Discretization of Blood-Based Infrared Fingerprints Reveals Biomarker-like Variables for Efficient Multi-omics Integration

Kosmas V. Kepesidis,[1,2,3,*] Nico Feiler,[1,2] Vasileios Papalampropoulos,[1] Michael Trubetskov,[2] Andreas Döpp,[1,2] and Ferenc Krausz[1,2,3,†]

[1]*Fakultät für Physik, Ludwig-Maximilians-Universität München (LMU), Garching, Germany*
[2]*Laboratory for Attosecond Physics, Max Planck Institute of Quantum Optics (MPQ), Garching, Germany*
[3]*Center for Molecular Fingerprinting (CMF), Frontiers Foundation, Budapest, Hungary*

(Received 11 August 2025; accepted 9 March 2026; published 1 April 2026)

We present a methodology that leverages the zero-crossing time gaps of electric-field molecular fingerprints from liquid biopsies to extract information about their molecular composition. By analyzing zero-crossing timings, we demonstrate that these temporal features encode all medically relevant information contained in the original time traces. A clinical study targeting lung cancer detection using human blood samples reveals that aberrations in zero-crossing time gaps are associated with the presence of the disease and its progression. The proposed approach to discretizing continuous molecular signals through zero-crossing analysis enables compact representation of physiological signatures and can facilitate integration with other omics technologies, opening pathways for comprehensive, multidimensional biological insights.

DOI: 10.1103/8x56-xvyh

I. INTRODUCTION

Pioneering research over the past decade has provided growing evidence that individual-based health monitoring through molecular profiling is a promising path toward preventive medicine and early disease detection. A diverse array of "omics" data—spanning genomics, transcriptomics, proteomics, metabolomics, epigenomics, microbiomics, and exposomics—along with data from biosensors, imaging technologies, and electronic health records, is generating an ever-expanding reservoir of health-related information [1–13]. These datasets hold immense potential for deepening our understanding of human biology, predicting disease risk, and enabling earlier detection of pathological states. Realizing this potential in clinical practice depends critically on our ability to quantitatively compare and combine candidate data sources based on their sensitivity to health perturbations.

High-throughput *in vitro* diagnostic technologies are central to identifying disease-related molecular aberrations, particularly in conditions such as cancer, thereby supporting early intervention and improved prognoses [14,15]. These platforms rely on the extraction of low-dimensional representations that capture physiologically and pathologically relevant signals. Such compact data representations enhance disease diagnostics, facilitate population-scale screening, and enable personalized health monitoring [16–24]. By mapping complex biological data into interpretable and scalable formats, these approaches also reduce computational burden and increase compatibility with clinical workflows.

Inspired by these advances, we are motivated to explore whether similarly compact and informative representations could be derived from molecular signals captured through ultrafast infrared spectroscopy [25–30]. In particular, we hypothesize that low-dimensional representations containing diagnostically relevant information could be extracted from the electric-field molecular fingerprints (EMFs) of human blood samples [31,32]. EMF profiles encode the molecular composition of complex biological fluids and reflect systemic physiological and pathological changes. A prior study has shown that EMFs can differentiate between healthy and diseased states by capturing subtle molecular variations, particularly in large-scale studies of cancer detection [33].

Here, we introduce a method that analyzes zero-crossing time gaps in EMF time traces to discretize molecular signals through local phase encoding. We show that this phase information alone is sufficient for medical diagnostic applications while providing a simplified and compact signal representation. In a clinical study of lung cancer detection, we demonstrate that deviations in these time gaps are associated with the presence and progression of the disease. In addition, we show that these deviations occur consistently on the subfemtosecond scale. These results underscore the importance of high temporal resolution in capturing diagnostically relevant features of EMF signals.

Moreover, the compact representation generated through zero-crossing analysis facilitates the comparison of EMF data with other omics platforms such as proteomics [34–37] and metabolomics [17,38–40]. Transforming continuous EMF signals into structured formats supports direct integration with existing multi-omics pipelines. This interoperability opens the door to more holistic diagnostics and biomarker discovery,

*Contact author: kosmas.kepesidis@lmu.de
†Contact author: ferenc.krausz@mpq.mpg.de

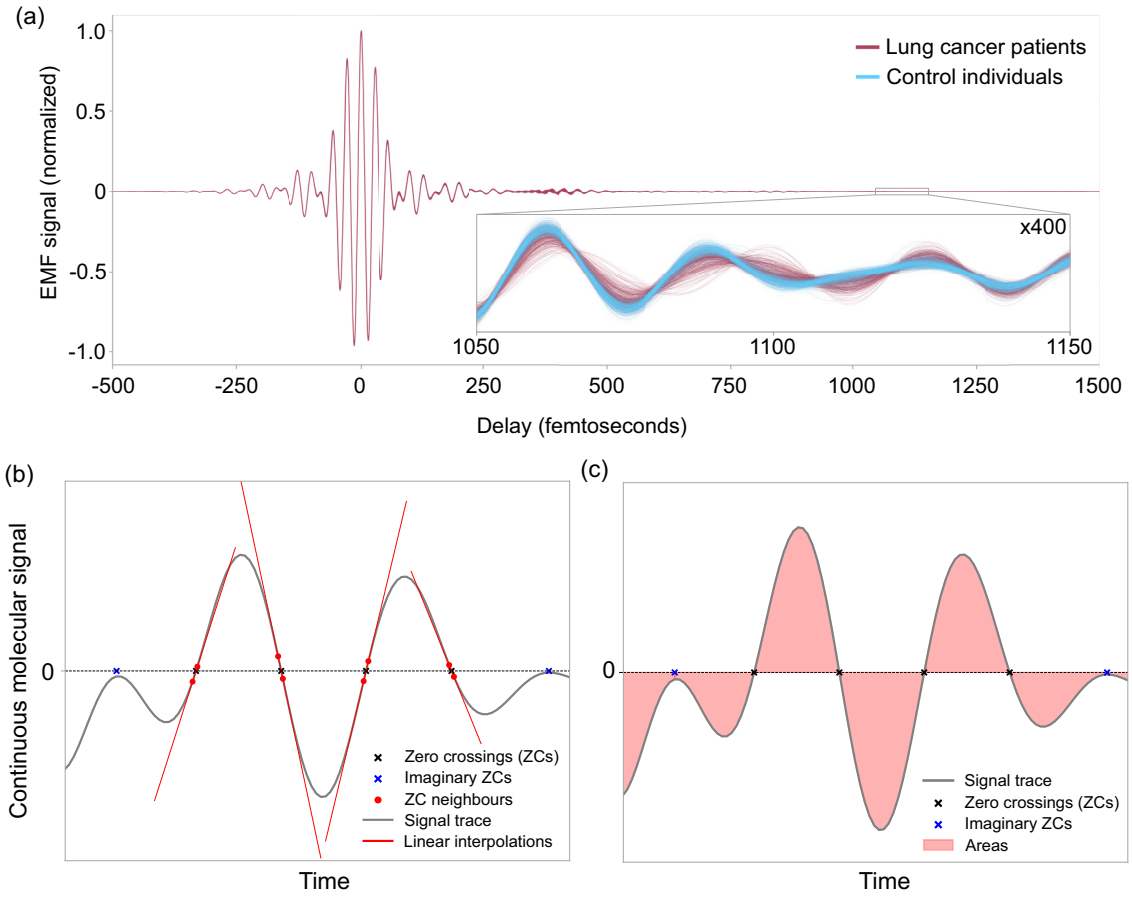

FIG. 1. Representations of electric-field molecular fingerprints. (a) Examples of infrared electric-field molecular fingerprints illustrating EMF signals derived from blood plasma samples of lung cancer patients (magenta) and control individuals (light blue). (b) Zero crossings of the EMF signals are identified through local linear interpolation, with time-gap values between consecutive zero crossings extracted at attosecond precision. (c) Areas under the EMF signals are calculated between consecutive zero crossings.

advancing a systems-level understanding of disease and contributing to the realization of precision medicine.

II. ZERO-CROSSING-BASED REPRESENTATIONS OF ELECTRIC-FIELD MOLECULAR FINGERPRINTS

Infrared vibrational spectroscopy probes molecular bond vibrations, capturing their frequency, phase, and oscillator strength. This approach enables molecular identification and quantification [41] and demonstrates potential as a minimally invasive method for characterizing physiological and biochemical states [14,42–46]. By capturing the vibrational signatures of biomolecules, the established method of Fourier-transform infrared (FTIR) spectroscopy offers a comprehensive molecular snapshot of blood plasma's composition, positioning it as a promising tool for personalized health monitoring and disease prediction [45,47–57]. Unlike FTIR spectroscopy, which uses continuous infrared irradiation, EMFs leverage ultrashort laser pulses for impulsive excitation, followed by time-resolved sampling of the infrared electric field emitted by the sample [27,31]. This method enhances sensitivity by isolating coherent molecular responses and eliminating noise from the excitation source [31].

For complex biofluids such as blood plasma, infrared electric fields emitted by various molecular classes (e.g., proteins, carbohydrates) combine coherently, forming a cross-molecular fingerprint. In a prior study, the use of EMFs demonstrated its potential for *in vitro* blood plasma profiling in cancer diagnostics [33]. Figure 1(a) shows examples of normalized infrared fingerprints of blood plasma samples from lung cancer patients (magenta) and controls (light blue), with an inset showing an expanded view of the delay range from 1050 to 1150 fs.

Building on this approach, we propose that the zero-crossing time gaps in EMF time traces offer a compact, attosecond-precise representation of the molecular information encoded in EMFs, which we term zero-crossing encoding (ZCE). These time gaps encapsulate phase information of the molecular signal. In particular, the spacing Δt between zero crossings is inversely related to the local instantaneous frequency $f(t)$, i.e.,

$$\Delta t \approx \frac{1}{2f(t)}, \tag{1}$$

where $f(t)$ is evaluated near the zero crossing. This phase information is linked to the molecular composition of blood plasma and can enable robust disease phenotyping under real-world conditions.

The present study leverages the phase information encoded in the time gaps of zero crossings in EMF time traces to create a compact representation of the information carried by EMFs. This compact representation consists of a small set of latent variables, referred to as ZCE. We explore two distinct approaches for constructing zero-crossing-based latent variables.

The first approach directly uses the time-gap values between zero crossings as features to construct a tabular dataset. Examples of these zero crossings are illustrated in Fig. 1(b). The method for extracting their precise locations and corresponding time gaps is described in detail in Sec. III B.

The second approach focuses on using the areas enclosed by the EMF time trace between two consecutive zero crossings. This representation captures information about both the phase and amplitude of the EMF time traces, offering a richer data encoding. Figure 1(c) illustrates examples of such enclosed areas. Details regarding the extraction of these zero-crossing-based areas are also provided in Sec. III B.

It is worth noting that latent variables can alternatively be extracted using a data-driven approach. Unsupervised deep learning methods, such as autoencoders, are capable of deriving condensed, information-rich representations [58–61]. However, machine-learning-based techniques are highly dependent on the data used for training. Consequently, these methods often yield latent features that may not generalize well to external datasets, especially in multicentric clinical studies involving diverse populations and varying measurement conditions. This limitation underscores the potential superiority of a physics-driven approach, such as ZCE, for deriving latent variables in such contexts.

To evaluate the clinical utility of the ZCE approach, we apply it to blood-based spectra obtained within the framework of the Lasers4Life (L4L) study, which focuses on the detection of lung cancer [33]. A detailed description of this case-control study is provided in the following section.

III. CASE-CONTROL STUDY SETUP AND DATA ANALYSIS PIPELINE

This work aims to evaluate the potential of ZCE in supporting cancer diagnostics. The analysis was performed using blood samples collected in the framework of the multicenter L4L clinical study. This clinical study was conducted in the Munich area and is registered (ID DRKS00013217) at the German Clinical Trials Register (DRKS). The study cohort includes patients with various types of cancer and asymptomatic control individuals. The population cohort and EMF measurements analyzed here represent a subset of the data utilized in prior research [33]. The subset of study participants analyzed in this work is described in the following section.

A. Study cohort

Figure 2(a) outlines the study design. In particular, the current work focuses on comparing therapy-naïve lung cancer patients (cases) to statistically matched asymptomatic control individuals. Table I provides a detailed breakdown of both groups. Participants, including patients and control individuals, are randomly divided into training and test sets. The training set comprises 942 individuals, representing approximately 81% of the total cohort. EMF measurements for these individuals are conducted in a fully randomized manner over 19 weeks. The remaining 19% (219 individuals) form the test set, measured in randomized order over 2 weeks, following a 10-week gap introduced to ensure robust testing and to account for potential spectrometer performance drifts.

TABLE I. Study cohort characteristics.

Group	No. individuals	Age (years)	% Female	Body mass index (kg/m^2)
Train set				
Cases	471	68 ± 9	46	26 ± 5
Controls	471	62 ± 10	59	26 ± 5
Test set				
Cases	57	68 ± 9	42	26 ± 5
Controls	162	66 ± 10	28	27 ± 5

Binary classification models tailored to the specific medical question are trained using the training dataset. To minimize the influence of obvious confounding factors, the case-control design that makes up the training set is constructed through pairwise statistical matching of controls to cases based on age and sex. This procedure is standard in observational studies. The methodology used here is based on optimal pair matching with Mahalanobis distance within propensity score calipers and is described in relevant textbooks [62].

B. Data acquisition

Figure 2(b) illustrates the data acquisition pipeline. Venous blood samples are processed into plasma according to established standard operating procedures to minimize preanalytical variability [47]. Transmission-mode spectroscopic measurements are performed using an automated sample delivery system. The plasma samples are excited by broadband midinfrared laser pulses (910–1530 cm^{-1} at -20 dB intensity) with a pulse duration of 60 fs (full width at intensity half maximum). Molecular responses, in the form of EMF time traces, are recorded over 40 s using dual-oscillator electro-optic sampling with attosecond precision [25]. Zero crossings are identified using linear interpolation between consecutive data points. Subsequently, the time intervals between consecutive zero crossings, referred to as zero-crossing time gaps, are extracted. To ensure feature consistency across all data, so-called imaginary zero crossings are introduced in specific cases. These cases are defined at locations where a maximum or minimum occurred near zero. In some traces, two actual zero crossings are present at these locations, thereby contributing a corresponding feature to the zero-crossing time-gap representation. For traces lacking such zero crossings at these positions, an imaginary zero crossing is inserted by duplicating the time point at which the local maximum or minimum occurs. This approach results in a zero-valued entry in the zero-crossing time-gap array. Additionally, the areas are extracted by integrating the trace between two consecutive zero crossings and taking the absolute value of the resulting integral.

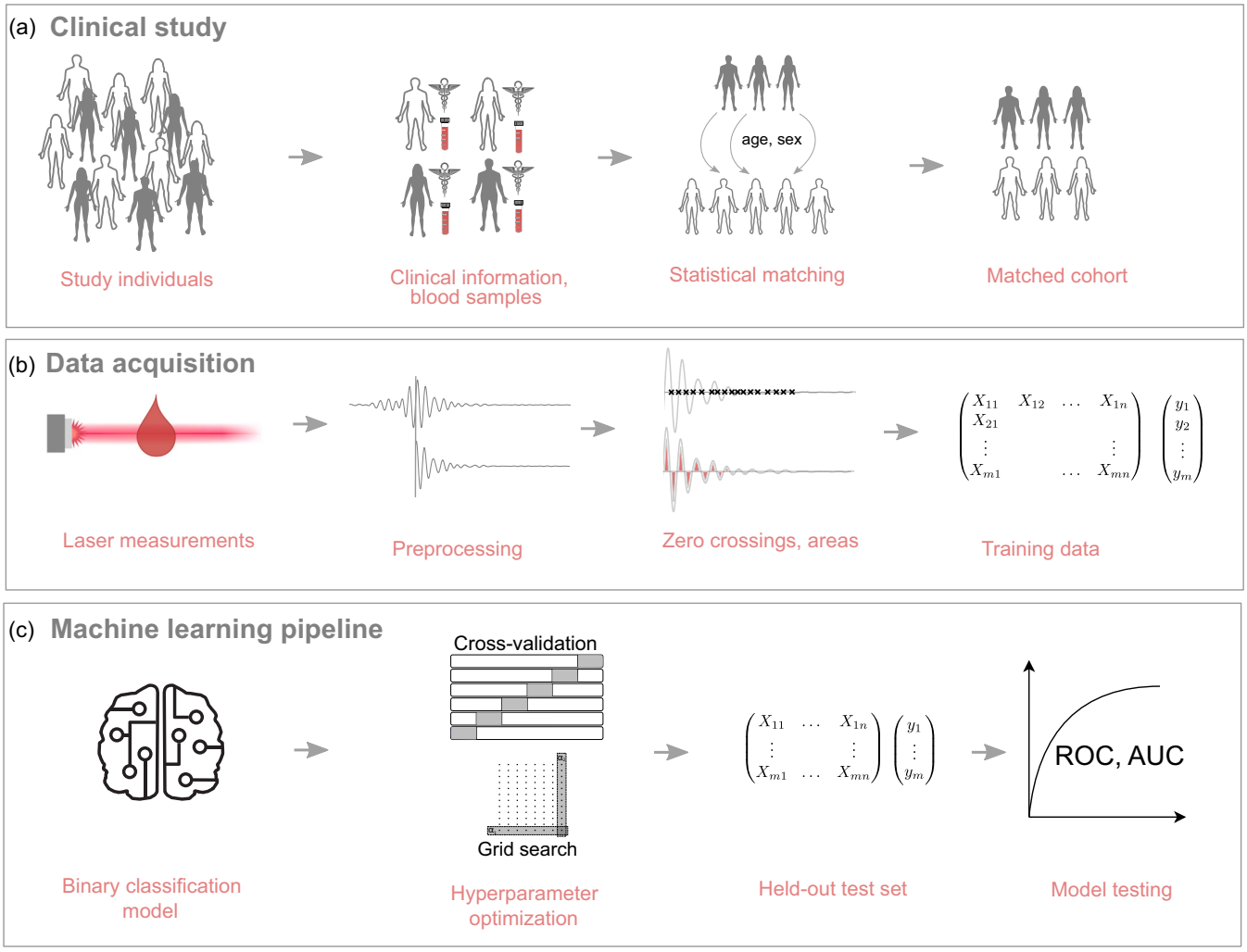

FIG. 2. Case-control study design and workflow. (a) Therapy-naïve lung cancer patients (cases) and statistically matched asymptomatic controls are recruited through the multicenter Lasers4Life clinical study. All participants underwent medical characterization and provided venous blood samples. (b) Blood plasma samples are processed and analyzed using an EMF instrument, generating infrared electric-field molecular fingerprints. The resulting EMF time traces are standardized using a previously described method [32]. Key features, such as zero-crossing time gaps and areas, are extracted for the development of a classification model. (c) Binary classification models are trained using logistic regression with repeated stratified ten-fold cross-validation. Model performance is evaluated using descriptive statistics, reporting the mean and standard deviation of values of area under the curve (AUC) alongside mean receiver operating characteristic (ROC) curves. Optimized models are further validated on a separate held-out test set.

C. Data analysis pipeline

Figure 2(c) summarizes the data analysis methodology. The analysis begins by evaluating the feasibility of ZCE in distinguishing therapy-naïve lung cancer patients (cases) from age- and sex-matched asymptomatic controls within the training set. EMF time traces and ZCE variables (zero-crossing time gaps and areas) are utilized to train machine learning models for binary classification, distinguishing between cancer and noncancer reference groups. To train classification models, the logistic regression algorithm is applied to the ZCE data (and EMF data for comparison). Hyperparameter optimization is performed via a five-fold grid search cross-validation, nested within a ten-fold cross-validation, repeated five times. Classification effectiveness is evaluated based on the area under the receiver operating characteristic (ROC) curve (ROC AUC).

Appendix A offers an introduction to ROC curves and the AUC as performance metrics, highlighting their importance in medical diagnostics. Additionally, Appendix B provides more details on the machine learning methods utilized in this study, including regularized logistic regression; nested cross-validation, used for model optimization and evaluation; and recursive feature elimination, used for feature selection.

Optimized models are validated on an independent held-out test set, which is not statistically matched. Statistical matching for age and sex is limited to the training set.

In addition to the overall analysis, an additional evaluation is conducted within the training set to investigate the relation-

TABLE II. Characteristics of the cohort of lung cancer patients with known TNM* stage.

Group	No. individuals	Age (years)	% Female	Body mass index (kg/m^2)
Stage I				
Cases	74	70 ± 9	45	26 ± 5
Controls	74	70 ± 9	45	26 ± 5
Stage II				
Cases	47	67 ± 10	43	26 ± 4
Controls	47	67 ± 10	43	27 ± 5
Stage III				
Cases	92	68 ± 10	43	26 ± 5
Controls	92	68 ± 10	43	27 ± 7
Stage IV				
Cases	142	68 ± 9	52	25 ± 5
Controls	142	68 ± 9	52	26 ± 4

*The TNM cancer staging system is a standardized, internationally recognized method for classifying solid tumors based on the anatomic extent of disease. It defines cancer stages through three key components tumor size (T), nodal involvement (N), and distant metastasis (M), typically classifying cases from stage I to stage IV.

ship between ZCE and disease progression. Table II details the distribution of lung cancer stages among study participants.

IV. RESULTS

This section evaluates the proposed framework based on a large case-control study targeting lung cancer. In particular, we examine the information encoded in ZCE related to the presence of the disease and its progression.

A. ZCE captures lung-cancer-specific signatures

To assess the ability of ZCE to detect lung cancer, we compare signals from diagnosed individuals to those from age- and sex-matched asymptomatic controls (see Table I). To visualize the information patterns that distinguish cases from controls, we employ the concept of differential fingerprints, as described in previous studies [43,45,47,51]. Differential fingerprints represent the mean difference of each feature between cases (e.g., lung cancer patients) and control individuals. Figure 3 shows such a differential fingerprint for the zero-crossing time gaps. The plot reveals that the observed differences occur predominantly and consistently at the subfemtosecond scale, underscoring the importance of attosecond precision in EMF signal acquisition.

To evaluate the predictive power of ZCE-based features, we train classification models using ZCE-derived features. Figure 4(a) presents results from recursive feature elimination with cross-validation (RFECV) [63], which identifies the most informative subsets of features. Using zero-crossing gaps, model performance peaks with 305 features, though a plateau is reached with fewer than 100. For zero-crossing areas, peak performance is achieved with only 48 features, indicating a more compact yet effective disease-specific signal representation. Details of the RFECV procedure are provided in Sec. 3 of Appendix B, and the distribution of selected zero-crossing time gaps after feature selection is shown in Fig. 8 of the same Appendix.

Importantly, we further explore compressed feature sets that achieve performance within 1 percentage point of the best models. We find that subsets containing only 23 time-gap features or 28 area features maintain high classification accuracy in cross-validation. These compact feature panels align with prior findings from FTIR studies, which suggest optimal class separation can be achieved using infrared biomarkers comprising $m \geqslant 20$ independent variables [64].

While it is notable that a model trained on just 48 zero-crossing area features performs comparably to one trained on the full time trace [Fig. 4(b)], what is more significant is that reducing the feature count below this threshold leads to a measurable decline in performance. This decline suggests that these 48 features capture nonredundant, potentially independent signal components relevant to the emergence of lung cancer. As such, they may not only form the basis of an efficient biomarker assay but also offer high specificity, potentially enabling the stratification of molecularly distinct disease subtypes. This has critical implications for future applications in diagnostics and personalized treatment strategies.

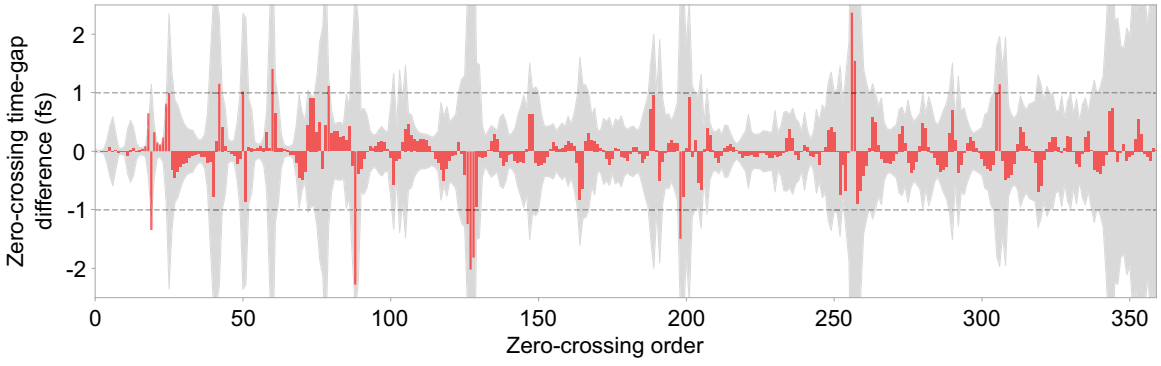

FIG. 3. Differential fingerprint illustrating the average differences in the zero-crossing time gaps between lung cancer patients from matched controls. The shaded area on the background corresponds to the standard deviation of the controls. A larger magnitude of the differential fingerprint indicates a greater average difference between zero-crossing time gaps of cases and controls. The characteristics of the cohort used in this analysis are detailed in Table I.

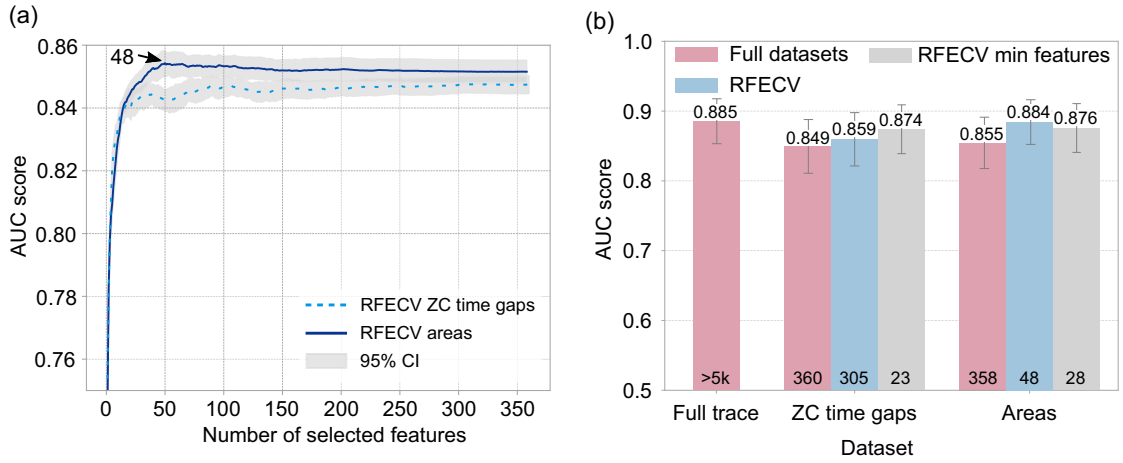

FIG. 4. Performance of ZCE-derived variables in detecting lung cancer. (a) Feature selection using recursive feature elimination with cross-validation (RFECV). (b) Cross-validated AUC on the training set using each feature type.

To provide a comparative perspective on the diagnostic information contained in the spectral representation of the data, we repeat the above analysis using frequency-domain spectra. These spectra are obtained by applying a Fourier transform to the time-domain EMF traces. The resulting spectra and the performance evaluation are presented in Fig. 9 (see Appendix C).

For completeness, the ROC curves associated with all AUC values presented in this work are summarized in Fig. 10 (see Appendix D).

B. Testing ZCE performance under nonidentical conditions

To validate these results and assess generalizability, we evaluate the optimized models on an independent held-out test set (Fig. 5). This dataset was collected 10 weeks after the original measurements used for training, providing a statistically independent sample that more accurately reflects real-world deployment conditions than conventional data-splitting strategies.

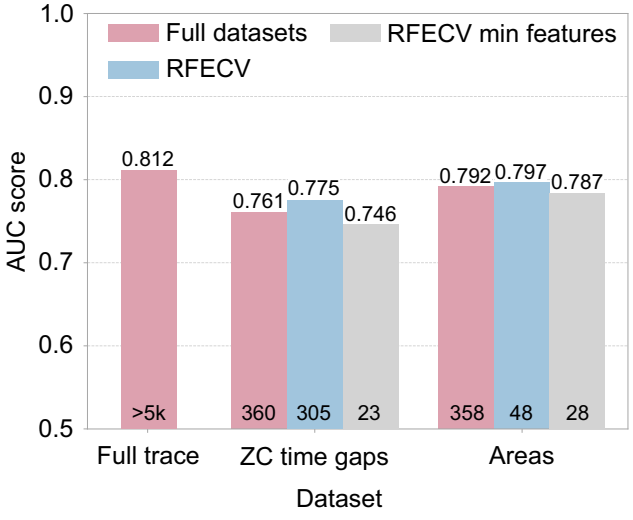

FIG. 5. Test set performance for optimized models trained on the three ZCE-derived variables in detecting lung cancer.

As expected, all models exhibit a modest performance drop, likely attributable to measurement drift between acquisition campaigns, which introduces partial distributional shifts. Despite this drop in performance, models using reduced ZCE-derived feature sets remain competitive with full-feature counterparts. In particular, the model based on 48 selected zero-crossing area features maintains performance comparable to that of the EMF-based model, highlighting the diagnostic strength and robustness of this compact feature panel.

C. ZCE correlates with lung cancer progression

An effective, minimally invasive diagnostic tool must not only distinguish between diseased and healthy individuals but also detect early-stage cancer, when interventions are most effective. Having demonstrated that ZCE can distinguish lung cancer patients from controls, we next investigate whether ZCE variables correlate with disease progression, as defined by the TNM Classification of Malignant Tumors (Union for International Cancer Control) [65]. Patient and control group characteristics by cancer stage are summarized in Table II.

Figure 6(a) presents the cross-validated classification performance for models trained on case-control subsets stratified by lung cancer stage. We evaluate three ZCE-derived signal types: full-time traces (EMF), zero-crossing time gaps (using 305 RFECV-selected features), and zero-crossing areas (using 48 RFECV-selected features). Across all feature types, classification performance increases with advancing disease stage, consistent with the expected dose-response relationship observed in prior studies [33,47,51].

Although early-stage (Stage I and II) classification performance is lower than for late-stage cases, all models perform significantly above chance, indicating a detectable disease-specific ZCE signal even in early disease.

The progressive increase in model performance with advancing disease stage, consistently observed across all ZCE-derived feature types, supports the hypothesis that ZCE captures tumor-specific physiological alterations that intensify

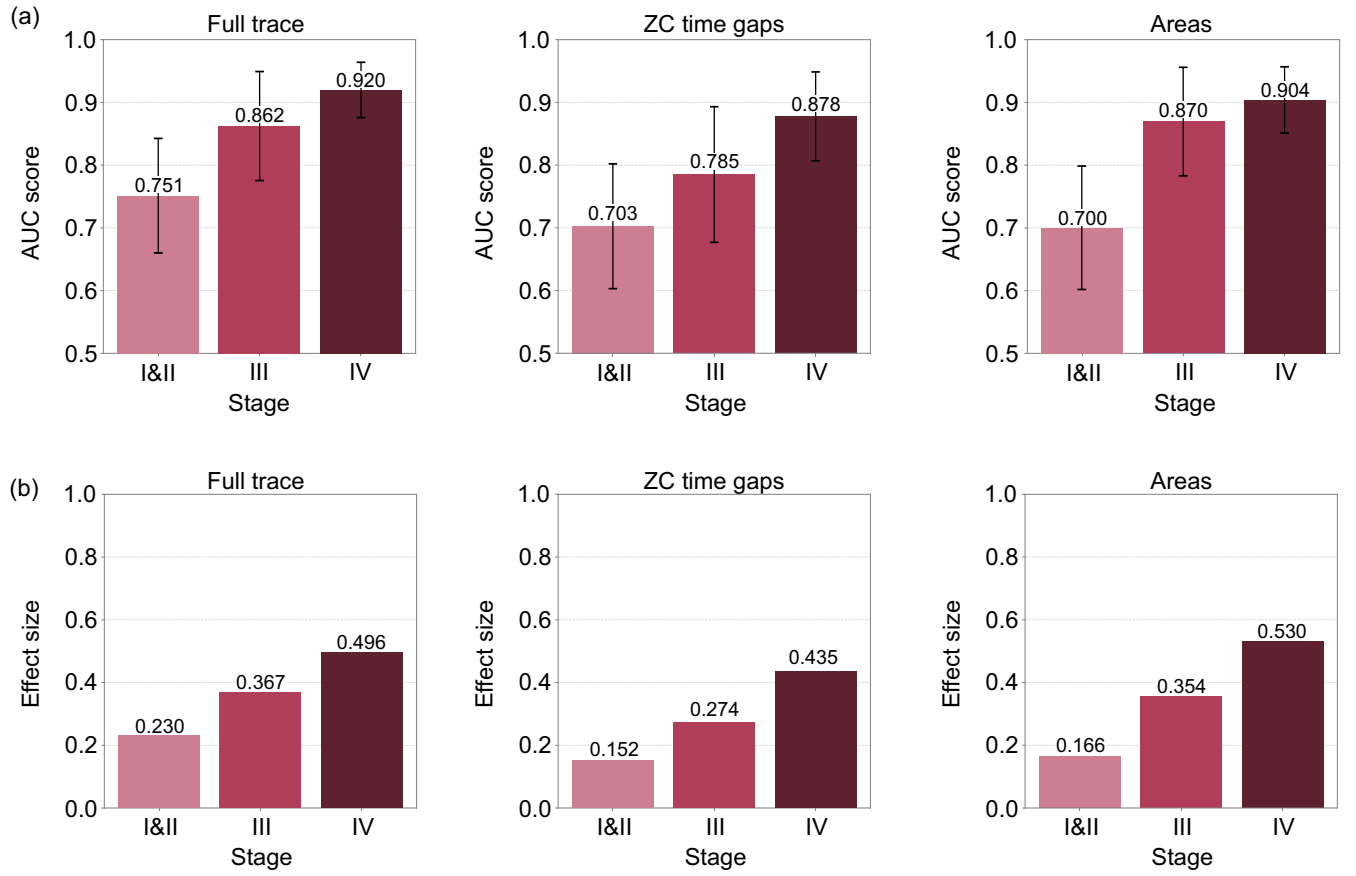

FIG. 6. ZCE signals reflect lung cancer progression. (a) Cross-validated classification performance on the training set for models based on full-time traces (EMF), zero-crossing time gaps (using 305 RFECV-selected features), and zero-crossing areas (using 48 RFECV-selected features), stratified by lung cancer stage. (b) Average effect size (Cohen's d) estimated on the training set for all three cases described in (a).

as cancer progresses. To further quantify the separability between patient and control groups based on these feature sets, we compute the effect size using Cohen's d, defined as

$$d = \frac{|M_1 - M_2|}{s_c}, \tag{2}$$

where M_1 and M_2 represent the group means, and s_c is the standard deviation of the control group [66]. As illustrated in Fig. 6(b), this analysis reveals a growing effect size with increasing disease severity, indicating a widening physiological divergence between patient and control groups in later cancer stages.

V. DISCUSSION

This study presents ZCE as a method for extracting compact sets of medically meaningful variables from EMF time traces. Rather than aiming to greatly boost the diagnostic accuracy of EMF itself, ZCE provides a streamlined collection of clinically relevant features, analogous to the feature sets used in other omics disciplines. By leveraging zero-crossing time gaps and areas measured with attosecond precision, we show that ZCE variables capture disease-specific molecular signatures and offer strong potential for seamless integration with other omics technologies in disease diagnostics and health monitoring.

Our findings show that ZCE identifies subtle, temporally encoded molecular variations associated with lung cancer, even at early stages. The ability to detect these variations at attosecond timescales emphasizes the unique value of time-resolved EMF data, which provides access to previously inaccessible diagnostic features. The dose-response relationship observed between ZCE signal intensity and disease stage further underscores the tumor-specific nature of these temporal biomarkers, aligning with previous findings [47].

Importantly, we demonstrate that ZCE achieves robust classification performance comparable to that of full EMF time traces, while requiring significantly fewer features. This compact representation streamlines the integration of ZCE into existing diagnostic workflows, improving scalability and efficiency. Furthermore, its simplicity facilitates direct comparison and integration with omics datasets, such as proteomics and metabolomics, laying the groundwork for multimodal biomarker discovery and comprehensive disease profiling. Compared to conventional spectroscopic techniques, the ZCE approach captures fewer but informative features, reducing redundancy and noise. This makes it a highly effective tool for feature extraction and model training.

Despite the promising results, this study highlights certain limitations. The observed reduction in model performance on

the independent test set underscores the challenges posed by dataset variability and measurement drifts. These issues are expected to be amplified in real-world scenarios where data are acquired using different electro-optical sampling devices. In such cases, domain adaptation becomes critical. To address this concern, we have previously developed and validated a robust domain adaptation framework based on data augmentation. This method simulates realistic distribution shifts during training and has been shown to improve cross-device generalization [67]. Given its generality, it can be readily applied to the ZCE setting to mitigate device-related biases and enhance model robustness across heterogeneous acquisition platforms.

Finally, while our results demonstrate ZCE's efficacy in lung cancer detection, its performance across other diseases and populations remains to be systematically investigated. Future work should explore ZCE's integration with multiomics and clinical data in broader disease contexts, including metabolic disorders, neurodegenerative diseases, and other cancers, to validate its generalizability and systemic relevance.

In summary, ZCE introduces a novel and scalable modality for capturing ultrafast molecular signatures, with potentially significant implications for precision medicine. Its compact features not only enable efficient diagnostics but also provide a crucial link to broader multitechnology health datasets. By transforming continuous signals into structured data, this approach fosters interoperability across molecular datasets, paving the way for comprehensive diagnostics and deeper, systems-level insights into health and disease.

ACKNOWLEDGMENTS

We thank Mihaela Žigman and all contributors involved in the design and execution of the Lasers4Life clinical study. We also acknowledge insightful discussions with Nicholas Karpowicz, Vladislav Yakovlev, and Alexander Weigel. This work was supported by the Centre for Advanced Laser Applications (CALA) at LMU Munich, the Max Planck Institute of Quantum Optics (MPQ), the Center for Molecular Fingerprinting Research Nonprofit LLC (CMF), and the Frontiers Foundation. The work is part of Project No. 2020-2.1.1-ED-2022-00213 that has been implemented with the support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, financed under the 2020-2.1.1-ED funding scheme.

DATA AVAILABILITY

The data that support the findings of this article are not publicly available because they contain sensitive personal information. The data are available from the authors upon reasonable request.

APPENDIX A: PERFORMANCE METRICS IN MEDICAL DIAGNOSTICS

This appendix summarizes the key performance metrics used to evaluate binary decision systems in medical diagnostics, with a focus on receiver operating characteristic (ROC) curves and the area under the curve (AUC).

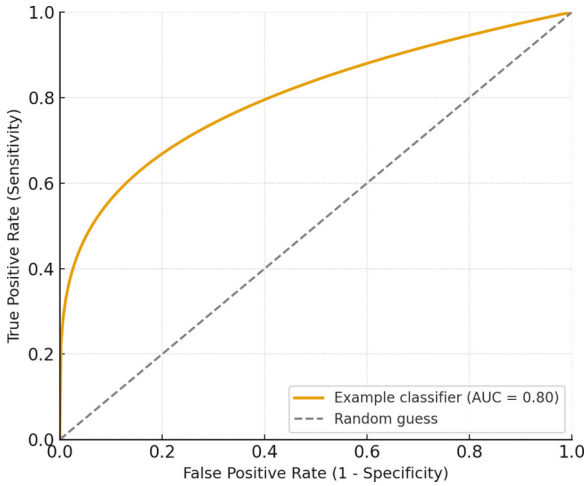

FIG. 7. Receiver operating characteristic (ROC) curve illustrating a classifier with AUC = 0.8. The dashed diagonal corresponds to random guessing (AUC = 0.5). Higher curves indicate improved discriminative ability.

1. Receiver operating characteristic curves

ROC curves are a standard graphical tool for evaluating the performance of binary classification systems. In a binary decision problem, each test or model produces a continuous score or probability reflecting the likelihood that a sample belongs to the positive class (e.g., diseased) rather than the negative class (e.g., healthy). By varying the decision threshold used to classify samples, different pairs of true positive rate (TPR) and false positive rate (FPR) can be obtained. The true positive rate (also referred to as sensitivity or recall) is defined as

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{A1}$$

where TP and FN denote the number of true positives and false negatives, respectively. The false positive rate (also referred to as 1− specificity) is defined as

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}, \tag{A2}$$

where FP and TN denote the number of false positives and true negatives, respectively. Plotting TPR against FPR for all possible thresholds yields the ROC curve, which illustrates the trade-off between sensitivity and specificity of a classifier. A curve closer to the upper-left corner of the plot indicates better discrimination between the two classes. An example of an ROC curve is shown in Fig. 7.

2. The area under the curve (AUC)

The area under the ROC curve (AUC) provides a single scalar measure of classification performance that is independent of the chosen threshold. The AUC represents the probability that the classifier assigns a higher score to a randomly selected positive sample instead of to a randomly

selected negative sample. The value of the AUC ranges from 0.5 (random performance, equivalent to chance) to 1.0 (perfect classification). Values below 0.5 indicate systematic misclassification, where the test performs worse than random guessing. In practical applications, AUC values above 0.8 are considered good, whereas values above 0.9 are often considered excellent, depending on the problem and dataset complexity.

3. Role in medical diagnostics

In medical diagnostics, ROC analysis provides a robust framework for assessing the accuracy of tests that produce continuous or probabilistic outputs, such as biomarker levels, imaging scores, or molecular signatures. Unlike a single threshold-dependent metric (e.g., accuracy), the ROC curve characterizes diagnostic performance across the full range of potential operating points, enabling clinicians or researchers to select thresholds appropriate for specific clinical contexts. The AUC is particularly valuable for comparing diagnostic models or biomarkers, as it summarizes overall discriminative ability independent of disease prevalence or threshold choice. High AUC values indicate strong separation between diseased and nondiseased populations, suggesting high diagnostic utility. In this study, ROC curves and AUC values are used to quantify the discriminative power of features derived from electric-field molecular fingerprints for distinguishing healthy individuals from patients with lung cancer.

APPENDIX B: MACHINE LEARNING CONCEPTS AND METHODS

This appendix summarizes the machine learning methods employed in this study to analyze features derived from EMF time traces. We describe the logistic regression classifier, the selection of the regularization coefficient, the nested cross-validation strategy for model optimization and evaluation, and the recursive feature elimination (RFE) procedure for feature selection.

1. Logistic regression classifier

Logistic regression is a widely used linear classification method in biomedical and physical sciences. Given an input vector of features, $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, the model estimates the probability that a sample belongs to the positive class (e.g., diseased) as

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{n} \beta_i x_i)}}, \tag{B1}$$

where β_0 is the intercept and β_i are model coefficients learned from the data. To improve generalization and reduce overfitting—especially relevant in high-dimensional biomedical datasets—the model is trained with an L2 regularization term (ridge regularization). With L2 regularization, the objective function becomes

$$\mathcal{L}(\boldsymbol{\beta}) = -\sum_{j=1}^{m} [y_j \log \hat{p}_j + (1 - y_j) \log(1 - \hat{p}_j)] + \lambda \sum_{i=1}^{n} \beta_i^2, \tag{B2}$$

where $\hat{p}_j = P(y_j = 1|\mathbf{x}_j)$, m is the number of training samples, and $\lambda \geqslant 0$ controls the strength of regularization. Larger λ values shrink coefficients toward zero, reducing variance but retaining all features. The model defines a linear decision boundary, and samples with $P(y = 1|\mathbf{x})$ above a chosen threshold are assigned to the positive class. Logistic regression remains highly interpretable: the sign and magnitude of each β_i indicate both the direction and strength of each feature's contribution.

2. Nested cross-validation

Nested cross-validation is used to obtain an unbiased estimate of model performance while simultaneously tuning the regularization coefficient λ. The procedure consists of an outer ten-fold cross-validation loop for performance assessment and an inner five-fold cross-validation loop for hyperparameter selection.

In the *outer* loop, the dataset is partitioned into ten folds. In each iteration, one fold serves as the outer validation fold, and the remaining nine folds constitute the outer training set. This loop provides an unbiased estimate of the model's generalization performance, as every sample is used exactly once as held-out validation data.

Within each outer training set, an inner five-fold cross-validation is conducted to identify the optimal value of the L2 regularization coefficient. A predefined grid of candidate values ($\lambda \in \{10^{-4}, 10^{-3}, \ldots, 10^{2}\}$) is evaluated by training the model on four inner folds and validating it on the remaining fold. After all five folds have been used as validation data, the average inner-loop performance (e.g., ROC AUC) is computed. The value of λ yielding the highest mean inner-loop performance is selected as optimal.

The logistic regression model is then retrained on the entire outer training set using the selected λ and evaluated on the corresponding outer validation fold. Repeating this process across all ten outer folds yields a robust estimate of the expected performance on unseen data. Hyperparameter tuning thus remains confined to inner-loop training data, preventing information leakage and ensuring that outer-loop validation metrics reflect true generalization capability. After nested cross-validation is complete, the final model is retrained on the full training set using the selected regularization strength and evaluated on the held-out test set.

3. Recursive feature elimination

Feature selection aims to identify the subset of features that most effectively contribute to accurate classification while reducing dimensionality and overfitting. In this study, we employ recursive feature elimination (RFE), a wrapper method that iteratively removes the least informative features based on model-derived importance scores. RFE proceeds by training a model (e.g., logistic regression) on all features and computing feature importance values, typically the absolute magnitudes of model coefficients. The least important features are removed, and the model is retrained on the reduced feature set. This process is repeated until a specified number of features remain or performance no longer improves. By focusing the classifier on the most informative features,

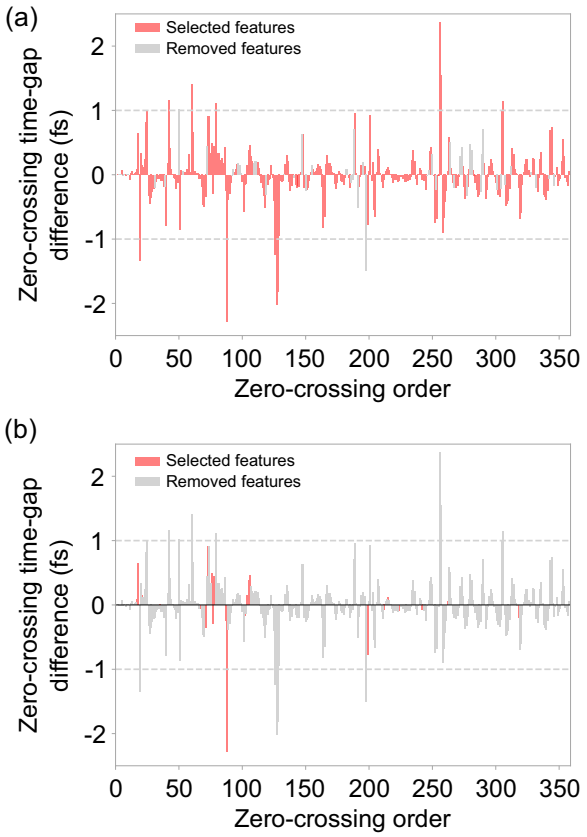

FIG. 8. Differential fingerprints showing the mean zero-crossing time-gap differences between lung cancer patients and matched controls for features selected via RFE: (a) using 305 selected zero-crossing features and (b) using a reduced set of 23 features.

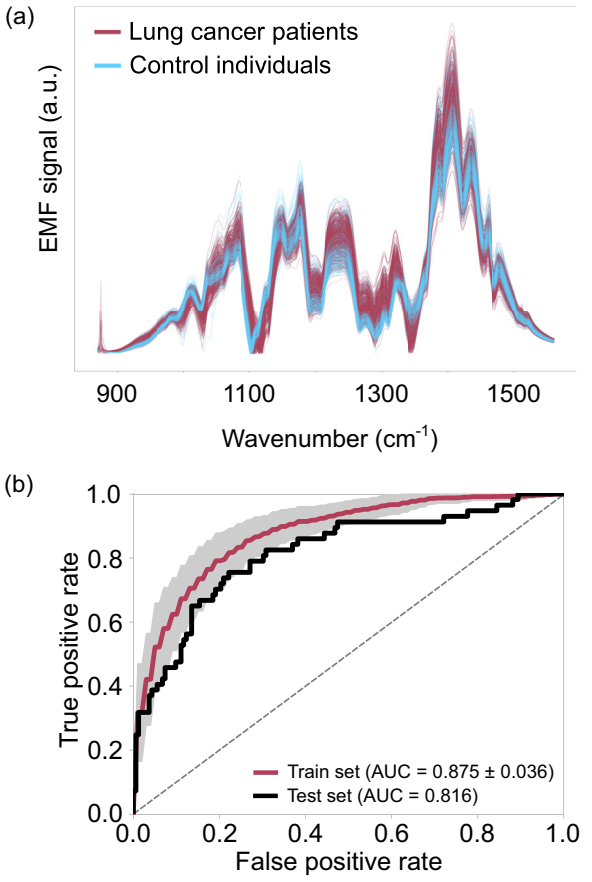

FIG. 9. (a) Absolute value of Fourier-transformed infrared electric-field molecular fingerprints derived from blood plasma samples of lung cancer patients (magenta) and control individuals (light blue). (b) Nested cross-validated training-set classification performance using the absolute value of the Fourier-transformed EMF signals, together with corresponding performance on the held-out test set.

RFE enhances interpretability and can improve predictive performance, particularly in high-dimensional settings such as molecular fingerprint analysis. In this work, RFE helps identify zero-crossing time gaps and areas most strongly associated with disease-related molecular differences. Figure 8 illustrates the distribution of selected zero-crossing time gaps, showing that they are dispersed across the full time trace rather than concentrated in isolated intervals.

APPENDIX C: FREQUENCY DOMAIN ANALYSIS

This Appendix presents an analysis of the electric-field molecular fingerprint (EMF) signals in the frequency domain. It is included for completeness and to provide a comparative perspective on the diagnostic information contained in the spectral representation of the data. The frequency-domain spectra are obtained by applying a Fourier transform to the time-domain EMF traces. Figure 9(a) shows the resulting spectra for both lung cancer patients and matched healthy controls, equivalent to Fig. 1(a) in the main text. To evaluate whether the frequency-domain representation retains diagnostic information comparable to the time-domain data, we train logistic regression classifiers using frequency-domain features. Two ROC curves are provided in Fig. 9(b): one based on cross-validation within the training set and another showing model performance on the held-out test set. The corresponding AUCs indicate that classification performance in the frequency domain is similar to that achieved using the original time-domain data. This finding demonstrates that the essential diagnostic information encoded in the EMF signals is preserved across both temporal and spectral representations. While our main analysis focuses on time-domain features such as zero-crossing time gaps, the comparable performance of frequency-domain models supports the robustness of EMF-based disease detection.

APPENDIX D: RECEIVER OPERATING CHARACTERISTIC CURVES FOR ALL CLASSIFICATION TASKS

This Appendix presents the ROC curves for all classification analyses performed in this study. Figure 10 shows the ROC curves for all classification tasks discussed in the main text. The corresponding AUC values are indicated in the figure legends.

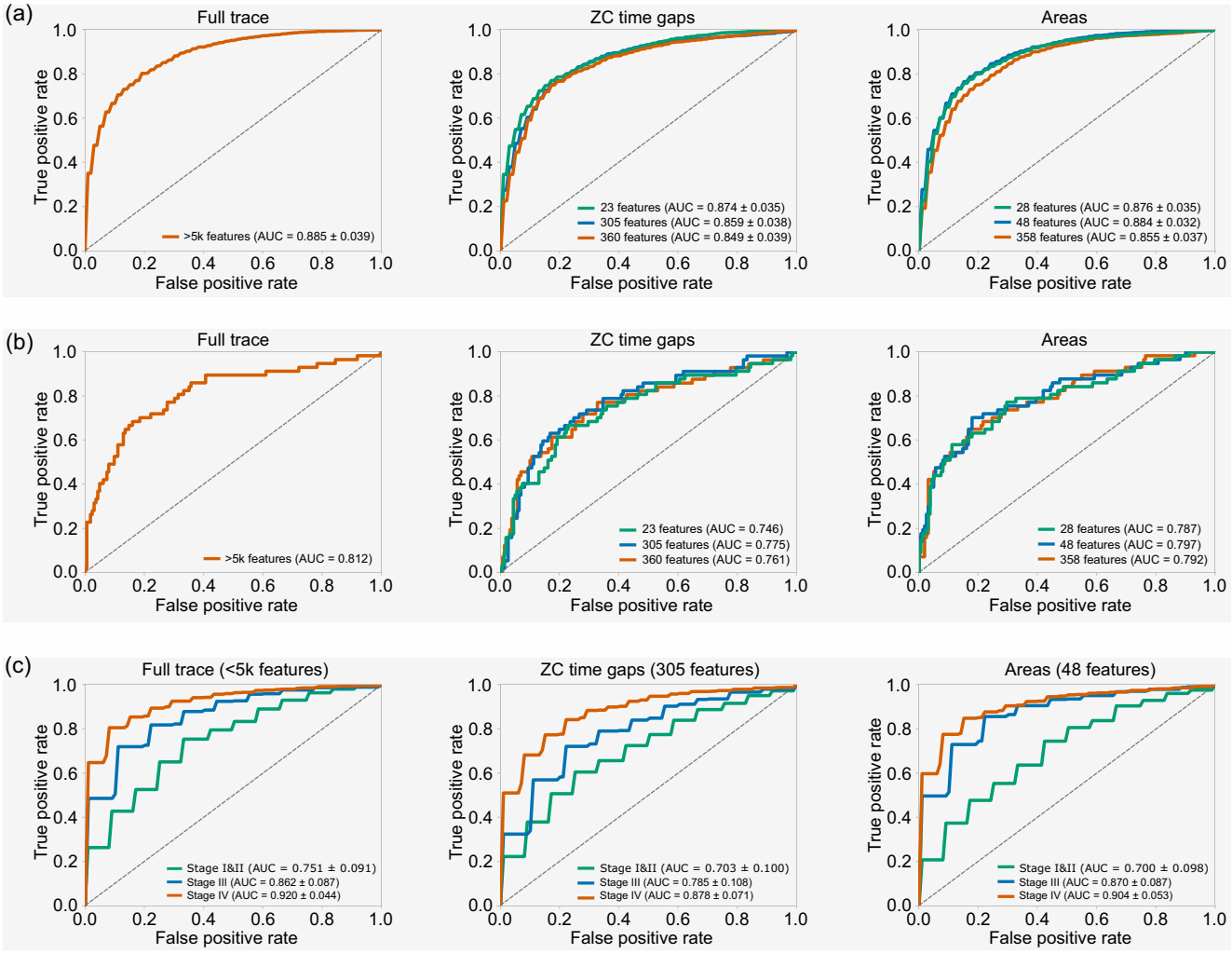

FIG. 10. ROC curves for the classification tasks discussed in the main text, corresponding to (a) Fig. 4(b), (b) Fig. 5, and (c) Fig. 6(a).

[1] R. Chen, G. I. Mias, J. Li-Pook-Than, L. Jiang, H. Y. Lam, R. Chen, E. Miriami, K. J. Karczewski, M. Hariharan, F. E. Dewey, *et al.*, Personal omics profiling reveals dynamic molecular and medical phenotypes, Cell **148**, 1293 (2012).

[2] F. S. Collins and H. Varmus, A new initiative on precision medicine, N. Engl. J. Med. **372**, 793 (2015).

[3] N. D. Price, A. T. Magis, J. C. Earls, G. Glusman, R. Levy, C. Lausted, D. T. McDonald, U. Kusebauch, C. L. Moss, Y. Zhou, *et al.*, A wellness study of 108 individuals using personal, dense, dynamic data clouds, Nat. Biotechnol. **35**, 747 (2017).

[4] B. D. Piening, W. Zhou, K. Contrepois, H. Röst, G. J. G. Urban, T. Mishra, B. M. Hanson, E. J. Bautista, S. Leopold, C. Y. Yeh, *et al.*, Integrative personal omics profiles during periods of weight gain and loss, Cell Syst. **6**, 157 (2018).

[5] S. M. Schüssler-Fiorenza Rose, K. Contrepois, K. J. Moneghetti, W. Zhou, T. Mishra, S. Mataraso, O. Dagan-Rosenfeld, A. B. Ganz, J. Dunn, D. Hornburg, *et al.*, A longitudinal big data approach for precision health, Nat. Med. **25**, 792 (2019).

[6] S. Ahadi, W. Zhou, S. M. Schüssler-Fiorenza Rose, M. R. Sailani, K. Contrepois, M. Avina, M. Ashland, A. Brunet, and M. Snyder, Personal aging markers and ageotypes revealed by deep longitudinal profiling, Nat. Med. **26**, 83 (2020).

[7] J. C. Earls, N. Rappaport, L. Heath, T. Wilmanski, A. T. Magis, N. J. Schork, G. S. Omenn, J. Lovejoy, L. Hood, and N. D. Price, Multi-omic biological age estimation and its correlation with wellness and disease phenotypes: A longitudinal study of 3,558 individuals, J. Gerontol. Ser. A **74**, S52 (2019).

[8] Y.-C. C. Hou, H.-C. Yu, R. Martin, E. T. Cirulli, N. M. Schenker-Ahmed, M. Hicks, I. V. Cohen, T. J. Jönsson, R. Heister, L. Napier, *et al.*, Precision medicine integrating whole-genome sequencing, comprehensive metabolomics, and advanced imaging, Proc. Natl. Acad. Sci. USA **117**, 3053 (2020).

[9] A. Tebani, A. Gummesson, W. Zhong, I. S. Koistinen, T. Lakshmikanth, L. M. Olsson, F. Boulund, M. Neiman, H. Stenlund, C. Hellström, *et al.*, Integration of molecular profiles in a longitudinal wellness profiling cohort, Nat. Commun. **11**, 4487 (2020).

[10] X. Chen, J. Gole, A. Gore, Q. He, M. Lu, J. Min, Z. Yuan, X. Yang, Y. Jiang, T. Zhang, *et al.*, Non-invasive early detection of cancer four years before conventional diagnosis using a blood test, Nat. Commun. **11**, 3475 (2020).

[11] E. J. Topol, Individualized medicine from prewomb to tomb, Cell **157**, 241 (2014).

[12] C. E. Mason, S. G. Porter, and T. M. Smith, Characterizing multi-omic data in systems biology, Syst. Anal. Hum. Multigene Disord. **799**, 15 (2013).

[13] S. Shilo, H. Rossman, and E. Segal, Axes of a revolution: Challenges and promises of big data in healthcare, Nat. Med. **26**, 29 (2020).

[14] 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, Nat. Commun. **10**, 4501 (2019).

[15] J. Gala de Pablo, M. Lindley, K. Hiramatsu, and K. Goda, High-throughput Raman flow cytometry and beyond, Acc. Chem. Res. **54**, 2132 (2021).

[16] V. Kulasingam and E. P. Diamandis, Strategies for discovering novel cancer biomarkers through utilization of emerging technologies, Nat. Clin. Pract. Oncol. **5**, 588 (2008).

[17] B. Roig, M. Rodríguez-Balada, S. Samino, E. W.-F. Lam, S. Guaita-Esteruelas, A. R. Gomes, X. Correig, J. Borràs, O. Yanes, and J. Gumà, Metabolomics reveals novel blood plasma biomarkers associated to the BRCA1-mutated phenotype of human breast cancer, Sci. Rep. **7**, 17831 (2017).

[18] X. Han, J. Wang, and Y. Sun, Circulating tumor DNA as biomarkers for cancer detection, Genomics, Proteomics Bioinf. **15**, 59 (2017).

[19] H. Schwarzenbach, D. S. Hoon, and K. Pantel, Cell-free nucleic acids as biomarkers in cancer patients, Nat. Rev. Cancer **11**, 426 (2011).

[20] A. Sudhindra, R. Ochoa, and E. S. Santos, Biomarkers, prediction, and prognosis in non–small-cell lung cancer: A platform for personalized treatment, Clin. Lung Cancer **12**, 360 (2011).

[21] L. M. Seijo, N. Peled, D. Ajona, M. Boeri, J. K. Field, G. Sozzi, R. Pio, J. J. Zulueta, A. Spira, P. P. Massion, *et al.*, Biomarkers in lung cancer screening: Achievements, promises, and challenges, J. Thorac. Oncol. **14**, 343 (2019).

[22] M. P. Davies, T. Sato, H. Ashoor, L. Hou, T. Liloglou, R. Yang, and J. K. Field, Plasma protein biomarkers for early prediction of lung cancer, eBioMedicine **93**, 104686 (2023).

[23] V. Bataille, B. S. Kato, M. Falchi, J. Gardner, M. Kimura, M. Lens, U. Perks, A. M. Valdes, D. C. Bennett, A. Aviv, *et al.*, Nevus size and number are associated with telomere length and represent potential markers of a decreased senescence *in vivo*, Cancer Epidemiol. Biomarkers Prev. **16**, 1499 (2007).

[24] H. Julkunen, A. Cichońska, M. Tiainen, H. Koskela, K. Nybo, V. Mäkelä, J. Nokso-Koivisto, K. Kristiansson, M. Perola, V. Salomaa, *et al.*, Atlas of plasma NMR biomarkers for health and disease in 118,461 individuals from the UK Biobank, Nat. Commun. **14**, 604 (2023).

[25] A. Weigel, P. Jacob, W. Schweinberger, M. Huber, M. Trubetskov, P. Karandušovský, C. Hofer, T. Buberl, T. Amotchkina, M. Högner, *et al.*, Dual-oscillator infrared electro-optic sampling with attosecond precision, Optica **11**, 726 (2024).

[26] M. Maiuri, M. Garavelli, and G. Cerullo, Ultrafast spectroscopy: State of the art and open challenges, J. Am. Chem. Soc. **142**, 3 (2020).

[27] A. S. Kowligy, H. Timmers, A. J. Lind, U. Elu, F. C. Cruz, P. G. Schunemann, J. Biegert, and S. A. Diddams, Infrared electric field sampled frequency comb spectroscopy, Sci. Adv. **5**, eaaw8794 (2019).

[28] C. Riek, D. V. Seletskiy, and A. Leitenstorfer, Femtosecond measurements of electric fields: From classical amplitudes to quantum fluctuations, Eur. J. Phys. **38**, 024003 (2017).

[29] R. Chikkaraddy, R. Arul, L. A. Jakob, and J. J. Baumberg, Single-molecule mid-infrared spectroscopy and detection through vibrationally assisted luminescence, Nat. Photon. **17**, 865 (2023).

[30] D. Adamou, L. Hirsch, T. Shields, S. Yoon, A. C. Dada, J. M. Weaver, D. Faccio, M. Peccianti, L. Caspani, and M. Clerici, Quantum-enhanced time-domain spectroscopy, Sci. Adv. **11**, eadt2187 (2025).

[31] I. Pupeza, M. Huber, M. Trubetskov, W. Schweinberger, S. A. Hussain, C. Hofer, K. Fritsch, M. Poetzlberger, L. Vamos, E. Fill, *et al.*, Field-resolved infrared spectroscopy of biological systems, Nature (London) **577**, 52 (2020).

[32] M. Huber, M. Trubetskov, W. Schweinberger, P. Jacob, M. Zigman, F. Krausz, and I. Pupeza, Standardized electric-field-resolved molecular fingerprinting, Anal. Chem. **96**, 13110 (2024).

[33] K. V. Kepesidis, P. Jacob, W. Schweinberger, M. Huber, N. Feiler, F. Fleischmann, M. Trubetskov, L. Voronina, J. Aschauer, T. Eissa, *et al.*, Electric-field molecular fingerprinting to probe cancer, ACS Cent. Sci. **11**, 560 (2025).

[34] 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 Mol. Med. **11**, e10427 (2019).

[35] P. E. Geyer, L. M. Holdt, D. Teupser, and M. Mann, Revisiting biomarker discovery by plasma proteomics, Mol. Syst. Biol. **13**, 942 (2017).

[36] A. C. Uzozie and R. Aebersold, Advancing translational research and precision medicine with targeted proteomics, J. Proteomics **189**, 1 (2018).

[37] J. M. Bader, V. Albrecht, and M. Mann, MS-based proteomics of body fluids: The end of the beginning, Mol. Cell. Proteomics **22**, 100577 (2023).

[38] J. Xia, D. I. Broadhurst, M. Wilson, and D. S. Wishart, Translational biomarker discovery in clinical metabolomics: An introductory tutorial, Metabolomics **9**, 280 (2013).

[39] S. Qiu, Y. Cai, H. Yao, C. Lin, Y. Xie, S. Tang, and A. Zhang, Small molecule metabolites: Discovery of biomarkers and therapeutic targets, Signal Transduction Targeted Ther. **8**, 132 (2023).

[40] J. D. Zhang, C. Xue, V. B. Kolachalama, and W. A. Donald, Interpretable machine learning on metabolomics data reveals biomarkers for Parkinson's disease, ACS Cent. Sci. **9**, 1035 (2023).

[41] P. R. Griffiths, Fourier transform infrared spectrometry, Science **222**, 297 (1983).

[42] 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, *et al.*, Using Fourier transform IR spectroscopy to analyze biological materials, Nat. Protoc. **9**, 1771 (2014).

[43] L. Voronina, C. Leonardo, J. B. Mueller-Reif, P. E. Geyer, M. Huber, M. Trubetskov, K. V. Kepesidis, J. Behr, M. Mann, F. Krausz, *et al.*, Molecular origin of blood-based infrared spectroscopic fingerprints, Angew. Chem. **133**, 17197 (2021).

[44] M. Paraskevaidi, B. J. Matthew, B. J. Holly, B. J. Hugh, C. P. Thulya, C. Loren, C. St John, G. Peter, G. Callum, K. G. Sergei, *et al.*, Clinical applications of infrared and Raman spectroscopy in the fields of cancer and infectious diseases, Appl. Spectrosc. Rev. **56**, 804 (2021).

[45] 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, Nat. Commun. **12**, 1511 (2021).

[46] K. V. Kepesidis, Z. I. Zarandy, F. B. Nemeth, L. Gigou, M. Žigman, and F. Krausz, Integration of infrared molecular fingerprinting data in a longitudinal health profiling cohort, in *International Conference on Artificial Intelligence in Medicine* (Springer, Berlin, 2025), pp. 180–190.

[47] M. Huber, K. V. Kepesidis, L. Voronina, F. Fleischmann, E. Fill, J. Hermann, I. Koch, K. Milger-Kneidinger, T. Kolben, G. B. Schulz, *et al.*, Infrared molecular fingerprinting of blood-based liquid biopsies for the detection of cancer, eLife **10**, e68758 (2021).

[48] H. Ghimire, C. Garlapati, E. A. Janssen, U. Krishnamurti, G. Qin, R. Aneja, and A. U. Perera, Protein conformational changes in breast cancer sera using infrared spectroscopic analysis, Cancers **12**, 1708 (2020).

[49] J. Ollesch, D. Theegarten, M. Altmayer, K. Darwiche, T. Hager, G. Stamatis, and K. Gerwert, An infrared spectroscopic blood test for non-small cell lung carcinoma and subtyping into pulmonary squamous cell carcinoma or adenocarcinoma, Biomed. Spectrosc. Imaging **5**, 129 (2016).

[50] T. Eissa, C. Leonardo, K. V. Kepesidis, F. Fleischmann, B. Linkohr, D. Meyer, V. Zoka, M. Huber, L. Voronina, L. Richter, *et al.*, Plasma infrared fingerprinting with machine learning enables single-measurement multi-phenotype health screening, Cell Rep. Med. **5**, 101625 (2024).

[51] K. V. Kepesidis, M.-G. Stoleriu, N. Feiler, L. Gigou, F. Fleischmann, J. Aschauer, S. Eiselen, I. Koch, N. Reinmuth, A. Tufman, *et al.*, Assessing lung cancer progression and survival with infrared spectroscopy of blood serum, BMC Med. **23**, 101 (2025).

[52] K. V. Kepesidis, M. Bozic-Iven, M. Huber, N. Abdel-Aziz, S. Kullab, A. Abdelwarith, A. Al Diab, M. Al Ghamdi, M. A. Hilal, M. R. Bahadoor, *et al.*, Breast-cancer detection using blood-based infrared molecular fingerprints, BMC Cancer **21**, 1287 (2021).

[53] J. Backhaus, R. Mueller, N. Formanski, N. Szlama, H.-G. Meerpohl, M. Eidt, and P. Bugert, Diagnosis of breast cancer with infrared spectroscopy from serum samples, Vib. Spectrosc. **52**, 173 (2010).

[54] F. Elmi, A. F. Movaghar, M. M. Elmi, H. Alinezhad, and N. Nikbakhsh, Application of FT-IR spectroscopy on breast cancer serum analysis, Spectrochim. Acta, Part A **187**, 87 (2017).

[55] J. Ollesch, M. Heinze, H. M. Heise, T. Behrens, T. Brüning, and K. Gerwert, It's in your blood: Spectral biomarker candidates for urinary bladder cancer from automated FTIR spectroscopy, J. Biophotonics **7**, 210 (2014).

[56] D. Anderson, R. Anderson, S. Moug, and M. Baker, Liquid biopsy for cancer diagnosis using vibrational spectroscopy: Systematic review, BJS Open **4**, 554 (2020).

[57] L. Gigou, K. V. Kepesidis, and F. Krausz, Towards precision medicine with infrared molecular profiles: Identifying and explaining subgroups, in *International Conference on Artificial Intelligence in Medicine* (Springer, Berlin, 2025), pp. 176–180.

[58] G. E. Hinton and R. R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science **313**, 504 (2006).

[59] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, Cambridge, MA, 2016).

[60] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, Extracting and composing robust features with denoising autoencoders, in *Proceedings of the 25th International Conference on Machine Learning* (ACM Press, New York, 2008), pp. 1096–1103.

[61] C. Wegner, Z. I. Zarandy, N. Feiler, L. Gigou, T. Halenke, N. Leopold-Kerschbaumer, M. Krusche, W. Skibicka, and K. V. Kepesidis, Toward informative representations of blood-based infrared spectra via unsupervised deep learning, J. Biophoton. **18**, e70011 (2025).

[62] P. R. Rosenbaum, P. B. Rosenbaum, and Briskman, *Design of Observational Studies* (Springer, Berlin, 2010), Vol. 10.

[63] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, Scikit-learn: Machine learning in python, J. Mach. Learn. Res. **12**, 2825 (2011).

[64] T. Eissa, K. V. Kepesidis, M. Zigman, and M. Huber, Limits and prospects of molecular fingerprinting for phenotyping biological systems revealed through *in silico* modeling, Anal. Chem. **95**, 6523 (2023).

[65] A. G. Nicholson, M. S. Tsao, M. B. Beasley, A. C. Borczuk, E. Brambilla, W. A. Cooper, S. Dacic, D. Jain, K. M. Kerr, S. Lantuejoul, *et al.*, The 2021 WHO classification of lung tumors: Impact of advances since 2015, J. Thorac. Oncol. **17**, 362 (2022).

[66] S. S. Sawilowsky, New effect size rules of thumb, J. Mod. App. Stat. Meth. **8**, 597 (2009).

[67] F. B. Nemeth, N. Leopold-Kerschbaumer, D. Debreceni, F. Fleischmann, K. Borbely, D. Mazurencu-Marinescu-Pele, T. Bocklitz, M. Zigman, and K. V. Kepesidis, Bridging spectral gaps: Cross-device model generalization in blood-based infrared spectroscopy, Anal. Chem. **97**, 10264 (2025).