How to cite: *Angew. Chem. Int. Ed.* **2021**, *60*, 17060–17069
International Edition: doi.org/10.1002/anie.202103272
German Edition: doi.org/10.1002/ange.202103272

Analytical Methods

Molecular Origin of Blood-Based Infrared Spectroscopic Fingerprints**

Liudmila Voronina, Cristina Leonardo, Johannes B. Mueller-Reif, Philipp E. Geyer, Marinus Huber, Michael Trubetskov, Kosmas V. Kepesidis, Jürgen Behr, Matthias Mann, Ferenc Krausz, and Mihaela Žigman**

Abstract: *Infrared spectroscopy of liquid biopsies is a time- and cost-effective approach that may advance biomedical diagnostics. However, the molecular nature of disease-related changes of infrared molecular fingerprints (IMFs) remains poorly understood, impeding the method's applicability. Here we probe 148 human blood sera and reveal the origin of the variations in their IMFs. To that end, we supplemented infrared spectroscopy with biochemical fractionation and proteomic profiling, providing molecular information about serum composition. Using lung cancer as an example of a medical condition, we demonstrate that the disease-related differences in IMFs are dominated by contributions from twelve highly abundant proteins—that, if used as a pattern, may be instrumental for detecting malignancy. Tying proteomic to spectral information and machine learning advances our understanding of the infrared spectra of liquid biopsies, a framework that could be applied to probing of any disease.*

Introduction

Infrared spectroscopy is a well-established method of studying chemical substances via analyzing the vibrational transitions that are characteristic of their molecular structure.[1] In particular, infrared molecular fingerprinting of human biofluids has the potential to provide information about the health state of individuals when combined with appropriate machine learning algorithms.[2–14] The idea behind is to record an infrared absorption spectrum of the whole molecular ensemble composing a biofluid using Fourier-transform infrared (FTIR) spectroscopy and pinpoint the deviations, associated with a given pathophysiological condition. However, the molecular origin of such changes in infrared molecular fingerprints (IMFs) is poorly understood.[15,16] The interpretation of the infrared absorption spectra is currently largely restricted to the characteristic spectral signatures of various functional groups.[17–19] However, these are contained in many different types of biomolecules, their spectral features in aqueous environment are broad and strongly overlapping, and the molecular complexity of biofluids is extremely high. Therefore, the understanding of the underlying molecular changes of the IMFs has so far been limited.[20,21]

Thorough exploration of the molecular origin of IMFs would be instrumental for successful application and verification of molecular fingerprinting in clinical settings.[3] It would allow for improved sample preparation, ensure that the spectral features used for building the computational models are indeed caused by a medical condition and not by confounding factors and help define the possible limitations of blood-based IMFs' applicability.[22] In this study we focus on human blood serum analysis as an example of minimally invasive and cost-effective biofluid probing procedure. Several studies measured the concentrations of a range of analytes in human blood serum using conventional biochemical methods and demonstrated that IMFs can be used to retrieve these concentrations using multivariate regression or consecutive spectral subtraction approaches.[14,23–28] However, they come up short in determining how exhaustive the list of

[*] Dr. L. Voronina, C. Leonardo, Dr. M. Huber, Dr. K. V. Kepesidis, Prof. F. Krausz, Dr. M. Žigman
Department of Physics, Ludwig Maximilian University of Munich
85748 Garching (Germany)
E-mail: Liudmila.voronina@mpq.mpg.de
Mihaela.zigman@mpq.mpg.de

Dr. L. Voronina, C. Leonardo, Dr. M. Huber, Dr. M. Trubetskov, Prof. F. Krausz, Dr. M. Žigman
Max Planck Institute of Quantum Optics
85748 Garching (Germany)

J. B. Mueller-Reif, Dr. P. E. Geyer, Prof. M. Mann
Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry
82152 Martinsried (Germany)

Dr. P. E. Geyer, Prof. M. Mann
Novo Nordisk Foundation Center for Protein Research, Faculty of Health Sciences, University of Copenhagen
2200 Copenhagen (Denmark)

Prof. J. Behr
Comprehensive Pneumology Center, Department of Internal Medicine V, Clinic of the Ludwig Maximilians University Munich (LMU), Member of the German Center for Lung Research (Germany)

J. B. Mueller-Reif, Dr. P. E. Geyer
New address: OmicEra Diagnostics GmbH
82152 Planegg (Germany)

[**] A previous version of this manuscript has been deposited on a preprint server (http://arxiv.org/abs/2102.00765).

Supporting information and the ORCID identification number(s) for the author(s) of this article can be found under: https://doi.org/10.1002/anie.202103272.

molecular constituents is and connecting disease-related changes in the molecular composition of biofluids to the changes in the corresponding IMFs.[26] Moreover, majority of previous studies reported measurements of dry samples,[14,23,24,26,28] which decreases the unwanted water background. However, the infrared spectra of dried compounds can be very different from those in the native environment as dehydration distorts the spectral contributions of hydrophilic molecules.[29] We avoid this issue by measuring the samples in their native liquid state and analytically subtracting the water absorption background.[11]

It had been suggested that large variations in blood-based IR spectra may be caused by a varying albumin-to-globulin ratio.[30] Indeed, the spectroscopic signature of human blood serum is vastly dominated by a few highly abundant molecular components, such as human serum albumin (HSA) and immunoglobulins.[31] To overcome the challenge of strong molecular signals that overshadow the signals from less abundant molecules, splitting complex biological samples into several fractions of different chemical nature is beneficial.[28,32,33] Previously, ultrafiltration has been used to fractionate human blood serum based on molecular weight of the components.[15,24,28,34,35] However, commercially available centrifugal filters introduce unwanted chemicals and require additional washing steps.[36] In this study, we chose to adapt a combination of solvent-extraction sample preparation protocols, which are typically used in metabolomics[37] and proteomics,[38] because of their robustness and speed.[39]

In order to explore the dependence of the IMF of human blood serum on its molecular composition, spectroscopic molecular fingerprinting should be ultimately combined with a technique that is able to provide molecular-specific information over a high dynamic range.[40] Recently, a high-throughput mass spectrometry (MS)-based proteomic workflow has been established for the analysis of human blood plasma.[41] We adapted this technology for human blood serum and applied it to our sample set in order to model the IMFs of hydrated biofluids as a linear combination of molecular components. Although FTIR has been integrated with proteomics to study tissue thin-sections,[42,43] such a parallelized approach for molecular annotation of disease-relevant vibrational fingerprints of human blood derivatives has been lacking this far.

With the gained understanding of the molecular composition underlying the IMFs of human blood serum, we compare the samples of lung cancer patients (tumor node metastasis (TNM) clinical stages II and III) with reference individuals matched in age, gender and smoking status. We focused on lung cancer as a prototypical disease for which non-invasive early detection from blood profiling would be highly beneficial.[44,45] The ability of FTIR spectroscopy of blood serum to discriminate lung cancer cases from controls has been previously shown in several studies.[46,47] Pattern recognition algorithms were used to identify non-small cell lung carcinoma and subtype the disease conditions.[46] Independently, the ratio between intensities at 1080 and 1170 cm^{-1} was put forward as the most informative for disease detection, and it was suggested that changes in the protein secondary structure might be correlated with lung cancer.[47] Other types of cancer have also been detected with various efficiencies using blood-based IMFs, with little insight into molecular changes for the reasons stated above.[10,48–52]

In this study, we obtain reproducible, cost- and time-efficient IMFs of human sera and use proteomic measurements to facilitate their understanding at a molecular-level. In particular, we reveal a pattern of changes of human blood serum composition, which correlates with the presence of lung cancer and results in an observable difference between IMFs of blood sera of lung cancer patients compared to the reference group. Both spectral and molecular information was used to build explainable classification models for lung cancer detection.[53] This paradigm can be applied to possibly any other health phenotypes in order to develop efficient and explainable diagnostic tools.

Results

Decomposing the Complexity of Human Blood Sera Using Biochemical Fractionation

We recorded infrared absorption spectra of liquid human blood sera in the range from 1000 to 3000 cm^{-1} (Figure 1 A,B). The spectra are dominated by amide bands that are attributed to the vibrations of protein backbone.[54] In particular, the most prominent feature between 1600 and 1700 cm^{-1} (Amide I band) is characteristic of the secondary structure of the proteins.[54] The region on the red side of the spectrum (1000–1200 cm^{-1}) is often referred to as "carbohydrate region", because of the typical absorption patterns that glycans exhibit here.[18] Finally, lipids produce several absorption bands around 1735 cm^{-1}, 2852 cm^{-1} and 2926 cm^{-1}.[55]

Attributing the distinct features of the mid-infrared absorption spectrum of human blood serum to a specific molecular class is somewhat oversimplified, since absorption spectra of various biological molecules often overlap. In order to gain deeper insight into the origins of different spectral features, we built a comprehensive model of the human blood serum absorption. To this end, we used a set of 148 blood serum samples (Figure 1 A).

As a first step, we recorded the IMFs of each full intact, fluid, serum sample using high-throughput automated FTIR spectrometer in transmission mode (black line in Figure 1 B).[11] After every sample, a reference measurement of water-filled cuvette is performed and used to subtract the water background from the sample spectra such that the first derivative of the resulting curve is minimal from 1800 to 2200 cm^{-1} (SI Materials and Methods, section 4). Next, we biochemically fractionated each sample into three fractions (metabolites, human serum albumin (HSA)-depleted proteins and HSA-enriched proteins) and recorded their IMFs (colored lines in Figure 1 B) in order to assess the relative contributions of roughly defined molecular classes. In parallel, we used proteomic analysis of the crude sera and HSA-depleted fractions to characterize the efficiency of HSA depletion and the molecular composition of each protein fraction.

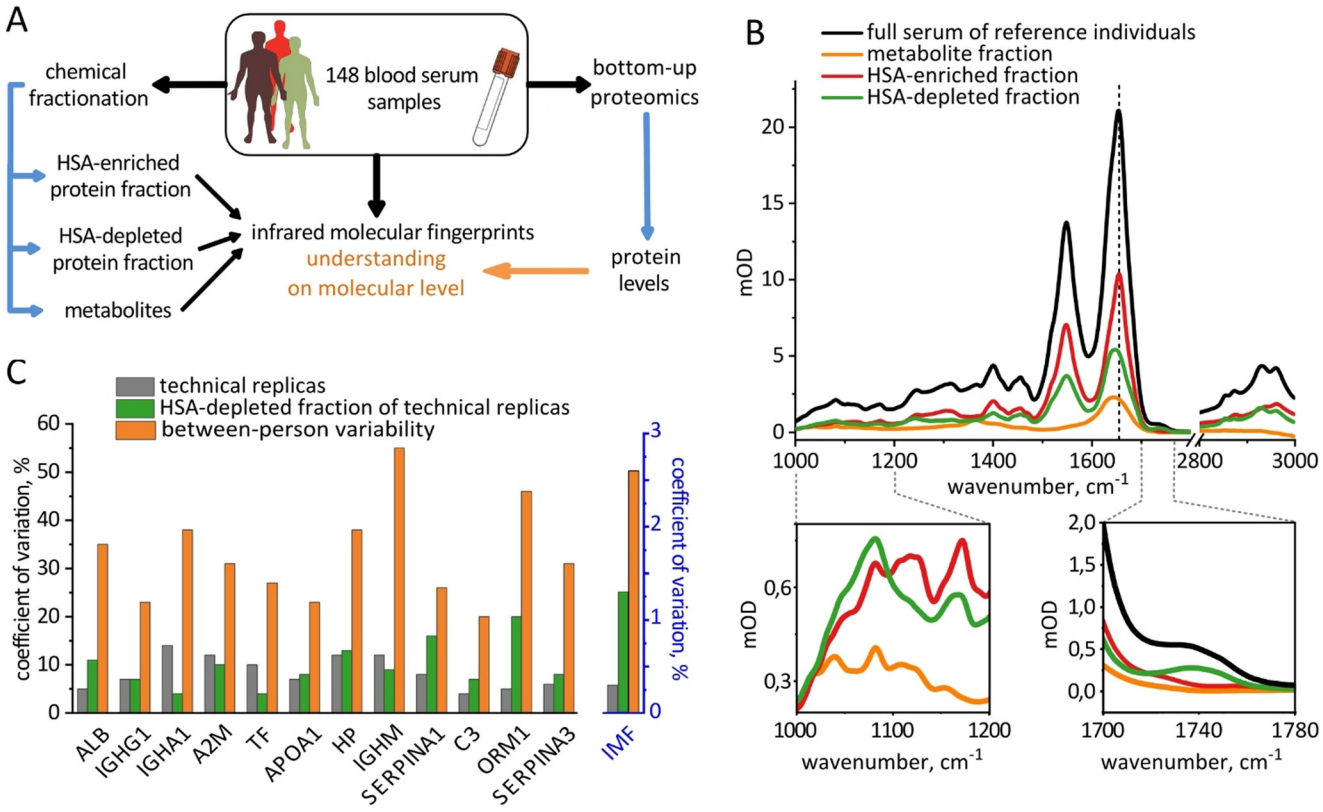

Figure 1. Decomposing the complexity of human blood sera using chemical fractionation. A) Overview of the workflow of the study. B) Average infrared molecular fingerprint (IMF) of human blood serum of 93 reference individuals and the corresponding IMFs of 3 fractions. The dashed vertical line shows the position of the Amide I band in the HSA-enriched fraction. The two lower inserts highlight the regions with the largest relative differences between the fractions. C) Reproducibility of the fractionation protocol assessed with proteomic and FTIR measurements. Left axis: coefficients of variation for the levels of 12 proteins considered in this study for the same 8 serum samples with and without fractionation as well as their between-person variability in 93 control individuals. Right axis: the corresponding variations in the IMFs, averaged across wavenumbers.

Human serum albumin is the most abundant serum protein and constitutes about a half of total protein mass.[31] It is helpful to separate HSA away from other proteins, because its intense absorption potentially obscures the signals from other molecules.[32] For this purpose, we first precipitated most of the proteins using cold ethanol.[38] The supernatant was enriched in HSA, which we precipitated in the next step to separate it from metabolites.[56] The latter fraction was dried in vacuum and all three of them were re-dissolved in water prior to spectroscopic measurements.

We assessed the reproducibility of our fractionation protocol both with FTIR spectroscopy and proteomic analyses (Figure 1C). First, we estimated the measurement uncertainty of the proteomic workflow as the coefficient of variation (CV) in repeated measurements of the same single human blood plasma sample. The average CV for the 12 proteins considered in this study (see below) in the crude plasma samples is 9%, and it rises to 10% in the HSA-depleted fraction of the same sample, suggesting that the process of fractionation adds only minor error compared to the instrumental one. The CV measured for 93 reference individuals provides a rough estimate for the between-person variability, which is higher than the instrumental error for all considered proteins (33% on average). The analysis based on IMFs leads to similar conclusions (Figure 1C, right axis).

We further compared the spectral intensities of each of the fractions (Figure 1B). This procedure facilitates several unexpected conclusions about the nature of the IMFs of crude blood sera: Firstly, the signals between 1000 and 1200 cm^{-1} are typically attributed to carbohydrates.[18] Indeed, we detected the metabolite fraction containing free carbohydrates, exhibiting characteristic pattern in this region of the spectra. However, the intensity of the signals from both two protein fractions combined is an order of magnitude higher than that of metabolite fraction in this spectral region. We attribute this effect to glycosylation of proteins and further demonstrate it below. Additionally, we show that metabolites exhibit an absorption band that overlaps with Amide I of the proteins and reaches 10% of its intensity.

Altogether, our fractionation workflow enabled us to disentangle the quantitative contributions of metabolites and proteins to the IMF of crude blood sera. Since the absorption of proteins fractions is, as expected, significantly higher than that of metabolites, in the next step we focused on understanding and modeling the contribution of protein absorption to the overall fingerprints.

We demonstrated that the IR spectrum of blood serum mostly exhibits signals originating from the protein absorption. It is therefore important to understand how various proteins of blood sera contribute to the overall IR absorption spectra of this biofluid. To that end, we performed bottom-up proteomic analysis of the same samples. They were subjected to an established mass-spectrometry based proteomics pipeline.[41] In brief, proteins in the sample are denatured and disulfide bonds reduced and quenched. Proteins are then digested into tryptic peptides and desalted. The peptides are separated by reversed phase chromatography coupled online to the mass spectrometer to detect the mass to charge ratios of peptides and their fragments in a quantitative manner. This enables software-dependent peptide identification and subsequently quantitative protein assembly from detected peptides.[57,58]

The first ten proteins listed in Figure 1C are the ten most abundant proteins in human blood serum (Table S1). The quantitative values for each protein (so called "label-free quantification" or LFQ values) provided by proteomic measurements are suited to characterize the differences between subjects in a study, but not directly proportional to the absolute concentrations of proteins,[59] as revealed by Table S1. To obtain the actual protein concentrations, we re-scaled the LFQ values using the average reference concentrations of these proteins in healthy subjects.

To be able to link the actual individual protein levels directly to the IMFs of blood sera, we measured IR absorption spectra of each of the 10 most abundant proteins separately, dissolved in phosphate-buffered saline (PBS). Figure 2A demonstrates the IR spectra of 5 highly abundant proteins (Figure S2 for all proteins). The position and shape of the Amide I band is characteristic for their secondary structure and qualitatively corresponds to the known β-sheet and α-helix content of proteins.[54] As expected, alpha-1-acid glycoprotein (ORM1 in Figure 2A) shows particularly high absorption in the region of 1000–1200 cm^{-1}, because about 45% of its dry mass is comprised of carbohydrates.[60]

In order to estimate the contribution of each protein to the IMF of blood serum, we modeled the absorption spectra of every individual's serum as a sum of IR absorption spectra of proteins multiplied by their respective concentrations, measured by proteomics [Eq. (1)]:

$$IMF(\tilde{\nu}) = \sum_i C_i \times S_i(\tilde{\nu}), \quad (1)$$

where $\tilde{\nu}$ represents wavenumber, C_i—concentration of the protein i in $mg\,mL^{-1}$, $S_i(\tilde{\nu})$—absorption spectrum of the protein i for 1 $mg\,mL^{-1}$.

We started by taking into account the spectral contribution of HSA only ($i=1$) and building complexity by adding proteins one by one, in the order as listed in Table S1. Figure 2C shows how the model becomes closer to the experimentally measured IMFs with every additional protein. Adding further lower abundant proteins to the model is expected to yield only small improvements, since the total concentration of remaining proteins that are beyond the ten molecules considered here is about the same order of magnitude as the level of complement component C3.

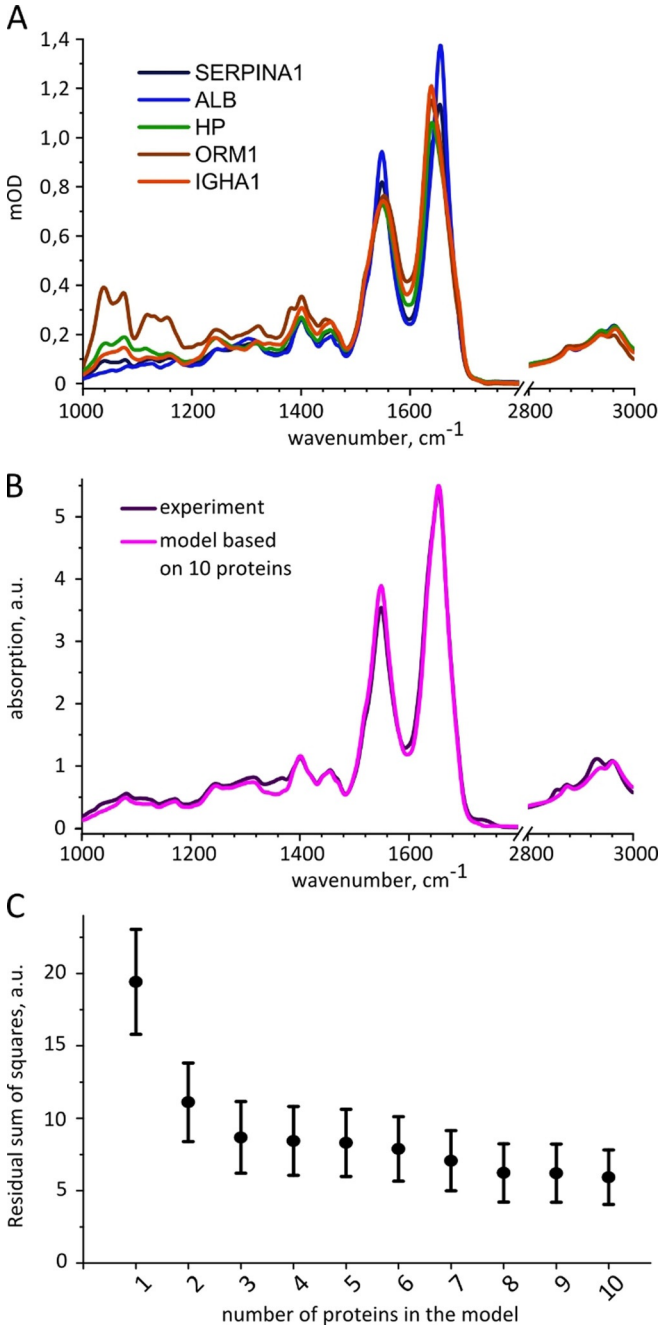

Figure 2. Molecular modeling of infrared fingerprints based on serum proteomic profiling. A) Examples of infrared absorption spectra of human serum proteins at the same concentration, 5 $mg\,mL^{-1}$, labeled according to the corresponding genes: SERPINA1, alpha-1-antitrypsin; ALB, human serum albumin; HP, haptoglobin; ORM1, alpha-1-acid glycoprotein 1; IGHA1, immunoglobulin A. B) Average IMF of 148 human blood sera, each modelled as a sum of contributions of 10 proteins compared to the average experimentally measured IMF. C) Average vector distance between the model and experimental spectra for all 148 samples depending on the number of proteins introduced into the model.

In Figure 2B we compare the average modeled and experimental absorption spectra of human blood serum. Given the linear character of the model and the limited number of considered components, the matching is remarkably high. The only prominent peaks missing from the modeled spectra are the C=O (at 1735 cm^{-1}) and C–H stretches (at 2852 cm^{-1} and 2926 cm^{-1}) known to be unique for lipids.[55] Indeed, the average concentration of cholesterol in human blood serum is of the same order of magnitude as the last proteins we considered.[61] The model can, therefore, be further refined by including cholesterol and other metabolites, such as ATP, melanin, glucose and urea. In fact, adding the entire metabolite fraction to the model further reduces the RSS between the model and the experiment by 50% (Figure S3).

Combining MS-Based Proteomics and IR Fingerprinting Reveals Lung Cancer-Related Molecular Changes in Blood Serum

Having obtained a simple model of the IR absorption of human blood serum, we can address the question how this absorption changes as a consequence of a disease. In this study we focused on lung cancer, as the most common cause of cancer-related deaths worldwide.[44] We compare the IMFs of sera between two cohorts: 55 lung cancer patients (therapy naïve, prior to any cancer-related therapy, at TNM clinical stages II and III) with 93 reference individuals. In the latter cohort we gathered non-symptomatic individuals ("healthy"), patients with chronic pulmonary obstructive disease (COPD) and individuals with lung hamartoma, to challenge our detection regime by non-cancerous lung diseases. Importantly, to avoid possible confounding bias the cohorts are gender, age and smoking-status matched (Table S2).

We find that infrared molecular fingerprints of lung cancer patients clearly differ from that of reference individuals. The black line in Figure 3A shows the difference between the average IMF of lung cancer patients and those of references as a function of wavenumber, which we specify as "differential fingerprint". The p-values of the most prominent spectral peaks are below 10^{-6} (Table S3), strongly suggesting that the differences between the IMFs of two cohorts are statistically significant. To further quantify these differences, we applied support vector machine (SVM) algorithm to classify the samples into two classes—cancer cases and reference individuals. To that end, the data were split into train and test sets, employing 10-times repeated 10-fold cross-validation. The area under the curve (AUC) of the receiver operating characteristics (ROC) curve was used as a measure of classification efficiency. For the classification of lung cancer patients versus references, the model reveals an AUC of 0.85 ± 0.1, implying that the SVM model can, in principle, be trained to distinguish between the two cohorts.

We find that the differential fingerprint of lung cancer has a specific shape, with prominent features around 1000–1200 cm^{-1}, as well as in the Amide I and Amide II regions. Such shape could result from changes in the proteins secondary structure, as previously suggested[47] or, alternatively, from the changes in their concentration.[22] The distinction between the two possibilities can only be obtained by comparison of two sample sets with a technique that provides information about molecular concentrations.

The HSA-enriched and HSA-depleted fractions reflect the largest differences between lung cancer and reference samples with p-values below 10^{-6} (Table S3), while the metabolite fraction is not significantly different in the samples from reference individuals versus these of the lung cancer patients. This finding is confirmed by the AUC values: for the metabolite fraction the AUC is 0.62 ± 0.2, while for the HSA-enriched fraction it is 0.82 ± 0.1, and for the HSA-depleted fraction 0.75 ± 0.1. Thus, we turned to the proteomic measurements of the same sample set—aiming for the identification of individual proteins responsible for the observed changes in the IMFs.

In line with previous research,[45,62–68] we find a number of proteins that demonstrate p-values below 0.0005 (Table S4). However, the purpose of this study is not the search for specific biomarking candidates; instead, we wish to evaluate whether lung cancer results in a pattern of changes in protein concentrations responsible for its IR signature.

The first question we have addressed is: which proteins do we have to consider in order to model the differences in the IMFs between the lung cancer patients and reference individuals. The differential fingerprint is affected by the disease-related absolute change in the protein concentration due to the linear character of the absorption measurement. Therefore, we ranked all detected proteins according to the absolute difference in average concentration between lung cancer and reference samples, as measured by MS (Table S5). Out of ten proteins that are most extensively changing, eight are also among the ten most abundant proteins in the blood sera.

We further identify other proteins reflecting the differences between the two sample sets, such as alpha-1-acid glycoprotein-1 and alpha-1-antichymotrypsin: although their concentrations in non-symptomatic subjects are below the ten most abundant proteins, they are changing significantly in lung cancer patients and thus have to be taken into account to accurately model the disease differential fingerprint. In total, we considered twelve proteins for the model of lung cancer differential fingerprint, as shown in Figure 3B: ten most abundant ones and two additional ones that are changing most significantly.

The change in the concentrations of some proteins, for example, HSA, does not reach statistical significance ($p = 0.1$), being not sufficiently large (−9%) as compared to between-person variability (reference range ±45%[69]). However, in absolute terms, the concentration of HSA changes the most (−0.4 $g\,dL^{-1}$) due to its initially high abundance leading to an observable change in the infrared absorption. It is therefore important to take the albumin level into account when modelling the impact of a disease on an IMF. Moreover, previous studies have demonstrated lower level of HSA in lung cancer patients than in general population.[70–72]

After we have modelled the IMF of every individual as described above, the differential fingerprint of lung cancer was calculated as the difference between the average fingerprint of lung cancer patients and reference individuals. The

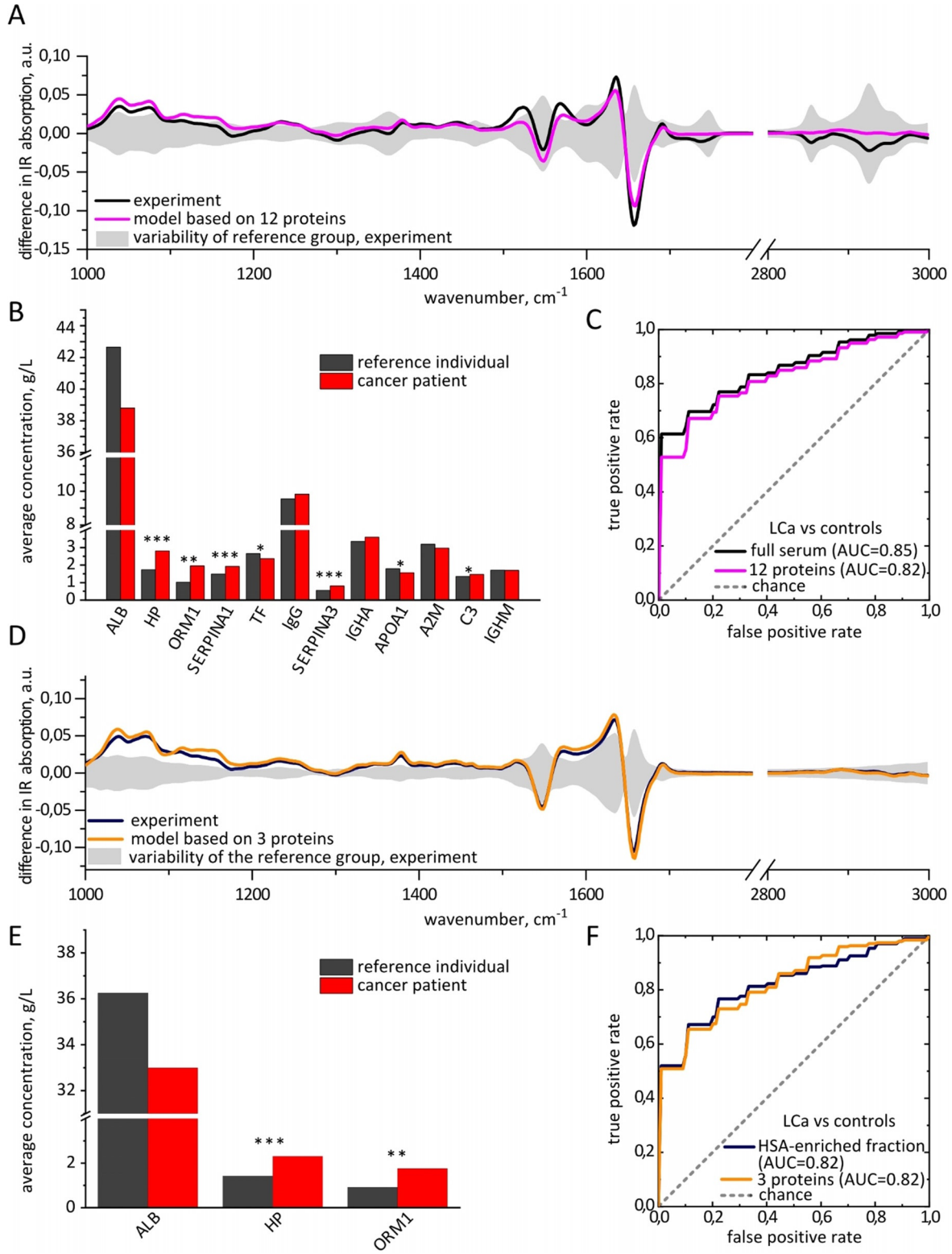

Figure 3. Lung cancer-related molecular changes in blood serum, based on comparison between 55 lung cancer patients and 93 reference individuals. A) Differential fingerprints of lung cancer in full sera: experimentally measured and modeled based on the levels of 12 proteins. The shaded area shows the standard deviation of the IMFs of the reference group. B) Change in the concentrations of proteins in blood serum caused by lung cancer, measured by proteomics. The proteins are ordered according the absolute difference in the concentrations in lung cancer and control individuals. *, p-value below 0.05; **, p-value below 0.0005; ***, p-value below 10^{-6}; no star, p-value above 0.05. C) ROC curves based on the experimental measurement of IMF of full serum and the set of 12 proteins measured by proteomics. The STDs are 0.1 for AUC in panels (C) and (F). D) Differential fingerprints of lung cancer in HSA-enriched fraction: experimentally measured and modeled based on the levels of 3 proteins. E) Change in the concentrations of proteins in HSA-enriched fraction caused by lung cancer, measured by proteomics. F) Comparison between the ROC curves based on the experimental measurement of IMF of HSA-enriched fraction and the corresponding set of 3 proteins.

resulting curve of this twelve-protein model very closely resembles the measured differential fingerprint, reflecting all the important features (pink line in Figure 3A). Moreover, the binary classification of lung cancer cases versus reference individuals based on the concentrations of the twelve identified proteins produces an AUC of 0.82 ± 0.1, which is close to the value for experimentally measured serum spectra (0.85 ± 0.1). These findings suggest that most of the information in IMFs regarding lung cancer status stems from the molecular changes in these twelve proteins. Moreover, such kind of information can be measured in time- and cost-efficient manner by applying FTIR, without the need to measure the concentrations of each of the protein separately.

Interestingly, the three proteins that change the most between the lung cancer patients and the reference group (namely, HSA, haptoglobin and alpha-1-acid glycoprotein 1, Figure 3B,E) remain predominantly contained in the HSA-enriched fraction during the fractionation procedure. This explains the high AUC obtained for this protein fraction: 0.82 ± 0.1, blue line in Figure 3F. It further suggests that most of the molecular information about the presence of lung cancer is encoded in the concentrations of the three proteins named above, out of all twelve proteins analyzed. Indeed, the SVM binary classification based on the concentrations of these three proteins reveales the AUC of 0.82 ± 0.1, the same as based on all 12 proteins considered above.

We modeled the IMFs of the HSA-enriched fraction as detailed above, taking into account the proportion of each protein in HSA-enriched fraction compared to full serum (Table S1, Figure S1 and S4). In line with only a minor contribution of low-abundant proteins and metabolites to the IR spectra of HSA-enriched fraction, we find that the model very well reproduces the experimental differential fingerprint (Figure 3D).

In summary, we observe statistically significant differences between the IMFs of blood serum of lung cancer pateints when compared to the IMFs of reference individuals. Biochemical fractionation and proteomic profiling of the very same sample set facilitated identification of the compounds responsible for these differences and revealed previously unappreciated pattern of changes in the concentrations of well known proteins that we find to be characteristic of lung cancer.

Discussion

Although FTIR has been used over decades and blood-based studies suggested the applicability of this approach to disease diagnostics, the molecular nature of blood-based infrared molecular fingerprints (IMFs) and changes therein has not been well understood. Being cost- and time-efficient, suitable for high-throughput approaches, IMFs could greatly contribute to clinical diagnostics if their robust correlation with any given condition is reproducibly demonstrated. Molecular understanding of the IMFs along with computational models may open up a path towards informed choice of biofluid (e.g. serum vs. plasma), improved sample preparation and possibly even initial steps of the biomarker identification. Here we examined the samples with two independent techniques—IR spectroscopy and mass spectrometry (MS)-based proteomics—with the goal to elucidate the molecular entities dominating human blood-based IMFs.

As a first step to decompose chemical complexity of IMFs, we established a protocol for highly reproducible fractionation of crude human blood sera into three fractions: human serum albumin (HSA)-enriched proteins, HSA-depleted proteins, and metabolites. The strongest IR absorption signal in human blood serum arises from proteins. We therefore measured their relative concentrations in the samples using MS-based proteomic profiling and used the concentrations of ten most abundant proteins to reconstruct individual spectra of the human blood serum. This concept is shown in the bottom part of Figure 4 for the general case of any 'omic technology. Indeed, the model built in this study can be further developed by adding highly abundant metabolites and additional proteins until the model reproduces measured IMFs within their noise limit. In particular, it has been shown previously that in addition to the proteins discussed here, FTIR spectra of blood plasma provide information about the levels of lactate, urea, apolipoproteins B and C, as well as immunoglobulin D.[26] However, the data presented here suggest that our 10-protein-based approach leaves little room for improvement in modelling IMFs measured by FTIR spectroscopy. The ultimate limitation of such modeling lies in the linearity of the model, disregarding any interaction between different blood components. In general, this approach is facilitated by the measurements of fluid samples, as performed in this study, where all blood serum compounds are interrogated in their native aqueous environment. It remains to be tested if similar modelling could be performed relying on the spectra of dry films, which are commonly measured in the field.[10,17]

Infrared molecular fingerprints acquired by field-resolved spectroscopy[73] may drastically increase the precision of infrared molecular fingerprinting by reducing the noise limit. This will render smaller molecular contributions significant, uncovering thereby more molecular information just as the combination of further biochemical fractionation (e.g. by liquid chromatography) with field-resolved spectroscopy will do. Both may allow more lower-abundance molecules to contribute to the identification of a pathophysiological condition.

In this study we use lung cancer as a case scenario of a medical condition, the outcome of which could significantly benefit from early detection. We find that IMFs of sera samples of lung cancer patients differ significantly from that of reference individuals. Using MS-based proteomics, we identify a pattern of known highly-abundant proteins that determine the observed change in the IMFs of blood sera. Some of them have been previously linked to cancer: unexplained hypoalbuminaenia has been associated with increased cancer risk,[74] and low pre-treatment albumin level—with poor survival rate.[75] Moreover, in line with our findings, the levels of haptoglobin, complement component C3, alpha-1 antytrypsin and alpha-1-acid glycoprotein were previously shown to rise in blood of lung cancer patients.[63–65,68] Due to differences in tumor growth rate,

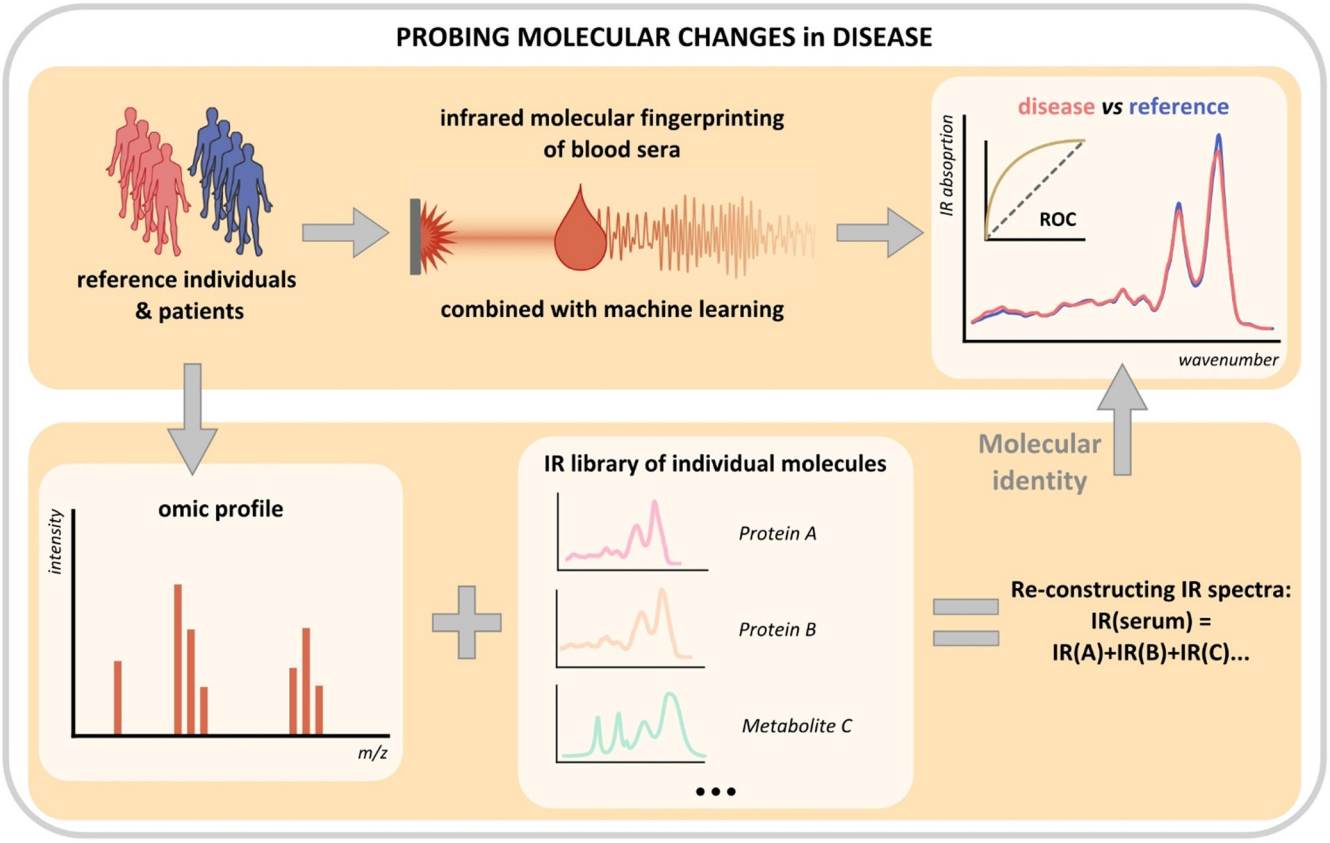

Figure 4. General workflow for probing molecular changes in disease. The infrared absorption spectra of blood sera are reconstructed as a linear combination of the spectra from individual molecular constituents, while the concentrations of the latter are measured using an 'omics technology. The resulting model is compared to the measured IMFs of blood sera and used to explain disease-related features therein. A similar workflow can potentially be applied to the detection of any phenotype in human biofluids.

invasiveness and other factors, the exact pattern of changes might be organ- or cancer type-specific, potentially providing a way to distinguish between different cancer entities in a single IMF measurement. This exciting prospect could be tested using the proposed general workflow (Figure 4).

Importantly, although these proteins are not specifically challenging to detect and measure, they have previously not been used in a combined fashion to help detect or diagnose lung cancer. It is meanwhile widely accepted that using multiple biomarking molecules together, as a pattern, is more effective and robust for detecting a particular health condition than using a single biomarker.[31,67,76,77] Infrared fingerprinting of human blood serum takes this approach to a new level: here we effectively combine a wide range of molecules into a single IR spectrum that can be easily measured and interpreted. To illustrate that, we considered the levels of all 114 proteins detected by proteomics in every sample. Importantly, the binary classification efficiency based on all these proteins measured separately is not higher than the efficiency based on a single IMF measurement (Table S6).

Lung cancer induces a number of changes in the levels of blood serum proteins that have been previously linked to acute-phase response, and it is well-known that cancer is often associated with inflammatory states.[44,78] In line with the general discussion in the field,[22] our findings underscore the need for additional clinical studies that would look into the specificity of IMFs. A well-designed reference cohort should include individuals with potentially similar pattern of changes in the blood composition: for example, in the case of lung cancer, with chronic or acute inflammation. Due to cost-efficiency and rapidity of blood-based infrared molecular fingerprinting, it could still find a wide range of applications, even if its specificity proves insufficient for screening applications.

This study featured case-control design, which is limited in that between-person variability potentially masks the disease-related signals. It has been recently demonstrated that IMFs of healthy individuals remain stable over clinically relevant periods of time,[11] opening up an avenue of health-monitoring applications. In such settings, molecular-level understanding of the disease-related changes in IMFs will become even more important, helping establish better clinical study design, and ultimately leading to improved diagnostics to probe human health and disease.

Conclusion

As the focus of future healthcare is shifting from treatment to early detection and prevention, such rapid, cost-

effective and holistic approaches as infrared molecular fingerprinting of body liquids will become ever more relevant. So far, infrared spectral changes in complex bioliquids were linked to multiple diseases but have remained uninterpretable with regard to which specific molecule accounts for a spectral change. In this study we looked systematically into the contributions of different constituents of blood serum to the overall IMF. In particular, we showed that the IMFs of blood serum can be to a high extent modelled using the concentrations of the ten most abundant proteins. With non-metastatic lung cancer as an example of a medical condition, we showed that a number of highly abundant acute-phase proteins are up- and down-regulated in cancer patients compared to the reference group, leading to an observable change in the IMFs of blood serum. Accompanied by a meaningful molecular annotation, this change is more likely to find its use in everyday clinical practice.

The paradigm presented here could in principle be used for any pathophysiological condition. After having recorded the IMFs of patients and compared them to matched reference individuals, one could use biochemical fractionation to determine which molecular class is responsible for the disease-related differences and perform in-depth omics profiling of the identified fraction (Figure 4). This would provide insights into the nature of information that infrared molecular fingerprinting is able to provide and into its additional value compared to well-established clinical tests. Moreover, combining biochemical fractionation with field-resolved spectroscopy-based infrared molecular fingerprinting[73] might yield deeper molecular insight along with higher specificity and sensitivity for disease detection. Ultimately, the larger clinical studies with purposefully chosen reference groups, stratified and controlled for comorbidities, may bring IMF—an inexpensive and time-efficient method—closer to everyday clinical use.

Acknowledgements

This work was funded by the Center for Advanced Laser Applications (CALA) of the Ludwig Maximilians University Munich (LMU), Department of Laser Physics, and the Max Planck Institute of Quantum Optics (MPQ), Laboratory for Attosecond Physics, Germany. We also thank the Asklepios Biobank for Lung Diseases, member of the German Center for Lung Research (DZL), for providing clinical samples and data. We would like to thank Dr. Frank Fleischmann, Dr. Catherine Vasilopoulou, Dr. Ina Koch, Jacqueline Hermann, Katja Leitner, Dr. Sigrid Auweter, Daniel Meyer, Beate Rank, Sabine Eiselen, Christina Mihm and Dr. Incinur Zellhuber for their help with this study. In particular, we wish to acknowledge the efforts of many individuals who participated as volunteers in the clinical study reported here. We also thank Prof. A. Barth for his insightful suggestions. Open access funding enabled and organized by Projekt DEAL.

Conflict of interest

The authors declare no conflict of interest.

Keywords: cancer · infrared molecular fingerprinting · IR spectroscopy · liquid biopsy · proteomics

[1] P. R. Griffiths, J. A. de Haseth, in *Fourier Transform Infrared Spectrom.*, Wiley, Hoboken, **2007**, pp. 1–18.
[2] L. Lechowicz, M. Chrapek, J. Gaweda, M. Urbaniak, I. Konieczna, *Mol. Biol. Rep.* **2016**, *43*, 1321–1326.
[3] J. Titus, E. Viennois, D. Merlin, A. G. Unil Perera, *J. Biophotonics* **2017**, *10*, 465–472.
[4] F. Elmi, A. F. Movaghar, M. M. Elmi, H. Alinezhad, N. Nikbakhsh, *Spectrochim. Acta Part A* **2017**, *187*, 87–91.
[5] X. Yang, T. Fang, Y. Li, L. Guo, F. Li, F. Huang, L. Li, *Optik* **2019**, *180*, 189–198.
[6] H. J. Byrne, I. Behl, G. Calado, O. Ibrahim, M. Toner, S. Galvin, C. M. Healy, S. Flint, F. M. Lyng, *Spectrochim. Acta Part A* **2021**, *252*, 119470.
[7] I. Maitra, C. L. M. Morais, K. M. G. Lima, K. M. Ashton, R. S. Date, F. L. Martin, *Analyst* **2019**, *144*, 7447–7456.
[8] A. L. Mitchell, K. B. Gajjar, G. Theophilou, F. L. Martin, P. L. Martin-Hirsch, *J. Biophotonics* **2014**, *7*, 153–165.
[9] P. Carmona, M. Molina, M. Calero, F. Bermejo-Pareja, P. Martínez-Martín, A. Toledano, *J. Alzheimer's Dis.* **2013**, *34*, 911–920.
[10] H. J. Butler, P. M. Brennan, J. M. Cameron, D. Finlayson, M. G. Hegarty, M. D. Jenkinson, D. S. Palmer, B. R. Smith, M. J. Baker, *Nat. Commun.* **2019**, *10*, 1–9.
[11] M. Huber, K. V. Kepesidis, L. Voronina, M. Božić, M. Trubetskov, N. Harbeck, F. Krausz, M. Žigman, *Nat. Commun.* **2021**, *12*, 1511.
[12] 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, *Proc. Natl. Acad. Sci. USA* **2017**, *114*, E7929–E7938.
[13] T. G. Mayerhöfer, S. Pahlow, J. Popp, *Spectrochim. Acta Part A* **2021**, *251*, 119411.
[14] S. Roy, D. Perez-Guaita, D. W. Andrew, J. S. Richards, D. McNaughton, P. Heraud, B. R. Wood, *Anal. Chem.* **2017**, *89*, 5238–5245.
[15] 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. Sylvén, G. Werner, R. Zerback, *Analyst* **2009**, *134*, 1092–1098.
[16] B. Bird, M. Miljkovi, S. Remiszewski, A. Akalin, M. Kon, M. Diem, *Lab. Invest.* **2012**, *92*, 1358–1373.
[17] 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 Lett.* **2020**, *477*, 122–130.
[18] M. J. Baker, S. R. Hussain, L. Lovergne, V. Untereiner, C. Hughes, R. A. Lukaszewski, G. Thiéfin, G. D. Sockalingum, *Chem. Soc. Rev.* **2016**, *45*, 1803–1818.
[19] V. Balan, C. Mihai, F. Cojocaru, C. Uritu, G. Dodi, D. Botezat, I. Gardikiotis, *Materials* **2019**, *12*, 2884.
[20] L. M. Rodrigues, T. D. M. Alva, H. D. Martinho, J. D. Almeida, *Vib. Spectrosc.* **2019**, *100*, 195–201.
[21] C. Paluszkiewicz, E. Pięta, M. Woźniak, N. Piergies, A. Koniewska, W. Ścierski, M. Misiołek, W. M. Kwiatek, *J. Mol. Liq.* **2020**, *307*, 112961.
[22] M. Diem, *J. Biophotonics* **2018**, *11*, e201800064.
[23] R. A. Shaw, S. Kotowich, M. Leroux, H. H. Mantsch, *Ann. Clin. Biochem.* **1998**, *35*, 624–632.

[24] I. Elsohaby, J. T. McClure, C. B. Riley, J. Bryanton, K. Bigsby, R. A. Shaw, *J. Pharm. Biomed. Anal.* **2018**, *150*, 413 – 419.
[25] K. Spalding, F. Bonnier, C. Bruno, H. Blasco, R. Board, I. Benz-de Bretagne, H. J. Byrne, H. J. Butler, I. Chourpa, P. Radhakrishnan, M. J. Baker, *Vib. Spectrosc.* **2018**, *99*, 50 – 58.
[26] C. Petibois, G. Cazorla, A. Cassaigne, G. Déléris, *Clin. Chem.* **2001**, *47*, 730 – 738.
[27] G. Hoşafçi, O. Klein, G. Oremek, W. Mäntele, *Anal. Bioanal. Chem.* **2007**, *387*, 1815 – 1822.
[28] H. J. Byrne, F. Bonnier, J. McIntyre, D. R. Parachalil, *Clin. Spectrosc.* **2020**, *2*, 100004.
[29] J. Grdadolnik, Y. Maréchal, *Biopolym. Biospectroscopy Sect.* **2001**, *62*, 54 – 67.
[30] H. Fabian, P. Lasch, D. Naumann, *J. Biomed. Opt.* **2005**, *10*, 031103.
[31] N. L. Anderson, N. G. Anderson, *Mol. Cell. Proteomics* **2002**, *1*, 845 – 867.
[32] F. Bonnier, H. Blasco, C. Wasselet, G. Brachet, R. Respaud, L. F. C. S. Carvalho, D. Bertrand, M. J. Baker, H. J. Byrne, I. Chourpa, *Analyst* **2017**, *142*, 1285 – 1298.
[33] F. Bonnier, G. Brachet, R. Duong, T. Sojinrin, R. Respaud, N. Aubrey, M. J. Baker, H. J. Byrne, I. Chourpa, *J. Biophotonics* **2016**, *9*, 1085 – 1097.
[34] D. R. Parachalil, C. Bruno, F. Bonnier, H. Blasco, I. Chourpa, J. McIntyre, H. J. Byrne, *Analyst* **2019**, *144*, 4295 – 4311.
[35] C. Hughes, M. Brown, G. Clemens, A. Henderson, G. Monjardez, N. W. Clarke, P. Gardner, *J. Biophotonics* **2014**, *7*, 180 – 188.
[36] F. Bonnier, M. J. Baker, H. J. Byrne, *Anal. Methods* **2014**, *6*, 5155 – 5160.
[37] E. J. Want, G. O'Maille, C. A. Smith, T. R. Brandon, W. Uritboonthai, C. Qin, S. A. Trauger, G. Siuzdak, *Anal. Chem.* **2006**, *78*, 743 – 752.
[38] D. A. Colantonio, C. Dunkinson, D. E. Bovenkamp, J. E. Van Eyk, *Proteomics* **2005**, *5*, 3831 – 3835.
[39] S. Tulipani, R. Llorach, M. Urpi-Sarda, C. Andres-Lacueva, *Anal. Chem.* **2013**, *85*, 341 – 348.
[40] D. Perez-Guaita, S. Garrigues, M. de la Guardia, *TrAC Trends Anal. Chem.* **2014**, *62*, 93 – 105.
[41] P. E. Geyer, N. A. Kulak, G. Pichler, L. M. Holdt, D. Teupser, M. Mann, *Cell Syst.* **2016**, *2*, 185 – 195.
[42] F. Großerueschkamp, T. Bracht, H. C. Diehl, C. Kuepper, M. Ahrens, A. Kallenbach-Thieltges, A. Mosig, M. Eisenacher, K. Marcus, T. Behrens, T. Brüning, D. Theegarten, B. Sitek, K. Gerwert, *Sci. Rep.* **2017**, *7*, 44829.
[43] J. H. Rabe, D. A. Sammour, S. Schulz, B. Munteanu, M. Ott, K. Ochs, P. Hohenberger, A. Marx, M. Platten, C. A. Opitz, D. S. Ory, C. Hopf, *Sci. Rep.* **2018**, *8*, 6361.
[44] S. Blandin Knight, P. A. Crosbie, H. Balata, J. Chudziak, T. Hussell, C. Dive, *Open Biol.* **2017**, *7*, 170070.
[45] J. M. Ahn, J. Y. Cho, *J. Mol. Biomark. Diagn.* **2013**, *S4*, 001.
[46] J. Ollesch, D. Theegarten, M. Altmayer, K. Darwiche, T. Hager, G. Stamatis, K. Gerwert, *Biomed. Spectrosc. Imaging* **2016**, *5*, 129 – 144.
[47] X. Wang, X. Shen, D. Sheng, X. Chen, X. Liu, *Spectrochim. Acta Part A* **2014**, *122*, 193 – 197.
[48] D. K. R. Medipally, D. Cullen, V. Untereiner, G. D. Sockalingum, A. Maguire, T. N. Q. Nguyen, J. Bryant, E. Noone, S. Bradshaw, M. Finn, M. Dunne, A. M. Shannon, J. Armstrong, A. D. Meade, F. M. Lyng, *Ther. Adv. Med. Oncol.* **2020**, *12* https://doi.org/10.1177/1758835920918499.
[49] H. Ghimire, C. Garlapati, E. A. M. Janssen, U. Krishnamurti, G. Qin, R. Aneja, A. G. Unil Perera, *Cancers* **2020**, *12*, 1708.
[50] C. Hughes, G. Clemens, B. Bird, T. Dawson, K. M. Ashton, M. D. Jenkinson, A. Brodbelt, M. Weida, E. Fotheringham, M. Barre, J. Rowlette, M. J. Baker, *Sci. Rep.* **2016**, *6*, 20173.
[51] J. Ollesch, M. Heinze, H. M. Heise, T. Behrens, T. Brüning, K. Gerwert, *J. Biophotonics* **2014**, *7*, 210 – 221.
[52] K. Gajjar, J. Trevisan, G. Owens, P. J. Keating, N. J. Wood, H. F. Stringfellow, P. L. Martin-Hirsch, F. L. Martin, *Analyst* **2013**, *138*, 3917 – 3926.
[53] R. Roscher, B. Bohn, M. F. Duarte, J. Garcke, *IEEE Access* **2020**, *8*, 42200 – 42216.
[54] A. Barth, *Biochim. Biophys. Acta Bioenerg.* **2007**, *1767*, 1073 – 1101.
[55] K. Z. Liu, R. A. Shaw, A. Man, T. C. Dembinski, H. H. Mantsh, *Clin. Chem.* **2002**, *48*, 499 – 506.
[56] D. Vuckovic, *Anal. Bioanal. Chem.* **2012**, *403*, 1523 – 1548.
[57] R. Aebersold, M. Mann, *Nature* **2003**, *422*, 198 – 207.
[58] R. Aebersold, M. Mann, *Nature* **2016**, *537*, 347 – 355.
[59] J. Cox, M. Y. Hein, C. A. Luber, I. Paron, N. Nagaraj, M. Mann, *Mol. Cell. Proteomics* **2014**, *13*, 2513 – 2526.
[60] T. Fournier, N. Medjoubi-N, D. Porquet, *Biochim. Biophys. Acta Protein Struct. Mol. Enzymol.* **2000**, *1482*, 157 – 171.
[61] 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*, https://doi.org/10.1371/journal.pone.0016957.
[62] P. Zhao, J. Wu, F. Lu, X. Peng, C. Liu, N. Zhou, M. Ying, *BMC Cancer* **2019**, *19*, 201.
[63] P. Dowling, C. Clarke, K. Hennessy, B. Torralbo-Lopez, J. Ballot, J. Crown, I. Kiernan, K. J. O'Byrne, M. J. Kennedy, V. Lynch, M. Clynes, *Int. J. Cancer* **2012**, *131*, 911 – 923.
[64] W. M. Gao, R. Kuick, R. P. Orchekowski, D. E. Misek, J. Qiu, A. K. Greenberg, W. N. Rom, D. E. Brenner, G. S. Omenn, B. B. Haab, S. M. Hanash, *BMC Cancer* **2005**, *5*, 1 – 10.
[65] P. A. Ganz, M. Baras, P. Y. Ma, R. M. Elashoff, *Cancer Res.* **1984**, *44*, 5415 – 5421.
[66] R. Gasparri, G. Sedda, R. Noberini, T. Bonaldi, L. Spaggiari, *Proteomics Clin. Appl.* **2020**, *14*, 1900138.
[67] Y. I. Kim, J. M. Ahn, H. J. Sung, S. S. Na, J. Hwang, Y. Kim, J. Y. Cho, *J. Proteomics* **2016**, *148*, 36 – 43.
[68] T. N. Zamay, G. S. Zamay, O. S. Kolovskaya, R. A. Zukov, M. M. Petrova, A. Gargaun, M. V. Berezovski, A. S. Kichkailo, *Cancers* **2017**, *9*, 155.
[69] R. F. Ritchie, G. E. Palomaki, L. M. Neveux, O. Navolotskaia, T. B. Ledue, W. Y. Craig, *J. Clin. Lab. Anal.* **1999**, *13*, 273 – 279.
[70] Y. N. Lee, *J. Surg. Oncol.* **1977**, *9*, 179 – 187.
[71] Y. Jin, L. Zhao, F. Peng, *Clinics* **2013**, *68*, 686 – 693.
[72] H. R. Scott, D. C. McMillan, L. M. Forrest, D. J. F. Brown, C. S. McArdle, R. Milroy, *Br. J. Cancer* **2002**, *87*, 264 – 267.
[73] 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.
[74] F. Hamilton, R. Carroll, W. Hamilton, C. Salisbury, *Br. J. Cancer* **2014**, *111*, 1410 – 1412.
[75] D. Gupta, C. G. Lis, *Nutr. J.* **2010**, *9*, 69.
[76] S. Ma, W. Wang, B. Xia, S. Zhang, H. Yuan, H. Jiang, W. Meng, X. Zheng, X. Wang, *EBioMedicine* **2016**, *11*, 210 – 218.
[77] Y. Fan, S. Wang, F. Zhang, *Angew. Chem. Int. Ed.* **2019**, *58*, 13208 – 13219; *Angew. Chem.* **2019**, *131*, 13342 – 13353.
[78] J. Watson, C. Salisbury, J. Banks, P. Whiting, W. Hamilton, *Br. J. Cancer* **2019**, *120*, 1045 – 1051.

Manuscript received: March 5, 2021
Revised manuscript received: March 30, 2021
Accepted manuscript online: April 21, 2021
Version of record online: May 26, 2021

