PNAS nexus

PNAS Nexus, 2024, **3**, pgae449

https://doi.org/10.1093/pnasnexus/pgae449

Advance access publication 15 October 2024

Research Report

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

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

[a]Chair of Experimental Physics - Laser Physics, Ludwig-Maximilians-Universität München, Bavaria 85748, Germany
[b]Laboratory for Attosecond Physics, Max Planck Institute of Quantum Optics, Bavaria 85748, Germany
[c]School of Computation, Information and Technology, Technical University of Munich, Bavaria 85748, Germany
[d]Department of Internal Medicine, Division of Endocrinology and Diabetology, Medical University, Styria 8010, Austria
[e]Institute of Epidemiology, Helmholtz Zentrum München, Bavaria 85764, Germany
[f]Chair of Epidemiology, Institute for Medical Information Processing, Biometry and Epidemiology, Medical Faculty, Ludwig-Maximilians-Universität München, Bavaria 81377, Germany
*To whom correspondence should be addressed: Email: tarek.eissa@mpq.mpg.de (T.E.); mihaela.zigman@mpq.mpg.de (M.Ž.)

Edited By: Lydia Kavraki

Abstract

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

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

Significance Statement

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

Introduction

Technological advances in molecular analytics increasingly enable the probing of biological systems. Distinguishing between physiologically relevant states from quantitative molecular fingerprints presents a new opportunity for *in vitro* phenotyping. Extensive efforts are thus dedicated to developing standardized procedures involving streamlined biological sampling, post-collection handling, and sensitive quantitative measurements. Nevertheless, empirical datasets are susceptible to diverse sources of variability, both analytical and inherently biological (1–9). Obtaining a dataset that reflects a realistic data distribution is often resource-intensive, costly, and, in some cases, impossible. This applies especially in the context of clinical studies, covering all pathophysiological strata, studying rare disease, or

Competing Interest: The authors declare no competing interests.
Received: April 9, 2024. **Accepted:** September 7, 2024

OXFORD UNIVERSITY PRESS

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

Downloaded from https://academic.oup.com/pnasnexus/article/3/10/pgae449/7823116 by guest on 19 February 2025

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

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

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

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

Downloaded from https://academic.oup.com/pnasnexus/article/3/10/pgae449/7823116 by guest on 19 February 2025

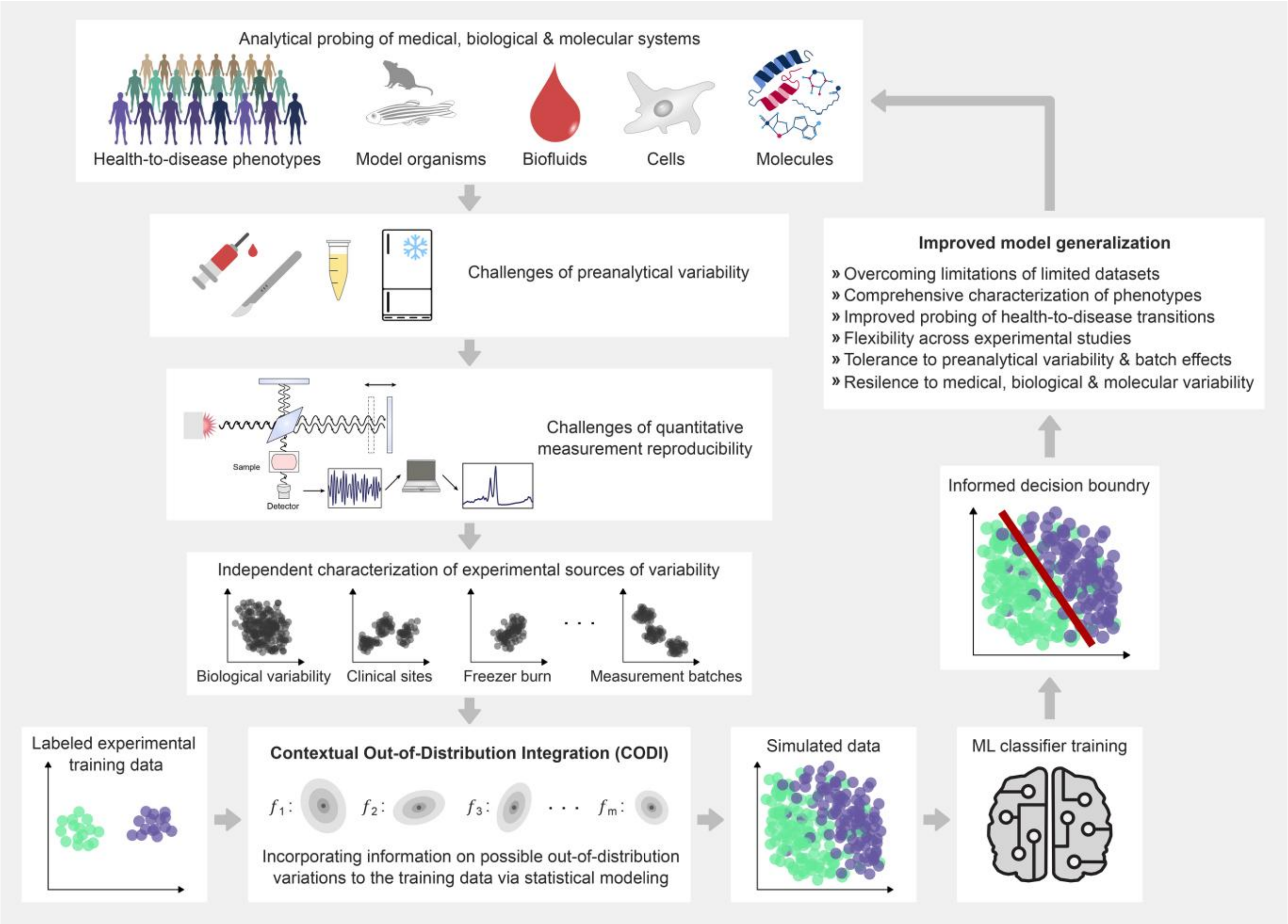
Analytical probing of medical, biological & molecular systems
Health-to-disease phenotypes
Model organisms
Biofluids
Cells
Molecules
Challenges of preanalytical variability
Challenges of quantitative measurement reproducibility
Sample
Detector
Independent characterization of experimental sources of variability
Biological variability
Clinical sites
Freezer burn
. . .
Measurement batches
Labeled experimental training data
Contextual Out-of-Distribution Integration (CODI)
f_1:
f_2:
f_3:
· · ·
f_m:
Incorporating information on possible out-of-distribution variations to the training data via statistical modeling
Simulated data
ML classifier training
Informed decision boundry
Improved model generalization
» Overcoming limitations of limited datasets
» Comprehensive characterization of phenotypes
» Improved probing of health-to-disease transitions
» Flexibility across experimental studies
» Tolerance to preanalytical variability & batch effects
» Resilence to medical, biological & molecular variability

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

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

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

Results

Characterizing empirically observed variability

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

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

After the characterization of measurement deviations, these differences ($\boldsymbol{u}_i - \boldsymbol{v}_j$) are scaled by a random variable that assumes a Gaussian distribution and then combined over the entire variability calibration set. This aggregation may then be added onto an independent experimental training seed measurement $\boldsymbol{x}_k$ to create a new simulated measurement that is now modeled as a statistical outcome. Such a simulation approach can be repeatedly applied to generate a cohort of simulated measurements in arbitrary size. The generated cohort, as a whole, would reflect the variability properties observed between $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$ onto $\boldsymbol{x}_k$. If the set of calibration measurements would reflect a source of variability that was unobserved in a given training set of measurements $\{\boldsymbol{x}_k \mid k = 1, \ldots, n\}$, a new level of variability would be introduced onto the training set of measurements. This strategy allows for the creation of realistic synthetic data without needing to fine tune free parameters controlling the data generation. Further detailed descriptions are in the Materials and methods and Supplementary Information.

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

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

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

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

Introducing experimental variability *in silico*

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

Downloaded from https://academic.oup.com/pnasnexus/article/3/10/pgae449/7823116 by guest on 19 February 2025

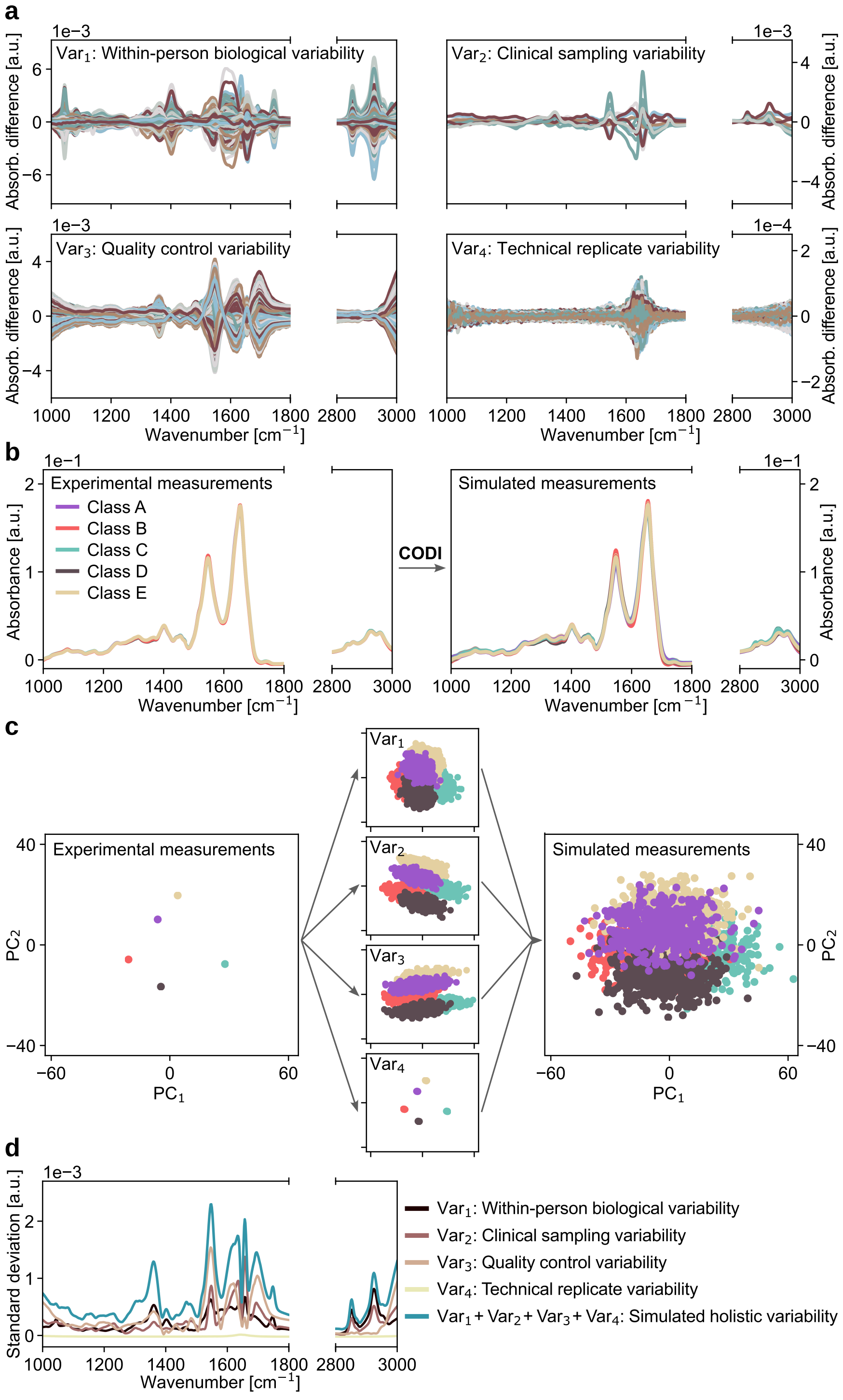

Fig. 2. Applying CODI to introduce measurement variability onto illustrative experimental IR spectra. a) Four distinct sources of possible variability were characterized from calibration sets of experimental observations. The curves depict how each measurement differs from its expected mean. b) By applying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

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

Principal component analysis (PCA) applied to the original and simulated measurements reveals that each source of measurement variability affected the spatial distribution of the seed data differently across the first two components (Fig. 2c). In other words, each set of calibration measurements—modeling different variability properties—affected linearly independent data features. Once all four sources of measurement variability were incorporated into

Downloaded from https://academic.oup.com/pnasnexus/article/3/10/pgae449/7823116 by guest on 19 February 2025

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

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

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

Application to longitudinal study settings

In longitudinal clinical studies that involve the collection of biological specimens tainted by attrition and loss-to-follow-up over time, great efforts are required to gather sufficiently large datasets. Typically, individuals participate in an initial baseline sample collection, followed by extended waiting periods for subsequent collections from the same individuals. In situations where only few samples are initially available per individual, the challenge arises in extrapolating meaningful insights to later collected and measured follow-ups—owing to the dynamic nature of the empirical procedure as previously described. To examine whether our proposed approach offers added value when severely limited samples are available for analysis, we first employ CODI in the context of longitudinal analyses (Fig. 3).

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

Empowering long-term molecular profiling

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

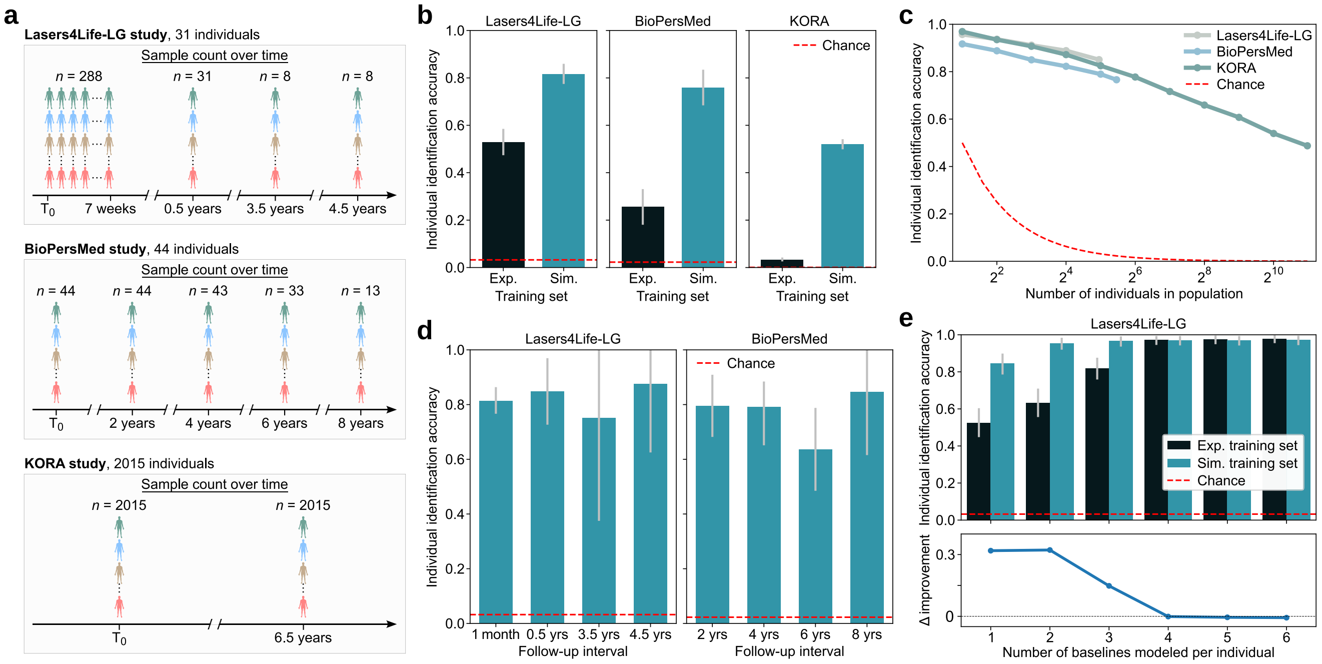

Fig. 3. CODI enhances personalized fingerprinting through more accurate long-term molecular profiling. a) Setup of three independent longitudinal clinical studies in which same individuals repeatedly participated in venous blood sampling over time. Experimental IR spectroscopic measurements were performed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

Downloaded from https://academic.oup.com/pnasnexus/article/3/10/pgae449/7823116 by guest on 19 February 2025

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

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

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

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

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

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

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

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

Personalized multibaseline modeling

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

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

Downloaded from https://academic.oup.com/pnasnexus/article/3/10/pgae449/7823116 by guest on 19 February 2025

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

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

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

Comparison to domain-agnostic augmentation schemes

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

Impact of increased variance on classification efficacy

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

Cross-specimen generalization

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

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

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

Downloaded from https://academic.oup.com/pnasnexus/article/3/10/pgae449/7823116 by guest on 19 February 2025

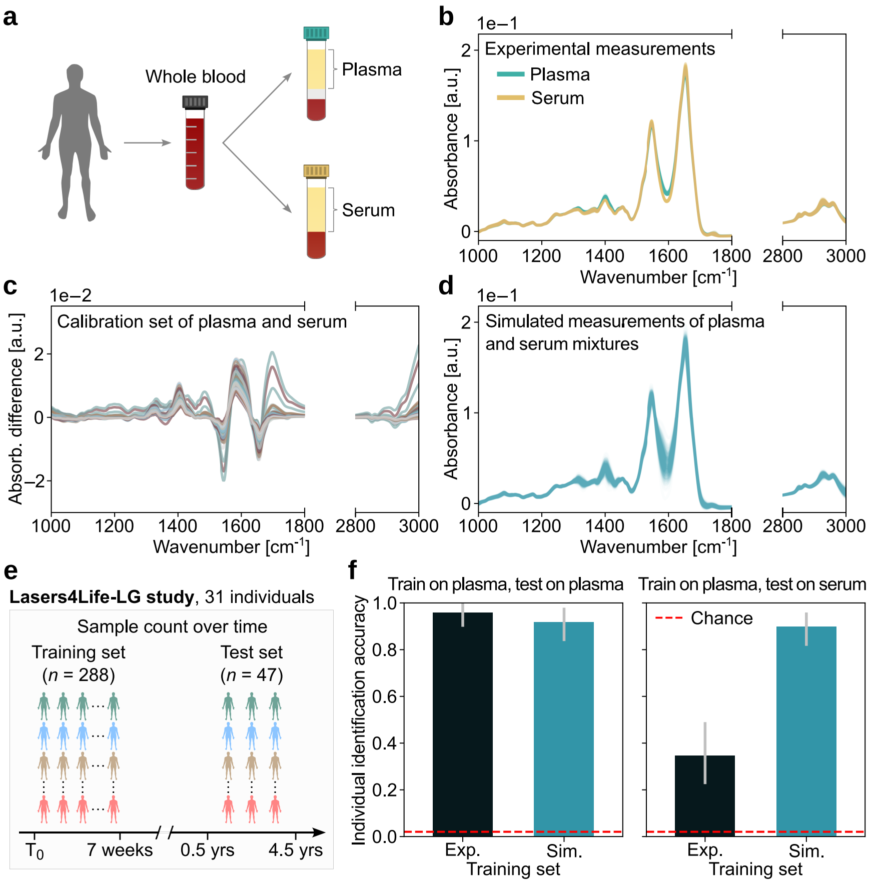

Fig. 4. CODI enables classification flexibility across biological specimen variants. a) Plasma and serum were collected as cell-free products of whole venous blood. b) Experimental spectra were measured from several plasma and serum samples of the same individuals. c) Differences between spectra of plasma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

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

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

Altogether, this proof-of-principle analysis further demonstrated the potential of CODI to overcome analytical limitations, enabling a classification transfer despite significant measurement deviations. For one, there is no need to re-collect a large number of specimens when deviations occur in the sample collection procedure. One can leverage a limited set of measurements that characterize differences between specimens, collected in a class-independent fashion. CODI may then extend ML applications to different specimen variations. Nevertheless, here we only demonstrated such potential on serum and plasma. If the specimens widely vary in their molecular composition and reflection of physiology (e.g. blood-based vs. urine- or saliva-based media), this approach may not perform as effectively as demonstrated here. A promising avenue for future exploration may involve adapting a classifier trained on EDTA plasma for use with citrate samples (43).

Generalization to independently acquired datasets

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

Downloaded from https://academic.oup.com/pnasnexus/article/3/10/pgae449/7823116 by guest on 19 February 2025

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

As a benchmark, we first investigated the performance of classifying each cancer entity, relying on a cross-validation procedure (Fig. 5b, gray curves/bars). The cross-validation was performed exclusively on the training set of experimental samples and the receiver operating characteristics (ROC) curve of the validation splits was examined for each cancer entity. Next, we trained a classifier on the training set of experimental samples, testing it on the test set of experimental samples (Fig. 5b, black curves/bars). This investigation revealed that the classification efficiency

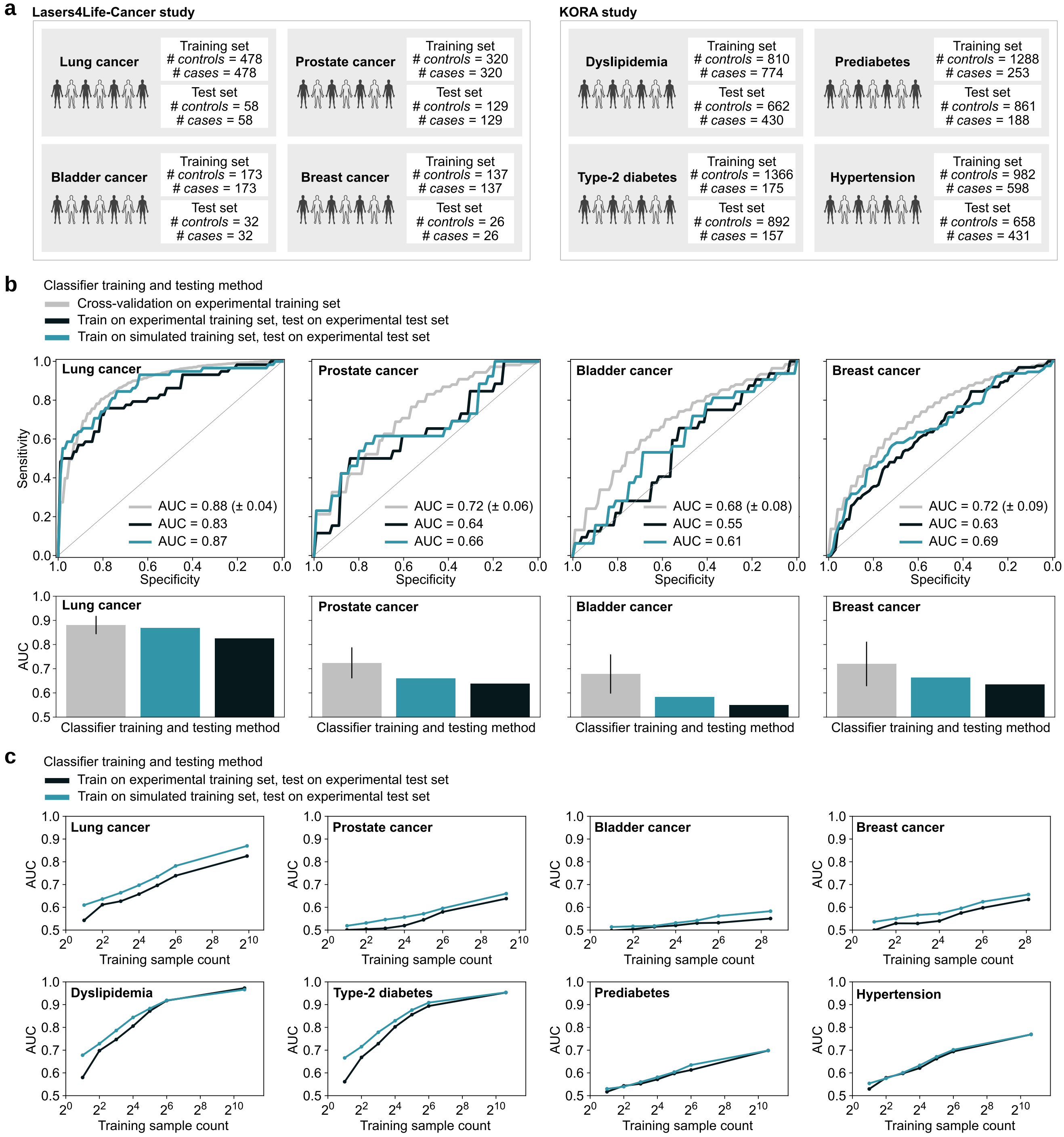

Fig. 5. CODI recovers lost classification efficacy on independently measured case–control test sets. a) Setup of eight binary classifications, spanning diverse health conditions, in two independent clinical studies. IR spectroscopy of blood plasma was performed on all samples. Training and test sample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

Downloaded from https://academic.oup.com/pnasnexus/article/3/10/pgae449/7823116 by guest on 19 February 2025

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

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

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

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

Influence of experimental training cohort size

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

We then randomly selected samples from the training sets at varying cohort sizes to train several classifiers on each subset of selected samples (Fig. 5c). First, we trained directly on the experimental observations, always testing on the held-out test sets (Fig. 5c, black curves). Unsurprisingly, the smaller the training set was, the worse the classifier performed on the experimental test sets. We then employed CODI to generate simulated datasets that utilized the experimental observations at each sample count as seed input (Fig. 5c, blue curves). Here, it was revealed the classification with CODI almost consistently outperformed the experimental modeling approach across the varying sample counts available as a basis for training. Notably, for the detection of dyslipidemia and type-2 diabetes, two conditions with strong molecular deviations reflected in IR fingerprints, CODI provided the largest advance when smaller training sample counts were available.

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

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

Computational efficiency

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

Discussion

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

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

Downloaded from https://academic.oup.com/pnasnexus/article/3/10/pgae449/7823116 by guest on 19 February 2025

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

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

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

Further applications of CODI to IR spectroscopic fingerprinting showcased its versatility and potential impact in aiding model generalization. We observed a remarkable level of comparability across experimental data collected over almost a decade, underlining the method's ability to improve classification efficacy on independently measured test sets. The adaptability of the framework extended to a proof-of-principle application that involved training a classifier on one sample medium (plasma) and applying it to another (serum). This application demonstrated the potential of CODI to streamline cross-specimen dataset analyses in various biological and biomedical applications. Such a strategy may be particularly valuable when gathering data sets from retrospective studies or online repositories to help ensure specimen comparability to another envisioned application—i.e. domain adaptation applications (72, 73). Another promising use of CODI would be to harmonize data obtained from different measurement devices (e.g. several spectrometers made by the same or different manufacturers), potentially improving ML model transferability between them (5).

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

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

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

Materials and methods

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

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

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

Downloaded from https://academic.oup.com/pnasnexus/article/3/10/pgae449/7823116 by guest on 19 February 2025

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

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

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

Four independent study cohorts were utilized in this study: Lasers4Life-LG (30), BioPersMed (28), KORA (29, 33), and Lasers4Life-Cancer (31). Lasers4Life-LG involved the collection of blood serum and plasma from 31 healthy individuals initially sampled up to 13 times over a 7-week period, with an additional follow-up after 6 months as detailed in a previous publication (30). Since this initial publication, the same individuals were invited to participate in two additional sample donations at 3.5 and 4.5 years post their initial involvement. The Lasers4Life-LG study was approved by the Ethics Committee of the Ludwig-Maximillian-University (LMU) of Munich and all participants provided written informed consent (research study protocol #17-532). BioPersMed is an ongoing population-based cohort performed at the Medical University Graz, Austria (28). Repetitive examinations of participants were conducted in 2-year intervals. In the current study, we utilized blood plasma samples and medical data from a subset of 44 healthy individuals (out of 1,022 participants). The BioPersMed study was approved by the Ethics Committee of the Medical University of Graz, Austria (EC Nr. 24-224 ex 11/12; project application number 4008_22). KORA is a population-based cohort in Southern Germany (29). The cohort comprised of an age- and gender-stratified sample of participants randomly drawn from the resident registration offices within the study area. In the current study, we utilized blood plasma samples and medical data from the second and third visits (named KORA-F4 and KORA-FF4, respectively). The available KORA-F4 data consisted of 3,044 samples, while the KORA-FF4 data consisted of 2,140 samples. A subset of 2015 individuals participated in both samplings, while 1,154 individuals participated in only one of the samplings. The KORA-F4 and KORA-FF4 study methods were approved by the ethics committee of the Bavarian Chamber of Physicians, Munich (EC No. 06068). Lasers4Life-Cancer is a case–control study cohort involving several cancer entities where both serum and plasma are collected (31). The samples utilized in this study largely overlapped with samples from our previous study (31)—although the measurement procedures differed (Supplementary Information). Since the original publication, blood plasma and serum samples from different individuals were collected and included in this study. Case samples were collected prior to cancer-related treatment. Nonsymptomatic controls were pair-matched to cancer cases by age, gender, and body mass index. All participants in the Lasers4Life-Cancer study provided written informed consent for the study under research study protocol #17-141 and under research study protocol #17-182, both of which were approved by the Ethics Committee of the LMU of Munich. The clinical trial is registered at the German Clinical Trials Register (ID DRKS00013217).

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

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

Acknowledgments

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

Supplementary Material

Supplementary material is available at *PNAS Nexus* online.

Funding

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

Downloaded from https://academic.oup.com/pnasnexus/article/3/10/pgae449/7823116 by guest on 19 February 2025

Author Contributions

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

Preprint

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

Data Availability

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

References

1 Batool SM, *et al.* 2023. Extrinsic and intrinsic preanalytical variables affecting liquid biopsy in cancer. *Cell Rep Med*. 4(10):101196.

2 Bowen RA, Remaley AT. 2014. Interferences from blood collection tube components on clinical chemistry assays. *Biochem Med (Zagreb)*. 24(1):31–44.

3 Čuklina J, *et al.* 2021. Diagnostics and correction of batch effects in large-scale proteomic studies: a tutorial. *Mol Syst Biol*. 17(8): e10240.

4 Dvinge H, *et al.* 2014. Sample processing obscures cancer-specific alterations in leukemic transcriptomes. *Proc Natl Acad Sci U S A*. 111(47):16802–16807.

5 Guo S, *et al.* 2020. Comparability of Raman spectroscopic configurations: a large scale cross-laboratory study. *Anal Chem*. 92(24): 15745–15756.

6 Kwak JT, Reddy R, Sinha S, Bhargava R. 2011. Analysis of variance in spectroscopic imaging data from human tissues. *Anal Chem*. 84(2):1063–1069.

7 Morais CLM, Lima KMG, Singh M, Martin FL. 2020. Tutorial: multivariate classification for vibrational spectroscopy in biological samples. *Nat Protoc*. 15(7):2143–2162.

8 Perez-Guaita D, Ventura-Gayete J, Pérez-Rambla C, Sancho-Andreu M, Garrigues S. 2013. Evaluation of infrared spectroscopy as a screening tool for serum analysis. *Microchem J*. 106:202–211.

9 Yin P, Lehmann R, Xu G. 2015. Effects of pre-analytical processes on blood samples used in metabolomics studies. *Anal Bioanal Chem*. 407(17):4879–4892.

10 Check E. 2004. Proteomics and cancer: running before we can walk? *Nature*. 429(6991):496–497.

11 Cohen JP, *et al.* 2021. Problems in the deployment of machine-learned models in health care. *Can Med Assoc J*. 193(35): E1391–E1394.

12 Goetz L, Seedat N, Vandersluis R, van der Schaar M. 2024. Generalization—a key challenge for responsible AI in patient-facing clinical applications. *NPJ Digit Med*. 7(1):126.

13 Moreno-Torres JG, Raeder T, Alaiz-Rodríguez R, Chawla NV, Herrera F. 2012. A unifying view on dataset shift in classification. *Pattern Recognit*. 45(1):521–530.

14 Obermeyer Z, Emanuel EJ. 2016. Predicting the future—big data, machine learning, and clinical medicine. *N Engl J Med*. 375(13): 1216–1219.

15 Zech JR, *et al.* 2018. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med*. 15(11):e1002683.

16 Gagnon-Bartsch JA, Speed TP. 2011. Using control genes to correct for unwanted variation in microarray data. *Biostatistics*. 13(3):539–552.

17 Livera AMD, *et al.* 2015. Statistical methods for handling unwanted variation in metabolomics data. *Anal Chem*. 87(7):3606–3615.

18 Molania R, *et al.* 2022. Removing unwanted variation from large-scale RNA sequencing data with PRPS. *Nat Biotechnol*. 41(1):82–95.

19 Molania R, Gagnon-Bartsch JA, Dobrovic A, Speed TP. 2019. A new normalization for nanostring nCounter gene expression data. *Nucleic Acids Res*. 47(12):6073–6083.

20 Peng M, Li Y, Wamsley B, Wei Y, Roeder K. 2021. Integration and transfer learning of single-cell transcriptomes via cFIT. *Proc Natl Acad Sci U S A*. 118(10):e2024383118.

21 Chong Y, *et al.* 2023. Machine learning of spectra-property relationship for imperfect and small chemistry data. *Proc Natl Acad Sci U S A*. 120(20):e2220789120.

22 Liu J, *et al.* 2023. Towards out-of-distribution generalization: a survey, arXiv, arXiv:2108.13624, preprint: not peer reviewed. https://doi.org/10.48550/arXiv.2108.13624

23 Mirkes EM, *et al.* 2022. Domain adaptation principal component analysis: base linear method for learning with out-of-distribution data. *Entropy*. 25(1):33.

24 Li X, *et al.* 2022. Uncertainty modeling for out-of-distribution generalization. In: International Conference on Learning Representations (ICLR).

25 Zhang X, *et al.* 2022. Towards unsupervised domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). p. 4910–4920.

26 Mikolajczyk A, Grochowski M. 2018. Data augmentation for improving deep learning in image classification problem. In: 2018 International Interdisciplinary PhD Workshop (IIPhDW). IEEE.

27 Shorten C, Khoshgoftaar TM, Furht B. 2021. Text data augmentation for deep learning. *J Big Data*. 8(1):101.

28 Haudum CW, *et al.* 2022. Cohort profile: 'biomarkers of personalised medicine' (BioPersMed): a single-centre prospective observational cohort study in Graz/Austria to evaluate novel biomarkers in cardiovascular and metabolic diseases. *BMJ Open*. 12(4):e058890.

29 Holle R, Happich M, Löwel H, Wichmann H. 2005. KORA - a research platform for population based health research. *Das Gesundheitswesen*. 67(S 01):19–25.

30 Huber M, *et al.* 2021. Stability of person-specific blood-based infrared molecular fingerprints opens up prospects for health monitoring. *Nat Commun*. 12(1):1511 .

31 Huber M, *et al.* 2021. Infrared molecular fingerprinting of blood-based liquid biopsies for the detection of cancer. *eLife*. 10:e68758.

Downloaded from https://academic.oup.com/pnasnexus/article/3/10/pgae449/7823116 by guest on 19 February 2025

32 Eissa T, Kepesidis KV, Zigman M, Huber M. 2023. Limits and prospects of molecular fingerprinting for phenotyping biological systems revealed through in silico modeling. *Anal Chem*. 95(16): 6523–6532.

33 Eissa T, *et al*. 2024. Plasma infrared fingerprinting with machine learning enables single-measurement multi-phenotype health screening. *Cell Rep Med*. 5(7):101625.

34 Assfalg M, *et al*. 2008. Evidence of different metabolic phenotypes in humans. *Proc Natl Acad Sci U S A*. 105(5):1420–1424.

35 Wallner-Liebmann S, *et al*. 2016. Individual human metabolic phenotype analyzed by 1H-NMR of saliva samples. *J Proteome Res*. 15(6):1787–1793.

36 Yousri NA, *et al*. 2014. Long term conservation of human metabolic phenotypes and link to heritability. *Metabolomics*. 10(5): 1005–1017.

37 Moqri M, *et al*. 2023. Biomarkers of aging for the identification and evaluation of longevity interventions. *Cell*. 186(18): 3758–3775.

38 Raudys S, Jain A. 1991. Small sample size effects in statistical pattern recognition: recommendations for practitioners. *IEEE Trans Pattern Anal Mach Intell*. 13(3):252–264.

39 Baker MJ, *et al*. 2014. Using fourier transform IR spectroscopy to analyze biological materials. *Nat Protoc*. 9(8):1771–1791.

40 Chetwynd AJ, Dunn WB, Rodriguez-Blanco G. 2017. Collection and preparation of clinical samples for metabolomics. In: Advances in experimental medicine and biology. Springer International Publishing. p. 19–44.

41 Yu Z, *et al*. 2011. Differences between human plasma and serum metabolite profiles. *PLoS One*. 6(7):e21230.

42 Eissa T, Voronina L, Huber M, Fleischmann F, Žigman M. 2024. The perils of molecular interpretations from vibrational spectra of complex samples. *Angewandte Chemie International Edition*, pages Accepted Manuscript, In Press.

43 Staniszewska E, Bartosz AK, Malek K, Baranska M. 2013. An effect of anticoagulants on the FTIR spectral profile of mice plasma. *Biomed Spectrosc Imaging*. 2(4):317–330.

44 Soneson C, Gerster S, Delorenzi M. 2014. Batch effect confounding leads to strong bias in performance estimates obtained by cross-validation. *PLoS One*. 9(6):e100335.

45 Cameron JM, *et al*. 2020. Exploring pre-analytical factors for the optimisation of serum diagnostics: progressing the clinical utility of ATR-FTIR spectroscopy. *Vib Spectrosc*. 109: 103092.

46 González-Domínguez R, González-Domínguez Á, Sayago A, Fernández-Recamales Á. 2020. Recommendations and best practices for standardizing the pre-analytical processing of blood and urine samples in metabolomics. *Metabolites*. 10(6):229.

47 Pérez-Guaita D, Quintás G, Farhane Z, Tauler R, Byrne HJ. 2022. Combining pharmacokinetics and vibrational spectroscopy: MCR-ALS hard-and-soft modelling of drug uptake in vitro using tailored kinetic constraints. *Cells*. 11(9):1555.

48 Eling N, Morgan MD, Marioni JC. 2019. Challenges in measuring and understanding biological noise. *Nat Rev Genet*. 20(9): 536–548.

49 Hawkridge AM, Muddiman DC. 2009. Mass spectrometry–based biomarker discovery: toward a global proteome index of individuality. *Annu Rev Anal Chem*. 2(1):265–277.

50 López-Otín C, Kroemer G. 2021. Hallmarks of health. *Cell*. 184(1): 33–63.

51 López-Otín C, Blasco MA, Partridge L, Serrano M, Kroemer G. 2013. The hallmarks of aging. *Cell*. 153(6):1194–1217.

52 Rose SMS-F, *et al*. 2019. A longitudinal big data approach for precision health. *Nat Med*. 25(5):792–804.

53 Guo S, Popp J, Bocklitz T. 2021. Chemometric analysis in Raman spectroscopy from experimental design to machine learning–based modeling. *Nat Protoc*. 16(12):5426–5459.

54 Haghverdi L, Lun ATL, Morgan MD, Marioni JC. 2018. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol*. 36(5): 421–427.

55 Arjovsky M, Bottou L, Gulrajani I, Lopez-Paz D. 2019. Invariant risk minimization, arXiv, arXiv:1907.02893, preprint: not peer reviewed. https://doi.org/10.48550/arXiv.1907.02893

56 Bellot A, van der Schaar M. 2022. Accounting for unobserved confounding in domain generalization, arXiv, arXiv:2007.10653, preprint: not peer reviewed. https://doi.org/10.48550/arXiv.2007.10653

57 Krueger D, *et al*. 2021. Out-of-distribution generalization via risk extrapolation (rex). In: Proceedings of the 38th International Conference on Machine Learning of Proceedings of Machine Learning Research. vol. 139. PMLR. p. 5815–5826.

58 Sun B, Saenko K. 2016. Deep coral: correlation alignment for deep domain adaptation. In: Hua G, Jégou H, editors. Computer Vision – ECCV 2016 Workshops. Cham: Springer International Publishing. p. 443–450.

59 Teney D, Abbasnejad E, Lucey S, Van den Hengel A. 2022. Evading the simplicity bias: training a diverse set of models discovers solutions with superior OOD generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. p. 16761-16772.

60 Li C, Zhang D, Huang W, Zhang J. 2023. Cross contrasting feature perturbation for domain generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. p. 1327-1337.

61 Mazaheri B, Mastakouri A, Janzing D, Hardt M. 2023. Causal information splitting: engineering proxy features for robustness to distribution shifts. In: Shpitser I ERJ, editor. Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence of Proceedings of Machine Learning Research. vol. 216. PMLR. p. 1401-1411.

62 Yao H, *et al*. 2022. Improving out-of-distribution robustness via selective augmentation. In: International Conference on Machine Learning. PMLR. p. 25407-25437.

63 Bjerrum EJ, Glahder M, Skov T. 2017. Data augmentation of spectral data for convolutional neural network (CNN) based deep chemometrics, arXiv, arXiv:1710.01927, preprint: not peer reviewed. https://doi.org/10.48550/arXiv.1710.01927

64 Freer D, Yang G-Z. 2020. Data augmentation for self-paced motor imagery classification with c-LSTM. *J Neural Eng*. 17(1): 016041.

65 Guo S, *et al*. 2018. Model transfer for Raman-spectroscopy-based bacterial classification. *J Raman Spectrosc*. 49(4):627–637.

66 Lebrun A, *et al*. 2022. Pushing the limits of surface-enhanced Raman spectroscopy (SERS) with deep learning: identification of multiple species with closely related molecular structures. *Appl Spectrosc*. 76(5):609–619.

67 Lotte F. 2015. Signal processing approaches to minimize or suppress calibration time in oscillatory activity-based brain–computer interfaces. *Proc IEEE*. 103(6):871–890.

68 Tsinganos P, Cornelis B, Cornelis J, Jansen B, Skodras A. 2020. Data augmentation of surface electromyography for hand gesture recognition. *Sensors*. 20(17):4892.

69 Wang F, Zhong S, Peng J, Jiang J, Liu Y. 2018. Data augmentation for EEG-based emotion recognition with deep convolutional neural networks. In: MultiMedia modeling. Springer International Publishing. p. 82–93.

Downloaded from https://academic.oup.com/pnasnexus/article/3/10/pgae449/7823116 by guest on 19 February 2025

70 Zanini RA, Colombini EL. 2020. Parkinson's disease EMG data augmentation and simulation with DCGANs and style transfer. *Sensors*. 20(9):2605.

71 Sohn K, *et al*. 2020. Fixmatch: simplifying semi-supervised learning with consistency and confidence. *Adv Neural Inf Process Syst*. 33:596–608.

72 Bareinboim E, Pearl J. 2016. Causal inference and the data-fusion problem. *Proc Natl Acad Sci U S A*. 113(27): 7345–7352.

73 Kyono T, van der Schaar M. 2021. Exploiting causal structure for robust model selection in unsupervised domain adaptation. *IEEE Trans Artif Intell*. 2(6):494–507.

74 Argelaguet R, Cuomo ASE, Stegle O, Marioni JC. 2021. Computational principles and challenges in single-cell data integration. *Nat Biotechnol*. 39(10):1202–1215.

75 Dorkenwald S, *et al*. 2023. Multi-layered maps of neuropil with segmentation-guided contrastive learning. *Nat Methods*. 20(12): 2011–2020.

Downloaded from https://academic.oup.com/pnasnexus/article/3/10/pgae449/7823116 by guest on 19 February 2025