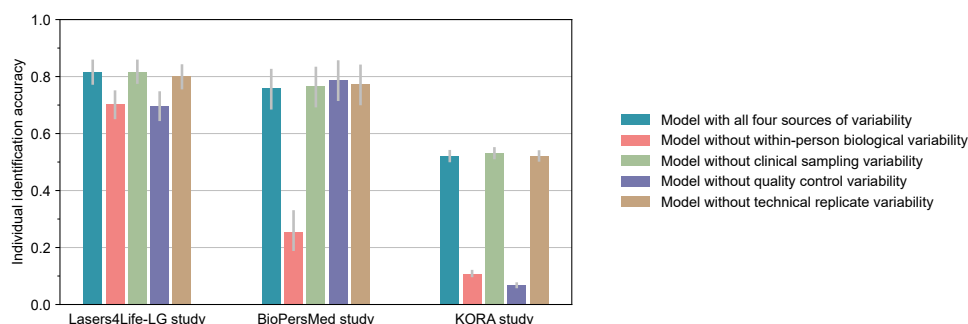


Figure S1. Impact of the modeled variability sources on the classification accuracy in longitudinal molecular monitoring. Each bar illustrates the classification accuracy achieved by simulating data through CODI with different combinations of the variability sources included.

3 COMPARATIVE ANALYSIS OF CODI TO DOMAIN-AGNOSTIC AUGMENTATION TECHNIQUES

The CODI framework inherently relies on *a priori* information on possible sources sample/measurement variability. In contrast, domain-agnostic augmentation methods employ generic transformations (e.g., additive or multiplicative noise) on the input data to simulate new measurements, eliminating the need for *a priori* information. Here, we provide a comparative analysis between CODI and other domain-agnostic methods of generating simulated training sets. Specifically, for spectral datasets, domain-agnostic augmentation methods often include the introduction of additive white noise, random scaling of intensities by multiplicative factors, linear slope variations, and vertical offset shifts [10, 11] (Fig. S2A).

In Fig. S2B, several domain-agnostic augmentation methods were employed to generate simulated training data for the classification task of identifying individuals based on a single baseline measurement per class (similar to Fig. 3B in the main text). As benchmarks, we include the performance of classifiers trained directly on the experimental data and data generated through CODI. Each augmentation strategy, including CODI, was applied to the same seed experimental training set, generating 1000 measurements per class, and tested on the same follow-up experimental set. Unlike CODI, domain-agnostic augmentation methods require the tuning of parameters that control the extent to which they affect the seed input. Across the three longitudinal cohorts this investigation was carried out, we found that the data generated through CODI had a clear advantage over all other methods.

While the domain-agnostic augmentation techniques are simpler to implement and do not require *a priori* knowledge, overall they yielded minimal-to-no improvement in the classification accuracy when compared to training directly on the experimental data. Unlike the domain-agnostic augmentation methods, CODI's targeted integration of unrepresented variations allowed for generating more representative simulated data that enhanced the classifier's ability to generalize beyond the original experimental training set.

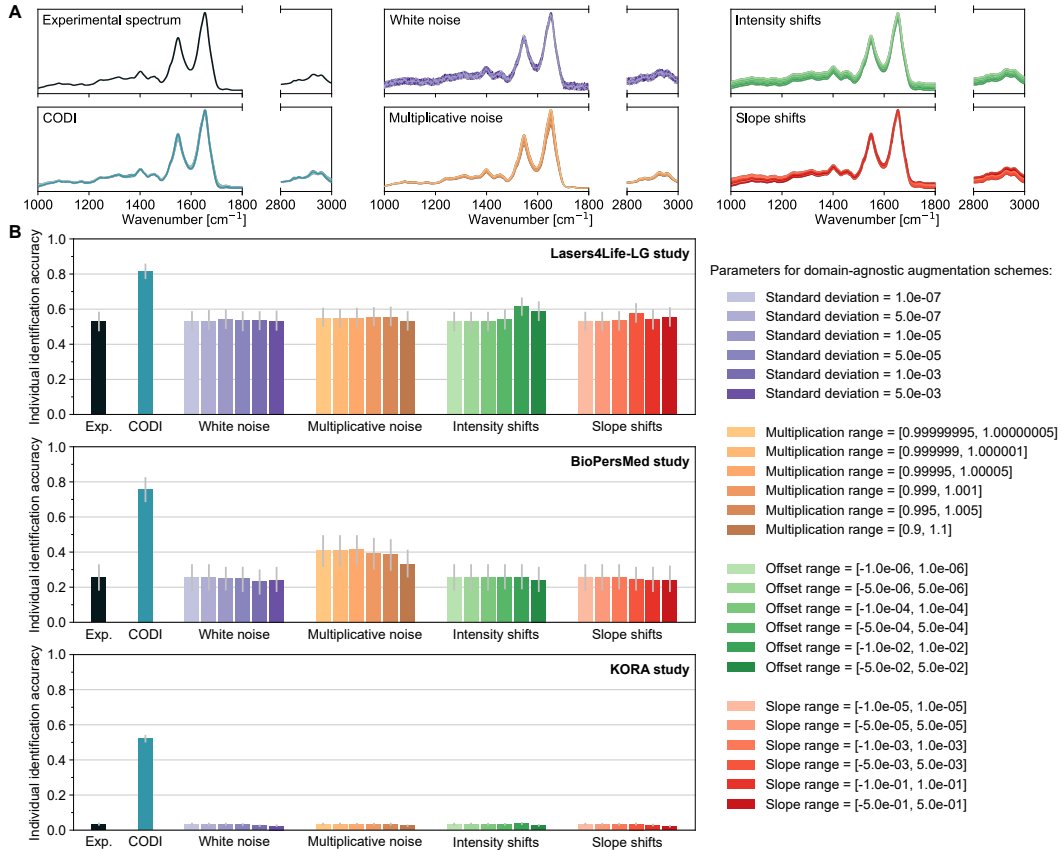


Figure S2. Comparison of CODI to domain-agnostic augmentation schemes. **(A)** Given an input seed experimental spectrum (upper left panel), several methods were employed to generate spectra with added variance. These methods include CODI (bottom left panel), as well as other domain-agnostic augmentation schemes. White noise (upper middle panel) was introduced by repeatedly generating a random Gaussian vector and adding it to the input seed. Multiplicative noise (lower middle panel) was introduced by repeatedly scaling the input seed with random scaling factors. Intensity shifts (upper right panel) were introduced to the input seed by vertically shifting the intensity with random factors. Slope shifts (lower right panel) were introduced by repeatedly manipulating the linear slope of the input seed. **(B)** The individual identification classification task was investigated across three study cohorts. The baseline measurement of each individual in the cohort served as the training seed measurements. Classifiers were trained on the experimental data (black bars), data generated through CODI (blue bars), and data generated through the domain-agnostic augmentation schemes (remaining bars). For each domain-agnostic augmentation scheme, several parameters controlling the data generation were examined, as listed in the legend to the right.

4 EFFECTS OF INTRODUCING AN EXCESSIVE AMOUNT OF VARIANCE

As the basis of CODI is to introduce variance to a given dataset, this approach carries certain risks. A key concern is that the simulated variance might significantly exceed what is typically expected from the domain of possible empirical observations. When training a classifier, this excessive variance increases the potential of obscuring the underlying signals of interest, resulting in poor performance on unseen data.

To investigate the extent of this risk, we increased the level of variance introduced by CODI and simultaneously adjusted the number of generated samples (Fig. S3). We revisited the classification task of identifying individuals based on a single experimental baseline measurement per individual involving our three longitudinal clinical studies, similar to Fig. 3B in the main text. We modeled the same four sources of variability as in the main text – biological, clinical, quality control, and technical replicate variability – and adjusted the factors by which the standard deviation of these sources were multiplied. Classifier training was performed on the simulated data and testing was performed on the held-out experimental follow-ups.

Without added variance (variability intensity = 0), the CODI procedure generates data that are exact replicates of the experimental seeds which resulted in the same classification efficacy as training directly on experimental data (Fig. S3). When the variance matched what was observed in calibration datasets (variability intensity = 1), the classification efficacy reached its maximum potential with our approach. When the variance increased beyond what was present in the calibration datasets, the classification efficacy progressively worsened. This decline was particularly pronounced when fewer simulated data were generated, which caused the classifier to perform even worse than what can be achieved by directly training on the (unrepresentative) experimental training seed. However, once a sufficiently large amount of samples were generated per class, the classification performed well despite the increased variability. These results align with the well-known principle that noisier datasets require larger sample sizes to effectively capture underlying signals. Therefore, when employing CODI, we recommend generating larger datasets to minimize the risks associated with introducing excessive variance.

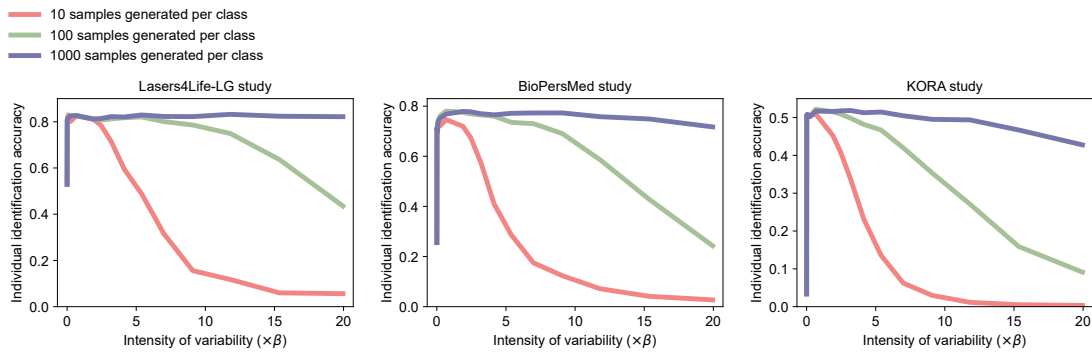


Figure S3. Effects of introducing increasing variance to the training seed. Simulated datasets were generated by CODI at varying intensities of variance and different numbers of generated samples. The x-axis depicts the scalars used to multiply the standard deviation of the random Gaussian variable (β in Eqs. 2 and 3) to model an increasing level of variability. The y-axis depicts the classification accuracy for the task of longitudinally identifying individuals in a population, testing on unseen experimental samples.

5 COMPUTATIONAL EFFICIENCY

The computational efficiency of CODI is based on three primary operations: random number generation, matrix multiplication, and matrix addition. These operations are highly optimized in modern computational libraries, which enables a fast data generation procedure. The expected runtime of the data generation is influenced by three key parameters: (1) the number of training seed observations; (2) the size of the variability calibration dataset; and (3) the number of simulated samples to be generated.

In Fig. S4, we varied these three parameters and report the runtime of our implementation of CODI in Python (see section “Analysis software” above). The seed, calibration, and simulated sample measurements all exist in a $\mathbb{R}^{519}$ space – i.e., each sample measurement is a vector with 519 discrete features. We repeated this analysis 20 times for each combination of parameters and calculated the average runtime. The analysis was conducted on a Dell XPS laptop with an Intel(R) Core(TM) i9-11900H CPU and 32 GB of RAM. The results indicate that, for any reasonable application, the runtime of the data generation can be expected to be within a few seconds on standard hardware.

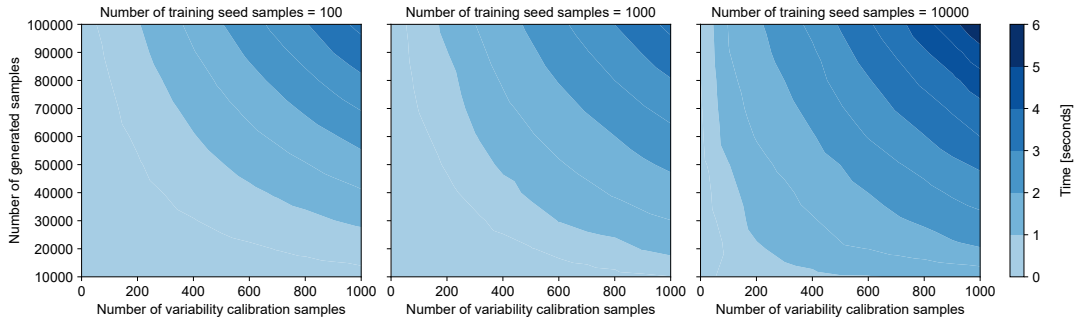


Figure S4. Computational runtime of the CODI data generation procedure. The color scale indicates the amount of time taken to retrieve the simulated measurements, varying the number of input and output samples.

6 EFFECTS OF PRINCIPAL COMPONENT ANALYSIS ON CLASSIFICATION EFFICIENCY

As detailed in the “Simulation and classification analysis” section of our Extended Methods, PCA was applied prior to passing the data to the multi-class classification algorithm. To investigate PCA’s impact on these classification tasks, we eliminated PCA and re-performed the analyses we investigated in the main text (Fig. S5). Specifically, we evaluated the multi-class classifications aimed to longitudinally identify individuals from a given population with limited experimental data (Fig. 3B in the main text) and transfer classifications between serum and plasma samples (Fig. 4F in the main text).

Overall, the PCA step helped improve the predictive accuracy across several of our classification problems, for both experimental and simulated training sets. Importantly, the overall trend remained consistent, with simulated training sets achieving higher accuracy compared to the experimental sets, regardless of the PCA application. The motivation behind CODI is to aid the downstream classifier in optimizing feature coefficients to capture underlying discriminative signals while being informed of variance that was missing from the experimental training seed. Thus, this analysis demonstrates that CODI effectively achieves this optimization, whether in the PCA projection or in the original format of the data.

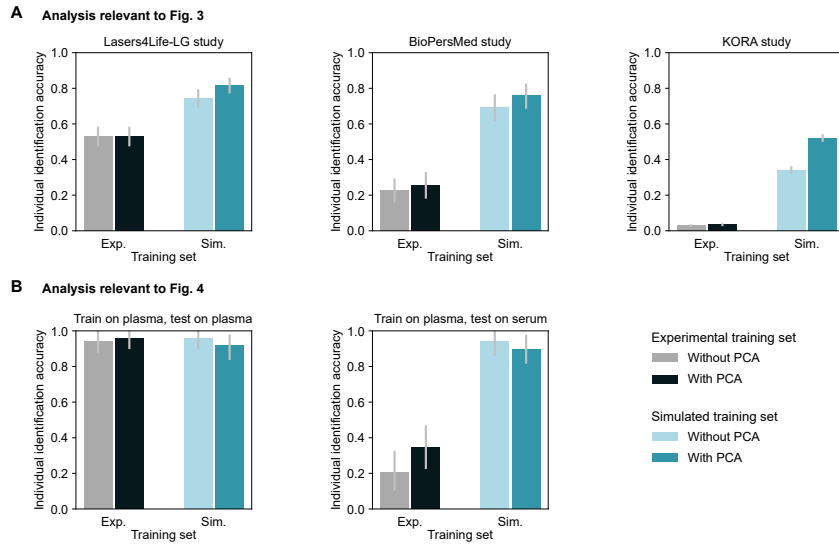


Figure S5. Impact of PCA removal on the multi-class classification tasks. **(A)** Analysis relevant to Fig. 3, in which the individual identification accuracy was examined in three longitudinal study cohorts. of PCA on experimental and simulated multi-class classifications. **(B)** Analysis relevant to Fig. 4, in which the classification accuracy was examined when training on plasma samples and testing on both plasma and serum samples. Accuracy is compared with PCA (black/dark blue bars) and without PCA (gray/light blue bars) for both experimental and simulated training sets.

REFERENCES

- [1] 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," *Analytical Chemistry*, vol. 95, no. 16, pp. 6523–6532, 2023.
- [2] M. Huber, K. V. Kepesidis, L. Voronina, F. Fleischmann, E. Fill, J. Hermann, I. Koch, K. Milger-Kneidinger, T. Kolben, G. B. Schulz, F. Jokisch, J. Behr, N. Harbeck, M. Reiser, C. Stief, F. Krausz, and M. Zigman, "Infrared molecular fingerprinting of blood-based liquid biopsies for the detection of cancer," *eLife*, vol. 10, Oct. 2021.
- [3] C. W. Haudum, E. Kolesnik, C. Colantonio, I. Mursic, M. Url-Michitsch, A. Tomaschitz, T. Glantschnig, B. Hutz, A. Lind, N. Schweighofer, C. Reiter, K. Ablasser, M. Wallner, N. J. Tripolt, E. Pieske-Kraigher, T. Madl, A. Springer, G. Seidel, A. Wedrich, A. Zirlik, T. Krahn, R. Stauber, B. Pieske, T. R. Pieber, N. Verheyen, B. Obermayer-Pietsch, and A. Schmidt, "Cohort profile: 'biomarkers of personalised medicine' (BioPersMed): a single-centre prospective observational cohort study in Graz/Austria to evaluate novel biomarkers in cardiovascular and metabolic diseases," *BMJ Open*, vol. 12, p. e058890, Apr. 2022.
- [4] R. Holle, M. Happich, H. Löwel, and H. Wichmann, "KORA - a research platform for population based health research," *Das Gesundheitswesen*, vol. 67, pp. 19–25, July 2005.
- [5] H.-E. Wichmann, C. Gieger, and T. Illig, "KORA-gen - resource for population genetics, controls and a broad spectrum of disease phenotypes," *Das Gesundheitswesen*, vol. 67, pp. 26–30, July 2005.
- [6] C. Sujana, J. Seissler, J. Jordan, W. Rathmann, W. Koenig, M. Roden, U. Mansmann, C. Herder, A. Peters, B. Thorand, and C. Then, "Associations of cardiac stress biomarkers with incident type 2 diabetes and changes in glucose metabolism: KORA F4/FF4 study," *Cardiovascular Diabetology*, vol. 19, Oct. 2020.
- [7] W. Rathmann, B. Haastert, A. Icks, H. Löwel, C. Meisinger, R. Holle, and G. Giani, "High prevalence of undiagnosed diabetes mellitus in southern germany: Target populations for efficient screening. the KORA survey 2000," *Diabetologia*, vol. 46, pp. 182–189, Feb. 2003.
- [8] M. Huber, K. V. Kepesidis, L. Voronina, M. Božić, M. Trubetskov, N. Harbeck, F. Krausz, and M. Zigman, "Stability of person-specific blood-based infrared molecular fingerprints opens up prospects for health monitoring," *Nature Communications*, vol. 12, Mar. 2021.
- [9] B. Sanchez-Lengeling, J. N. Wei, B. K. Lee, R. C. Gerkin, A. Aspuru-Guzik, and A. B. Wiltschko, "Machine learning for scent: Learning generalizable perceptual representations of small molecules," *arXiv*, 2019.
- [10] E. J. Bjerrum, M. Glahder, and T. Skov, "Data augmentation of spectral data for convolutional neural network (CNN) based deep chemometrics," *arXiv*, 2017.
- [11] A. Lebrun, H. Fortin, N. Fontaine, D. Fillion, O. Barbier, and D. Boudreau, "Pushing the limits of surface-enhanced raman spectroscopy (SERS) with deep learning: Identification of multiple species with closely related molecular structures," *Applied Spectroscopy*, vol. 76, p. 609–619, Mar. 2022.

Appendix E

Copyright notices



Plasma infrared fingerprinting with machine learning enables single-measurement multi-phenotype health screening

Author:
Tarek Eissa, Cristina Leonardo, Kosmas V. Kepesidis, Frank Fleischmann, Birgit Linkohr, Daniel Meyer, Viola Zoka, Marinus Huber, Liudmila Voronina, Lothar Richter, Annette Peters, Mihaela Žigman

Publication: Cell Reports Medicine

Publisher: Elsevier

Date: 16 July 2024

© 2024 The Authors. Published by Elsevier Inc.

Creative Commons

This is an open access article distributed under the terms of the [Creative Commons CC-BY](#) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

You are not required to obtain permission to reuse this article.

To request permission for a type of use not listed, please contact [Elsevier](#) Global Rights Department.

Are you the [author](#) of this Elsevier journal article?



The Perils of Molecular Interpretations from Vibrational Spectra of Complex Samples

Author: Tarek Eissa, Liudmila Voronina, Marinus Huber, et al

Publication: Angewandte Chemie International Edition

Publisher: John Wiley and Sons

Date: Nov 7, 2024

© 2024 The Authors. Angewandte Chemie International Edition published by Wiley-VCH GmbH

Open Access Article

This is an open access article distributed under the terms of the [Creative Commons CC BY](#) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

You are not required to obtain permission to reuse this article.

For an understanding of what is meant by the terms of the Creative Commons License, please refer to [Wiley's Open Access Terms and Conditions](#).

Permission is not required for this type of reuse.

Wiley offers a professional reprint service for high quality reproduction of articles from over 1400 scientific and medical journals. Wiley's reprint service offers:

- Peer reviewed research or reviews
- Tailored collections of articles
- A professional high quality finish
- Glossy journal style color covers
- Company or brand customisation
- Language translations
- Prompt turnaround times and delivery directly to your office, warehouse or congress.

Please contact our Reprints department for a quotation. Email corporatesaleseurope@wiley.com or corporatesalesusa@wiley.com or corporatesalesDE@wiley.com.

Appendix E

Limits and Prospects of Molecular Fingerprinting for Phenotyping Biological Systems Revealed through In Silico Modeling



Author: Eissa, Tarek; Kepesidis, Kosmas V.; et al

Publication: Analytical Chemistry

Publisher: American Chemical Society

Date: Apr 12, 2023

Copyright © 2023, Copyright © 2023 The Authors. Published by American Chemical Society

Creative Commons

This is an open access article distributed under the terms of the [Creative Commons CC BY](#) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

You are not required to obtain permission to reuse this article.

Please note the following:

- If the requested figure/table/micrograph is attributed to an ACS journal and is to be republished in an ACS journal, permission to reuse is not necessary. Credit should be given to the ACS source.
- Under ACS Policy, RightsLink cannot be used for institutional repository requests. Please send your request to the appropriate ACS journal editor.
- If you seek to reuse a portion from an ACS publication, it is your responsibility to examine each portion as published to determine whether a credit to, or copyright notice of, a third party owner was published adjacent to the item. You may only obtain permission via RightsLink to use material owned by ACS. Permission to use any material published in an ACS publication, journal, or article which is reprinted with permission of a third party must be obtained from the third party owner. ACS disclaims any responsibility for any use you make of items owned by third parties without their permission.

CODI: Enhancing machine learning-based molecular profiling through contextual out-of-distribution integration



Author: Eissa, Tarek; Huber, Marinus; et al

Publication: PNAS Nexus

Publisher: Oxford University Press

Date: 2024-10-15

Copyright © 2024, © The Author(s) 2024. Published by Oxford University Press on behalf of National Academy of Sciences.

Creative Commons

This is an open access article distributed under the terms of the [Creative Commons CC BY](#) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

You are not required to obtain permission to reuse this article.



ATTRIBUTION 4.0 INTERNATIONAL

Deed

Canonical URL : <https://creativecommons.org/licenses/by/4.0/>

[See the legal code](#)

You are free to:

Share — copy and redistribute the material in any medium or format for any purpose, even commercially.

Adapt — remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

Attribution — You must give [appropriate credit](#), provide a link to the license, and [indicate if changes were made](#). You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

No additional restrictions — You may not apply legal terms or [technological measures](#) that legally restrict others from doing anything the license permits.

Notices:

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable [exception or limitation](#).

No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as [publicity, privacy, or moral rights](#) may limit how you use the material.

Sama Aljarhi

<https://www.samaalj.com/>

sama.alj@outlook.com

22 January 2025

Subject: Permission to Use Artwork

Dear Tarek Eissa,

I, Sama Aljarhi, grant you permission to use my artwork featured on the cover of *Cell Reports Medicine*, Volume 5, Issue 7, in your PhD thesis.

Thank you for your interest in my work. Please feel free to reach out if further clarification is needed.

Sincerely,



Sama Aljarhi