analytical chemistry

pubs.acs.org/ac
Article

Bridging Spectral Gaps: Cross-Device Model Generalization in Blood-Based Infrared Spectroscopy

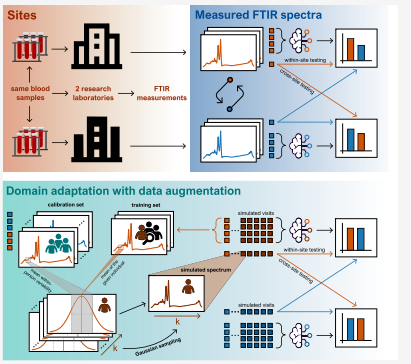
Flora B. Nemeth,[#] Niklas Leopold-Kerschbaumer,[#] Diana Debreceni, Frank Fleischmann, Krisztian Borbely, David Mazurencu-Marinescu-Pele, Thomas Bocklitz, Mihaela Žigman, and Kosmas V. Kepesidis[*]

Cite This: *Anal. Chem.* 2025, 97, 10264−10272

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: This paper presents a solution to the challenge of cross-device model generalization in blood-based infrared spectroscopy. As infrared spectroscopy becomes increasingly popular for analyzing human blood, ensuring that machine learning models trained on one device can be effectively transferred to others is essential. However, variations in device characteristics often reduce model performance when applied across different devices. To address this issue, we propose a straightforward domain adaptation method based on data augmentation incorporating device-specific differences. By expanding the training data to include a broader range of nuances, our approach enhances the model's ability to adapt to the unique characteristics of various devices. We validate the effectiveness of our method through experimental testing on two Fourier-Transform Infrared (FTIR) spectroscopy devices from different research laboratories, demonstrating improved prediction accuracy and reliability.



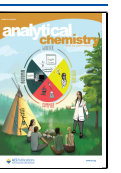
INTRODUCTION

Infrared spectroscopy has emerged as a powerful tool in medical diagnostics, particularly in the analysis of human blood samples.[1−7] It has shown its ability to differentiate between benign and malignant tissues,[8,9] and identify unique spectral fingerprints in biofluids,[4,5] showcasing its potential to revolutionize disease detection. Specifically, Fourier-transform infrared (FTIR) spectroscopy achieves this by capturing the absorption patterns of infrared light across different frequencies,[10] and serves as a tool in the identification and analysis of a wide spectrum of biomolecules, such as proteins, lipids, carbohydrates, and nucleic acids. This capability not only can aid in the detection of diseases but also their ongoing monitoring, underpinning significant advancements in medical diagnostics.[11−19]

With the advent of large-scale observational studies involving thousands of individuals and multiple follow-up visits,[20−22] the potential of infrared spectroscopy in understanding disease progression and personalized medicine can be significantly amplified.[23] By capturing the unique spectral fingerprints of molecular components within blood samples, such studies hold immense promise for advancing our understanding of various health conditions. However, the practical deployment of infrared spectroscopy in large observational studies can be hindered by the challenge of ensuring the generalizability of machine learning models across different instrumentation setups. In particular, variations in device characteristics, such as spectral resolution, noise levels, and instrumental drift, pose significant hurdles to achieving robust and reliable predictive performance across devices, especially when dealing with data from large cohorts.[9,24] In the context of FTIR spectroscopy, the problem of cross-device model generalization becomes particularly pronounced. Even subtle differences between FTIR devices can result in discrepancies in the acquired spectra, leading to a degradation in model performance when applied to data from a different device.

Domain adaptation techniques are becoming a standard strategy across various fields, demonstrating their versatility and effectiveness.[25] In medical image segmentation, deep stacked transformations augmentation techniques have established a strong benchmark for domain generalization, rivaling fully supervised methods in accuracy while enhancing the feasibility of deep learning segmentation in practice.[26] Similarly, in human activity recognition, data augmentation significantly boosts model performance, particularly when target data are unlabeled, offering practical solutions for real-life scenarios such as assistive living monitoring with varying viewpoints.[27] The approach also extends to enhancing hate speech detection by generating domain-adapted training data,

Received: January 9, 2025
Revised: April 13, 2025
Accepted: April 16, 2025
Published: May 7, 2025



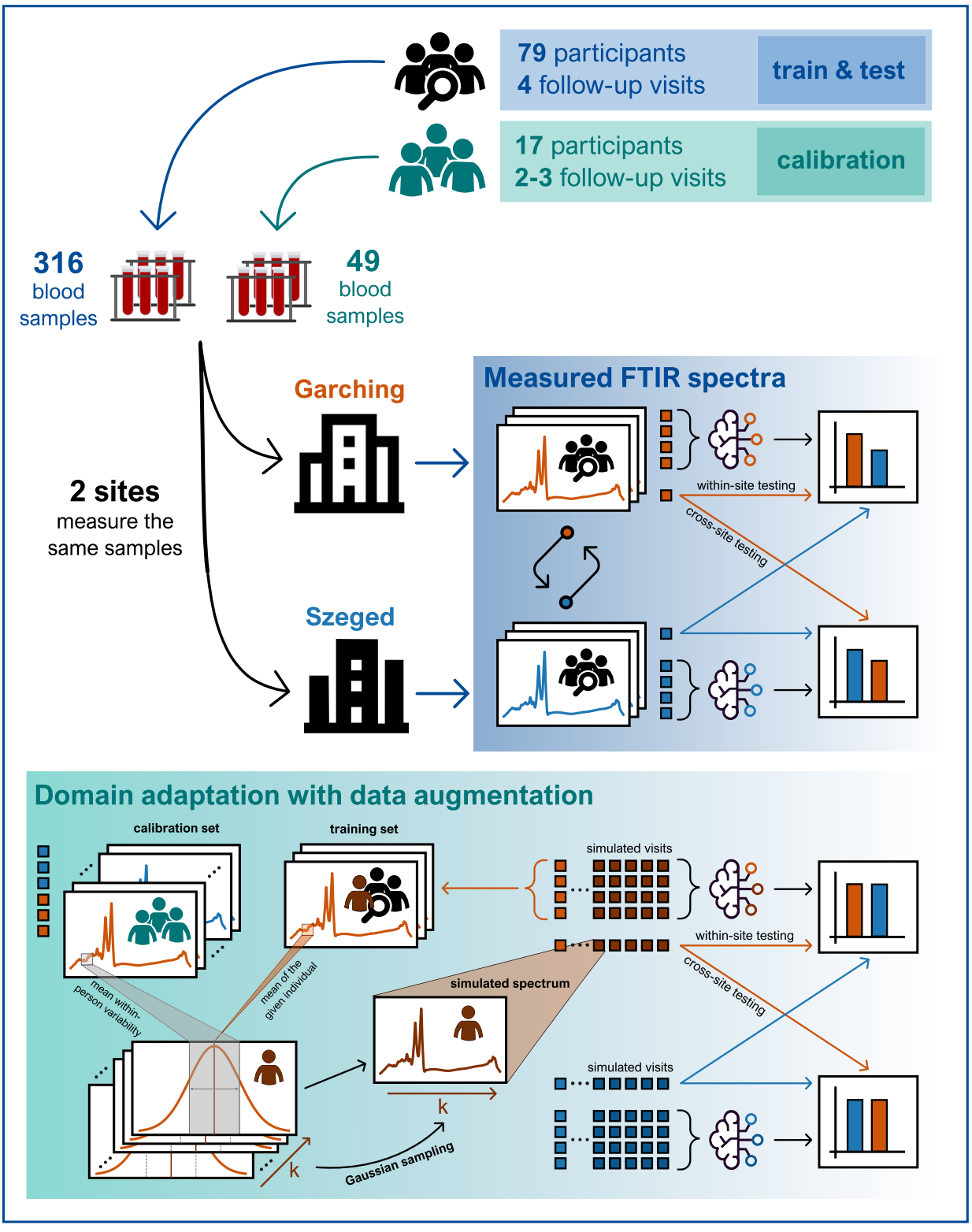


Figure 1. Overview. This study utilizes blood sample measurements from a longitudinal sample collection framework to analyze and compare the FTIR spectra acquired at two laboratory sites—one in Germany and one in Hungary—using instruments with largely identical nominal specifications. Further details can be found in the Methods section. Blood plasma samples were collected from 96 participants, with 17 of these samples used to calibrate the domain adaptation approach. Domain adaptation is achieved through data augmentation using a multivariate Gaussian fit that incorporates the differences between the two devices. The augmented training data enhances the prediction models' accuracy and reliability when applied to spectra measured across different devices. These findings highlight the importance of domain adaptation and introduce a data augmentation strategy that improves the generalization capability of classification models across various measurement sites.

thereby improving the model's accuracy over different domains.[28] In the field of sound source localization, novel data augmentation and weakly-supervised domain adaptation methods address the challenge of scarce labeled training data,

https://doi.org/10.1021/acs.analchem.5c00185
Anal. Chem. 2025, 97, 10264−10272

achieving comparable results to models trained on fully-labeled real data.[29] Furthermore, in face recognition, replacing traditional face alignment with aggressive data augmentation has led to remarkable accuracy, underscoring the method's potential to prevent overfitting in deep learning models.[30] These examples collectively highlight the transformative impact of domain generalization and data augmentation techniques across diverse research areas. In contrast to the above-mentioned domain adaptation techniques, for spectral data also model transfer techniques have been developed. In a similar approach, Tikhonov regularization-based model transfer tries to fine-tune a classical ML/chemometric model to work better with data from a secondary condition,[31] augmenting or transforming the training dataset, so a transfer to the secondary condition does not introduce errors[32] or EMSC-based model transfer[33] generating adapted data preprocessing techniques to mitigate the differences between conditions, e.g., between setups.

Figure 1 provides an overview of the challenge of cross-device model generalization in blood-based infrared spectroscopy, along with the proposed solution. Our longitudinal study involved samples from 96 fully healthy individuals, each with 2−4 follow-up visits. Our analysis synthesizes additional follow-up visits for each subject to expand the dataset. We propose a data augmentation technique that accounts for differences across devices, enhancing the robustness and transferability of machine learning models from one device to another. By enriching the training dataset with synthetically generated spectral variations reflecting the intricacies of both devices, our method enables models to learn device-agnostic representations of blood spectra, thereby improving generalization across devices. We experimentally validate our approach using real-world blood samples measured on two FTIR devices of the same type, located at different research laboratories. Our results demonstrate improved prediction accuracy and reliability when models are deployed on the second device. It should be noted that although this study uses liquid samples, the proposed method is sufficiently general to apply to measurements on dried samples. This work enhances the practical applicability of infrared spectroscopy in blood analysis for large-scale clinical studies, addressing a fundamental challenge in model deployment across different instrumentation setups.

METHODS

Sample Collection. In this investigation, we gathered blood sample data from a longitudinal sample collection framework of healthy individuals, with the study code: H4H_HU_2020_Sample Collection (Study approval reference number: 2754-11/2020/EÜIG). Specifically for our analysis, we employed FTIR spectroscopy to scrutinize blood samples from individuals deemed fully healthy. In this context, the spectra obtained from each individual can be interpreted as repeated measurements, representing their unique "fingerprint" of their healthy status. The analysis was conducted using two FTIR instruments from the same manufacturer, located in Garching, Germany, and Szeged, Hungary, respectively (see details in the following subsection). To evaluate potential differences between these instruments, we conducted parallel analyses on blood samples collected from overlapping subsets of participants using both devices. Precisely, in this study, we examined blood samples from 79 participants across four distinct visits and an additional 17 individuals across 2−3 visits. For our subsequent analysis, we utilized the dataset from the 79 participants with four visits each as our training and test set, resulting in a total of 632 spectra (2 devices × 79 participants × 4 visits). The remaining 49 spectra from the 17 individuals with 2−3 visits, also measured with both devices, were used as a calibration set for domain adaptation.

Sample Handling and FTIR Measurements. In the frame of this study, EDTA plasma samples are collected in 8 study centers all over Hungary. All 8 centers use the same standardized procedures for blood draws and processing. The obtained plasma is split into several 0.5 mL aliquots and stored at −80 °C. To prepare samples for FTIR measurements, one 0.5 mL aliquot of plasma per sample was thawed in the Garching lab, carefully mixed for 30 s by shaking and centrifuged for 10 min at 2000*g*. Subsequently, four 90 μL aliquots were generated and refrozen at −80°C. Hence, the FTIR measurements were performed upon two freeze-thaw cycles. Measurement aliquots for the Szeged lab were shipped on dry ice from Germany to Hungary. After thawing, the measurement aliquots were mixed again by shaking and subsequently centrifuged. In Garching again 10 min at 2000 g were applied here, while in Szeged aliquots were centrifuged for 5 min at 2500*g*. However, that should not result in measurable differences in the spectra. The samples were measured in a fully randomized order. The samples were aliquoted and measured in a blinded fashion, that is, the person performing the measurements had no access to the clinical information on the study participants, or which samples belonged to the same individual.

For infrared spectroscopic measurements, a commercial FTIR device specialized in the analysis of liquid samples in transmission mode was used (MIRA Analyzer, CLADE GmbH, formally known as Micro-Biolytics GmbH). Although the Garching device is older (instrument version MA6) than the Szeged device (instrument version MA7) the optical parts of both devices are identical and so are the software versions used. The main differences between the devices refer to the auto-sampler unit, only. The flow-through transmission cuvette was made of CaF_2 windows with 8 μm optical path length. The spectra were acquired with a resolution of 4 cm^{-1} in a spectral range between 930 cm^{-1} and 3050 cm^{-1}. A water reference spectrum was recorded automatically after each sample measurement to reconstruct the IR absorption spectra. The manufacturer does not disclose the applied algorithm but tends to overcompensate for highly concentrated sample types such as human plasma. The actual path length was also determined automatically at each measurement, and the spectra were adjusted accordingly. Note that the MIRA Analyzer does not allow access to the raw spectra but only the water-compensated absorbance spectra of the sample. FTIR measurements were performed in batches of 25 samples with a quality control serum (pooled human serum, BioWest, Nuaillé, France) measured at the beginning of the batch and after five samples, each, resulting in a batch size of 31. The quality control samples allowed tracking of potential technical errors and drift over the entire measurement period.

Data Preprocessing. The preprocessing of raw spectra involved a three-step protocol. First, we truncated the spectral regions below 1000 cm^{-1} and above 3000 cm^{-1}. These boundaries were chosen to exclude regions that consistently lacked informative absorbance features and where signal intensity typically declines. The cutoffs at 1000 and 3000 cm^{-1} also provided practical, standardized limits for array

https://doi.org/10.1021/acs.analchem.5c00185
Anal. Chem. 2025, 97, 10264−10272

alignment across samples, ensuring consistent input dimensions for subsequent analysis. Second, we applied L2 normalization to standardize the spectra. This involves scaling each spectrum so that the sum of squared absorbance values equals one. This approach preserves the relative intensity patterns within each spectrum. Finally, we excluded the spectral regions devoid of relevant peaks. Consequently, the analysis focused on spectral data within the wavenumber ranges of 1000−1750 and 2800−3000 cm^{-1}.

Visualization of Spectrally Resolved Differences. To visualize spectrally resolved differences between two groups of measurements, we employed the concept of differential fingerprints, as described in previous studies.[4,7,19] Additionally, we incorporated a widely used statistic known as the effect size (Cohen's d), which quantifies the magnitude of the difference between two group means in units of pooled standard deviation. This provides a standardized measure of how distinct the groups are. The effect size is defined as

$$d = \frac{M_1 - M_2}{s_p}$$

where M_1 and M_2 are the group means, and s_p is the pooled standard deviation.[34]

Description of the Data and Classification Tasks. Our initial training and test datasets comprised 632 preprocessed FTIR spectra, each containing 493 wavenumber-absorbance pairs obtained after preprocessing (as elaborated in the Data Preprocessing section). These spectra were evenly distributed between two distinct devices located in Garching, Germany, and Szeged, Hungary, reflecting four separate visits involving 79 individuals (for information on cohort characteristics refer to Figure S1/a in Supporting Information). We utilized these 493 standard-scaled absorbance values from each spectrum as features for our analysis. To characterize the domain shift between the two devices we developed logistic regression models for two distinct classification challenges. In the multi-class classification task, we aimed to differentiate individual participants as classes, utilizing data from the fourth visit as the test set. Conversely, for the binary classification model, we trained it on data from 60 participants to predict the sex of the remaining 19 individuals, thereby ensuring the model did not overfit to individual characteristics (for basic demographic distributions of participants within the training and test set refer to Figure S1/b in Supporting Information).

Machine Learning Analysis. The machine learning analysis commenced with preprocessed FTIR spectrum data as the foundational dataset. Before utilizing these data as features standard scaling was applied to all spectrum points to ensure uniformity. This step is crucial, especially for logistic regression models, as it normalizes feature scales, enhancing model convergence and prediction accuracy. We consistently used the logistic regression algorithm from Scikit-learn for all classification tasks.[35] An L2 penalty was applied, which regularizes the model by discouraging large coefficient values through the addition of their squared sum to the loss function. We used the liblinear solver for optimization.[36] The regularization parameter was fine-tuned through a 3-fold grid search with stratified splits of the training set over the values $C = [0.001, 1, 10]$—resulting in $C = 10$ in most of the cases. This approach yielded robust performance in the train set in both binary and multi-class classification scenarios.

Model-Based Data Augmentation. To improve the performance and robustness of our trained models, we introduced model-based data augmentation to artificially expand the number of repeated measures per individual. The augmentation method is based on fitting a multivariate Gaussian (MVG) distribution. Such an approach for simulating spectral data has been previously used in the context of Raman spectra of single cells.[37] Initially, we explored incorporating an MVG fitted to the training spectra of each individual (model 1). Additionally, we investigated MVGs for each individual, consisting of the person-specific mean and the element-wise averaged covariance matrix over all individuals (model 2). However, as illustrated in Figure 3, these augmentation approaches exhibited negligible impact on performance. In contrast, our proposed augmentation method involved utilizing an MVG that integrates within-person variability, extracted from the calibration set, into its covariance matrix (model 3). For this model, the mean spectrum $\overline{p}_i$ was individually extracted for each participant, while the covariance matrix remained consistent across all subjects. This constant covariance matrix was computed by averaging element-wise over all 17 person-specific covariance matrices, derived from the two to three measured spectra of each individual in the calibration set. The covariance matrices of model 2 and model 3 are shown in Figure S3 in Supporting Information. Formally, model 3 is described as follows

$$\mathcal{G}(x|\overline{p}_i, \overline{\Sigma}_{cal}) = \frac{1}{(2\pi)^{W/2}|\overline{\Sigma}_{cal}|^{1/2}} \exp\left(-\frac{1}{2}(x - \overline{p}_i)^T \overline{\Sigma}_{cal}^{-1}(x - \overline{p}_i)\right) \quad (1)$$

$$[\overline{\Sigma}_{cal}]_{mn} := \frac{1}{H}\sum_{i=1}^{H} [\Sigma(p_i)]_{mn} \;\forall\; m, n \in \{1, ..., W\} \quad (2)$$

where W stands for the number of grid points on which the spectrum was measured and H denotes the number of study participants in the calibration set. In our study $W = 1037$ and $H = 17$. p_i is the matrix containing measurements as its rows. A schematic overview of all three augmentation models is provided in Figure S4, with explicit descriptions available in Section C in Supporting Information. For an explanatory 2D visualization of the augmentation models refer to Figure S5 in Supporting Information. The source code for model 3 is available on Github.[38] Also, it is important to acknowledge that a comparable methodology, grounded in prior research,[23] has been formulated to integrate various types of unrepresented sources of variation commonly encountered in real-world applications into a given dataset.[39] For results using this methodology, see Figure S6 in Supporting Information.

RESULTS AND DISCUSSION

In the following, we provide a descriptive analysis of the spectral differences between the two sites in our study. Building on this, we perform domain adaptation using tailored augmentation approaches, which are then tested on real infrared samples measured with two infrared spectroscopy devices from different research laboratories.

Cross-Device Spectral Differences. Substantial spectral differences between the two sites are evident in the measured data, with certain wavenumber regions showing markedly different absorbance distributions, indicative of device-specific differences (Figure 2/a). Given that the samples from the same

https://doi.org/10.1021/acs.analchem.5c00185
Anal. Chem. 2025, 97, 10264−10272

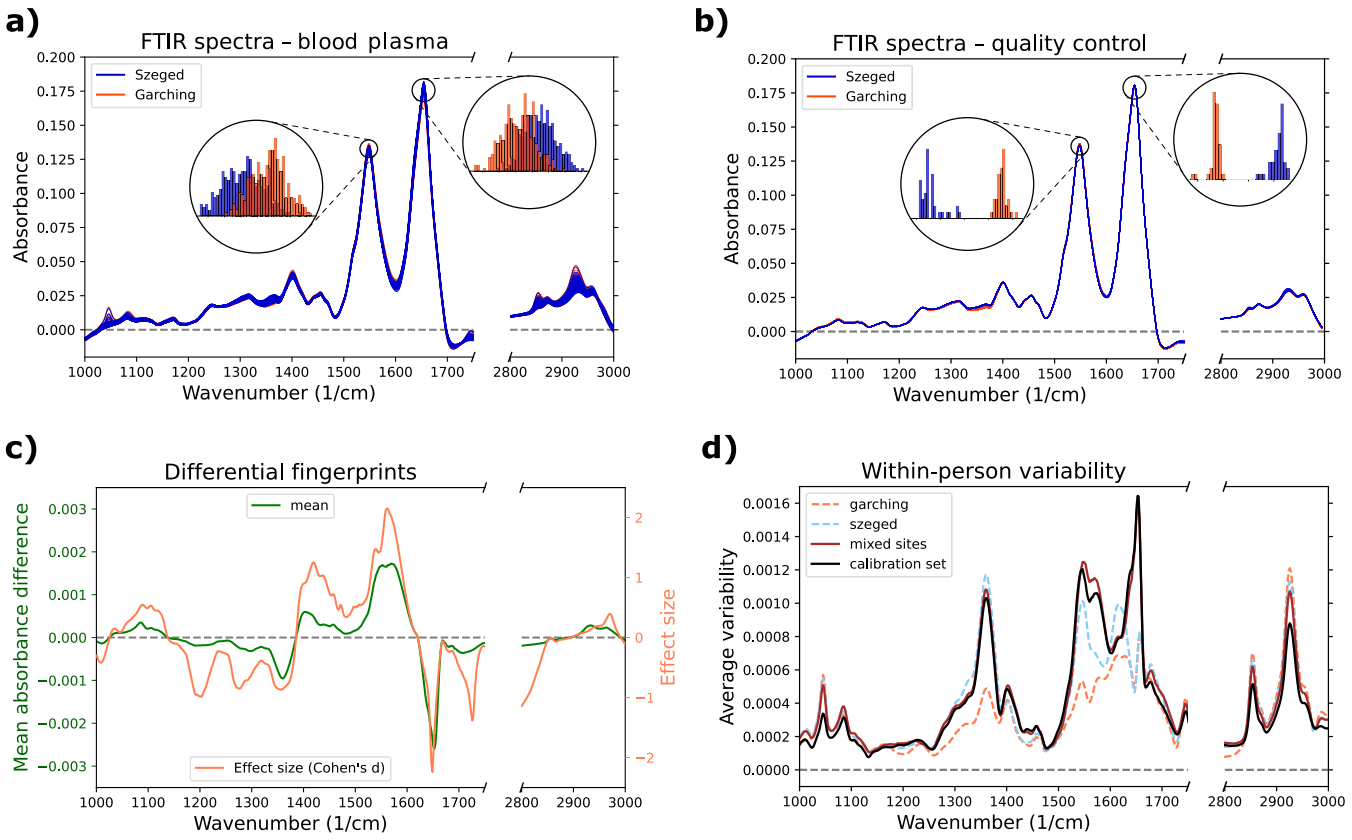


Figure 2. Measured data: (a) Preprocessed FTIR spectra of blood plasma samples, with data from the two sites differentiated by color. (b) FTIR spectra of 40 aliquots of the same quality control sample (Human Serum AB male HIV tested − S4190 − BioWest; sourced from a pool of AB male blood) measured at each site. (c) Differential fingerprints are obtained by subtracting paired spectra from each site and then taking the average. Additionally, the effect size (Cohen's *d*) is plotted alongside. Generating this analysis for each of the four visits separately reveals a consistent pattern, although both the average differences and the effect size diminish as the number of visits (and thus, time) increases. (Refer to Figure S7 in Supporting Information for more details.) (d) Average within-person variability using the train and test sets for the two sites separately (dashed lines) and combined (solid dark red line), and using the calibration set from both sites (solid dark blue line).

individuals were measured using both instruments, these discrepancies must stem from differences in instrumentation or sample handling. In addition, we analyzed quality control (QC) samples—pooled human blood sera aliquots measured at both sites—under identical conditions. The corresponding spectral distributions (Figure 2/b) show similar inter-site discrepancies as the investigated samples. These discrepancies are further quantified through differential fingerprints and effect sizes, where several regions exhibit very large to huge differences based on Cohen's *d* thresholds[34] (Figure 2/c), highlighting that the domain shift is both broad and substantial. Notably, while within-person variability (WPV) differs significantly between the two sites when assessed separately, the WPV values for both the calibration set (49 paired spectra measured at both sites) and the combined dataset (632 paired spectra from both sites) are highly consistent (Figure 2/d). This consistency supports the suitability of the calibration set for domain adaptation, as it effectively captures inter-site variability while smoothing over device-specific characteristics.

The existence of these variations facilitated the development of a logistic regression model capable of distinguishing between the two instruments with 100% accuracy, thereby highlighting their unique characteristics. However, while this finding is notable, the primary objective of our investigation was to determine whether these disparities significantly influence clinically relevant classification tasks. To address this, we trained both binary and multi-class classification models using data exclusively from one device and subsequently evaluated their performance using data from both instruments. This approach enabled us to directly evaluate how device-specific spectral differences influence model performance across devices, providing insight into their impact on key classification tasks.

Cross-Device Evaluations with Measured Data. To investigate how spectral variances unique to each measurement tool affect predictions across devices, we constructed and assessed binary and multi-class classifiers. These classifiers were trained exclusively with data from a single instrument and then evaluated on test datasets collected from both devices. For multi-class classification, where classes were defined by individual participants, the accuracy scores for data from Szeged and Garching were notably high, at 0.84 and 0.76, respectively, significantly surpassing random chance levels (1/79 ≈ 0.013). However, applying these models to cross-site data revealed a pronounced reduction in prediction accuracies, falling to 0.65 and 0.66. In the scenario of binary classification for determining the sex of participants not seen during training, the within-site Area Under the Curve (AUC) values were 0.92 for both Szeged and Garching data. These values declined to 0.86 and 0.81, respectively, when the models were applied to cross-site data. These findings highlight the critical impact of

https://doi.org/10.1021/acs.analchem.5c00185
Anal. Chem. 2025, 97, 10264−10272

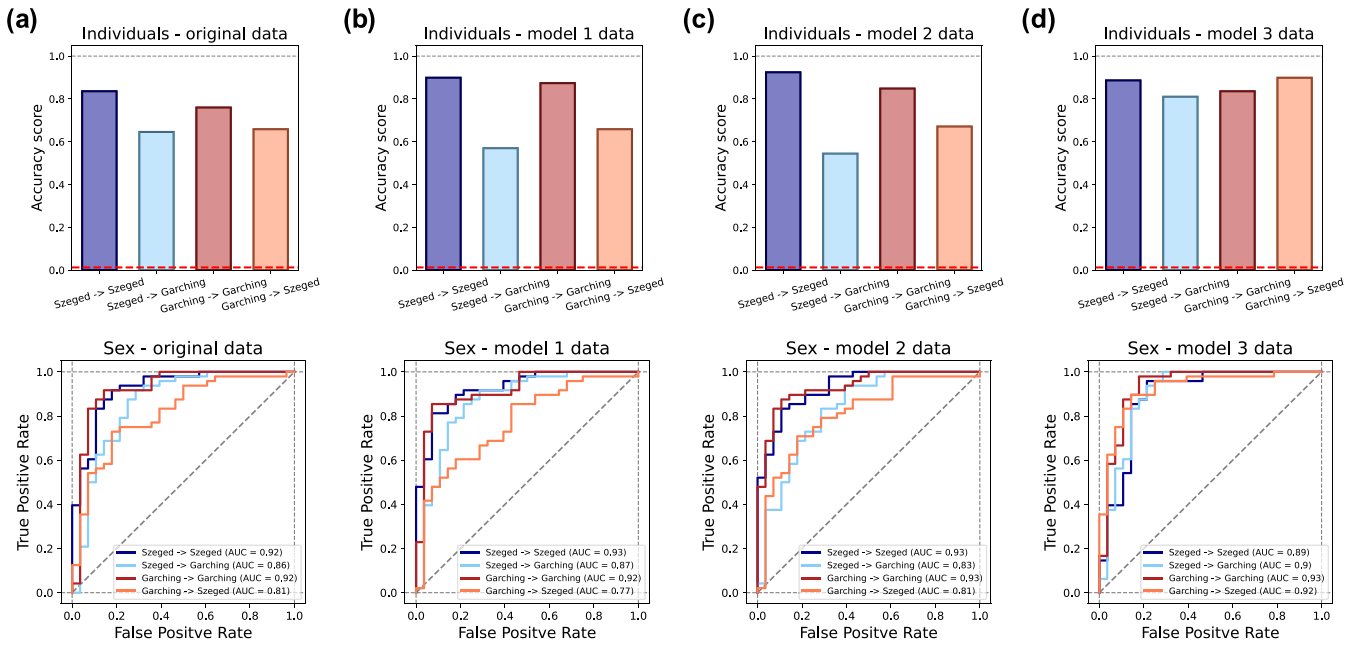


Figure 3. Comprehensive visualization of the model performances for multi-class (upper row) and binary classification (lower row) tasks, using original (a), model 1 augmented (b), model 2 augmented (c) and model 3 augmented (d) datasets. In the case of multi-class classification, the red dashed line shows the random chance accuracy level. Table 1 summarizes the performance of the original and the augmented models on the test set. Additional ROC curves when using a stratified separation of the datasets with respect to the sex can be found in Figure S2 in Supporting Information.

device-specific variations also on relevant classification tasks (see Figure 3/a).

Cross-Device Evaluations with Augmented Data. Following this, we enriched our dataset by simulating spectra for each site, artificially increasing our dataset by 100 additional visits per individual. We explored the impact of three data augmentation techniques: the first model applied the Multivariate Gaussian (MVG) distribution fit to each individual's training spectra separately (model 1), the second model leveraged the full training set of the given site to capture within-person variability (model 2), while the third model used the covariance matrix gained from the calibration set comprising of spectra from both sites and thus capturing cross-device domain shifts (model 3).

Incorporating the simulated spectra from model 1 markedly improved prediction accuracy for within-site evaluations. Specifically, multi-class classification accuracies, predicting individuals, rose from 0.84 and 0.76 to 0.90 and 0.87. Yet, this improvement did not extend to cross-site predictions, which altered from 0.65 and 0.66 to 0.57 and 0.66. A similar pattern was observed for classification tasks regarding sex, where within-site AUC scores remained constant, but there was no consistent improvement in cross-site scores. Hence, while this method enhanced or maintained within-site prediction accuracy, it did not bridge the gap between within- and cross-site prediction accuracies (refer to Figure 3/b).

Utilizing model 2 also markedly enhanced the accuracy of within-site predictions. However, compared to model 1, using model 2 showed no improvement regarding the prediction accuracy in any of the classification tasks.

Incorporating new spectra through model 3 improved the predictions for both binary and multi-class classification tasks. For multi-class classification, this approach not only improved within-site predictions but also significantly enhanced cross-site predictions. The accuracy scores for cross-site predictions rose from 0.65 and 0.66 to 0.81 and 0.90, respectively. These results indicate that prediction accuracy is driven by the test site rather than the training site, as models trained on either Garching or Szeged data achieve 0.81−0.84 accuracy on Garching test data and 0.89−0.90 on Szeged test data. For binary classifications, the AUC for cross-site predictions rose above 0.90, essentially matching the AUC for within-site predictions. Ultimately, this approach successfully addressed the differences in prediction accuracies between within-site and cross-site evaluations, as shown in Figure 3/c. For a detailed comparison between the within and cross-site performance of the original and augmented data refer to Table 1.

These results underscore the critical importance of incorporating cross-device differences in the data augmentation for simulating new spectra. In particular, the average within-person variation (WPV) exhibits significant differences when considering the sites separately versus together. Moreover, the WPV of the calibration set effectively captures the WPV of

Table 1. Multi-class Accuracies and Binary Classification AUC

multi-class classification for individuals				
accuracy scores	Sz → Sz	Sz → G	G → G	G → Sz
original data	0.84	0.65	0.76	0.66
aug model 1	0.90	0.57	0.87	0.66
aug model 2	0.92	0.54	0.85	0.67
aug model 3	0.89	0.81	0.84	0.90
binary classification for sex				
AUC	Sz → Sz	Sz → G	G → G	G → Sz
original data	0.92	0.86	0.92	0.81
aug model 1	0.93	0.87	0.92	0.77
aug model 2	0.93	0.83	0.93	0.81
aug model 3	0.89	0.90	0.93	0.92

https://doi.org/10.1021/acs.analchem.5c00185
Anal. Chem. 2025, 97, 10264−10272

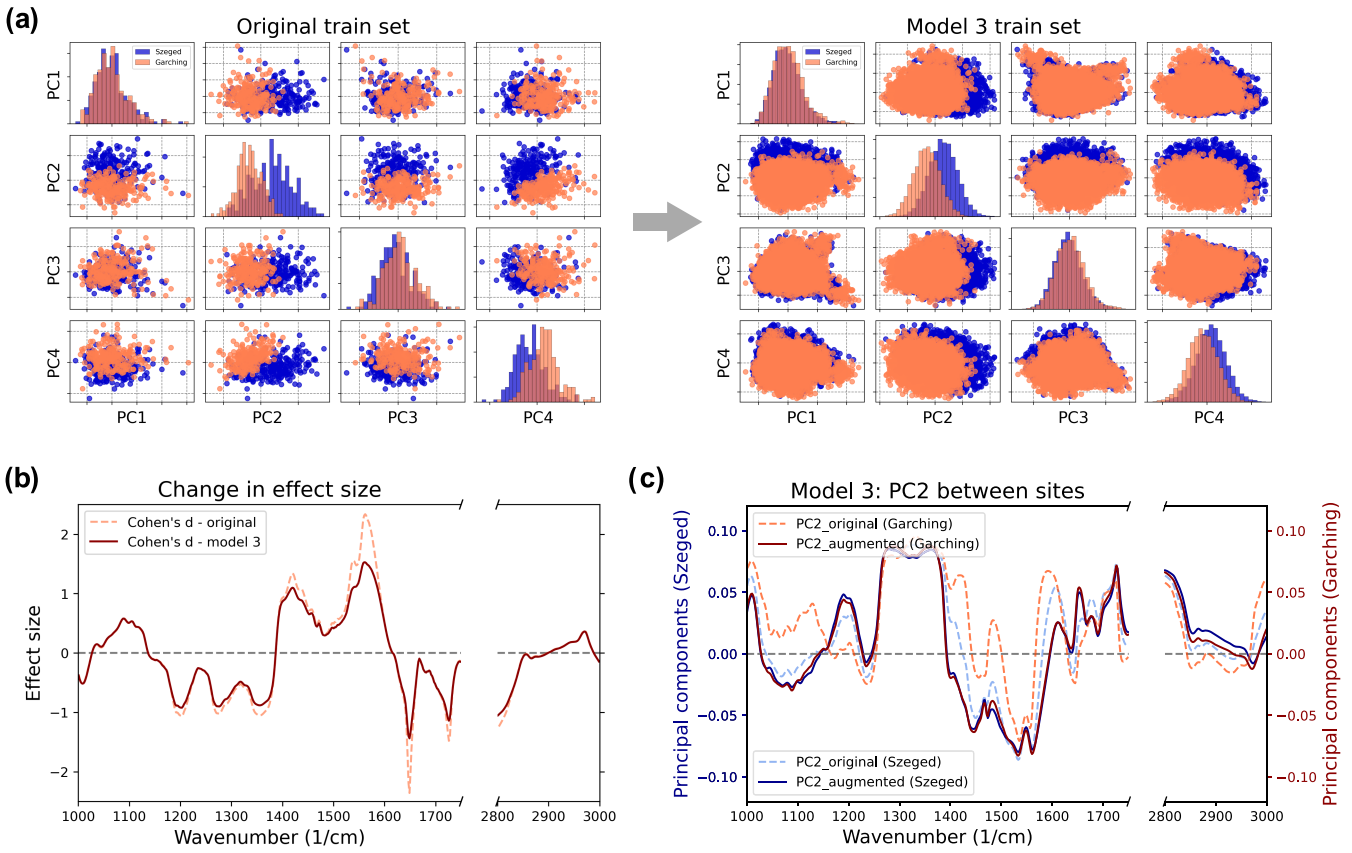


Figure 4. Impact of data augmentation with model 3: (a) Scatter matrices illustrate the comparison of the first four principal components pre- and post-data augmentation, revealing a diminished distinction between data from the two sites following the inclusion of simulated visits. Furthermore, the data distributions along the principal components have grown more similar in shape. (b) The change in effect size (Cohen's d) before and after the inclusion of simulated data. The effect size diminished significantly, with the notable reduction of peaks observed with the original data. (c) PC2 for both sites before and after data augmentation, highlighting how the addition of simulated data aids in aligning this principal component. A similar pattern is observed with the other principal components as well (refer to Section H in Supporting Information for further details regarding PCA for models 1−3).

both the train and test sets combined across sites (see Figure 2/c). Models 1 and 2, which utilized WPV from the sites separately, were outperformed by model 3, which incorporated the WPV most representative of the combined sites, specifically the WPV of the calibration set for the simulation of additional visit data for the individuals. This enhancement is due to model 3's ability to simulate spectra that incorporate information on domain shifts, significantly improving both within-site and cross-site prediction accuracies. This demonstrates model 3's superior capability in bridging the gap between within-site and cross-site predictive performance, highlighting its enhanced utility in managing device-specific spectral variations. Given its evident superiority, we exclusively adopted model 3 for further analysis.

In addition, we explored how many simulated visits were necessary to make the two sites appear almost identical. In multi-class classification, which focuses on identifying individual identities, adding about 10 simulated visits to each site was enough to effectively equalize the accuracy between cross-site and within-site prediction. On the other hand, for binary classification targeting prediction of sex, about 60 simulated visits were necessary to bridge the AUC gap between the predictions on the same sites, and additional simulated visits did not lead to any notable further improvements in performance (see Figure S8 in the Supporting Information for more details).

Following data augmentation with model 3, the datasets from the two sites exhibited reduced separation within the principal component space (refer to Figure 4/a), and the peaks previously observed in the effect size were notably reduced (refer to Figure 4/b). For PCA scatter matrices of the first four principal components of model 1 and model 2 refer to Figure S9 in Supporting Information.

To deepen our understanding of the impact of data augmentation, PCA was conducted independently for both sites, initially with the original dataset and subsequently with the model 3 augmented dataset. We then examined the first four principal components—which accounted for approximately 82% of the explained variance in the original dataset and 83% in the augmented dataset—across both sites before and after the inclusion of simulated data. This comparison revealed a notable alignment of these principal components across the two sites (refer to Figure 4/c). The alignment of the principal components with the inclusion of simulated data indicates a harmonization in the data distributions between the two sites, making them more similar and less distinguishable (see Figure S10, Figure S11, and Figure S12 in the Supporting Information). This outcome is consistent with the closing of the accuracy gap observed between cross-device and within-device predictions. Essentially, data augmentation not only aligns key features across devices but also enhances the model's cross-site predictive capabilities.

https://doi.org/10.1021/acs.analchem.5c00185
Anal. Chem. 2025, 97, 10264−10272

CONCLUSIONS

Our study addressed the critical challenge of achieving cross-device model generalization in a longitudinal dataset of 79 individuals across multiple follow-up visits. Variations in instrument setups, such as differences in spectral resolution and noise levels, introduced significant dataset shifts. These shifts create major obstacles to deploying infrared spectroscopy in clinical settings, particularly in large-scale studies involving data collected from multiple laboratories.

To address this, we implemented a data augmentation strategy that encapsulates cross-device differences, presenting a straightforward domain adaptation approach designed to enhance the generalizability and robustness of machine learning models across various FTIR devices. By synthesizing spectral variations, this method enables the development of domain-generalized models capable of making accurate predictions on new devices, regardless of the device used for initial training. Our experimental validation on real-world blood samples measured with two distinct FTIR devices confirms the efficacy of our approach, demonstrating improved prediction accuracy and reliability when models are applied to data from an alternative device.

The implications of our work extend beyond improving cross-device generalization in FTIR spectroscopy for blood analysis. It facilitates the broader adoption and practical application of infrared spectroscopy in large multi-center clinical studies, aiming to address early-stage disease detection, screening, and health monitoring. By overcoming device variability and model generalization challenges, it is possible to enhance the accuracy and reliability of disease diagnostics and significantly contribute to the evolving field of medical diagnostics.

In conclusion, our research underscores the critical importance of developing robust, device-agnostic machine learning models or methods for adapting and refining trained models to advance infrared spectroscopy in medical applications. Incorporating sophisticated domain adaptation techniques, such as data augmentation and cross-device generalization strategies, will be crucial for unlocking the full potential of infrared spectroscopy, particularly in large clinical studies involving multiple laboratories. Future research should extend this approach to address variations related to sample handling, experimental workflows, and study protocols. Insights from these variations could inform the development of ML-grade calibration procedures and improved preprocessing methods, further minimizing setup-related discrepancies.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.5c00185.

(A) Cohort characteristics. (B) Within-person covariance matrices. (C) Explicit definition of augmentation approaches. (D) Differences between augmentation models. (E) CODI for cross-device generalization. (F) Differential Fingerprints between sites for each visit. (G) Performance of augmentation model approaches at varying augmentation sizes. (H) Aligned principal components for real data and all models (PDF)

AUTHOR INFORMATION

Corresponding Author

Kosmas V. Kepesidis − *Center for Molecular Fingerprinting (CMF), 1093 Budapest, Hungary; Department of Laser Physics, Ludwig Maximilian University of Munich (LMU), 85748 Garching, Germany; Laboratory for Attosecond Physics, Max Planck Institute of Quantum Optics (MPQ), 85748 Garching, Germany;* orcid.org/0000-0002-6391-7743; Email: kosmas.kepesidis@cmf.hu

Authors

Flora B. Nemeth − *Center for Molecular Fingerprinting (CMF), 1093 Budapest, Hungary; Department of Laser Physics, Ludwig Maximilian University of Munich (LMU), 85748 Garching, Germany; Laboratory for Attosecond Physics, Max Planck Institute of Quantum Optics (MPQ), 85748 Garching, Germany*

Niklas Leopold-Kerschbaumer − *Department of Laser Physics, Ludwig Maximilian University of Munich (LMU), 85748 Garching, Germany; Laboratory for Attosecond Physics, Max Planck Institute of Quantum Optics (MPQ), 85748 Garching, Germany*

Diana Debreceni − *Center for Molecular Fingerprinting (CMF), 1093 Budapest, Hungary*

Frank Fleischmann − *Center for Molecular Fingerprinting (CMF), 1093 Budapest, Hungary; Department of Laser Physics, Ludwig Maximilian University of Munich (LMU), 85748 Garching, Germany; Laboratory for Attosecond Physics, Max Planck Institute of Quantum Optics (MPQ), 85748 Garching, Germany*

Krisztian Borbely − *Center for Molecular Fingerprinting (CMF), 1093 Budapest, Hungary; Department of Laser Physics, Ludwig Maximilian University of Munich (LMU), 85748 Garching, Germany; Laboratory for Attosecond Physics, Max Planck Institute of Quantum Optics (MPQ), 85748 Garching, Germany*

David Mazurencu-Marinescu-Pele − *Department of Laser Physics, Ludwig Maximilian University of Munich (LMU), 85748 Garching, Germany; Laboratory for Attosecond Physics, Max Planck Institute of Quantum Optics (MPQ), 85748 Garching, Germany*

Thomas Bocklitz − *Leibniz Institute of Photonic Technology (Leibniz-IPHT), 07745 Jena, Germany; Institute of Physical Chemistry and Abbe Center of Photonics (IPC/ACP), Friedrich-Schiller-University (FSU), 07745 Jena, Germany;* orcid.org/0000-0003-2778-6624

Mihaela Žigman − *Department of Laser Physics, Ludwig Maximilian University of Munich (LMU), 85748 Garching, Germany; Center for Molecular Fingerprinting (CMF), 1093 Budapest, Hungary; Laboratory for Attosecond Physics, Max Planck Institute of Quantum Optics (MPQ), 85748 Garching, Germany;* orcid.org/0000-0001-8306-1922

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.5c00185

Author Contributions

#F.B.N. and N.L.-K. contributed equally. F.B.N.: investigation; formal analysis; visualization; methodology; writing − original draft preparation; and writing—review & editing. N.L.-K.: investigation; formal analysis; visualization; methodology; writing—original draft preparation; and writing—review & editing. D.D.: investigation; methodology; project administration; and writing—review & editing. F.F.: investigation;

https://doi.org/10.1021/acs.analchem.5c00185
Anal. Chem. 2025, 97, 10264−10272

methodology; project administration; and writing—review & editing. K.B.: investigation; data curation; and writing—review & editing. D.M.-M.-P.: formal analysis; methodology; and writing— review & editing. T.B.: methodology and writing—review & editing. M.Ž.: project administration and writing—review & editing. K.V.K.: conceptualization; supervision; project administration; methodology; writing—original draft preparation; and writing—review & editing.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We express our heartfelt gratitude to Ferenc Krausz and all employees of the Center for Molecular Fingerprinting (CMF) for fostering the research environment that enabled the realization of this work. We also wish to recognize the efforts of all individuals who participated in the study reported here.

■ REFERENCES

(1) Baker, M. J.; Trevisan, J.; Bassan, P.; Bhargava, R.; Butler, H. J.; Dorling, K. M.; et al. *Nat. Protoc.* **2014**, *9* (8), 1771−1791.
(2) Butler, H. J.; Brennan, P. M.; Cameron, J. M.; Finlayson, D.; Hegarty, M. G.; Jenkinson, M. D.; et al. *Nat. Commun.* **2019**, *10* (1), No. 4501.
(3) Paraskevaidi, M.; Matthew, B. J.; Holly, B. J.; Hugh, B. J.; Thulya, C. P.; Loren, C.; et al. *Appl. Spectrosc. Rev.* **2021**, *56* (8-10), 804−868.
(4) Huber, M.; Kepesidis, K. V.; Voronina, L.; Fleischmann, F.; Fill, E.; Hermann, J.; et al. *eLife* **2021**, *10*, No. e68758.
(5) Kepesidis, K. V.; Bozic-Iven, M.; Huber, M.; Abdel-Aziz, N.; Kullab, S.; Abdelwarith, A.; et al. *BMC Cancer* **2021**, *21*, 1−9.
(6) Huber, M.; Kepesidis, K. V.; Voronina, L.; Božić, M.; Trubetskov, M.; Harbeck, N.; et al. *Nat. Commun.* **2021**, *12* (1), No. 1511.
(7) Voronina, L.; Leonardo, C.; Mueller-Reif, J. B.; Geyer, P. E.; Huber, M.; Trubetskov, M.; et al. *Angew. Chem.* **2021**, *133* (31), 17197−17206.
(8) Tian, P.; Zhang, W.; Zhao, H.; Lei, Y.; Cui, L.; Wang, W.; et al. *Int. J. Clinical Experimental Med.* **2015**, *8* (1), 972.
(9) Mittal, S.; Wrobel, T. P.; Walsh, M.; Kajdacsy-Balla, A.; Bhargava, R. *Clinical Spectroscopy* **2021**, *3*, No. 100006.
(10) Griffiths, P. R. *Science* **1983**, *222* (4621), 297−302.
(11) Martin, F. L.; Kelly, J. G.; Llabjani, V.; Martin-Hirsch, P. L.; Patel, II; Trevisan, J.; et al. *Nat. Protoc.* **2010**, *5* (11), 1748−1760.
(12) Elmi, F.; Movaghar, A. F.; Elmi, M. M.; Alinezhad, H.; Nikbakhsh, N. *Spectrochimica Acta A: Molecular Biomolecular Spectroscopy* **2017**, *187*, 87−91.
(13) Ghimire, H.; Garlapati, C.; Janssen, E. A.; Krishnamurti, U.; Qin, G.; Aneja, R.; Perera, A. G. U. *Cancers* **2020**, *12* (7), 1708.
(14) Zelig, U.; Barlev, E.; Bar, O.; Gross, I.; Flomen, F.; Mordechai, S.; et al. *BMC Cancer* **2015**, *15*, 1−10.
(15) Ollesch, J.; Heinze, M.; Heise, H. M.; Behrens, T.; Brüning, T.; Gerwert, K. *J. Biophotonics* **2014**, *7* (3-4), 210−221.
(16) Ollesch, J.; Theegarten, D.; Altmayer, M.; Darwiche, K.; Hager, T.; Stamatis, G.; Gerwert, K. *Biomed. Spectroscopy Imaging* **2016**, *5* (2), 129−144.
(17) Anderson, D. J.; Anderson, R.; Moug, S.; Baker, M. *BJS open* **2020**, *4* (4), 554−562.
(18) Eissa, T.; Leonardo, C.; Kepesidis, K. V.; Fleischmann, F.; Linkohr, B.; Meyer, D.; et al. *Cell Rep. Med.* **2024**, *5*, 7.
(19) Kepesidis, K. V.; Stoleriu, M. G.; Feiler, N.; Gigou, L.; Fleischmann, F.; Aschauer, J.; et al. *BMC Med.* **2025**, *23* (1), 101.
(20) Reiner, M.; Niermann, C.; Jekauc, D.; Woll, A. *BMC Public Health* **2013**, *13*, 1−9.
(21) Siegrist, M.; Visschers, V. H.; Hartmann, C. *Food Quality Preference* **2015**, *46*, 33−39.
(22) Zheng, M.; Zhang, X.; Chen, S.; Song, Y.; Zhao, Q.; Gao, X.; et al. *Circ. Res.* **2020**, *127* (12), 1491−1498.
(23) Eissa, T.; Kepesidis, K. V.; Zigman, M.; Huber, M. *Anal. Chem.* **2023**, *95* (16), 6523−6532.
(24) Hofko, B.; Porot, L.; Falchetto Cannone, A.; Poulikakos, L.; Huber, L.; Lu, X.; et al. *Materials Structures* **2018**, *51*, 1−16.
(25) Quiñonero-Candela, J.; Sugiyama, M.; Schwaighofer, A.; Lawrence, N. D. *Dataset shift in machine learning*; MIT Press, 2022.
(26) Zhang, L.; Wang, X.; Yang, D.; Sanford, T.; Harmon, S.; Turkbey, B.et al. When unseen domain generalization is unnecessary? rethinking data augmentation2019, arXiv:190603347. arXiv.org e-Print archive. https://arxiv.org/abs/190603347.
(27) Spyrou, E.; Mathe, E.; Pikramenos, G.; Kechagias, K.; Mylonas, P. *Technologies* **2020**, *8* (4), 55.
(28) Sarwar, S. M.; Murdock, V. Unsupervised domain adaptation for hate speech detection using a data augmentation approach. In *Proceedings of the International AAAI Conference on Web and Social Media* 2022; pp 852−862.
(29) He, W.; Motlicek, P.; Odobez, J. M. *IEEE/ACM Transactions Audio, Speech, Language Processing* **2021**, *29*, 1303−1317.
(30) Wen, G.; Chen, H.; Cai, D.; He, X. *Neurocomputing* **2018**, *287*, 45−51.
(31) Guo, S.; Heinke, R.; Stöckel, S.; Rösch, P.; Popp, J.; Bocklitz, T. *J. Raman Spectrosc.* **2018**, *49* (4), 627−637.
(32) Guo, S.; Heinke, R.; Stöckel, S.; Rösch, P.; Bocklitz, T.; Popp, J. *Vib. Spectrosc.* **2017**, *91*, 111−118.
(33) Guo, S.; Kohler, A.; Zimmermann, B.; Heinke, R.; Stöckel, S.; Rösch, P.; et al. *Anal. Chem.* **2018**, *90* (16), 9787−9795.
(34) Sawilowsky, S. S. *J. Modern Applied Statistical Methods* **2009**, *8* (2), 26.
(35) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; et al. *J. Machine Learning Res.* **2011**, *12*, 2825−2830.
(36) Fan, R. E.; Chang, K. W.; Hsieh, C. J.; Wang, X. R.; Lin, C. J. *J. Machine Learning Res.* **2008**, *9*, 1871−1874.
(37) Beleites, C.; Neugebauer, U.; Bocklitz, T.; Krafft, C.; Popp, J. *Anal. Chim. Acta* **2013**, *760*, 25−33.
(38) Leopold-Kerschbaumer, N. *AdaptFTIR*, GitHub, 2024, https://github.com/Niklas-LK/AdaptFTIR.
(39) Eissa, T.; Huber, M.; Obermayer-Pietsch, B.; Linkohr, B.; Peters, A.; Fleischmann, F.; Zigman, M. *PNAS nexus* **2024**, *3* (10), 449.

https://doi.org/10.1021/acs.analchem.5c00185
Anal. Chem. 2025, 97, 10264−10272