DISS. ETH NO. 28718

Artificial Intelligence for Long-Term Reproducible High-Throughput Untargeted Metabolomics

A thesis submitted to attain the degree of
DOCTOR OF SCIENCES
(Dr. sc. ETH Zurich)

presented by
ANDREI DMITRENKO
Master of Science, Peter the Great St. Petersburg Polytechnic University
born on 27.03.1994

accepted on the recommendation of

Prof. Dr. Nicola Zamboni
Prof. Dr. Uwe Sauer
Prof. Dr. Bernd Wollscheid
Prof. Dr. Juho Rousu

2022

In memory of my father

Contents

ABSTRACT **7**

ZUSAMMENFASSUNG **11**

CHAPTER 1 — GENERAL INTRODUCTION **16**

CHAPTER 2 — A SYSTEM SUITABILITY TESTING PLATFORM FOR UNTARGETED, HIGH-RESOLUTION MASS SPECTROMETRY **40**

CHAPTER 3 — REGULARIZED ADVERSARIAL LEARNING FOR NORMALIZATION OF MULTI-BATCH UNTARGETED METABOLOMICS DATA **71**

CHAPTER 4 — PREDICTION OF RELATIVE METABOLITE CONCENTRATIONS IN HUMAN SERUM ANALYZED BY FLOW INJECTION TIME-OF-FLIGHT MASS SPECTROMETRY **111**

CHAPTER 5 — COMPARING REPRESENTATIONS OF BIOLOGICAL DATA LEARNED WITH DIFFERENT AI PARADIGMS, AUGMENTING AND CROPPING STRATEGIES **132**

CHAPTER 6 — SELF-SUPERVISED LEARNING FOR ANALYSIS OF TEMPORAL AND MORPHOLOGICAL DRUG EFFECTS IN CANCER CELL IMAGING DATA **159**

CHAPTER 7 — CONCLUDING REMARKS **185**

ACKNOWLEDGMENTS **194**

CURRICULUM VITAE **195**

Abstract

Mass spectrometry (MS)-based assays suffer from the inherent variability of measurements across instruments and over time, caused by multiple sources of variation, such as differences in sample preparation and system setups, biological matrix effects, acquisition batch effects and so on. Across -omics, reproducibility of quantitative experiments is a well-known issue. Untargeted metabolomics is the method of choice for comprehensive characterization of all chemical compounds that occur in a cell of a biological sample. With growing demand for untargeted metabolomics in personalized health applications, it is crucial to achieve the level of reproducibility enabling robust sample quantification in longitudinal clinical studies.

In this PhD thesis, we aim at improving reproducibility of untargeted metabolomics leveraging the most recent developments in AI. We build a platform for continuous system suitability testing (SST), develop a batch correction method and investigate calibration strategies for a high-throughput acquisition method. Complementary to these efforts, we investigate representation learning approaches across data modalities and develop an explainable deep learning application to demonstrate exciting opportunities for multi-modal biomedical research.

In **Chapter 2**, we develop an SST platform to report on a mass spectrometer state. For that, we design a QC sample that produces 37 robust ion peaks, including isotopes, fragments and adducts. We develop an acquisition method to measure the QC sample repeatedly and collect detailed spectral information related to the chemical background, QC mix and detector noise scans of ion chronogram. After the measurement, the raw profile data is processed by the software to extract 2850 numerical features reflecting different aspects of the system, e.g., resolution, mass accuracy, ionization efficiency, levels of dirt and detector noise, etc. Those are further used to engineer the 16 quality indicators visualized for users through a web-service. We analyze in depth the QC data systematically acquired within 2 years, investigate the relationships between the QC features and the instrument settings and discuss the potentials of automatic retuning and MS diagnostics applications.

In **Chapter 3**, we develop RALPS (regularized adversarial learning preserving similarity) to correct for batch effects in untargeted metabolomics data. We propose a loss function that consists of three terms: the one penalizing batch separation to mitigate batch effects, another one promoting tight clustering of replicates to retain the biological information, and the last one penalizing sample-wise variance increase to stabilize the training process. We test RALPS extensively on several multi-batch datasets and compare to state-of-the-art methods. In addition, we run series of ablation experiments to demonstrate flexibility, scalability and robustness of our method. One of the datasets was generated specifically for the purpose of benchmarking data normalization methods. It has a total of 2750 samples acquired in 7 batches over a course of 2 months. The samples comprise human serum and ^{13}C-labeled E. coli extracts with three classes of different spike-ins and dilution series. The benchmarking dataset is largely affected by batch effects and presents a major challenge for a number of recently proposed normalization approaches.

In **Chapter 4**, we use the benchmarking dataset of flow injection time-of-flight mass spectrometry (FIA-TOF-MS) for analysis of calibration curves. We formulate three calibration-related tasks and approach them with machine learning. We train models to predict relative concentrations, absolute and relative ion abundances in human serum extracts. Then, we carefully assess generalizing capabilities of the models relevant for applications. Specifically, we test the ability of each model to extrapolate across dilution factors and structurally similar compounds. Finally, we discuss current limitations and next steps to improve the calibration of FIA-TOF-MS.

The final two chapters describe collaboration projects on the metabolism of cancer cell lines under drug pressure. These works are not directly related to metabolomics, since my role was to develop AI frameworks for the analysis of time-resolved images of tissue cultures.

In **Chapter 5**, we investigate and compare representation learning approaches using cancer cell imaging data. We implement four different models having the same CNN backbone for downstream feature extraction. We train them under identical conditions testing four strategies of random augmentations and crops for each model. To compare the learned representations, we formulate three independent tasks and evaluate multiple metrics for each. Among other empirical results, we confirm the efficiency of multi-task representation learning approaches across data modalities.

In **Chapter 6**, we develop an explainable deep learning model to classify drugs based on cells images of different cancer types. The model features intrinsic local interpretability, i.e., it allows to visualize regions of an image driving the classifier decision. We present many examples shedding light on drug-specific morphological features of cells and discuss the potential to extend this approach to a multi-modal setting, e.g., to explain image classification with the corresponding metabolomics data.

Finally, in **Chapter 7**, we summarize the key contributions of this thesis. We address the reproducibility issue of metabolomics from multiple standpoints. We propose means of quality assurance and quality control for data acquisition, investigate calibration strategies and develop a batch correction method to improve comparability of FIA-TOF-MS measurements across acquisition batches. We demonstrate how these individual contributions advance reproducibility of FIA-TOF-MS and discuss their potential to synergize to enable reliable implementation of untargeted metabolomics into clinics. Additionally, we investigate deep representation learning approaches and demonstrate advantages of multi-task learning across data modalities. We present an explainable AI application for analysis of anti-cancer drugs and discuss particular ways to integrate mass spectrometry, microscopy imaging or other data modalities for fundamental and clinical research using cutting-edge AI technology.

Zusammenfassung

Massenspektrometrie (MS)-basierte Assays leiden unter der inhärenten Variabilität von Messungen zwischen verschiedenen Instrumenten und im Laufe der Zeit, die durch mehrere Variationsquellen verursacht wird, wie z.B. Unterschiede bei der Probenvorbereitung und Systemeinrichtung, biologische Matrixeffekte, Chargeneffekte bei der Datenerfassung usw. In allen Omic-Bereichen ist die Reproduzierbarkeit quantitativer Experimente ein bekanntes Problem. Ungezielte Metabolomik ist die Methode der Wahl für die umfassende Charakterisierung aller chemischen Verbindungen, die in einer Zelle einer biologischen Probe vorkommen. Angesichts der wachsenden Nachfrage nach ungezielter Metabolomik in personalisierten Gesundheitsanwendungen ist es von entscheidender Bedeutung, ein Niveau der Reproduzierbarkeit zu erreichen, das eine robuste Probenquantifizierung in klinischen Längsschnittstudien ermöglicht.

In dieser Doktorarbeit zielen wir darauf ab, die Reproduzierbarkeit der ungezielten Metabolomik zu verbessern, indem wir die neuesten Entwicklungen in der KI nutzen. Wir bauen eine Plattform für kontinuierliche Systemeignungstests (SST) auf, entwickeln eine Batch-Korrekturmethode und untersuchen Kalibrierstrategien für eine Hoch-Durchsatz-Erfassungsmethode. Ergänzend zu diesen Bemühungen untersuchen wir die Methoden von Repräsentationslernen über Datenmodalitäten hinweg und entwickeln eine erklärbare Deep-Learning-Anwendung, um spannende Möglichkeiten für die multimodale biomedizinische Forschung aufzuzeigen.

In **Kapitel 2** entwickeln wir eine SST-Plattform, um den Zustand eines Massenspektrometers zu ermitteln. Dazu entwerfen wir eine QC-Probe, die 37 robuste Ionenpeaks erzeugt, einschließlich Isotopen, Fragmenten und Addukten. Wir entwickeln eine Erfassungsmethode, um die QC-Probe wiederholt zu messen und detaillierte spektrale Informationen in Bezug auf den chemischen Hintergrund, den QC-Mix und die Detektorrauschscans des Ionenchronogramms zu sammeln. Nach der Messung werden die rohen Profildaten von der Software verarbeitet, um 2850 numerische Merkmale zu extrahieren, die verschiedene Aspekte des Systems

widerspiegeln, z.B. Auflösung, Massengenauigkeit, Ionisationseffizienz, Grad der Verschmutzung und des Detektorrauschens usw. Aus diesen Merkmalen werden die 16 Qualitätsindikatoren erstellt, die den Benutzern über einen Webdienst angezeigt werden. Wir analysieren eingehend die innerhalb von 2 Jahren systematisch erfassten QC-Daten, untersuchen die Beziehungen zwischen den QC-Merkmalen und den Geräteeinstellungen und erörtern die Möglichkeiten der automatischen Neueinstellung und MS-Diagnoseanwendungen.

In **Kapitel 3** entwickeln wir RALPS (regularized adversarial learning preserving similarity) zur Korrektur von Batch-Effekten in ungezielten Metabolomik-Daten. Wir schlagen eine Verlustfunktion vor, die sich aus drei Begriffen zusammensetzt: Ein Begriff bestraft die Batch-Trennung, um Batch-Effekte abzuschwächen, ein weiterer fördert die enge Clusterung von Replikaten, um die biologische Information zu erhalten, und der letzte bestraft die probenweise Varianzzunahme, um den Trainingsprozess zu stabilisieren. Wir testen RALPS ausgiebig an verschiedenen Multi-Batch-Datensätzen und vergleichen die Ergebnisse mit modernsten Methoden. Darüber hinaus führen wir eine Reihe von Ablationsexperimenten durch, um die Flexibilität, Skalierbarkeit und Robustheit unserer Methode zu demonstrieren. Einer der Datensätze wurde speziell zum Zwecke des Benchmarkings von Datennormalisierungsmethoden generiert. Es besteht aus insgesamt 2750 Proben, die in 7 Batches über einen Zeitraum von 2 Monaten gemessen wurden. Die Proben bestehen aus Humanserum und ^{13}C-markierten E. coli-Extrakten mit drei Klassen unterschiedlicher Spike-Ins und Verdünnungsreihen. Der Benchmarking-Datensatz wird in hohem Maße von Batch-Effekten beeinflusst und stellt eine große Herausforderung für eine Reihe kürzlich vorgeschlagener Normalisierungsmethode dar.

In **Kapitel 4** verwenden wir den Benchmarking-Datensatz der Fließinjektions-Flugzeit-Massenspektrometrie (FIA-TOF-MS) zur Analyse von Kalibrierkurven. Wir formulieren drei kalibrierungsbezogene Aufgaben und gehen diese mit dem maschinellen Lernen an. Wir trainieren Modelle zur Vorhersage von relativen Konzentrationen, absoluten und relativen Ionenhäufigkeiten in Humanserumextrakten. Danach bewerten wir sorgfältig die Verallgemeinerungsfähigkeiten der Modelle, die relevant für weitere Anwendungen sind. Insbesondere testen wir die Fähigkeit jedes Modells, über

Verdünnungsfaktoren und strukturell ähnliche Verbindungen hinweg zu extrapolieren. Schließlich diskutieren wir aktuelle Einschränkungen und nächste Schritte zur Verbesserung der Kalibrierung von FIA-TOF-MS.

In den letzten beiden Kapiteln werden Zusammenarbeitsprojekte zum Stoffwechsel von Krebszelllinien unter Medikamentendruck beschrieben. Diese Arbeiten stehen nicht in direktem Zusammenhang mit der Metabolomik, da meine Aufgabe darin bestand, KI-Frameworks für die Analyse von zeitaufgelösten Bildern von Gewebekulturen zu entwickeln.

In **Kapitel 5** untersuchen und vergleichen wir Ansätze zum Repräsentationslernen von Krebszellenbilddaten. Wir implementieren vier verschiedene Modelle mit demselben CNN-Backbone für die nachgelagerte Merkmalsextraktion. Wir trainieren sie unter identischen Bedingungen und testen vier Strategien für zufällige Augmentationen und Ausschnitte für jedes Modell. Um die erlernten Repräsentationen zu vergleichen, formulieren wir drei unabhängige Aufgaben und bewerten jeweils mehrere Metriken. Neben anderen empirischen Ergebnissen bestätigen wir die Effizienz von Multi-Task-Ansätzen zum Repräsentationslernen für verschiedene Datenmodalitäten.

In **Kapitel 6** entwickeln wir ein erklärbares Deep-Learning-Modell zur Klassifizierung von Medikamenten basierend auf Zellbildern verschiedener Krebsarten. Das Modell zeichnet sich durch eine intrinsische lokale Interpretierbarkeit aus, d.h. es ermöglicht die Visualisierung von Bildregionen, die die Klassifizierungsentscheidung beeinflussen. Wir stellen viele Beispiele vor, die medikamentspezifische morphologische Merkmale von Zellen beleuchten, und erörtern das Potenzial, diesen Ansatz auf ein multimodales Umfeld auszuweiten, z.B. um die Bildklassifizierung mit den entsprechenden Metabolomik-Daten zu erklären.

Abschließend fassen wir in **Kapitel 7** die wichtigsten Beiträge dieser Doktorarbeit zusammen. Wir betrachten das Reproduzierbarkeitsproblem der Metabolomik aus mehreren Blickwinkeln. Wir schlagen Mittel zur Qualitätssicherung und Qualitätskontrolle der Datenerfassung vor, untersuchen Kalibrierstrategien und entwickeln eine Batch-Korrekturmethode, um die Vergleichbarkeit von FIA-TOF-MS-Messungen der verschiedenen Batches zu verbessern. Wir zeigen, wie diese

einzelnen Beiträge die Reproduzierbarkeit von FIA-TOF-MS erhöhen und diskutieren ihr Synergiepotenzial, um eine zuverlässige Implementierung ungezielter Metabolomik in Kliniken zu ermöglichen. Darüber hinaus untersuchen wir Ansätze zum tiefen Repräsentationslernen und demonstrieren die Vorteile der Multi-Task-Ansätzen für verschiedene Datenmodalitäten. Wir stellen eine erklärbare KI-Anwendung für die Analyse von Krebsmedikamenten vor und erörtern besondere Möglichkeiten zur Integration von Massenspektrometrie-, Mikroskopie- und andere Arten von Daten für die Grundlagen- und klinische Forschung mit modernster KI-Technologie.

Chapter 1

General introduction

The role of metabolomics in systems biology

Systems biology aims at characterizing a biological system as a whole, as opposed to focusing on its individual parts[1]. Much of this effort is based on the comprehensive analysis of molecular components: genes, transcripts, proteins, metabolites, lipids, etc. This has been made possible by steady improvements of -omics technologies. Powerful analytical methods were developed to unravel each layer of the central dogma of molecular biology[2]. Decoding the genome of an organism and being able to track its realization through RNA molecules down to protein levels was the first necessary step to understand biological function and its regulation. To date, DNA and RNA analyses are dominated by high throughput sequencing technologies[3,4]. Proteins are primarily analyzed by mass spectrometry. It allows to sequence the peptides that result upon enzymatic digestion, and thereby identify a multitude of proteoforms. Further, it allows to analyse intact soluble proteins to unravel complexes.

Genes, transcripts, and proteins are hierarchically linked. They all originate from a stretch of DNA and, therefore, their variety is determined by the genome. This, however, only relates to their sequence, which is the information that determines the function. The actual building blocks of all cellular compartments originate from nutrients that are taken up from the environment. Genome-encoded enzymes play a fundamental role in transforming variable nutrients into hundreds of building blocks that are necessary to grow and duplicate an entire cell. The ensemble of all chemical compounds that occur in a cell is the metabolome, that is the set of all metabolites. Lipids have similar origin and fate and, therefore, they can be considered metabolites even though they are hydrophobic and tend to aggregate in droplets and bilayers.

Conceptually, the metabolome results from the interaction of the environment and the intracellular configuration determined by the genome through the proteome. All changes in gene expression or environment will converge at the metabolic level and potentially induce a new steady-state. To ensure fitness in face of changing conditions or stress, cells have evolved regulatory mechanisms that sense metabolite levels and trigger adaptive programs to optimally adjust cellular or metabolic processes. These processes, in turn, are reflected in metabolic changes. Overall, measuring the metabolome provides unique information on cellular status. Compared to, e.g., proteomics, it allows to capture the integrated response of the metabolic network, and it is therefore best suited to investigate metabolic phenotypes and regulation.

Methods of metabolomics

In metabolomics by mass spectrometry, metabolites must be converted into charged species (ions) to be moved and isolated by electrical and magnetic fields. Also the quantification requires ions, which are detected as current by extremely fast and sensitive electronic devices. As the counting device is not able to discern the identity of ions generating the current, different means have been implemented to identify metabolites. The first one is to measure the molecular weight or, to be more precise, the mass-to-charge ratio of the molecular ion. The two frequently desirable analytical parameters in this process are resolution and mass accuracy. Resolution essentially represents the ability to distinguish between ions of close mass-to-charge ratios. It can be calculated as ion mass divided by the full width at half maximum of the ion intensity peak. Thus, the higher the resolution the better. Mass accuracy can be defined as proximity of experimentally measured ion mass to the theoretical one (calculated by adding up the atom masses of a molecule). Therefore, low values of mass accuracy are critical for metabolite identification.

To measure mass-to-charge ratios, the Orbitrap mass analyzers use frequency of harmonic oscillations of the orbitally trapped ions[5]. Despite the fact that Orbitraps offer the best resolution and mass accuracy, there are specific applications where other systems may be preferable. In particular, Time-of-Flight (TOF) mass analyzers measure the travel time of ions accelerated through the same electrical potential until they reach the detector[6]. There is evidence[7] suggesting that TOF systems are capable of detecting many more metabolites in complex samples with flow injection analysis (FIA). Other types of mass analyzers exist as well, such as Quadrupole Ion Trap, Ion Cyclotron Resonance and others[8].

The second technique to identify metabolites is to use multi-stage mass spectrometry and fragmentation. Most MS instruments have the capacity to induce the breakdown of ions by, e.g., collision at high kinetic energy or other forms of molecular excitation. Compounds will break depending on their structure and generate peculiar fragments. Fragments can be analysed by the MS analyzer to obtain a catalogue of all moieties holding any charge. This is the MS2 spectrum. On some instruments, it is possible to select fragments and perform additional cycles of fragmentation with increasing activation energy. This leads to MS3 or, in general, MSn spectra. All fragmentation spectra can be used to infer structural properties of the molecular ion, i.e., to determine what functional groups are present (or not). The inference of structure from MS2 or MSn is a non-trivial task which is solved either by comparing to MS2-libraries obtained for known compounds or by machine learning[9].

A third source for identification is the use of separation techniques. They aim at separating molecules in time on the basis of chemical or physical properties. The most common types of separations are gas and liquid chromatography (GC and LC), which adopt a column in front of the mass spectrometer. In metabolomics, liquid chromatography is more diffused because most metabolites are natively amenable to LC[10,11]. In the past decade, separation of molecules in the gas phase based on their cross-section has become also quite popular. Collectively, these methods add orthogonal information in the identification process. Similarly to MS2 data, interpretation of retention and mobility time can rely on known values for chemical standards or predictions by machine learning.

Two fundamental approaches to metabolomics exist: targeted and untargeted (or non-targeted). Targeted metabolomics is used to detect substances of particular interest. Typically, those are metabolites expected to appear within predefined mass ranges and measured to compare multiple experimental conditions to confirm or reject hypotheses. In targeted metabolomics, all protocols and acquisition methods are tailored to the compounds of interest. These include sample preparation, internal standards, chromatographic steps, MS acquisition, and data analysis. The main advantages of, and reasons to choose, targeted metabolomics are the best performance in quantification and best sensitivity.

Non-targeted metabolomics is better suited for discovery. Its fundamental goal is to detect and “profile” as many metabolites as possible. Such data are analyzed by statistical means and, therefore, don't require quantification to support hypothesis generation[12]. Non-targeted methods employ instruments with the highest resolution to capture tens of thousands of ions. Processing of untargeted data is challenging, in particular for large scale studies[13]. A problem of this approach is the identification. Typically, thousands of metabolites are identified putatively (based on machine learning predictions), and only a few hundreds are assigned with high confidence because of the availability of chemical standards. Also in the absence of unambiguous identification, pathway enrichment analysis allows to investigate cellular responses[14]. Overall, untargeted metabolomics is the method of choice for comprehensive characterization of biological samples.

The role of metabolomics in personalized health

Multiple clinical applications of metabolomics have emerged in the last decade, including newborn screening, identification of biomarkers of disease and novel therapeutic targets, as well as personalized phenotyping and drug-response monitoring[15]. A number of recent publications acknowledge the growing importance of metabolomics for personalized medicine[16–18]. However, several challenges exist hampering inter-operability and reusability of metabolomics data, which is crucial for frequently longitudinal clinical studies[19]. Lack of standardization and reproducibility is definitely one of them.

Reproducibility problems in LC-MS based metabolomics

Reproducibility of experiment is a cornerstone of the scientific method. In absence of it, any scientific conclusion is doubted. In the recent years, several publications discussed the phenomenon of reproducibility crisis in science[20,21]. A survey of 703 scientists on reproducibility of research in biology reported over 60% of respondents who were unable to replicate their own or someone else's experiment[22]. Additionally, the respondents estimated almost a half of pubslihed work in their field not reproducible. The practical implications of that are, of course, additional efforts required to confirm previously reported results, poorly informed follow-up study designs, growing skepticism of the scientific community to the kind of analysis and beyond. Although top factors contributing to the lack of reproducibility seem to be anthropological[23], there exists a purely technical aspect to the issue.

The reproducibility problems in metabolomics, for instance, are rooted in the nature of the process that underly ionization and detection of ions. There are multiple sources of variation affecting the measurements on different levels. Day-to-day variability of the instrument response in a multi-day experimental design results in batch effects, i.e., biases specific to the system setup at acquisition times. Those can be magnified by the differences in the sample preparation, i.e., by variability among the replicates of the same experimental conditions. Matrix effects introduce additional layer of variability due to co-eluting biological matrix components altering the ionization of target metabolites. On top of that, samples measured in long sequences often present signal drifts associated with accumulated dirt in the system and contamination of precursor ions during electrospray ionization. All of these factors are inherent to mass spectrometry based metabolomics and limit reproducibility of the measurements.

Quantification in metabolomics

Two common approaches to compensate for the above problems are external and internal standardization[24]. External standardization refers to a reference standard measured at a series of known concentrations to calibrate the instrument response. A calibration line describes the signal produced by the analyte as a function of concentration. As soon as this dependency is established, it can be further used to determine the unknown concentrations given the sample preparation procedure stays the same.

Alternatively, internal stardardization is used in cases when multiple sample preparation steps are involved that may be associated with increased variability of response[25]. Internal standards (IS) are added at the same concentration to each sample to further calculate ratios of concentrations between the analyte and the IS that are conserved better. Therefore, calibration with internal standards defines a relationship between ratios of concentrations and ratios of the responses between the analyte and the IS of the experiment. For application, the ratio of responses needs to be calculated first to determine the unknown sample concentration with a calibration curve.

Linearity of the calibration indicates the quality of a bioanalytical method, thus, serving as a mean of quality assurance and control[26,27]. However, this only applies to LC-MS methods, as opposed to the FIA-MS often referred to as semi-quantitative. Flow injection analysis coupled with electrospray ionization mass spectrometry (FIA-ESI-MS) often suffers from ion suppression driven by competitive ionization of compounds[28]. Moreover, ion suppression effects may differ among compound classes and biological matrices. In practice, it results in a non-linear relationship between compound concentration and detector response[29]. This is why a general practice with FIA-MS is to calculate and compare fold-changes between experimental conditions.

A few works tried to address the aforementioned limitation in the past. In 2015, a matrix-induced ion suppression method was proposed to predict relative concentrations in urinary samples[30]. The authors experimented with multiple compounds to select an ion suppression indicator and then used it to predict low, medium and high concentrations in diluted pooled urine samples (16-fold, 10-fold and 2-fold, respectively). They were able to achieve the accuracy ranging from 97.15% to 102.10% and successfully apply the method to metabolic profiling of breast cancer. In 2021, a DS-FIA (**d**ilute and **s**hoot) method was developed to accurately quantify amino acids in microbial cultivation supernatants[31]. The authors argue that with large sample dilutions and the use of labeled isotope standards the ion suppression and matrix effects can be minimized or even excluded, which is essential for reproducible FIA-MS.

Measurement of (ir)reproducibility

Quantification of reproducibility is an active area of research[32–34]. Simple, generic metrics exist and are well suited to benchmark -omics data. The coefficient of variation (CV), also known as relative standard deviation (RSD), is defined as the standard deviation divided by the mean of a sample. The CV provides a measure of variability of a sample in relation to its average value. Thus, lower CV values indicate better reproducibility with zero being equivalent to identical experimental outcomes. In an -omics data, the CV can be calculated for each detectable feature individually, and then aggregated by the mean.

Albeit simple, the CV is a potent indicator of reproducibility in -omics experiments, and can be used to measure the improvement of standardization or normalization procedures. For instance, several recent works used it in evaluation of reproducibility in lipidomics[35–38]. Multi-laboratory assessment of reproducibility for targeted proteomics using SWATH-MS has shown inter-site CV reaching 39.4% for the normalized data and median 57.6% for the unnormalized protein quantification[39]. Another study on reproducibility of targeted metabolomics using LC and FIA tandem mass spectrometry reported inter-laboratory CV per metabolite class reaching 28% for the normalized data and median 68% for the unnormalized data[40]. Even small-scale inter-laboratory studies of untargeted metabolomics using GC-MS communicate CVs of absolute spectra ion intensities reaching 30%[41], which is likely to increase drastically as the number of samples, instruments or labs grows.

As follows directly from the above, data normalization techniques can improve reproducibility of results by significantly reducing CVs. However, good quality of the measurements is still a prerequisite that can only be guaranteed by following the policy of quality assurance (QA) and quality control (QC).

Quality assurance and quality control

Quality assurance refers to all kinds of activities to ensure high quality of the data before actual acquisition[42]. Those include routine system suitability testing (SST), following standard operation procedures, observing regular instrument maintenance and calibration, etc. Meanwhile, quality control occurs after the data has been measured to assess its veracity[43] by, e.g., correlating standard reference materials or other quality control samples.

In 2018, a Quality Assurance and Quality Control Consortium (mQACC) was established to unite academic, industry and government institutions in addressing key

QA and QC challenges in untargeted metabolomics such that multi-laboratory studies become broadly possible[44]. In 2020, mQACC published a collaborative study between 23 laboratories from North and South America, Europe and Australasia describing commonly adopted QA and QC practices[45]. According to the results, all participants employed system suitability procedures (100%), most of the laboratories used QC samples (>86%) and internal standards (91%), and manually reviewed peak integration following data acquisition (91%) to verify quality control. The authors state, however, that the 23 contributors were current members of mQACC and, perhaps, not representative of the worldwide pool of practitioners.

Despite seemingly wide dissemination of SST practices and tools[46–49], there is hardly any public database of historical system suitability data acquired systematically over a long period of time, e.g., 12 months. Such datasets are absolutely necessary to evaluate instrument performance over time and to connect its actual settings with signal and background drifts in longitudinal studies. To the date, data normalization methods vastly ignore this information, which in reality presents an opportunity to systematically obtain more accurate measurements. Beyond normalization, however, characterization of the instrument performance over time is key for MS diagnostics and automatic retuning applications.

Correction for batch effects

Batch effects are certain biases in measurements driven by various factors, such as differences in sample preparation or system setup conditions at acquisition times[50]. In practice, batch effects lead to numerical discrepancies even between supposedly identical samples, making direct comparison of experimental conditions dubious. Therefore, raw measurements need a correction for batch effects before any downstream analysis[51].

The research on the optimal batch correction approach has been active for two decades, at least. It originated from genomics and the analysis of microarray data[52] but ultimately reached all the other omics fields, including untargeted metabolomics. A variety of methods available (TIC, log, median, quantile normalization[53], PQN[54], WaveICA[55], NormAE[56], to name just a few) suggests that no single solution fulfills the needs of the research community, on one hand. On the other hand, the evaluation criteria of batch correction procedures are debated[50,51,57]. Therefore, it remains unclear where the development of data normalization methods is headed, although every new approach is consistently shown superior to the predecessors. Nevertheless, two recent trends are observable, both associated with the growth of affordable computing power.

First, the tools optimizing the entire data processing workflows are newly developed, maintained and constantly updated[58–61]. Several processing steps are jointly optimized to achieve, e.g., the largest number of peaks detected or the highest annotation confidence in the untargeted LC-MS analysis. In terms of normalization, online and offline platforms evaluate tens and hundreds of batch correction approaches for a given dataset and compare them according to multiple criteria[62,63], thus, indirectly optimizing a joint loss function. A subtle risk of employing such “bulk” approaches for a single dataset is, of course, overfitting. Moreover, the more options (or parameters) available in the processing workflow are efficiently optimized, the higher the risk.

Second, deep learning approaches for multi-batch data normalization now appear more often. Several methods have been published in the last years for transcriptomics[64–66] and metabolomics[56,67] studies. The power of deep representation learning allows to surpass older statistical methods according to the same evaluation criteria, just like it happened in computer vision, text comprehension, etc. Despite significant efforts in developing theory of deep learning[68], many aspects of it are barely understood even within core AI community. Therefore, applied deep learning solutions must be treated carefully. In the context of data normalization, it is crucial to go beyond conventional evaluation criteria (e.g., reduced variation and reproducibility of QC samples) and verify consistency of the output on many levels. To support this, we demonstrate collapsed normalization solutions obtained with recently proposed deep learning methods later in this thesis.

Multiple evaluation criteria and at least two datasets constitute the minimum requirements to demonstrate the advantages of newly developed methods. A number of benchmarking datasets are publicly available for that purpose and have been used for systematic comparison of normalization approaches[69–71]. However, most of datasets are small (often below a hundred samples) and hardly include standard reference materials, internal standards, complex biological samples, spike-ins and dilution series all at once. As a result, normalization methods that excel on the benchmarking datasets fail on real data[57], which leads to development of custom or conceptually new batch correction procedures. To break this pattern and enable in-depth evaluation of normalization methods and their performance criteria, more benchmarking datasets of high quality is needed.

Promises of artificial intelligence

In the last decade, artificial intelligence (AI) has been growing rapidly; it penetrated and significantly advanced many fields of scientific research. In life sciences, for instance, the long-standing challenge of 3D protein structure prediction has recently been solved with AI using only amino acid sequence and limited meta information as inputs[72]. Beside successful applications of AI attracting a lot of public attention, there are numerous examples of domain-specific problems benefiting from AI. In untargeted metabolomics, for instance, deep learning frameworks have been proposed to improve peak integration[73], predict retention times[74], remove batch effects[67] and generate structures *de novo* using mass spectra[75]. Altogether, such developments enhance the capabilities of analytical methods and provide deeper insights into the data.

For system suitability testing and diagnostics, anomaly detection is perhaps the most relevant task. Anomaly (or outlier) detection refers to identifying data points that substantially deviate from the others in a given dataset. Initially, it has been viewed as a way to remove noisy data points[76]. With time and larger datasets available, it became clear that anomalies can be associated with rare events[77] (e.g., sudden drops in the instrument performance) and, hence, their analysis is prolific. Unsupervised machine learning methods are predominantly used to detect anomalies, since they do not require fully labelled datasets. Among those, Isolation Forest[78] (iForest) proved its superior efficiency, scalability and acceptable memory usage when applied to a variety of datasets[79]. Yet, it has not been frequently employed by the SST solutions.

Latest batch correction approaches heavily exploited multi-task representation learning[56,65–67], wherein two or more deep neural networks are trained simultaneously to solve multiple coherent tasks expressed as individual terms in the joint loss function. This approach is designed to learn such data representations that reflect the most important properties of the data (e.g., similarity of biological materials in spite of batch-related biases). In the context of batch correction, the learned representations are used to reconstruct the data without batch effects. Therefore, the main challenge is to set up multi-task representation learning in the way to remove batch effects while keeping the biological information intact. The exact mathematical formulation, however, remains an open research question.

One of the most exciting trends in biomedical AI nowadays is multi-modal learning[80]. It refers to applications of deep learning to a single research question using multiple data modalities. The reason it is particularly useful in biomedical applications is the complexity of biological systems that can often be described on multiple levels[81,82] (e.g., multi-omics data characterizing survival subtypes of cancer patients[83]). Therefore, combination of data modalities (also known as data fusion) augments the information flow and allows to explore interactions between different parts of the system[82]. The prospective of using deep learning to analyze the fusion of untargeted metabolomics and microscopy imaging (highly relevant for, e.g., drug screens) poses a question of efficient representation learning approaches across these modalities.

Another research direction critical for any healthcare application is explainable artificial intelligence[84]. The aspiration to deploy powerful decision-making systems based on deep learning is currently held back by the need to interpret their predictions, understand the underlying logic and make sure the behavior of AI system is robust and trustworthy[85,86].

A myriad of tools proposed to interpret the results of deep learning resulted in a taxonomy of methods listing local and global, intrinsic and post-hoc, model specific and model agnostic methods[87]. Local interpretability refers to the ability of the model to explain every single prediction during inference time, which is desirable for any clinical application. Therefore, research on deep learning models featuring intrinsic local interpretability is extremely topical.

Conclusion

In the above, we discussed the reproducibility issue of untargeted metabolomics using flow injection analysis with time-of-flight mass spectrometry. Several key components should synergize to tackle this problem: systematic quality assurance and control, data normalization strategy and instrument calibration. We briefly reviewed the existing tools and approaches with their limitations for each of the components, such as: i) lack of public system suitability datasets and limited awareness of the instrument state and how it affects the measurements; ii) trivial multi-batch benchmarking datasets, ambiguity of batch correction evaluation metrics and collapsing deep learning solutions; iii) non-linearity of calibration curves in LC-MS caused by ion suppression and only few attempts to overcome it. Acknowledging the current trends in AI, we also discussed opportunities of integrating metabolomics with other data modalities to enable multi-modal explainable deep learning applications for the benefit of fundamental and clinical biomedical research.

Aim of this thesis

In this PhD thesis, we aim at improving reproducibility of untargeted metabolomics leveraging the most recent developments in AI. Additionally, we investigate representation learning approaches across data modalities and develop an explainable deep learning application to demonstrate exciting opportunities for multi-modal biomedical research in the future.

Thesis outline

In **Chapter 2**, we develop an SST platform to report on a mass spectrometer state. For that, we design a QC sample that produces 37 robust ion peaks, including isotopes, fragments and adducts. We develop an acquisition method to measure the QC sample repeatedly and collect detailed spectral information related to the chemical background, QC mix and detector noise scans of ion chronogram. After the measurement, the raw profile data is processed by the software to extract 2850 numerical features reflecting different aspects of the system, e.g., resolution, mass accuracy, ionization efficiency, levels of dirt and detector noise, etc. Those are further used to engineer the 16 quality indicators visualized for users through a web-service. We analyze in depth the QC data systematically acquired within 2 years, investigate the relationships between the QC features and the instrument settings and discuss the potentials of automatic retuning and MS diagnostics applications.

In **Chapter 3**, we develop RALPS (regularized adversarial learning preserving similarity) to correct for batch effects in untargeted metabolomics data. We propose a loss function that consists of three terms: the one penalizing batch separation to mitigate batch effects, another one promoting tight clustering of replicates to retain the biological information, and the last one penalizing sample-wise variance increase to stabilize the training process. We test RALPS extensively on several multi-batch datasets and compare to state-of-the-art methods. In addition, we run series of ablation experiments to demonstrate flexibility, scalability and robustness of our method. One of the datasets was generated specifically for the purpose of benchmarking data normalization methods. It has a total of 2750 samples acquired in 7 batches over a course of 2 months. The samples comprise human serum and ^{13}C-labeled E. coli extracts with three classes of different spike-ins and dilution series. The benchmarking dataset is largely affected by batch effects and presents a major challenge for a number of recently proposed normalization approaches.

In **Chapter 4**, we use the benchmarking dataset of flow injection time-of-flight mass spectrometry for analysis of calibration curves. First, we investigate the possibility to predict relative concentrations of amino acids and nucleobases in human serum extracts. We test multiple machine learning models and assess their ability to generalize across compound classes and concentrations. Then, we evaluate the effect of batch correction methods on prediction of relative concentrations. Finally, we discuss prospective research directions to improve prediction of calibration curves for FIA-TOF-MS data.

The final two chapters are not related to metabolomics, but are projects that I worked on in collaboration with a colleague who had to analyze the metabolism of cancer cell lines under drug pressure. My role was to develop AI frameworks for the analysis of time-resolved images of tissue cultures.

In **Chapter 5**, we investigate and compare representation learning approaches using cancer cell imaging data. We implement four different models having the same CNN backbone for downstream feature extraction. We train them under identical conditions testing four strategies of random augmenting and cropping for each model. To compare the learned representations, we formulate three independent tasks and evaluate multiple metrics for each. Among other empirical results, we confirm the efficiency of multi-task representation learning approaches across data modalities.

In **Chapter 6**, we develop an explainable deep learning model to classify drugs based on cells images of different cancer types. The model features intrinsic local interpretability, i.e., it allows to visualize regions of an image driving the classifier decision. We present many examples shedding light on drug-specific morphological features of cells and discuss the potential to extend this approach to a multi-modal setting, e.g., to explain image classification with the corresponding metabolomics data.

References

(1) Kesić, S. Systems Biology, Emergence and Antireductionism. *Saudi J. Biol. Sci.* **2016**, *23* (5), 584–591. https://doi.org/10.1016/j.sjbs.2015.06.015.

(2) Cao, Z. J.; Gao, G. Multi-Omics Single-Cell Data Integration and Regulatory Inference with Graph-Linked Embedding. *Nat. Biotechnol.* **2022**, 23–25. https://doi.org/10.1038/s41587-022-01284-4.

(3) Mike May. Big Data, Big Picture: Metabolomics Meets Systems Biology. *Sci. Cust. Publ. Off.* **2017**.

(4) Zhu, C.; Preissl, S.; Ren, B. Single-Cell Multimodal Omics: The Power of Many. *Nat. Methods* **2020**, *17* (1), 11–14. https://doi.org/10.1038/s41592-019-0691-5.

(5) Hu, Q.; Noll, R. J.; Li, H.; Makarov, A.; Hardman, M.; Cooks, R. G. The Orbitrap: A New Mass Spectrometer. *J. Mass Spectrom.* **2005**, *40* (4), 430–443. https://doi.org/10.1002/jms.856.

(6) Singhal, N.; Kumar, M.; Kanaujia, P. K.; Virdi, J. S. MALDI-TOF Mass Spectrometry: An Emerging Technology for Microbial Identification and Diagnosis. *Front. Microbiol.* **2015**, *6* (AUG), 1–16. https://doi.org/10.3389/fmicb.2015.00791.

(7) Zamboni, N. Why do we prefer TOFs over Orbitraps for flow injection analysis? https://metabolomics.blog/2021/06/why-do-we-prefer-tofs-over-orbitraps-for-flow-injection-analysis/.

(8) Zubarev, R. A.; Makarov, A. Orbitrap Mass Spectrometry. *Anal. Chem.* **2013**, *85* (11), 5288–5296. https://doi.org/10.1021/ac4001223.

(9) Dührkop, K.; Fleischauer, M.; Ludwig, M.; Aksenov, A. A.; Melnik, A. V.; Meusel, M.; Dorrestein, P. C.; Rousu, J.; Böcker, S. SIRIUS 4: A Rapid Tool for Turning Tandem Mass Spectra into Metabolite Structure Information. *Nat. Methods* **2019**, *16* (4), 299–302. https://doi.org/10.1038/s41592-019-0344-8.

(10) T'Kindt, R.; Van Bocxlaer, J. LC-MS Based Metabolomics. *Handb. Mass Spectrom. Instrumentation, Data Anal. Appl.* **2010**, *8* (2), 39–73. https://doi.org/10.1007/978-1-4419-9863-7_1155.

(11) Perez de Souza, L.; Alseekh, S.; Scossa, F.; Fernie, A. R. Ultra-High-Performance Liquid Chromatography High-Resolution Mass Spectrometry Variants for Metabolomics Research. *Nat. Methods* **2021**, *18* (7), 733–746. https://doi.org/10.1038/s41592-021-01116-4.

(12) Schrimpe-Rutledge, Alexandra C. ; Codreanu, Simona G. ; Sherrod, Stacy D. ; McLean, J. A. . Untargeted Metabolomics Strategies – Challenges and Emerging Directions. *J Am Soc Mass Spectrom.* **2016**, *27* (12), 1897–1905. https://doi.org/10.1007/s13361-016-1469-y.

(13) Alonso, A.; Marsal, S.; Julià, A. Analytical Methods in Untargeted Metabolomics: State of the Art in 2015. *Front. Bioeng. Biotechnol.* **2015**, *3* (MAR), 1–20. https://doi.org/10.3389/fbioe.2015.00023.

(14) Tan, S. Z.; Begley, P.; Mullard, G.; Hollywood, K. A.; Bishop, P. N. Introduction to Metabolomics and Its Applications in Ophthalmology. *Eye* **2016**, *30* (6), 773–783. https://doi.org/10.1038/eye.2016.37.

(15) Wishart, D. S. Emerging Applications of Metabolomics in Drug Discovery and Precision Medicine. *Nat. Rev. Drug Discov.* **2016**, *15* (7), 473–484. https://doi.org/10.1038/nrd.2016.32.

(16) Koen, N.; Du Preez, I.; Loots, D. T. *Metabolomics and Personalized Medicine*, 1st ed.; Elsevier Inc., 2016; Vol. 102.

https://doi.org/10.1016/bs.apcsb.2015.09.003.

(17) Jacob, M.; Lopata, A. L.; Dasouki, M.; Abdel Rahman, A. M. Metabolomics toward Personalized Medicine. *Mass Spectrom. Rev.* **2019**, *38* (3), 221–238. https://doi.org/10.1002/mas.21548.

(18) Trivedi, D. K.; Goodacre, R. *The Role of Metabolomics in Personalized Medicine*; Elsevier Inc., 2020. https://doi.org/10.1016/b978-0-12-812784-1.00011-6.

(19) Castelli, F. A.; Rosati, G.; Moguet, C.; Fuentes, C.; Marrugo-Ramírez, J.; Lefebvre, T.; Volland, H.; Merkoçi, A.; Simon, S.; Fenaille, F.; Junot, C. *Metabolomics for Personalized Medicine: The Input of Analytical Chemistry from Biomarker Discovery to Point-of-Care Tests*; Analytical and Bioanalytical Chemistry, 2022; Vol. 414. https://doi.org/10.1007/s00216-021-03586-z.

(20) Fanelli, D. Is Science Really Facing a Reproducibility Crisis, and Do We Need It To? *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115* (11), 2628–2631. https://doi.org/10.1073/pnas.1708272114.

(21) Wilson, C. The Replication Crisis Has Spread through Science – Can It Be Fixed? *New Sci.* **2022**.

(22) Baker, M. 1,500 Scientists Lift the Lid on Reproducibility. *Nature* **2016**, *533*, 452–454. https://doi.org/10.1038/533452a.

(23) Baker, M.; Penny, D. Is There a Reproducibility Crisis? *Nature* **2016**, *533* (7604), 452–454. https://doi.org/10.1038/533452A.

(24) Dolan, J. W. Internal Standard Calibration Problems. *LCGC North Am.* **2015**, *33* (6), 396–400.

(25) Dolan, J. W. When Should an Internal Standard Be Used? *LCGC North Am.* **2012**, *30* (6), 474–480.

(26) Zabell, A. P. R.; Lytle, F. E.; Julian, R. K. A Proposal to Improve Calibration and Outlier Detection in High-Throughput Mass Spectrometry. *Clin. Mass Spectrom.* **2016**, *2* (2016), 25–33. https://doi.org/10.1016/j.clinms.2016.12.003.

(27) Moosavi, Seyed Mojtaba; Ghassabian, S. Linearity of Calibration Curves for Analytical Methods: A Review of Criteria for Assessment of Method Reliability. In *Calibration and Validation pf Analytical Methods*; 2018. https://doi.org/10.5772/intechopen.72932.

(28) Piovan, Anna; Filippini, Raffaella; Caniato, R. A Semi-Quantitative FIA-ESI-MS Method for the Rapid Screening of Hypericum Perforatum Crude Extracts. *Nat. Prod. Commun.* **2010**, *5* (3), 431–434.

(29) Martin, J.; Gracia, A. R.; Asuero, A. G. Fitting Nonlinear Calibration Curves: No Models Perfect. *J. Anal. Sci. Methods Instrum.* **2017**, *07* (01), 1–17. https://doi.org/10.4236/jasmi.2017.71001.

(30) Chen, G. yuan; Liao, H. wei; Tseng, Y. J.; Tsai, I. lin; Kuo, C. hua. A Matrix-Induced Ion Suppression Method to Normalize Concentration in Urinary Metabolomics Studies Using Flow Injection Analysis Electrospray Ionization Mass Spectrometry. *Anal. Chim. Acta* **2014**, *864* (33), 21–29. https://doi.org/10.1016/j.aca.2015.01.022.

(31) Reiter, A.; Herbst, L.; Wiechert, W.; Oldiges, M. Need for Speed: Evaluation of Dilute and Shoot-Mass Spectrometry for Accelerated Metabolic Phenotyping in Bioprocess Development. *Anal. Bioanal. Chem.* **2021**, *413* (12), 3253–3268. https://doi.org/10.1007/s00216-021-03261-3.

(32) Zhvansky, E. S.; Pekov, S. I.; Sorokin, A. A.; Shurkhay, V. A.; Eliferov, V. A.; Potapov, A. A.; Nikolaev, E. N.; Popov, I. A. Metrics for Evaluating the Stability and Reproducibility of Mass Spectra. *Sci. Rep.* **2019**, *9* (1), 1–8.

https://doi.org/10.1038/s41598-018-37560-0.

(33) Groff, L. C.; Grossman, J. N.; Kruve, A.; Minucci, J. M.; Lowe, C. N.; McCord, J. P.; Kapraun, D. F.; Phillips, K. A.; Purucker, S. T.; Chao, A.; Ring, C. L.; Williams, A. J.; Sobus, J. R. Uncertainty Estimation Strategies for Quantitative Non-Targeted Analysis. *Anal. Bioanal. Chem.* **2022**, 4919–4933. https://doi.org/10.1007/s00216-022-04118-z.

(34) Ghosh, T.; Philtron, D.; Zhang, W.; Kechris, K.; Ghosh, D. Reproducibility of Mass Spectrometry Based Metabolomics Data. *BMC Bioinformatics* **2021**, *22* (1), 1–25. https://doi.org/10.1186/s12859-021-04336-9.

(35) Thompson, J. W.; Adams, K. J.; Adamski, J.; Asad, Y.; Borts, D.; Bowden, J. A.; Byram, G.; Dang, V.; Dunn, W. B.; Fernandez, F.; Fiehn, O.; Gaul, D. A.; Hühmer, A. F. R.; Kalli, A.; Koal, T.; Koeniger, S.; Mandal, R.; Meier, F.; Naser, F. J.; O'Neil, D.; Pal, A.; Patti, G. J.; Pham-Tuan, H.; Prehn, C.; Raynaud, F. I.; Shen, T.; Southam, A. D.; St. John-Williams, L.; Sulek, K.; Vasilopoulou, C. G.; Viant, M.; Winder, C. L.; Wishart, D.; Zhang, L.; Zheng, J.; Moseley, M. A. International Ring Trial of a High Resolution Targeted Metabolomics and Lipidomics Platform for Serum and Plasma Analysis. *Anal. Chem.* **2019**. https://doi.org/10.1021/acs.analchem.9b02908.

(36) Bowden, J. A.; Heckert, A.; Ulmer, C. Z.; Jones, C. M.; Koelmel, J. P.; Abdullah, L.; Ahonen, L.; Alnouti, Y.; Armando, A. M.; Asara, J. M.; Bamba, T.; Barr, J. R.; Bergquist, J.; Borchers, C. H.; Brandsma, J.; Breitkopf, S. B.; Cajka, T.; Cazenave-Gassiot, A.; Checa, A.; Cinel, M. A.; Colas, R. A.; Cremers, S.; Dennis, E. A.; Evans, J. E.; Fauland, A.; Fiehn, O.; Gardner, M. S.; Garrett, T. J.; Gotlinger, K. H.; Han, J.; Huang, Y.; Neo, A. H.; Hyötyläinen, T.; Izumi, Y.; Jiang, H.; Jiang, H.; Jiang, J.; Kachman, M.; Kiyonami, R.; Klavins, K.; Klose, C.; Köfeler, H. C.; Kolmert, J.; Koal, T.; Koster, G.; Kuklenyik, Z.; Kurland, I. J.; Leadley, M.; Lin, K.; Maddipati, K. R.; McDougall, D.; Meikle, P. J.; Mellett, N. A.; Monnin, C.; Moseley, M. A.; Nandakumar, R.; Oresic, M.; Patterson, R.; Peake, D.; Pierce, J. S.; Post, M.; Postle, A. D.; Pugh, R.; Qiu, Y.; Quehenberger, O.; Ramrup, P.; Rees, J.; Rembiesa, B.; Reynaud, D.; Roth, M. R.; Sales, S.; Schuhmann, K.; Schwartzman, M. L.; Serhan, C. N.; Shevchenko, A.; Somerville, S. E.; St John-Williams, L.; Surma, M. A.; Takeda, H.; Thakare, R.; Thompson, J. W.; Torta, F.; Triebl, A.; Trötzmüller, M.; Ubhayasekera, S. J. K.; Vuckovic, D.; Weir, J. M.; Welti, R.; Wenk, M. R.; Wheelock, C. E.; Yao, L.; Yuan, M.; Zhao, X. H.; Zhou, S. Harmonizing Lipidomics: NIST Interlaboratory Comparison Exercise for Lipidomics Using SRM 1950-Metabolites in Frozen Human Plasma. *J. Lipid Res.* **2017**, *58* (12), 2275–2288. https://doi.org/10.1194/jlr.M079012.

(37) Ghorasaini, M.; Mohammed, Y.; Adamski, J.; Bettcher, L.; Bowden, J. A.; Cabruja, M.; Contrepois, K.; Ellenberger, M.; Gajera, B.; Haid, M.; Hornburg, D.; Hunter, C.; Jones, C. M.; Klein, T.; Mayboroda, O.; Mirzaian, M.; Moaddel, R.; Ferrucci, L.; Lovett, J.; Nazir, K.; Pearson, M.; Ubhi, B. K.; Raftery, D.; Riols, F.; Sayers, R.; Sijbrands, E. J. G.; Snyder, M. P.; Su, B.; Velagapudi, V.; Williams, K. J.; De Rijke, Y. B.; Giera, M. Cross-Laboratory Standardization of Preclinical Lipidomics Using Differential Mobility Spectrometry and Multiple Reaction Monitoring. *Anal. Chem.* **2021**, *93* (49), 16369–16378. https://doi.org/10.1021/acs.analchem.1c02826.

(38) Loef, M.; Von Hegedus, J. H.; Ghorasaini, M.; Kroon, F. P. B.; Giera, M.; Ioan-Facsinay, A.; Kloppenburg, M. Reproducibility of Targeted Lipidome Analyses (Lipidyzer) in Plasma and Erythrocytes over a 6-Week Period. *Metabolites* **2021**,

11 (1), 1–11. https://doi.org/10.3390/metabo11010026.

(39) Collins, B. C.; Hunter, C. L.; Liu, Y.; Schilling, B.; Rosenberger, G.; Bader, S. L.; Chan, D. W.; Gibson, B. W.; Gingras, A. C.; Held, J. M.; Hirayama-Kurogi, M.; Hou, G.; Krisp, C.; Larsen, B.; Lin, L.; Liu, S.; Molloy, M. P.; Moritz, R. L.; Ohtsuki, S.; Schlapbach, R.; Selevsek, N.; Thomas, S. N.; Tzeng, S. C.; Zhang, H.; Aebersold, R. Multi-Laboratory Assessment of Reproducibility, Qualitative and Quantitative Performance of SWATH-Mass Spectrometry. *Nat. Commun.* **2017**, *8* (1), 1–11. https://doi.org/10.1038/s41467-017-00249-5.

(40) Siskos, A. P.; Jain, P.; Römisch-Margl, W.; Bennett, M.; Achaintre, D.; Asad, Y.; Marney, L.; Richardson, L.; Koulman, A.; Griffin, J. L.; Raynaud, F.; Scalbert, A.; Adamski, J.; Prehn, C.; Keun, H. C. Interlaboratory Reproducibility of a Targeted Metabolomics Platform for Analysis of Human Serum and Plasma. *Anal. Chem.* **2017**, *89* (1), 656–665. https://doi.org/10.1021/acs.analchem.6b02930.

(41) Lin, Y.; Caldwell, G. W.; Li, Y.; Lang, W.; Masucci, J. Inter-Laboratory Reproducibility of an Untargeted Metabolomics GC–MS Assay for Analysis of Human Plasma. *Sci. Rep.* **2020**, *10* (1), 1–11. https://doi.org/10.1038/s41598-020-67939-x.

(42) Broadhurst, D.; Goodacre, R.; Reinke, S. N.; Kuligowski, J.; Wilson, I. D.; Lewis, M. R.; Dunn, W. B. Guidelines and Considerations for the Use of System Suitability and Quality Control Samples in Mass Spectrometry Assays Applied in Untargeted Clinical Metabolomic Studies. *Metabolomics* **2018**, *14* (6), 1–17. https://doi.org/10.1007/s11306-018-1367-3.

(43) Schulze, B.; Jeon, Y.; Kaserzon, S.; Heffernan, A. L.; Dewapriya, P.; O'Brien, J.; Gomez Ramos, M. J.; Ghorbani Gorji, S.; Mueller, J. F.; Thomas, K. V.; Samanipour, S. An Assessment of Quality Assurance/Quality Control Efforts in High Resolution Mass Spectrometry Non-Target Workflows for Analysis of Environmental Samples. *TrAC - Trends Anal. Chem.* **2020**, *133*, 116063. https://doi.org/10.1016/j.trac.2020.116063.

(44) Beger, R. D.; Dunn, W. B.; Bandukwala, A.; Bethan, B.; Broadhurst, D.; Clish, C. B.; Dasari, S.; Derr, L.; Evans, A.; Fischer, S.; Flynn, T.; Hartung, T.; Herrington, D.; Higashi, R.; Hsu, P. C.; Jones, C.; Kachman, M.; Karuso, H.; Kruppa, G.; Lippa, K.; Maruvada, P.; Mosley, J.; Ntai, I.; O'Donovan, C.; Playdon, M.; Raftery, D.; Shaughnessy, D.; Souza, A.; Spaeder, T.; Spalholz, B.; Tayyari, F.; Ubhi, B.; Verma, M.; Walk, T.; Wilson, I.; Witkin, K.; Bearden, D. W.; Zanetti, K. A. Towards Quality Assurance and Quality Control in Untargeted Metabolomics Studies. *Metabolomics* **2019**, *15* (1), 1–5. https://doi.org/10.1007/s11306-018-1460-7.

(45) Evans, A. M. . et al. Dissemination and Analysis of the Quality Assurance (QA) and Quality Control (QC) Practices of LC-MS Based Untargeted Metabolomics Practitioners. *Metabolomics* **2021**, *16* (10), 113. https://doi.org/10.1007/s11306-020-01728-5.

(46) Chiva, C.; Olivella, R.; Borràs, E.; Espadas, G.; Pastor, O.; Solé, A.; Sabidó, E. QCloud: A Cloud-Based Quality Control System for Mass Spectrometry-Based Proteomics Laboratories. *PLoS One* **2018**, *13* (1), 1–14. https://doi.org/10.1371/journal.pone.0189209.

(47) Stanfill, B. A.; Nakayasu, E. S.; Bramer, L. M.; Thompson, A. M.; Ansong, C. K.; Clauss, T. R.; Gritsenko, M. A.; Monroe, M. E.; Moore, R. J.; Orton, D. J.; Piehowski, P. D.; Schepmoes, A. A.; Smith, R. D.; Webb-Robertson, B. J. M.; Metz, T. O. Quality Control Analysis in Real-Time (QC-ART): A Tool for Real-Time Quality Control Assessment of Mass Spectrometry-Based Proteomics

Data. *Mol. Cell. Proteomics* **2018**, *17* (9), 1824–1836. https://doi.org/10.1074/mcp.RA118.000648.

(48) Morgenstern, D.; Barzilay, R.; Levin, Y. RawBeans: A Simple, Vendor-Independent, Raw-Data Quality-Control Tool. *J. Proteome Res.* **2021**, *20* (4), 2098–2104. https://doi.org/10.1021/acs.jproteome.0c00956.

(49) Stratton, K. G.; Webb-Robertson, B. J. M.; McCue, L. A.; Stanfill, B.; Claborne, D.; Godinez, I.; Johansen, T.; Thompson, A. M.; Burnum-Johnson, K. E.; Waters, K. M.; Bramer, L. M. PmartR: Quality Control and Statistics for Mass Spectrometry-Based Biological Data. *J. Proteome Res.* **2019**, *18* (3), 1418–1425. https://doi.org/10.1021/acs.jproteome.8b00760.

(50) Han, W.; Li, L. Evaluating and Minimizing Batch Effects in Metabolomics. *Mass Spectrom. Rev.* **2022**, *41* (3), 421–442. https://doi.org/10.1002/mas.21672.

(51) De Livera, A. M.; Dias, D. A.; De Souza, D.; Rupasinghe, T.; Pyke, J.; Tull, D.; Roessner, U.; McConville, M.; Speed, T. P. Normalizing and Integrating Metabolomics Data. *Anal. Chem.* **2012**, *84* (24), 10768–10776. https://doi.org/10.1021/ac302748b.

(52) Johnson, W. E.; Li, C.; Rabinovic, A. Adjusting Batch Effects in Microarray Expression Data Using Empirical Bayes Methods. *Biostatistics* **2007**, *8* (1), 118–127. https://doi.org/10.1093/biostatistics/kxj037.

(53) Wulff, J. E.; Mitchell, M. W. A Comparison of Various Normalization Methods for LC/MS Metabolomics Data. *Adv. Biosci. Biotechnol.* **2018**, *09* (08), 339–351. https://doi.org/10.4236/abb.2018.98022.

(54) Dieterle, F.; Ross, A.; Schlotterbeck, G.; Senn, H. Probabilistic Quotient Normalization as Robust Method to Account for Dilution of Complex Biological Mixtures. Application In1H NMR Metabonomics. *Anal. Chem.* **2006**, *78* (13), 4281–4290. https://doi.org/10.1021/ac051632c.

(55) Deng, K.; Zhao, F.; Rong, Z.; Cao, L.; Zhang, L.; Li, K.; Hou, Y.; Zhu, Z. J. WaveICA 2.0: A Novel Batch Effect Removal Method for Untargeted Metabolomics Data without Using Batch Information. *Metabolomics* **2021**, *17* (10), 1–8. https://doi.org/10.1007/s11306-021-01839-7.

(56) Rong, Z.; Tan, Q.; Cao, L.; Zhang, L.; Deng, K.; Huang, Y.; Zhu, Z. J.; Li, Z.; Li, K. NormAE: Deep Adversarial Learning Model to Remove Batch Effects in Liquid Chromatography Mass Spectrometry-Based Metabolomics Data. *Anal. Chem.* **2020**, *92* (7), 5082–5090. https://doi.org/10.1021/acs.analchem.9b05460.

(57) Yang, Q.; Wang, Y.; Zhang, Y.; Li, F.; Xia, W.; Zhou, Y.; Qiu, Y.; Li, H.; Zhu, F. NOREVA: Enhanced Normalization and Evaluation of Time-Course and Multi-Class Metabolomic Data. *Nucleic Acids Res.* **2021**, *48* (1), W436–W448. https://doi.org/10.1093/NAR/GKAA258.

(58) Pang, Z.; Zhou, G.; Ewald, J.; Chang, L.; Hacariz, O.; Basu, N.; Xia, J. Using MetaboAnalyst 5.0 for LC–HRMS Spectra Processing, Multi-Omics Integration and Covariate Adjustment of Global Metabolomics Data. *Nat. Protoc.* **2022**. https://doi.org/10.1038/s41596-022-00710-w.

(59) Delabriere, A.; Warmer, P.; Brennsteiner, V.; Zamboni, N. SLAW: A Scalable and Self-Optimizing Processing Workflow for Untargeted LC-MS. *Anal. Chem.* **2021**, *93* (45), 15024–15032. https://doi.org/10.1021/acs.analchem.1c02687.

(60) Gowda, H.; Ivanisevic, J.; Johnson, C. H.; Kurczy, M. E.; Benton, H. P.; Rinehart, D.; Nguyen, T.; Ray, J.; Kuehl, J.; Arevalo, B.; Westenskow, P. D.; Wang, J.; Arkin, A. P.; Deutschbauer, A. M.; Patti, G. J.; Siuzdak, G. Interactive XCMS Online: Simplifying Advanced Metabolomic Data Processing and Subsequent Statistical Analyses. *Anal. Chem.* **2014**, *86* (14), 6931–6939.

https://doi.org/10.1021/ac500734c.

(61) Franceschi, P.; Mylonas, R.; Shahaf, N.; Scholz, M.; Arapitsas, P.; Masuero, D.; Weingart, G.; Carlin, S.; Vrhovsek, U.; Mattivi, F.; Wehrens, R. MetaDB a Data Processing Workflow in Untargeted MS-Based Metabolomics Experiments. *Front. Bioeng. Biotechnol.* **2014**, *2* (DEC), 1–12. https://doi.org/10.3389/fbioe.2014.00072.

(62) Fu, J.; Zhang, Y.; Wang, Y.; Zhang, H.; Liu, J.; Tang, J.; Yang, Q.; Sun, H.; Qiu, W.; Ma, Y.; Li, Z.; Zheng, M.; Zhu, F. Optimization of Metabolomic Data Processing Using NOREVA. *Nat. Protoc.* **2022**, *17* (1), 129–151. https://doi.org/10.1038/s41596-021-00636-9.

(63) Chawade, A.; Alexandersson, E.; Levander, F. Normalyzer: A Tool for Rapid Evaluation of Normalization Methods for Omics Data Sets. *J. Proteome Res.* **2014**, *13* (6), 3114–3120. https://doi.org/10.1021/pr401264n.

(64) Wang, T.; Johnson, T. S.; Shao, W.; Lu, Z.; Helm, B. R.; Zhang, J.; Huang, K. BERMUDA: A Novel Deep Transfer Learning Method for Single-Cell RNA Sequencing Batch Correction Reveals Hidden High-Resolution Cellular Subtypes. *Genome Biol.* **2019**, *20* (1), 1–15. https://doi.org/10.1186/s13059-019-1764-6.

(65) Lakkis, J.; Wang, D.; Zhang, Y.; Hu, G.; Wang, K.; Pan, H.; Ungar, L.; Reilly, M.; Li, X.; Li, M. A Joint Deep Learning Model Enables Simultaneous Batch Effect Correction, Denoising and Clustering in Single-Cell Transcriptomics. *Genome Res.* **2021**, gr.271874.120. https://doi.org/10.1101/gr.271874.120.

(66) Li, X.; Wang, K.; Lyu, Y.; Pan, H.; Zhang, J.; Stambolian, D.; Susztak, K.; Reilly, M. P.; Hu, G.; Li, M. Deep Learning Enables Accurate Clustering with Batch Effect Removal in Single-Cell RNA-Seq Analysis. *Nat. Commun.* **2020**, *11* (1), 1–14. https://doi.org/10.1038/s41467-020-15851-3.

(67) Niu, J.; Xu, W.; Wei, D.; Qian, K.; Wang, Q. Deep Learning Framework for Integrating Multibatch Calibration, Classification, and Pathway Activities. *Anal. Chem.* **2022**. https://doi.org/10.1021/acs.analchem.2c00601.

(68) Roberts, Daniel A. ; Yaida, Sho; Hanin, B. *The Principles of Deep Learning Theory*; Cambridge University Press, 2022.

(69) Franceschi, P.; Masuero, D.; Vrhovsek, U.; Mattivi, F.; Wehrens, R. A Benchmark Spike-in Data Set for Biomarker Identification in Metabolomics. *J. Chemom.* **2012**, *26* (1), 16–24. https://doi.org/10.1002/cem.1420.

(70) Kirwan, J. A.; Weber, R. J. M.; Broadhurst, D. I.; Viant, M. R. Direct Infusion Mass Spectrometry Metabolomics Dataset: A Benchmark for Data Processing and Quality Control. *Sci. Data* **2014**, *1*, 1–13. https://doi.org/10.1038/sdata.2014.12.

(71) Li, B.; Tang, J.; Yang, Q.; Li, S.; Cui, X.; Li, Y.; Chen, Y.; Xue, W.; Li, X.; Zhu, F. NOREVA: Normalization and Evaluation of MS-Based Metabolomics Data. *Nucleic Acids Res.* **2017**, *45* (W1), W162–W170. https://doi.org/10.1093/nar/gkx449.

(72) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.

(73) Liebal, U. W.; Phan, A. N. T.; Sudhakar, M.; Raman, K.; Blank, L. M. Machine Learning Applications for Mass Spectrometry-Based Metabolomics. *Metabolites* **2020**, *10* (6), 1–23. https://doi.org/10.3390/metabo10060243.

(74) Giese, S. H.; Sinn, L. R.; Wegner, F.; Rappsilber, J. Retention Time Prediction Using Neural Networks Increases Identifications in Crosslinking Mass Spectrometry. *Nat. Commun.* **2021**, *12* (1), 1–11. https://doi.org/10.1038/s41467-021-23441-0.

(75) Stravs, M. A.; Dührkop, K.; Böcker, S.; Zamboni, N. MSNovelist: De Novo Structure Generation from Mass Spectra. *Nat. Methods* **2022**, *19* (July). https://doi.org/10.1038/s41592-022-01486-3.

(76) Goldstein, M.; Uchida, S. A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data. *PLoS One* **2016**, *11* (4), 1–31. https://doi.org/10.1371/journal.pone.0152173.

(77) Bahri, M.; Salutari, F.; Putina, A.; Sozio, M. AutoML: State of the Art with a Focus on Anomaly Detection, Challenges, and Research Directions. *Int. J. Data Sci. Anal.* **2022**, *14* (2), 113–126. https://doi.org/10.1007/s41060-022-00309-0.

(78) Liu, F. T.; Ting, K. M.; Zhou, Z. H. Isolation Forest. *Proc. - IEEE Int. Conf. Data Mining, ICDM* **2008**, 413–422. https://doi.org/10.1109/ICDM.2008.17.

(79) Domingues, R.; Filippone, M.; Michiardi, P.; Zouaoui, J. A Comparative Evaluation of Outlier Detection Algorithms: Experiments and Analyses. *Pattern Recognit.* **2018**, *74*, 406–421. https://doi.org/10.1016/j.patcog.2017.09.037.

(80) Ramachandram, D.; Taylor, G. W. Deep Multimodal Learning. *IEEE Signal Process. Mag.* **2017**, No. November, 96–108.

(81) Krassowski, M.; Das, V.; Sahu, S. K.; Misra, B. B. State of the Field in Multi-Omics Research: From Computational Needs to Data Mining and Sharing. *Front. Genet.* **2020**, *11* (December), 1–17. https://doi.org/10.3389/fgene.2020.610798.

(82) Stahlschmidt, S. R.; Ulfenborg, B.; Synnergren, J. Multimodal Deep Learning for Biomedical Data Fusion: A Review. *Brief. Bioinform.* **2022**, *23* (2), 1–15. https://doi.org/10.1093/bib/bbab569.

(83) Poirion, O. B.; Jing, Z.; Chaudhary, K.; Huang, S.; Garmire, L. X. DeepProg: An Ensemble of Deep-Learning and Machine-Learning Models for Prognosis Prediction Using Multi-Omics Data. *Genome Med.* **2021**, *13* (1), 1–15. https://doi.org/10.1186/s13073-021-00930-x.

(84) Minh, D.; Wang, H. X.; Li, Y. F.; Nguyen, T. N. *Explainable Artificial Intelligence: A Comprehensive Review*; Springer Netherlands, 2022; Vol. 55. https://doi.org/10.1007/s10462-021-10088-y.

(85) Thiebes, S.; Lins, S.; Sunyaev, A. Trustworthy Artificial Intelligence. *Electron. Mark.* **2021**, *31* (2), 447–464. https://doi.org/10.1007/s12525-020-00441-4.

(86) Zicari, R. V.; Brusseau, J.; Blomberg, S. N.; Christensen, H. C.; Coffee, M.; Ganapini, M. B.; Gerke, S.; Gilbert, T. K.; Hickman, E.; Hildt, E.; Holm, S.; Kühne, U.; Madai, V. I.; Osika, W.; Spezzatti, A.; Schnebel, E.; Tithi, J. J.; Vetter, D.; Westerlund, M.; Wurth, R.; Amann, J.; Antun, V.; Beretta, V.; Bruneault, F.; Campano, E.; Düdder, B.; Gallucci, A.; Goffi, E.; Haase, C. B.; Hagendorff, T.; Kringen, P.; Möslein, F.; Ottenheimer, D.; Ozols, M.; Palazzani, L.; Petrin, M.; Tafur, K.; Tørresen, J.; Volland, H.; Kararigas, G. On Assessing Trustworthy AI in Healthcare. Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls. *Front. Hum. Dyn.* **2021**, *3* (July), 1–24. https://doi.org/10.3389/fhumd.2021.673104.

(87) Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable Ai: A Review

of Machine Learning Interpretability Methods. *Entropy* **2021**, *23* (1), 1–45. https://doi.org/10.3390/e23010018.

Chapter 2

A system suitability testing platform for untargeted, high-resolution mass spectrometry

Andrei Dmitrenko, Michelle Reid and Nicola Zamboni

This manuscript was submitted for publication.

Author contributions:

AD, MR, NZ conceived the study. MR and NZ designed the QC mix and developed a data acquisition method. AD developed an SST platform software to process raw measurements, engineer QC features, manage QC databases, visualize and report on the instrument state over a web-service. AD performed the statistical data analysis. AD and NZ wrote the manuscript.

Abstract

The broad coverage of untargeted metabolomics poses fundamental challenges for the harmonization of measurements along time, even if they originate from the very same instrument. Internal isotopic standards can hardly cover the chemical complexity of study samples. Therefore, they are insufficient for normalizing data *a posteriori* as done for targeted metabolomics. Instead, it is crucial to verify instrument's performance *a priori*, that is, before samples are injected. Here, we propose a systems suitability testing platform for time-of-flight mass spectrometers. It includes a chemically defined quality control mixture, a fast acquisition method, software for extracting ca. 3000 numerical features from profile data, and a simple web-service for monitoring. We ran a pilot for 21 months and present illustrative results for anomaly detection or learning causal relationships between the spectral features and machine settings. Beyond mere detection of anomalies, our results highlight several future applications such as (i) recommending instrument retuning strategies to achieve desirable values of quality indicators, (ii) driving preventive maintenance, or (iii) using the obtained, detailed spectral features for posterior data harmonization.

Introduction

Reproducibility and replicability of experiments are essential mainstays of the scientific method[1]. Failure to reproduce measurements, computations, or results of a previous study is perceived as a lack of rigor and undermines the validity of study and its claims. Omics technologies are not immune to these challenges[2]. In fact, issues tend to increase with time and the steadily increasing number of features that every new technology allows to detect. This exacerbates the problems of small sample size[3] and overfitting. Mass spectrometry (MS)-based assays also suffer from the inherent variability of measurements across instruments and over time. In proteomics, lipidomics, metabolomics, etc., reproducibility of quantitative experiments is a well-known issue[4–6]. The common workaround to enable quantitation in MS-based assays is to add a known amount of isotopically labeled internal standards (IS), and quantify chemically similar compounds based on relative signals. This approach is limited by the availability of heavy standards and, therefore, is effective only in targeted studies or within compound classes.

In absence of heavy standards such as in untargeted metabolomics or label-free quantification in proteomics, it remains challenging to ensure reproducibility such that it would be possible to compare samples from different experiments. A steadily growing arsenal of normalization methods allows correcting for differences in feature intensities across different batches[7–9]. However, they only tackle one facet of the reproducibility challenge. They are, however, ineffective in the case features could not be detected or matched across batches. This problem is quite frequent as caused by a multitude of common issues: drifts in retention times, loss of sensitivity, differences in tuning, changes in the matrix, contaminations, etc. Such irreproducible behavior cannot be corrected by posterior data processing. This has led to the development of approaches for testing LC-MS instrument performance *a priori*. Samples are injected only after the test is passed. Testing relies on three pillars: standard quality control (QC) samples, a standard acquisition method, and software to discover deviations[10–12] from expectations[11,13]. The effort toward reproducible untargeted metabolomics is spearheaded by the metabolomics quality assurance and quality control consortium (mQACC[14,15]). A recent survey revealed that expert metabolomics labs employ procedures for system suitability testing[15] but the examples of literature discussing specific details or presenting solutions are very scarce[10].

Here we propose means for systematic quality control (QC) and monitoring of instrument performance and properties for high-resolution mass spectrometry, focusing on a time-of-flight (TOF) instrument. We have been using TOF-MS productively for more than 10 years and analyzed more than 1 Mio samples by untargeted metabolomics. On occasions, we had to reanalyze entire batches of samples because of major problems and biases that passed unnoticed during instrument tuning. Hence, this work originated from the need to verify system suitability before injecting precious samples, and to monitor performance drifts that might require user intervention. Our system includes a chemically defined QC sample, a short acquisition method, software for extracting detailed spectral information from profile data, and a simple visualization service for end users.

Results

We set out to implement a system suitability testing platform capable of quick, quantitative and comprehensive characterization of the state of a high-resolution ESI-TOF-MS instrument (Agilent 6550 iFunnel Q-TOF). On purpose, we omitted chromatography from testing. This decision was motivated by several practical reasons. Chromatography and detection are separated processes, and we frequently switch liquid chromatography systems depending on the type of separation needed (e.g., reversed phase, HILIC, ion-pairing). Chromatographic performance can be evaluated by means of retention time stability, height equivalent to a theoretical plate (HETP), peak tailing, etc., with samples that vary for the different methods. Here, we wanted to define a MS system suitability that is independent of a specific chromatographic setup and therefore more generally applicable. It was tailored to capture more information that relates to ionization, ion transmission and detection. We focused on negative mode ionization, which is predominant for metabolome profiling and less prone to adduct formation. The platform is composed of three core elements: a chemically defined quality control mix, an acquisition method, and a processing engine that extracts quantitative information from measured spectra to do analytics and reporting through a web-service.

Quality control mix

The first element is a chemically defined quality control (QC) sample. This should include analytes that allow testing the system, be stable over long periods, and ready to inject. For the purpose of testing an ESI-MS system, the QC analytes should span over the full mass range of interest (m/z 100 to 800 for us), be diverse in chemical properties (e.g., polarity, pKa), and in propensity of analytes to build adducts or fragment during ionization. Following these principles, we opted for a mix of nine compounds. We emphasize that there is likely ample room to further optimize the composition of the QC mix. In the current composition, it has been in use for almost two years with satisfactory results and, therefore, we have not tested different mixes. For each compound, the concentration was adjusted to obtain an intense monoisotopic peak. Typically, the nine analytes also produced 27 isotopic peaks, 1 adduct, and 9 fragments, for a total of 37 expected spectral peaks (**Table S1**). During processing, the expected peaks are analyzed individually to quantify peak and ionization properties. Any additional peak that is detectable but not part of the expected set is considered background. All background peaks are treated as a collective to quantify purity and dirt of the system.

Acquisition method

The acquisition method was designed to collect critical data in possibly short time. For the aforementioned reasons, we omitted a chromatographic separation and used an instrument method similar to flow injection analysis with a solvent flow of 150 μL/min. In principle, we were interested in capturing three types of scans: (i) full spectra for the chemical (solvent) background, (ii) full spectra for the QC mix, and (iii) full spectra in the absence of ionization. The latter was included to potentially assess background noise of the detector. It was obtained by stopping the liquid flow to the source and recording residual ion counts. The final profile is shown in **Figure 1**. The chemical background is acquired first to fully reflect the equilibrated LC-MS system and to avoid being affected by any tail of the QC mix peak. Given the absence of a column between autosampler and ionization chamber, the injection program was modified to introduce a temporal delay of about 25 sec between the start of MS acquisition and the injection of the QC mix. After the sample has cleared from the ionization chamber (about 80 sec), the flow is stopped to acquire detector background. Finally, the flow is ramped again. Throughout the period of 2 min, the MS acquires full scan profile data in negative ionization mode at a frequency of 4 GHz.

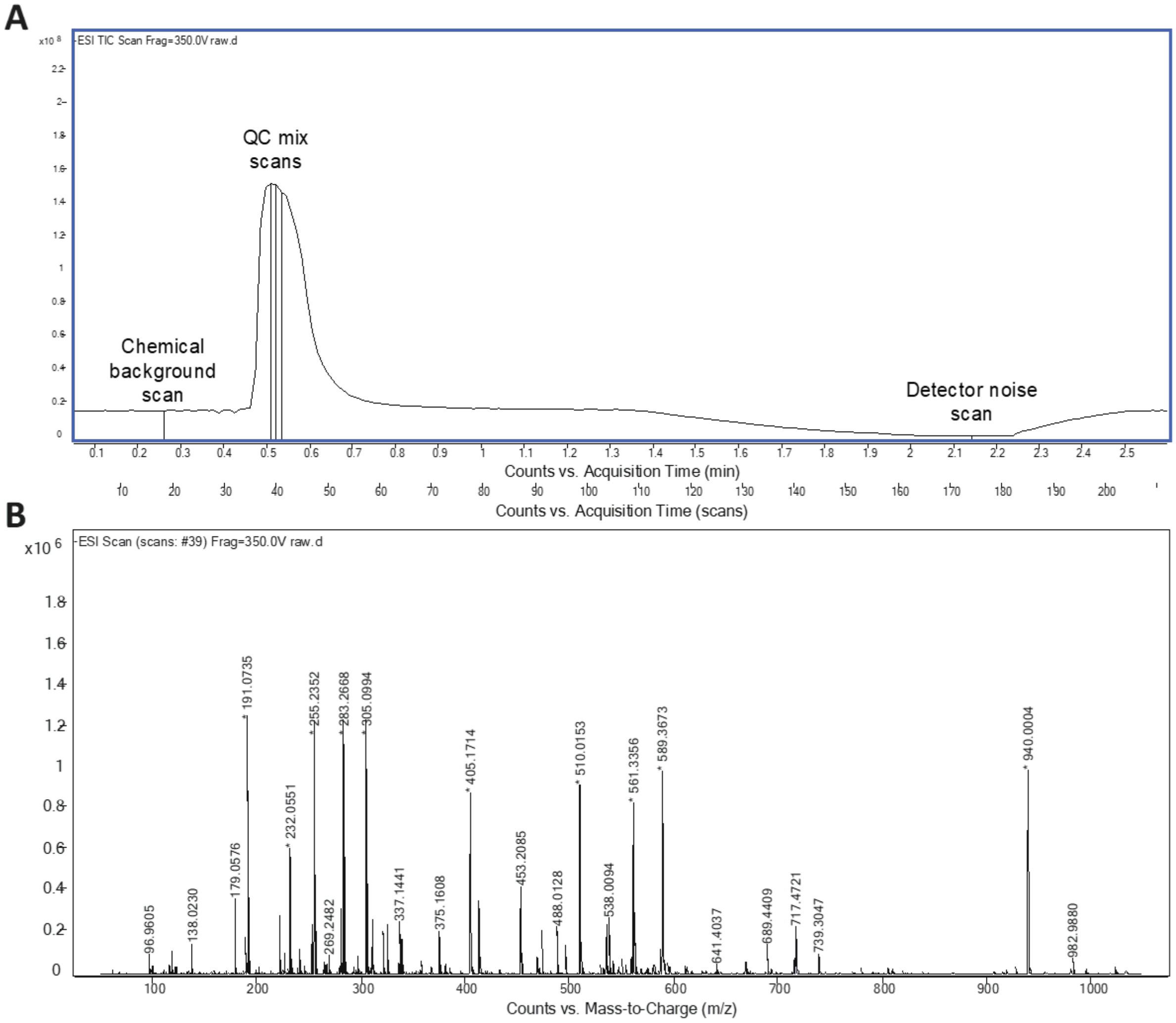


Figure 1. Total ion chromatogram of the acquisition method (**A**). Representative spectrum for QC mix scans (**B**).

Feature extraction

Upon acquisition, raw profile data is analyzed to extract quantitative information that describes spectral properties in much depth. Not knowing in advance which properties of a spectrum drift or shift over time, we designed a very inclusive analysis that extracts 2850 quantitative features for each QC sample injected. Ultimately, this information is obtained from a detailed analysis of five scans. The largest fraction of features is extracted from the chronogram peak related to the QC mix. The exact scan number is determined dynamically by picking the scan with the highest total ion current, and the analysis is extended to the two following scans to obtain an average value and a measure of deviation for each feature. For each of the 37 expected peaks of the QC mix, we collect intensity, absolute mass accuracy, factual ppm, multiple widths, area under peak tails, symmetry, goodness-of-fit with a Gaussian, number of subsequent peaks and their intensity ratios (**Figure S1**).

For each expected isotopologue, fragment, or adduct, we also measure the height relative to the deprotonated, monoisotopic peak, as well as the difference to the theoretical relative abundance. This allows diagnosing deviations from linear response, excessive in-source fragmentation, or increased salt contaminations, respectively. For each of the above features, we record the mean value and the standard deviation from three consecutive scans. Overall, the numeric features derived from the QC mix peaks are 1720.

To capture baseline properties, level and type of dirt, instead of focusing on a predefined list of m/z features, we segmented the mass axis in windows of 50 amu. For each window, we recorded number of peaks, intensity sum, intensity percentiles with all expected ions excluded, intensities of 10 most abundant peaks, as well as their intensity percentiles. This resulted in 720 more features, so the total of features coming from the QC mix is 2440.

Two other scans are analyzed to gain additional information about the system. The chemical background is characterized in a single scan before the QC mix peak, i.e., scan number 18 in our LC-MS setup. Similar to the QC mix scans, we extract 140 features related to two reference compounds that are co-sprayed and used as lock masses for intra-scan mass calibration (HOT and HEX in **Table S1**). 180 more features come from the windows of 100 amu, making a total of 320 background-related features. Finally, pure detector signal is characterized in a late scan in absence of ionization. We collect 90 more features from the windows of 200 amu for a grand total of 2850 for each injected QC sample.

Feature extraction was implemented in Python. In our environment, it is triggered automatically by the appearance of a QC acquisition file in a predefined location. The results are stored in a SQL database that constitutes the access point for all downstream analyses. Starting from raw MS profile data, out software extracts all features and updates the corresponding logs and databases within 20 sec. Including measurement, data logistics and web-service rendering time, the full process takes less than 5 minutes. In the following, we showcase and discuss the immediate and long-term benefits of using the SST platform.

Instrument monitoring

A primary goal of the aforementioned procedure is to verify system suitability before proceeding with data acquisition. For this purpose, we implemented a monitoring system to visualize and compare current and historical data. It consists of a web-service that pulls data in real-time from the database with extracted QC features and reports key information on a dashboard. For obvious reasons, including all 2850 features would have been problematic and inefficient. To favor visualization of accessible information over an overflow of data, we defined 16 quality indicators that report aspects of analytical relevance such as resolution, mass accuracy, accuracy of isotopic ratios, adduct formation, signal intensity, signal-to-noise, levels of dirt and detector noise. These indicators were calculated from the 2850 primary features (**Table S2**) and are presented for users on the dashboard.

Visualization was designed to inform on two types of patterns. First, we were interested in capturing particularly abnormal values of any of the quality indicators. Therefore, we integrated plots that visualize the distribution of quality indicator values in the past and the latest to be evaluated. To facilitate the analysis, the system also performs automatic outlier detection by the isolation forest algorithm[16]. The latter is based on an ensemble of decision trees, followed by a correction routine specific to the type of the indicator. In real-time, each QC sample is scored automatically by counting the number of outliers across the 16 quality indicators. As a rule of thumb, if more than 4 values are classified as outliers, the QC sample flagged with bad quality. The second type of pattern that we wanted to highlight is temporal trends. We hypothesized that factors such as detector aging or dirt accumulation could result in a subtle but continuous decay of instrument performance. Such drifts are slow and, therefore, they would not be recognized by outlier detection. We integrated a trend detection that uses linear models with empirical thresholds for R^2 and slope coefficients. Three temporal intervals are considered (two weeks, one month, and two months) and reported on the dashboard (**Figure S2**).

To assess the technical reproducibility of feature extraction, we analyzed 191 QC samples acquired on the same day without any modification of instrument parameters, i.e., tuning. The median intra-day coefficient of variation across the 2850 QC features was 27%, but with strong differences between the types of QC features. For example, noisier features were associated to the lock masses included in the buffer. Other noisy features described the tails of DC mix ions (i.e., ringing and baseline artefacts). For the 16 quality indicators, the median variation coefficient was 4%. This indicates that the setup is robust enough to capture shifts of about 10% or more.

The system has been operating in a pilot period of 21 months. During this period, at least one QC sample has been analyzed on 110 days constituting a total of 153 measurements. Among those, 37 QC samples featured four or more outlier values and were flagged as of bad quality. Based on automated outlier detection, *fragmentation_305* was found to be most out-of-order QC indicator (in 31% of QC samples), followed by *isotopic_presence* (28%) and *baseline_25_150* (25%). The most stable indicators were *average_accuracy* (only 8% of bad quality), *resolution_700* (6%) and *baseline_50_650* (5%). During the same pilot period, a total of 40 trends were automatically detected for the QC indicators within 2-month windows. For instance, an increasing *chemical_dirt* trend was detected between October and December 2019 (R^2 = 0.7071, n=16) and a decreasing *resolution_700* trend was detected between October and December 2020 (R^2 = 0.6873, n=14). Other examples of trend detection are given on **Figure S3**.

Upon detection of a bad quality QC sample, the user was prompted to take corrective actions to restore normal range. Thus far, corrective actions were suggested based on expert knowledge and the kind of outlier. For example, in the case of a loss of, e.g., resolution, mass accuracy, or ion transmission, the system was retuned focusing on the relevant section of the optics or opting for a general system tune in more extreme cases. In the case of increases of, e.g., chemical background signals, the primary response was to purge the system, replace buffers, or clean the source and front optics. In the case of a drop in signal or signal-to-noise, we evaluated whether sensitivity of the detector (i.e., the voltage of the multichannel plate detector, or the amp gain) had to be readjusted. These recommendations were adopted by users to prevent the injection of samples before normal operation was verified.

Analysis of QC features

We designed the feature extraction to be very inclusive and capture possibly granular information on spectral properties and, in turn, instrument characteristics. This resulted in a long vector with 2850 numerical values. By design, several of the features are likely correlated because they reflect similar aspects of the same peak, report the same property of different peaks of the QC mix, or simply relate to neighboring regions of the mass range. We therefore wondered about the actual information content: is there a benefit in collecting very detailed information or could one describe instrument state with much fewer features? To address this question, we analyzed the full matrix of values obtained in the initial 21 months of operations. We performed a principal component analysis (PCA) of the feature matrix and calculated a cumulative portion of total variance explained. We found that the first 10 components explained only about 52% of total variance. To capture 95%, 100 components were necessary. This highlights a substantial heterogeneity of the dataset. Rather than surfing through many principal components, we performed a hierarchical clustering to understand whether the list of QC features could be compressed without tangible losses in information (**Figure 2**). In line with the PCA, numerous subtrees with peculiar patterns over the measured QC samples emerged. To verify what kind of QC features correlated in these subtrees, we sought for enrichments in either the m/z range of the feature, its type, or the scan the feature was extracted from (encoded in the colors on the left of **Figure 2**).

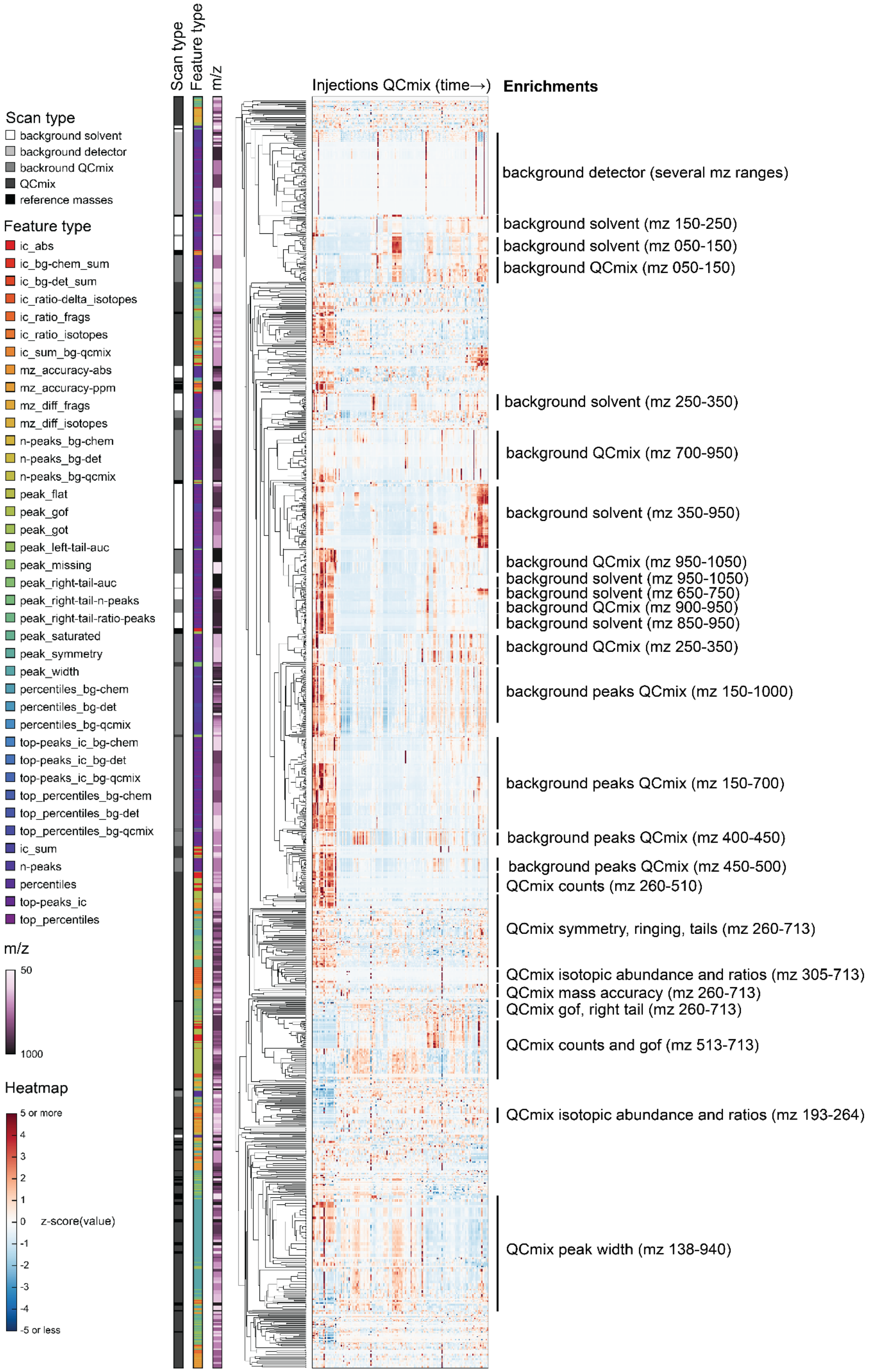


Figure 2. Hierarchical clustering of z-scored QC features with annotations by m/z, scan and feature types. See descriptions of feature types in **Table S3** and **Table S4**.

Tightest clustering was found for features linked to spectral background (labeled with "background" in **Figure 2**), pointing to some redundancy. These are the values that are not related to the chemicals spiked in the QC mix. They dominated the upper part of the clustered heatmap. Background features of similar type but different mass range were frequently adjacent. Background features extracted from the solvent and from the QC mix scans co-clustered frequently. Retrospectively, this was expected because apart from the regions populated by QC mix ions, the two scans are expected to be identical. In such cases, a single feature measured in a QC mix scan seems sufficient to recapitulate the principal drifts and shifts observed across scans and the mass range. In contrast, the features derived from ions deriving from the chemicals spiked in the QC mix were more heterogeneous (lower part of **Figure 2**). Albeit many features of similar type were close in the tree (e.g., symmetry, isotopic properties, mass accuracy, etc.), their distance was generally higher than observed for background features. This seems to reflect the fact that peak features tend to vary across the mass range, possibly because of differences in intensity or in the spectral neighborhood. The only exception was the peak width, which correlated well across the whole mass range.

To further investigate feature redundancy, we did a cross-correlation analysis between all 2417 continuous QC features. The resulting Pearson correlation coefficients were < 0.5 in 95% of the pairs. Only for 0.4%, the correlation was strong ($|r| > 0.9$). As expected, these cases were related either to the features of the same type or to "synonymic" features of the same m/z window. These results confirm that QC features describe many spectral properties, not likely to be fungible. It is further supported by a simple visual analysis of the differences between injections, i.e., the columns on the heatmap shown in **Figure 2**. Many vertical stripes emerge, which indicate sets of features with values at the far ends of the measured range. Importantly, the extreme values are not aligned vertically over a large fraction of features but tend to vary across samples. This heterogeneity indicates that fine-grained shifts are present in the data, even though they might have not been captured by the 16 quality indicators that were adopted for instrument monitoring. It remains to be tested if such drifts had a tangible effect on the measurement of studies that were acquired on the same day. This analysis would require expanding the outlier detection introduced for system monitoring to all measurable features. Whenever an extreme deviation was reported, a set of test samples ought to be run to evaluate the practical consequences.

Association analysis between instrument settings and QC features

We were wondering whether any of the aberrant behaviors detected during the pilot phase were associated to drifts in setpoints or readbacks of the instrument. Therefore, we studied the relationship between measured quality indicators (or QC features) and actual instrument settings, which include both tunable and non-tunable values. Tunable values affect ion optics and detection and are adjusted during instrument tuning or calibration according to the procedures that are implemented in the control software. Non-tunable values consist of readbacks of parameters such as pressures, currents, pump speeds, noise, and are collected for diagnostic purposes. In many cases, they are stored with each run or in the tuning reports. The number and type of accessible instrument settings varies across vendors and type of instruments. For the QTOF instrument described here, about ten non-tunable values and fifty tunable values were extracted for each QC run and stored.

The connection between instrument performance and instrument settings can be analyzed by different approaches. For example, a significant decay in resolution (at m/z 700) over the period of two months was once identified by the aforementioned trend detection (**Figure 3**). To find potential causes of this loss in performance, we sought for correlations with instrument settings during the same period. Significant associations were found for two parameters: the non-tunable pressure reading in the first quadrupole and the ion focus strength (**Figure 3B, C**). This suggests that the transient drop in resolution might be caused by an increased spread of trajectories or velocities of ions in the section preceding the TOF pulser.

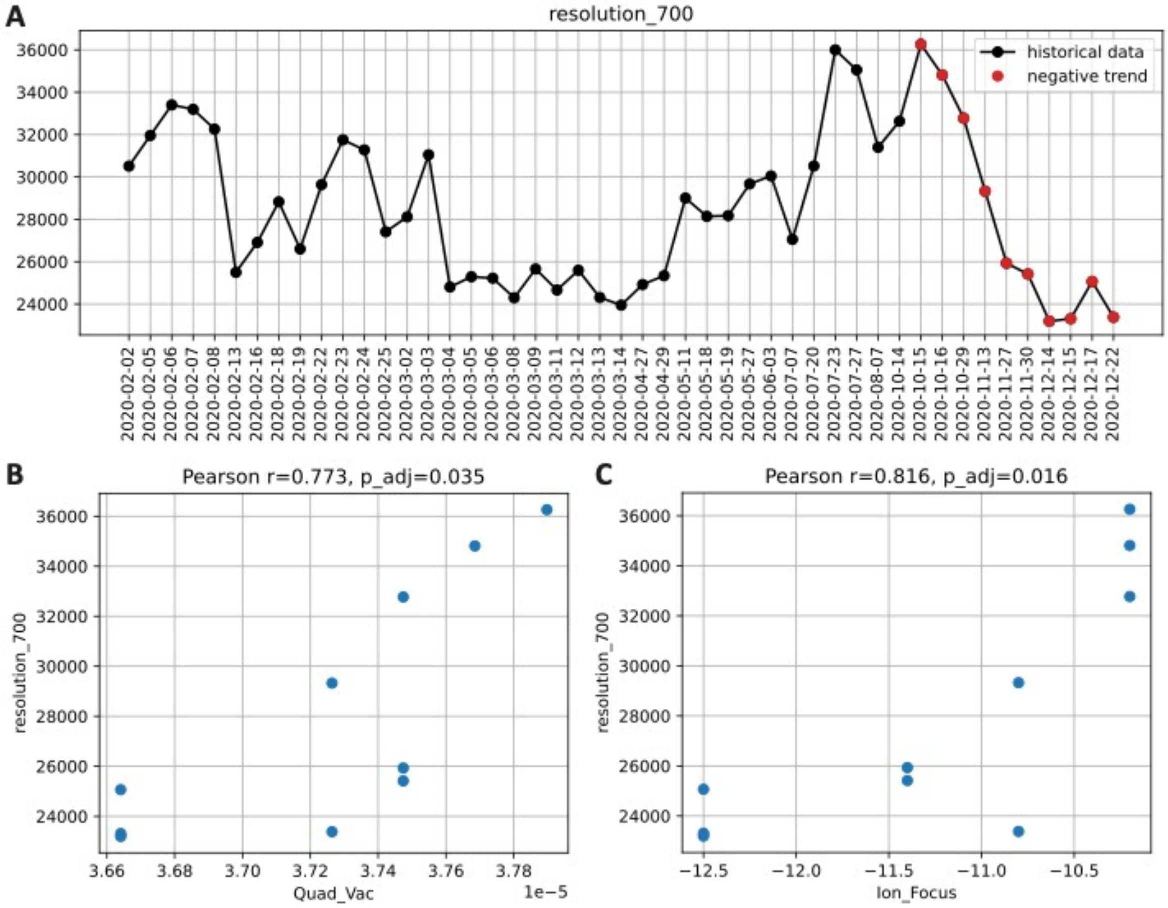


Figure 3. The decreasing *resolution_700* trend (**A**) and the corresponding correlations with the instruments settings: quadrupole vacuum (**B**) and ion focus (**C**). Pearson correlation coefficients and Bonferroni adjusted p-values are given.

An alternative approach is to seek for parameters that are associated with top performance. We illustrate this for the signal-to-noise, which reflects instrument sensitivity. In this case, we split all QC samples into groups of high (top 20th percentile) and low signal-to-noise. We then sought for instrument settings that were different between the two groups. The procedure was repeated for another quality indicator, the signal-to-background (**Table 1**). For both indicators, we found significant associations. Several expected associations were found. For instance, better vacuum (lower value of *TOF_Vac*) was linked to higher signal-to-noise and signal-to-background. Increased multichannel plate voltage (*MCP*), or instrument firmware (*InstrumentFW*) improved both indicators. We also found less intuitive associations, like the voltage of several lenses of the ion optics before the TOF section (e.g., *Top_Slit, Bottom_Slit, Lens_2*, *Oct_1_RF_Vpp_1*, etc.) which are likely to overall improve ion transmission.

Table 1. Statistical comparison of the machine tunes. Machine tunes significantly different between signal-to-noise groups are shown on the left. Machine tunes significantly different between signal-to-background groups are shown on the right. In both cases, three statistical tests were applied for each comparison (Kolmogorov-Smirnov, Mann-Whitney U, Kruskall), followed by FDR correction for multiple testing. The biggest significant p-value is reported. Sign of linear relationship is shown, where the medians of two distributions were different.

signal-to-noise			**signal-to-background**		
setting	**sign**	**p-value**	**setting**	**sign**	**p-value**
Acc_Focus	—	0.0118	*Amp_Offset*		0.0044
Amp_Offset		< 0.0001	*InstrumentFW*	+	0.0414
Bot_Slit	+	0.0229	*Length_of_Transients*		0.0443
Cell_Entr	+	0.0433	*Lens_2*		0.0414
Cell_Exit		0.0206	*Lens_2_RF_Ph*		0.0255
InstrumentFW	+	0.0001	*MCP*		0.0034
Lens_2		0.0007	*Oct_1_RF_Vpp_1*		0.0003
Lens_2_RF_Ph		0.0007	*Puller_Offset*		0.0044
MCP	+	0.0001	*TOF_Vac*	—	0.0091
Mirror_Mid	+	0.0156	*Top_Slit*	+	0.0044
Oct_1_RF_Vpp_1		< 0.0001			
Puller_Offset		0.0001			
TOF_Vac	—	0.0003			
Top_Slit	+	< 0.0001			

All identified associations may indicate what settings determine properties of measured spectra but were analyzed in isolation and for selected examples. In reality, settings and indicators are partly interdependent. For example, adjusting a voltage to increase sensitivity might negatively affect resolution. To go beyond individual correlations and statistical tests, we attempted to learn causal relationships between the tunable instrument settings and the observed quality indicators. Specifically, we attempted to learn the underlying structure from all available data using the PC[17,18] algorithm (named after Peter and Clark[19]) and conditional independence testing. The result is a directed acyclic graph that condenses statistical dependence between instrument settings and performance indicators (**Figure 4**). The results reveal that, for example, the spontaneous fragmentation of fluconazole (*fragmentation_305*) was associated to voltages in the section that precedes the collision cell (*Lens_2, Lens_2_RF_Ph, Oct_1_RD_Vpp_1, Cell_Entr*). Resolution (both at m/z 200 and 700) seems governed by the bottom slicer voltage. This is coherent with the function of the slicer, which flattens the ion beam before it is pulsed orthogonally in the flight tube. Suboptimal slicer settings increase differences in the length of the flight path which would result in different flight times even for ions of identical m/z. Thereby, it would worsen peak resolution. Instead, most settings related to the electrospray ionization process (aggregated on the right part of **Figure 4**) affect parameters related to signal intensity such as signal-to-baseline, isotopic accuracy, mass accuracy. With more data, it could be possible to build a reliable statistical model to also infer which settings should be adjusted to achieve or maintain a certain property of the measurement.

Conclusion

We present the concept of a system suitability testing platform for monitoring the status of a high-resolution QTOF mass spectrometer. The setup consists of a QC mixture, an acquisition method, software to extract a detailed ensemble of quantitative features describing spectral properties, and a simple R Shiny front-end for real-time visualization. We operated the testing platform in a pilot lasting for 21 months and including 153 individual measurements of the QC mixture. We demonstrated instrument monitoring by a small set of quality indicators (16 in our case) and the implementation of routines for trend and outlier detection. The platform, therefore, helps users in evaluating in depth the instrument readiness to measure biological samples. In most cases, unsatisfactory results could be effectively addressed by cleaning the ion optics or a thorough tuning/calibration of the instrument.

The presented setup offers ample room for further improvements. In particular, the long-term stability of the QC mixture should be verified. The feature set could also be optimized. The feature extraction is generic and easily transferrable to TOF instruments from other vendors. Adaptation to high-resolution instruments that use a

Fourier Transform to reconstruct spectra would require more work. Features related to peak symmetry, detector ringing, baseline shifts, etc. are irrelevant. In contrast, it would be important to include features that can capture artifacts of FT spectra: harmonic peaks, coalescence, etc.

Further, we illustrated how collection of instrument setpoints and readbacks allows to derive the causal relationships between instrument settings and instrument performance measured with the QC mixture. This highlights additional, potential applications of the testing platform. First, we envisage that the system could recommend instrument settings to maintain or attain a particular value of a quality indicator. Second, it could assist in timing preventive maintenance. We speculate that continuous QC data collection coupled with predictive models would be able to indicate when to replace wearable parts, clean specific parts of the ion path, or maybe even anticipate major failures such as a pump breakdown.

Thus far, the testing platform has been conceived to operate with data pertaining to a single instrument. Future work will explore the possibility of using the detailed information provided by the QC mix to harmonize data collected either on the same instrument but at different time points, or on different instruments of the same type. The canonical approach to normalize across batches or instruments is to include internal standards, or standard reference materials. This approach, however, works only for compounds that are present in the standard material and fails to capture non-linear effects. We hypothesize that capturing detailed information on, e.g., baseline, ion transmission, fragmentation, etc. might help to harmonize data before normalization by standards can be applied.

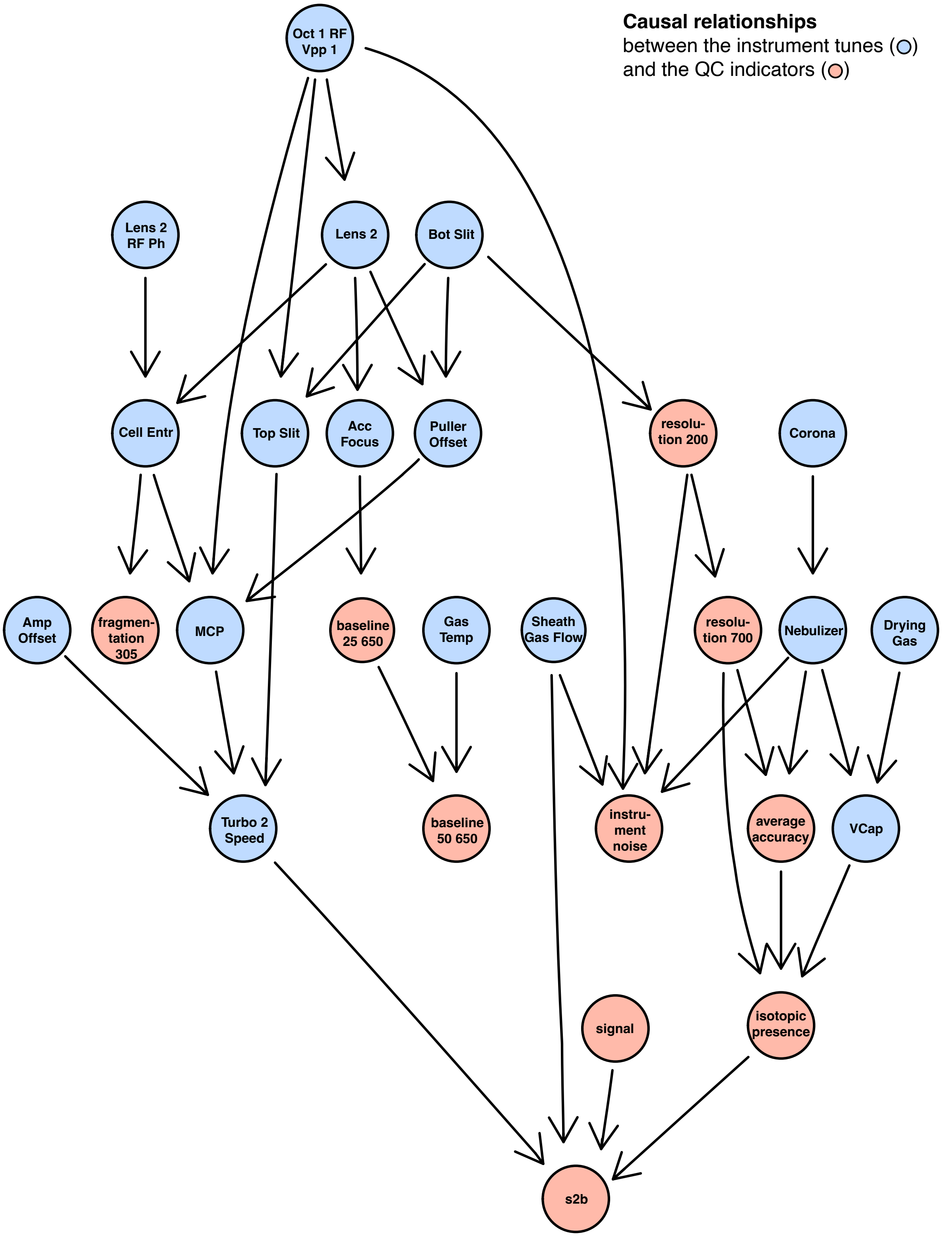


Figure 4. A reduced DAG reflecting the causal relationships between the instrument settings (in blue) and the QC indicators (in red). A significance level of 0.15 was used as the threshold for conditional independence testing.

Methods

Instrument details

All analyses were done on an Agilent 6550 QTOF instrument, operated in negative ionization and 4 GHz High-Resolution mode because it matches the configuration that we use in routine flow injection and LC-MS analysis. The mobile phase was 60% isopropanol in water (v/v) supplemented with homotaurine (Sigma-Aldrich, Germany) and Hexakis(1H, 1H, 3H-tetrafluoropropoxy)phosphazine (Agilent) as reference masses for m/z axis calibration. The solvent flow was 150 μL/min. All compounds included in the QC mixture (**Table S1**) were purchased from Sigma-Aldrich (Germany) at the highest purity available. The injection volume was 1 μL.

Anomaly detection and quality control

We implemented two types of anomaly detection: i) based on descriptive statistics, and ii) based on machine learning. Both approaches require some reference (or training) data to apply algorithms and determine whether a new quality indicator is likely to be an outlier or not. Each of the 16 quality indicators of the new run is evaluated individually, and the total number of non-outliers serves as the QC run score. Two different types of anomaly detection suggest different usage scenarios of the monitoring system.

The statistical approach assumes that the instrument preserves its properties within the period of the study. If the instrument performance remains the same with only little oscillations, appropriately, any quality indicator does as well. Thus, measuring quality indicators repeatedly over time makes it possible to derive confidence intervals for the expected mean, or the ranges that are considered as “good” or “bad” for each indicator. In this approach, we use quantiles to compute such ranges, as soon as enough data is generated and stored in the database. We set 60 measurements of the QC sample to be enough to classify further values of quality indicators as “good” or “bad”, i.e., within the expected interval or not. This number, however, is only empirical and remains a configurable parameter in the platform.

This approach may not be optimal for longitudinal studies, because it does not adapt to the changes in the instrument state over time. Intervals derived for the first *N* measurements will be applied to the data forever. Possible effects of instrument aging and hardware replacements will be ignored. To potentially account for them and to make the system adaptable, we implemented another approach based on machine learning. Isolation Forest, an unsupervised method for outlier detection, is used to re-evaluate all the entries in the database as soon as a new QC measurement is acquired. This way, the platform adapts to the gradual temporal drifts in quality indicators, while still being capable of detecting anomalies. Because of that, only *N*=20 measurements are set as a minimum number of entries to apply the method.

Both methods' predictions are corrected for the type of the quality indicator. For instance, low mass accuracy values are not treated as outliers, since a small difference between expected and measured *m/z* value for an ion is desired. Big signal-to-noise ratios are not treated as outliers, since high signal-to-noise ratio is preferable, in general. For other cases, adjustments are made on top of the aforementioned methods to compensate for artifacts caused by little data available (i.e., when the total number of QC samples in the database is still small).

In our experience, both methods to detect anomalies have shown similar results, when applied to the data systematically acquired within 21 months. However, in multi-day acquisitions, we see Isolation Forest to be preferable due to its adaptability and relative robustness. Isolation Forest is, therefore, a default method in the platform.

Web service

We used R Shiny framework to implement a web-service providing users with graphical representation of the data and analytics (**Figure S2**). The layout contains three tabs: *summary*, *trends* and *table* components. The *summary* tab allows users to select a QC run by date and to see how the corresponding quality indicators are aligned against the full dataset. Score of the selected run and the distributions for each metric are displayed. The *trends* tab depicts temporal progression for the selected indicator, marking overall run qualities. This allows to see and analyze each metric's behavior retrospectively. Linear trends are computed and visualized as well, which helps detecting gradual loss of sensitivity, gain of dirt in the system, etc. Finally, the *table* tab explicitly shows the values of the quality indicators from the database. Values classified as “good” or “bad” are colored in green and red, respectively.

Availability

The source code of the SST platform core (raw signal processing, feature extraction and engineering) is available at https://github.com/zamboni-lab/SST-platform-core. The QC database and the Shiny app of the web-service are available at https://github.com/zamboni-lab/SST-platform-shiny.

Acknowledgements

This project was supported by grants from the Strategic Focal Area Personalized Health and Related Technologies (PHRT) of the ETH Domain.

Supplementary material

Table S1. Expected ions of QC mix and buffer.

QC mix ions

#	compound	m/z	type
1	*Caffeine*	193.0725	isotope
2		194.0759	isotope
3		195.0792	isotope
4		179.0569	fragment
5	*Fluconazole*	305.0962	isotope
6		306.0995	isotope
7		307.1029	isotope
8		191.0681	fragment
9	*Albendazole*	264.0806	isotope
10		265.0840	isotope
11		266.0873	isotope
12		188.9996	fragment
13		232.0544	fragment
14	*Triamcinolone acetonide*	433.2026	isotope
15		434.2059	isotope
16		435.2093	isotope
17		453.2088	adduct
18		337.1439	fragment
19	*Pentadecafluoroheptyl*	368.9760	isotope
20		369.9794	isotope
21		370.9827	isotope
22	*3-Heptadecafluorooctylaniline*	510.0150	isotope
23		511.0184	isotope
24		512.0217	isotope
25	*Perfluorodecanoic acid*	512.9594	isotope
26		513.9628	isotope
27		514.9662	isotope
28		468.9696	fragment
29	*Tricosafluorododecanoic acid*	612.9531	isotope
30		613.9564	isotope
31		614.9598	isotope
32		568.9632	fragment
33	*Perfluorotetradecanoic acid*	712.9467	isotope
34		713.9500	isotope
35		714.9534	isotope
36		668.9568	fragment
37		646.9549	fragment

Buffer ions

#	compound	m/z	type
1	*HOT (Homotaurine)*	138.0224	isotope
2		139.0258	isotope
3		140.0291	isotope
4	*HEX (Hexakis (1H, 1H, 3H-tetrafluoropropoxy) phosphazine)*	940.0009	isotope
5		941.0042	isotope
6		942.0076	isotope

Figure S1. Examples of QC features related to the expected ion peaks.

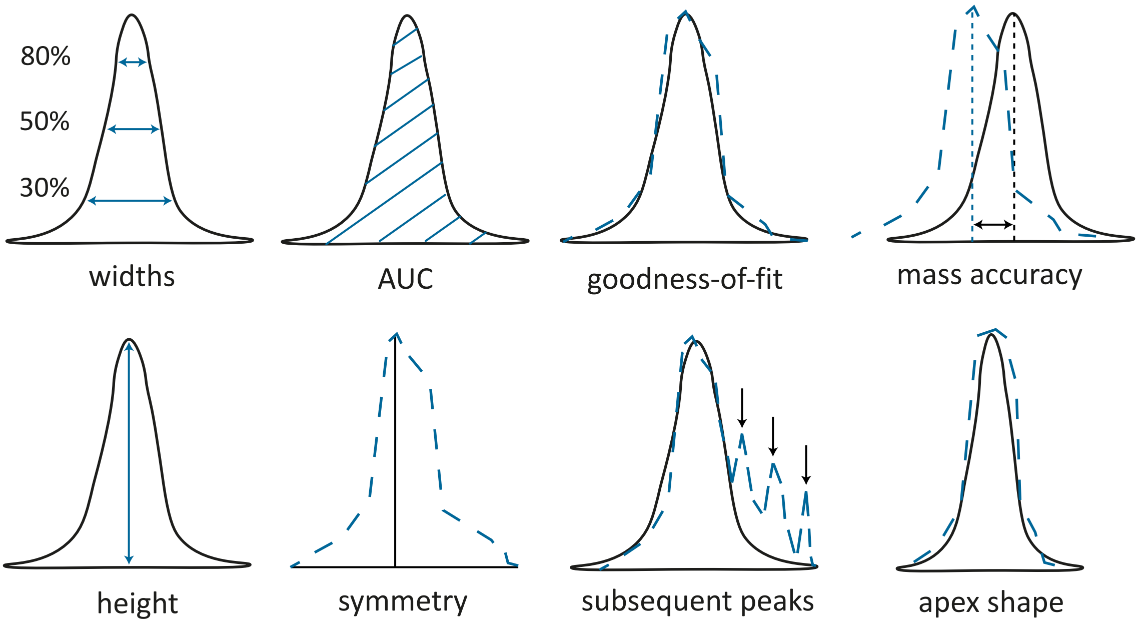
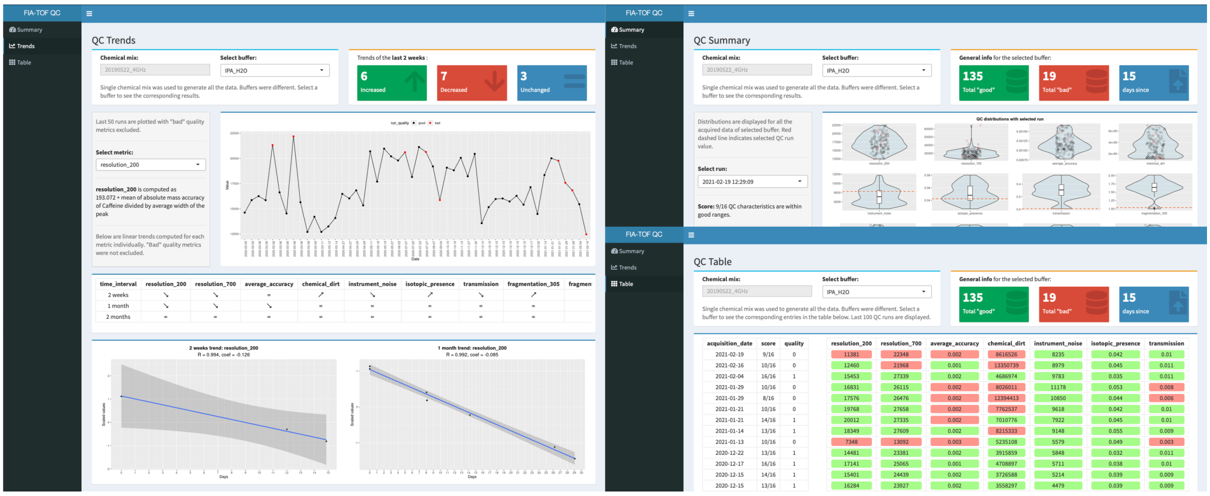


Table S2. QC indicators descriptions.

QC indicators

#	name	pseudo code	description
1	resolution_200	`mz_Caffeine / peak_width_Caffeine`	measured m/z of Caffeine divided by the average width of the peak
2	resolution_700	`mz_Perf_acid / peak_width_Perf_acid`	measured m/z of Perfluorotetradecanoic acid divided by the average width of the peak
3	average_accuracy	`sum(mean_abs_mass_accuracy_array) / n_ions`	sum of mean absolute mass accuracy for all 37 ions divided by its number
4	chemical_dirt	`sum(chem_bg_intensity_array)`	sum of all intensities in the chemical background scan
5	instument_noise	`sum(noise_intensity_array)`	sum of all intensities in the detector noise scan
6	isotopic_presence	`sum(abs(mean_iso_ratios_diffs_array)) / length(mean_iso_ratios_diffs_array)`	sum of all isotope ratios' diffs (in absolute numbers) divided by its number
7	transmission	`mean_intensity_Perf_acid / mean_intensity_Fluconazole`	mean of the Perfluorotetradecanoic acid intensity (m/z ~712) divided by the Fluconazole intensity (m/z ~305)
8	fragmentation_305	`mean_intensity_Fluconazole_fragment / mean_intensity_Fluconazole`	mean of Fluconazole fragment intensity (m/z ~191) divided by the Fluconazole intensity (m/z ~305)
9	fragmentation_712	`mean_intensity_Perf_acid_fragment / mean_intensity_Perf_acid`	mean of Perfluorotetradecanoic acid fragment intensity (m/z ~668) divided by the Perfluorotetradecanoic acid intensity (m/z ~712)
10	baseline_25_150	`percentile(chem_bg_intensities_150_250, 25)`	25th intensity percentile from a [150, 250] m/z range of a chemical background scan
11	baseline_50_150	`median(chem_bg_intensity_array_150_250)`	median intensity from a [150, 250] m/z range of a chemical background scan
12	baseline_25_650	`percentile(chem_bg_intensity_array_650_750, 25)`	25th intensity percentile from a [650, 750] m/z range of a chemical background scan
13	baseline_50_650	`median(chem_bg_intensities_650_750)`	median intensity from a [650, 750] m/z range of a chemical background scan
14	signal	`sum(mean_intensity_array)`	sum of mean intensities for all 37 ions
15	s2b	`mean_intensity_3Hepta / percentile(intensity_array_500_550, 25)`	mean intensity of 3-(Heptadecafluorooctyl)aniline (m/z ~510) divided by mean 25th intensity percentile from a [500, 550] m/z range
16	s2n	`mean_intensity_3Hepta / (median(intensity_array_500_550) - percentile(intensity_array_500_550, 25))`	mean intensity of 3-(Heptadecafluorooctyl)aniline (m/z 510) divided by the diff between the median and the 25th intensity percentile from a [500, 550] m/z range

Figure S2. Snapshots of the web-service. A *summary* tab (top right): distributions of the quality indicators are displayed with red dotted lines indicating the selected run. A *trends* tab (left): historical data is displayed for the selected quality indicator on top. A table with summary on the trends and two trend plots are displayed below. A *table* tab (bottom right): values of the quality indicators are displayed in a table for the last 100 QC runs. Indicators classified as 'good' and 'bad' (read "within a normal range" and "likely an outlier") are in green and red, respectively.



Chemical mix: 20190522_4GHz

Select buffer: IPA_H2O

Single chemical mix was used to generate all the data. Buffers were different. Select a buffer to see the corresponding results.

Trends of the **last 2 weeks**:

Last 50 runs are plotted with "bad" quality metrics excluded.

Select metric: resolution_200

resolution_200 is computed as 193.072 + mean of absolute mass accuracy of Caffeine divided by average width of the peak

Below are linear trends computed for each metric individually. "Bad" quality metrics were not excluded.

time_interval	resolution_200	resolution_700	average_accuracy	chemical_dirt	instrument_noise	isotopic_presence	transmission	fragmentation_305	fragmen
2 weeks	↘	↘	=	↗	↘	↗	↘	↗	
1 month	↘	↘	↘	=	=	=	=	=	
2 months	=	=	=	=	=	=	=	=	

Chemical mix: 20190522_4GHz

Select buffer: IPA_H2O

Single chemical mix was used to generate all the data. Buffers were different. Select a buffer to see the corresponding results.

General info for the selected buffer:

Distributions are displayed for all the acquired data of selected buffer. Red dashed line indicates selected QC run value.

Select run: 2021-02-19 12:29:09

Score: 9/16 QC characteristics are within good ranges.

Chemical mix: 20190522_4GHz

Select buffer: IPA_H2O

Single chemical mix was used to generate all the data. Buffers were different. Select a buffer to see the corresponding entries in the table below. Last 100 QC runs are displayed.

General info for the selected buffer:

acquisition_date	score	quality	resolution_200	resolution_700	average_accuracy	chemical_dirt	instrument_noise	isotopic_presence	transmission
2021-02-19	9/16	0	11381	22348	0.002	8616526	8235	0.042	0.01
2021-02-16	10/16	0	12460	21968	0.001	13350739	8979	0.045	0.011
2021-02-04	16/16	1	15453	27339	0.002	4686974	9783	0.035	0.011
2021-01-29	10/16	0	16831	26115	0.002	8026011	11178	0.053	0.008
2021-01-29	8/16	0	17576	26476	0.002	12394413	10850	0.044	0.006
2021-01-21	10/16	0	19768	27658	0.002	7762537	9618	0.042	0.01
2021-01-21	14/16	1	20012	27335	0.002	7010776	7922	0.045	0.01
2021-01-14	13/16	1	18349	27609	0.002	8215333	9148	0.055	0.009
2021-01-13	10/16	0	7348	13092	0.003	5235108	5579	0.049	0.003
2020-12-22	13/16	1	14481	23381	0.002	3915859	5848	0.032	0.011
2020-12-17	16/16	1	17141	25065	0.001	4708897	5711	0.038	0.01
2020-12-15	14/16	1	15401	24439	0.002	3726588	5214	0.039	0.009
2020-12-15	13/16	1	16284	23927	0.002	3558297	4479	0.039	0.009

Figure S3. Examples of trend detection for quality indicators. Two-weeks trends are shown on top and one-month trends are below. Linear regression coefficient defines the sign of the trend, whereas the R^2 score reflects its significance. Empirical thresholds allow to classify the quality indicators as increasing, decreasing or unchanged (no significant trend detected) within the corresponding time period.

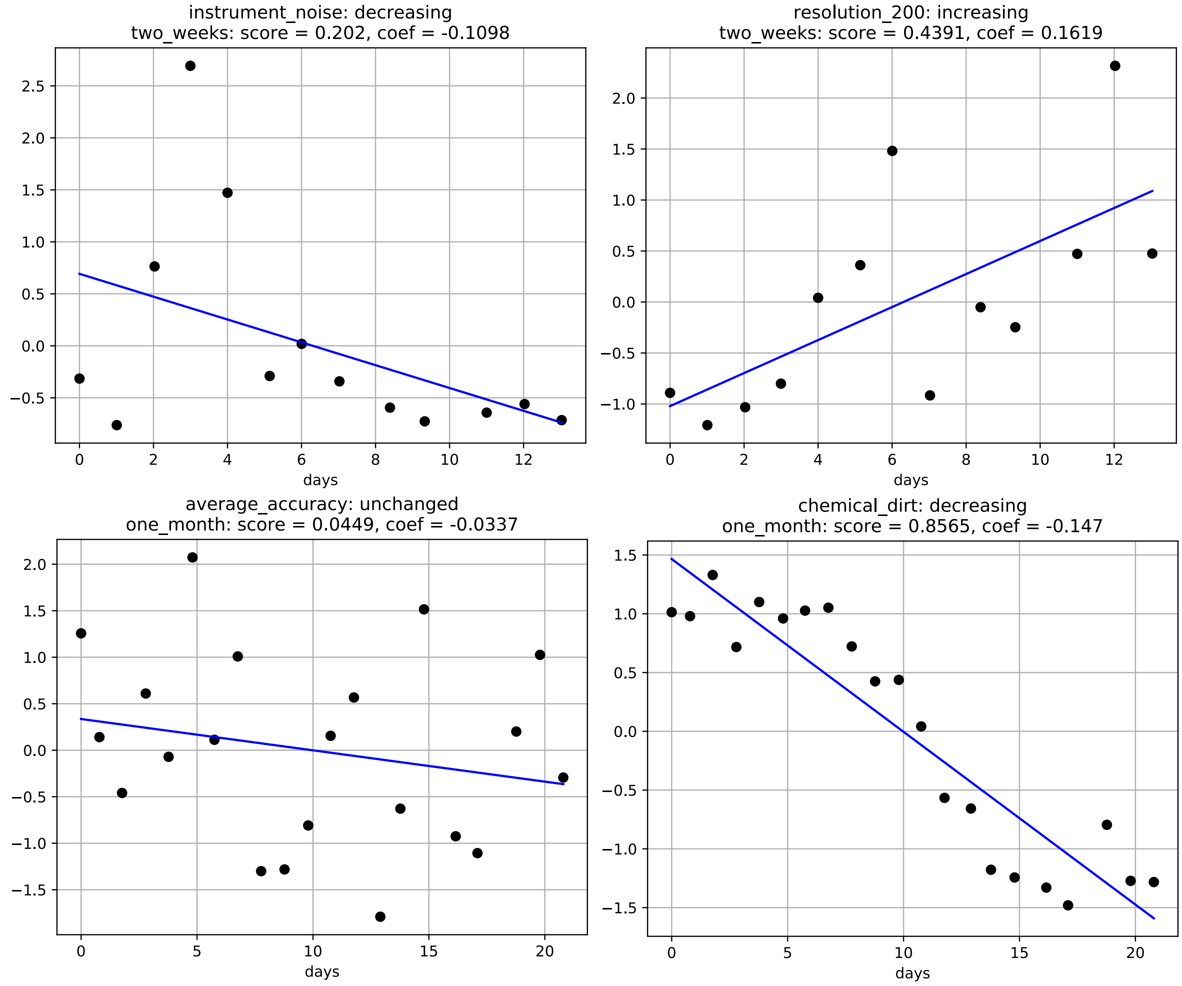


Table S3. Description of QC mixture feature types.

Types of QC mix features

#	type	feature	description
1	counts	`intensity`	measured intensity of the ion
2	peak width	`widths`	width of the peak (in amu units) at 20%, 50% and 80% of its total height
3	mass accuracy	`absolute_mass_accuracy`	absolute mass accuracy in amu units
		`ppm`	mass accuracy in ppm
4	peak vicinity	`subsequent_peaks_number`	a number of subsequent centroids within a window of 3 peak widths (at 50% of the height)
		`subsequent_peaks_ratios`	intensity ratios between the original peak and the subsequent peaks
5	peak shape	`left_tail_auc`	an integral of the difference between actual and fitted intensities on the left side of the ion peak
		`right_tail_auc`	an integral of the difference between actual and fitted intensities on the right side of the ion peak
		`symmetry`	a measure of peak symmetry defined as the sum of left and right tail aucs, divided by the two maximums of them
		`goodness_of_fit`	reduced chi-squared, AIC and BIC of the peak fit with a Gaussian model
6	isotopic abundance and ratios	`isotopes_ratios`	intensity ratios between the main ion peak and its isotopes
		`isotopes_ratios_diffs`	differences between the actual isotopes' ratios and the expected (theoretical) ones
		`isotopes_mass_diffs`	differences between the actual isotopes' m/z values and the expected (theoretical) ones
7	ion fragments	`fragments_ratios`	intensity ratios between the main ion peak and its fragments
		`fragments_ratios_diffs`	differences between the actual fragments' ratios and the expected (theoretical) ones
		`fragments_mass_diffs`	differences between the actual fragments' m/z values and the expected (theoretical) ones

Table S4. Description of background feature types.

Types of background features

#	type	feature	description
1	background QCmix	`number_of_peaks_norm`	a total number of peaks within a particular amu window of a QC mix scan with 37 expected ions excluded
		`intensity_sum_norm`	an intensity sum of all peaks within a particular amu window of a QC mix scan with 37 expected ions excluded
		`percentiles_norm`	25th, 50th and 75th percentiles of intensities within a particular amu window of a QC mix scan with 37 expected ions excluded
		`top_peaks_intensities_norm`	intensities of top 10 ion peaks within a particular amu window of a QC mix scan with 37 expected ions excluded
		`top_percentiles_norm`	25th, 50th and 75th percentiles of intensities of top 10 ion peaks within a particular amu window of a QC mix scan with 37 expected ions excluded
2	background solvent	`number_of_peaks_chem`	a total number of peaks within a particular amu window of a chemical background scan with 2 expected ions excluded
		`intensity_sum_chem`	an intensity sum of all peaks within a particular amu window of a chemical background scan with 2 expected ions excluded
		`percentiles_chem`	25th, 50th and 75th percentiles of intensities within a particular amu window of a chemical background scan with 2 expected ions excluded
		`top_peaks_intensities_chem`	intensities of top 10 ion peaks within a particular amu window of a chemical background scan with 2 expected ions excluded
		`top_percentiles_chem`	25th, 50th and 75th percentiles of intensities of top 10 ion peaks within a particular amu window of a chemical background scan with 2 expected ions excluded
3	background detector	`number_of_peaks_bg`	a total number of peaks within a particular amu window of a detector noise scan
		`intensity_sum_bg`	an intensity sum of all peaks within a particular amu window of a detector noise scan
		`percentiles_bg`	25th, 50th and 75th percentiles of intensities within a particular amu window of a detector noise scan
		`top_peaks_intensities_bg`	intensities of top 10 ion peaks within a particular amu window of a detector noise scan
		`top_percentiles_bg`	25th, 50th and 75th percentiles of intensities of top 10 ion peaks within a particular amu window of a detector noise scan

References

(1) Staddon, J. *Scientific Method: How Science Works, Fails to Work, and Pretends to Work*; 2017. https://doi.org/10.4324/9781315100708.

(2) Tarazona, S.; Balzano-Nogueira, L.; Gómez-Cabrero, D.; Schmidt, A.; Imhof, A.; Hankemeier, T.; Tegnér, J.; Westerhuis, J. A.; Conesa, A. Harmonization of Quality Metrics and Power Calculation in Multi-Omic Studies. *Nat. Commun.* **2020**, *11* (1), 1–13. https://doi.org/10.1038/s41467-020-16937-8.

(3) Inthout, J.; Ioannidis, J. P. A.; Borm, G. F.; Goeman, J. J. Small Studies Are More Heterogeneous than Large Ones: A Meta-Meta-Analysis. *J. Clin. Epidemiol.* **2015**, *68* (8), 860–869. https://doi.org/10.1016/j.jclinepi.2015.03.017.

(4) Beger, R. D.; Dunn, W. B.; Bandukwala, A.; Bethan, B.; Broadhurst, D.; Clish, C. B.; Dasari, S.; Derr, L.; Evans, A.; Fischer, S.; Flynn, T.; Hartung, T.; Herrington, D.; Higashi, R.; Hsu, P. C.; Jones, C.; Kachman, M.; Karuso, H.; Kruppa, G.; Lippa, K.; Maruvada, P.; Mosley, J.; Ntai, I.; O'Donovan, C.; Playdon, M.; Raftery, D.; Shaughnessy, D.; Souza, A.; Spaeder, T.; Spalholz, B.; Tayyari, F.; Ubhi, B.; Verma, M.; Walk, T.; Wilson, I.; Witkin, K.; Bearden, D. W.; Zanetti, K. A. Towards Quality Assurance and Quality Control in Untargeted Metabolomics Studies. *Metabolomics* **2019**, *15* (1), 1–5. https://doi.org/10.1007/s11306-018-1460-7.

(5) Benton, H. P.; Want, E.; Keun, H. C.; Amberg, A.; Plumb, R. S.; Goldfain-Blanc, F.; Walther, B.; Reily, M. D.; Lindon, J. C.; Holmes, E.; Nicholson, J. K.; Ebbels, T. M. D. Intra- and Interlaboratory Reproducibility of Ultra Performance Liquid Chromatography-Time-of-Flight Mass Spectrometry for Urinary Metabolic Profiling. *Anal. Chem.* **2012**, *84* (5), 2424–2432. https://doi.org/10.1021/ac203200x.

(6) Martin, J. C.; Maillot, M.; Mazerolles, G.; Verdu, A.; Lyan, B.; Migné, C.; Defoort, C.; Canlet, C.; Junot, C.; Guillou, C.; Manach, C.; Jabob, D.; Bouveresse, D. J. R.; Paris, E.; Pujos-Guillot, E.; Jourdan, F.; Giacomoni, F.; Courant, F.; Favé, G.; Le Gall, G.; Chassaigne, H.; Tabet, J. C.; Martin, J. F.; Antignac, J. P.; Shintu, L.; Defernez, M.; Philo, M.; Alexandre-Gouaubau, M. C.; Amiot-Carlin, M. J.; Bossis, M.; Triba, M. N.; Stojilkovic, N.; Banzet, N.; Molinié, R.; Bott, R.; Goulitquer, S.; Caldarelli, S.; Rutledge, D. N. Can We Trust Untargeted Metabolomics? Results of the Metabo-Ring Initiative, a Large-Scale, Multi-Instrument Inter-Laboratory Study. *Metabolomics* **2015**, *11* (4), 807–821. https://doi.org/10.1007/s11306-014-0740-0.

(7) Välikangas, T.; Suomi, T.; Elo, L. L. A Systematic Evaluation of Normalization Methods in Quantitative Label-Free Proteomics. *Brief. Bioinform.* **2018**, *19* (1), 1–11. https://doi.org/10.1093/bib/bbw095.

(8) Chawade, A.; Alexandersson, E.; Levander, F. Normalyzer: A Tool for Rapid Evaluation of Normalization Methods for Omics Data Sets. *J. Proteome Res.* **2014**, *13* (6), 3114–3120. https://doi.org/10.1021/pr401264n.

(9) Fu, J.; Zhang, Y.; Wang, Y.; Zhang, H.; Liu, J.; Tang, J.; Yang, Q.; Sun, H.; Qiu, W.; Ma, Y.; Li, Z.; Zheng, M.; Zhu, F. Optimization of Metabolomic Data Processing Using NOREVA. *Nat. Protoc.* **2022**, *17* (1), 129–151. https://doi.org/10.1038/s41596-021-00636-9.

(10) Broadhurst, D.; Goodacre, R.; Reinke, S. N.; Kuligowski, J.; Wilson, I. D.; Lewis, M. R.; Dunn, W. B. Guidelines and Considerations for the Use of System

Suitability and Quality Control Samples in Mass Spectrometry Assays Applied in Untargeted Clinical Metabolomic Studies. *Metabolomics* **2018**, *14* (6), 1–17. https://doi.org/10.1007/s11306-018-1367-3.

(11) Kuhring, M.; Eisenberger, A.; Schmidt, V.; Kränkel, N.; Leistner, D. M.; Kirwan, J.; Beule, D. Concepts and Software Package for Efficient Quality Control in Targeted Metabolomics Studies: MeTaQuaC. *Anal. Chem.* **2020**, *92* (15), 10241–10245. https://doi.org/10.1021/acs.analchem.0c00136.

(12) Dogu, E.; Taheri, S. M.; Olivella, R.; Marty, F.; Lienert, I.; Reiter, L.; Sabido, E.; Vitek, O. MSstatsQC 2.0: R/Bioconductor Package for Statistical Quality Control of Mass Spectrometry-Based Proteomics Experiments. *J. Proteome Res.* **2019**, *18* (2), 678–686. https://doi.org/10.1021/acs.jproteome.8b00732.

(13) Dogu, E.; Mohammad-Taheri, S.; Abbatiello, S. E.; Bereman, M. S.; MacLean, B.; Schilling, B.; Vitek, O. MSstatsQC: Longitudinal System Suitability Monitoring and Quality Control for Targeted Proteomic Experiments. *Mol. Cell. Proteomics* **2017**, *16* (7), 1335–1347. https://doi.org/10.1074/mcp.M116.064774.

(14) Beger, R. D. . et al. Towards Quality Assurance and Quality Control in Untargeted Metabolomics Studies. *Metabolomics* **2020**, *15* (1), 4. https://doi.org/10.1007/s11306-018-1460-7.

(15) Evans, A. M. . et al. Dissemination and Analysis of the Quality Assurance (QA) and Quality Control (QC) Practices of LC-MS Based Untargeted Metabolomics Practitioners. *Metabolomics* **2021**, *16* (10), 113. https://doi.org/10.1007/s11306-020-01728-5.

(16) Liu, F. T.; Ting, K. M.; Zhou, Z. H. Isolation Forest. *Proc. - IEEE Int. Conf. Data Mining, ICDM* **2008**, 413–422. https://doi.org/10.1109/ICDM.2008.17.

(17) Kalisch, M.; Mächler, M.; Colombo, D.; Maathuis, M. H.; Bühlmann, P. Causal Inference Using Graphical Models with the R Package Pcalg. *J. Stat. Softw.* **2012**, *47* (11). https://doi.org/10.18637/jss.v047.i11.

(18) Kalisch, M.; Hauser, A.; Maathuis, M. H.; Mächler, M. An Overview of the Pcalg Package for R. **2020**, 1–48.

(19) Spirtes, Peter; Glymour, Clark; Scheines, R. *Causation, Prediction and Search*; 1993. https://doi.org/10.1007/978-1-4612-2748-9.

Chapter 3

Regularized adversarial learning for normalization of multi-batch untargeted metabolomics data

Andrei Dmitrenko, Michelle Reid and Nicola Zamboni

This manuscript was submitted for publication.

Author contributions:

MR collected the dataset to test and benchmark batch correction methods. AD developed the new method, applied it to multiple datasets, systematically compared to existing methods and ran ablation experiments to test its scalability and robustness. AD, NZ wrote the manuscript.

Abstract

Untargeted metabolomics by mass spectrometry is the method of choice for unbiased analysis of molecules in complex samples of biological, clinical, or environmental relevance. The exceptional versatility and sensitivity of modern high-resolution instruments allows profiling of thousands of known and unknown molecules in parallel. Batch effects constitute a common and unresolved problem in untargeted metabolomics, however, and can bias data analysis and hinder long-term studies. Here, we present a new method, Regularized Adversarial Learning Preserving Similarity (RALPS), for the normalization of multi-batch untargeted metabolomics data. RALPS builds on deep adversarial learning with a three-term loss function that mitigates batch effects while preserving biological identity, spectral properties, and variation coefficients. Using two large metabolomics datasets, we showcase the superior performance of RALPS as compared with six state-of-the-art methods. Our results demonstrate that RALPS scales well, delivers robust results, deals with missing values, and can handle different experimental designs.

Introduction

Metabolomics is the method of choice for chemical characterization of biological, clinical, and environmental samples. When the aim of the analysis is to monitor many and potentially unexpected analytes, the traditional untargeted approach is to scan the full mass and dynamic range with high-resolving instruments. This strategy allows monitoring of virtually all compounds that can be ionized and are sufficiently abundant. Eventually, untargeted metabolomics experiments result in semi-quantitative data for thousands of detectable features and many more unknowns. A largely unsolved problem of such large metabolomics experiments, however, is data normalization. The sheer number of features and the extreme sensitivity of liquid chromatography–mass spectrometry (MS) instruments to multiple factors mean that ion-specific temporal drifts, day-to-day variability, or batch effects are common in metabolomics analyses. These problems increase with number of samples.

The most common way to account and correct for these issues is to employ isotopically labeled internal standards, which are added at a fixed amount to all samples and calibration standards. If the standard's elution times and ionization behavior are identical to those of the compound of interest, the standard can be used to correct for linear matrix effects. This approach is quite effective with a limited number of metabolites, as in targeted metabolomics analyses, but does not scale to untargeted metabolomics. The first problem is the limited availability of heavy standards, and one workaround is to use a few representatives for each class and extrapolate over structurally similar compounds. The second problem is the limit on the number of standards that can be spiked without introducing novel matrix effects, i.e., when using standards available in salt form.

In the absence of internal standards to assess and correct for experimental variations, drifts, batch effects, and so on, normalization needs to operate on the resulting data. Notable examples of batch normalization methods are ComBat[1], based on empirical Bayes frameworks, and EigenMS[2], using singular value decomposition to estimate and correct for bias trends in the data. These approaches, as well as probabilistic quotient normalization[3] (PQN), have been reported as some of the best options for untargeted metabolomics studies, when applied in combination[4]. In 2019, an algorithm based on the wavelet transform with independent component analysis, WaveICA, was proposed and showed superior performance in large-scale untargeted metabolomics studies[5].

A current trend is to employ deep learning to correct for batch effects, and several examples of this approach are available in single-cell RNA sequencing[6–8]. In metabolomics, the related state-of-the art method is NormAE[9]. All of these methods rely on self-supervised representation learning to project input features into a latent space and then apply a mechanism of merging similar patterns in that space. NormAE relies on adversarial learning in which two neural networks, an autoencoder and a classifier, are trained simultaneously to reconstruct the data and classify batches, respectively. The ultimate goal is to reproduce the data but remove differences between user-defined batches, so the autoencoder is trained with a loss function that includes two terms. The first term awards correct reconstruction, and the second term applies a penalty if samples from different batches are classified correctly. By using the latter term, the autoencoder is pushed to learn the data representations that make any user-defined batches indistinguishable. Decoding the learned representations results in a normalization effect on the batch. NormAE is a powerful method that excels in removing batch effects but carries important drawbacks. For example, it takes positive values of ion intensities as input, but outputs arbitrary units, undermining interpretability. NormAE also requires identical pooled study samples across batches for validation, which is not available in many studies[4,10]. Furthermore, the method suffers from issues common to all practical applications of deep learning, such as the non-trivial parameter optimization, computational complexity, and lack of reproducibility[11].

Here we present a new normalization method, Regularized Adversarial Learning Preserving Similarity (RALPS), for untargeted metabolomics that efficiently addresses all of these problems. RALPS builds on adversarial learning but implements a novel three-term loss function that suppresses batch effects while preserving biological information. Using several test sets, we show that RALPS outperforms state-of-the-art methods in terms of performance, scalability, usability, and robustness.

Results

Method overview

RALPS was inspired by NormAE and uses an autoencoder and a classifier to mitigate batch effects (**Figure 1**). In addition to the classifier discrimination loss (L_d), we wanted to introduce a mechanism to preserve characteristic differences of any set of supposedly similar samples across the whole sequence as a way to preserve biological information. For this purpose, we added a new regularization term to the autoencoder loss function to reward tight clustering of reference samples in the embedded space (r_g). To evaluate clustering, RALPS flattens the multidimensional space by uniform manifold approximation and projection[12] (UMAP) and performs unsupervised hierarchical density-based cluster analysis[13] (HDBSCAN). Eventually, we score the clustering by counting how frequently reference samples of the same type co-occur in the same cluster.

Information on reference samples is provided by the user to indicate all sample groups that are supposed to be similar. These groups can include biological replicates, technical replicates, pooled study samples, spike-ins, and dilution series. There is no formal limit on the number of reference groups, and although sample groups should span multiple batches, the same reference samples do not need to be present in all batches. This flexibility builds considerable freedom into the experimental design, as discussed below.

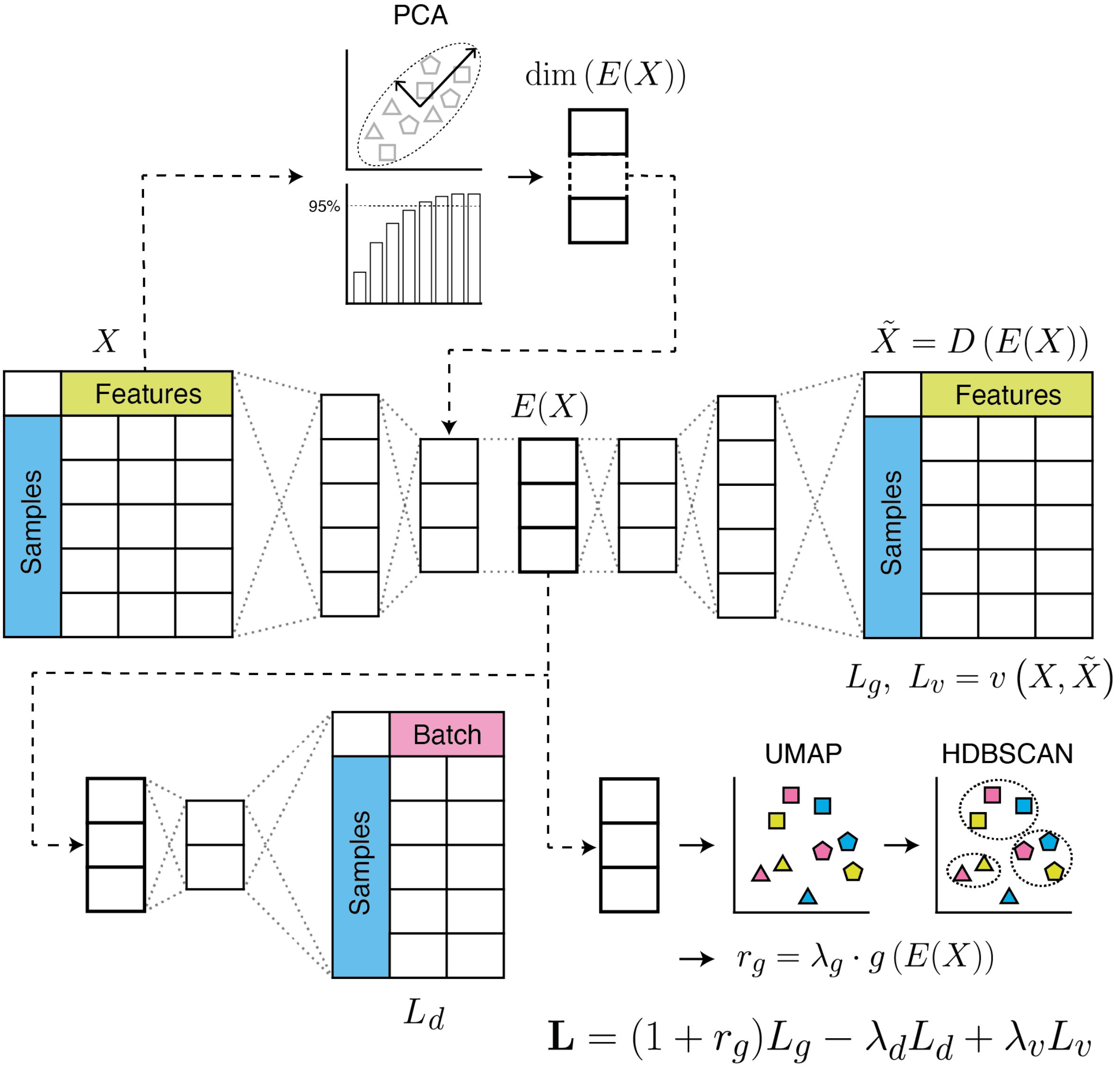

PCA
$\dim\left(E(X)\right)$
95%
X
Features
Samples
$E(X)$
$\tilde{X} = D\left(E(X)\right)$
Features
Samples
$L_g,\ L_v = v\left(X, \tilde{X}\right)$
Batch
Samples
UMAP
HDBSCAN
$\rightarrow r_g = \lambda_g \cdot g\left(E(X)\right)$
L_d
$\mathbf{L} = (1 + r_g)L_g - \lambda_d L_d + \lambda_v L_v$

Figure 1. Graphical overview of RALPS. The autoencoder takes measured data, *X,* as input and produces its reconstruction, *D(E(X))*, as output. The number of neurons in the bottleneck layer is determined by principal component analysis. The aggregated autoencoder loss **L** consists of three terms: the regularized autoencoder loss L_g, the classifier loss L_d, and the variation loss L_v with real coefficients λ_g, λ_d and λ_v, respectively.

Furthermore, we included a variation loss term (L_v) to encourage a decrease in batch variation coefficients (VCs). This addition was motivated by two observations. First, many normalization methods tend to inflate the VCs of replicate measurements. Second, deep learning models with encoder–decoder architectures are especially prone to producing outliers. Finally, we implemented several features aimed at improving scalability, usability, and robustness. First, RALPS adopts a flexible network architecture, in which the number of neurons in the model layers is automatically adjusted based on a principal component analysis. By default, RALPS sets the number of neurons equal to the number of principal components needed to describe at least 90% of the dataset variance. This modification eliminates many hyperparameters from the model and simplifies parameter optimization. Second, we introduced a randomized hyperparameter grid search and model selection logic, both of which enable finding multiple parameter sets that deliver top normalization results in an automated way. Third, we introduced input validation and a mechanism for early stopping to avoid collapsed normalization solutions that could arise because of inconsistent parametrization or increasing classifier loss. RALPS is implemented in Python programming language with a Torch deep learning framework. It requires a single configuration file containing the data and the batch information file paths, as well as a few other parameters to run normalization. RALPS is 100% open source.

Generation of a multi-batch benchmarking dataset

We faced the problem of finding suitable multi-batch datasets for testing and comparing normalization methods. Because such datasets are rare and often associated with clinical studies that preclude full disclosure and publication[9], we opted to generate a novel benchmarking dataset. We assembled a panel of 136 samples with a large variety in sample type, complexity, and concentrations (**Supplementary Figure 1**). To ensure a fair representation of complex samples, we prepared roughly half of them using human serum extracts (NIST SRM1950). We included spike-ins with selections of amino acids, fatty acids, and nucleobases, and a fully ^{13}C-labeled *E. coli* extract at different dilutions. The samples were distributed on two 96-well plates. Both plates were replicated several times and stored at –80 °C. The two plates were repeatedly analyzed in independent batches by untargeted metabolomics using flow injection–time-of-flight (TOF) MS on the same Agilent 6550 iFunnel Q-TOF system and in negative mode ionization[14]. The measurement of the 136 samples in technical triplicates and hundreds of intercalated blanks took ca. 12 hours. The same sequence was analyzed seven times over the span of about 2 months with time gaps of up to 3 weeks. Between batches, the instrument was used in different experiments and underwent routine maintenance as well as tuning procedures. From an analytical standpoint, all batches were acquired with an instrument that was operating normally, i.e., without any indication of problems that might affect measurement quality.

Each batch was analyzed independently. To focus on the subset of m/z peaks that possibly relate to metabolites, we selected features with m/z values that were matched to deprotonated compounds listed by the Human Metabolome Database (ver. 4.0, tolerance 0.001 Da). We intersected the putative peak lists obtained from each batch by m/z and retained those features that could be reproducibly detected. Low abundance ions with average intensity < 1000 counts were filtered. This procedure resulted in a consolidate data table with intensities for 170 putative deprotonated metabolites and 2856 files, divided in seven batches.

To visualize the extent of batch effects, we plotted the UMAP projections for all samples (**Supplementary Figure 2a**). We observed that samples grouped by batch, and almost no overlap between batches could be found. This result clearly indicates that batch effects dominate and confound chemically identical samples. Another way to quantify batch effects is to calculate the cross-correlation of identical samples within and between batches (**Supplementary Figure 2b**). Although Pearson correlation coefficients were >0.9 within the same batch, the distribution of inter-batch correlations shows a heavy tail, reaching values as low as $r = 0.4$, pointing to strong biases. Below, we use this dataset to benchmark RALPS against state-of-the-art approaches.

Normalization of the multi-batch dataset

As a first demonstration, we evaluated RALPS on our multi-batch benchmarking dataset. Initially, all sample labels were included as reference groups to maximize information available for training. We ran RALPS with default parameters and a randomized grid search of size 50. The best normalization solution selected by the model selection logic was compared to the initial dataset. First, we assessed the normalization effect of our method by highlighting batch labels on the UMAP embeddings plot (**Supplementary Figure 2d**). We observed that replicates from different batches were largely mixed for the normalized data compared with the initial data. The distribution of cross-correlations of all samples' replicates within and between batches shifted close to the ideal value of 1 (**Supplementary Figure 2e**). Moreover, VCs calculated for all intensities in every batch were consistently reduced in the normalized data (**Supplementary Figure 2f**) compared with the initial data

(**Supplementary Figure 2c**). The total runtime for the 50 independent runs summed to 291 minutes, i.e., about 6 minutes per run for a dataset with almost 3000 samples. Note that only a single CPU core was used for all computations. With an increasing number of CPU cores, it would be possible to evaluate much larger hyperparameter sets overnight, increasing the probability of finding an optimal normalization solution.

Normalization of the multi-batch dataset with limited reference groups

The test described above builds on a rather artificial scenario in which all samples are present in every batch. In practice, however, only a small subset of samples in each batch are repetitions of samples of different batches and can be used to correct for inter-batch effects. The most common scenario is inclusion of one or two reference samples in all batches. If a single sample is available, a frequent option is to spike in standards at a relevant concentration to ensure correct recovery of significant differences, including MS measurement and data processing. To mimic this realistic scenario, we attempted to normalize the benchmarking dataset with RALPS, using only a few reference sample groups each time. Different sets of groups were tested. For instance, Supplementary Figure 3 shows the normalization achieved by RALPS using exclusively an undiluted NIST1950 serum extract (P2_S_0001) and the same sample spiked with purines and pyrimidines (P2_S_PP_0001). The labels and relation of all remaining samples were neglected during training. For the normalized data, batches no longer appeared as isolated clouds of points, and instead, samples from different batches were well mixed. Qualitatively, the result was like the initial test in which all reference groups were provided during the training phase.

To demonstrate the flexibility of RALPS, we tested several combinations of up to four reference sample groups. For each combination, we applied RALPS with a randomized grid search of size 100. We used default parameters and set λ_v = 0 to loosen the constraints on the joint loss function optimized during training (**Figure 1**). We found several combinations of reference samples that produced good normalization effects (**Supplementary Table 1**). Comparing some of the evaluation metrics, using only undiluted NIST1950 serum extract (P2_S_0001) or using combinations of two or three samples yielded results similar to those from the examples described above (row #1 in **Supplementary Table 1**). As a negative control, we provide results obtained by training with a single reference group consisting of highly diluted fatty acids in water (row #10 in **Supplementary Table 1**). Because of matrix differences and the common presence of fatty acids in the background, these samples are unlikely to be sufficient to correct for complex batch effects, e.g., in serum samples.

All other tested configurations were comparably superior on all evaluation metrics, but not all were reproducible, as assessed by several repetitions of the training with identical reference groups but different random starts. Only half of the cases frequently reproduced comparably good results. For the other half that were not reproducible, including the case with undiluted serum as a unique reference group, multiple independent attempts would be necessary. Nevertheless, all the reference group configurations performed well in mixing batches on the UMAP embedding plot (**Supplementary Figure 4**). These results prove that RALPS is flexible in the choice of reference samples and, in principle, a few reference sample groups in triplicate can suffice.

Next, we compared RALPS to the state-of-the-art normalization approaches mentioned in the introduction. Of the several training scenarios listed above, RALPS was trained using two reference sample groups with undiluted serum and the derivative with spiked purines and pyrimidines (**Supplementary Figure 3**). The performance was evaluated using one qualitative and three quantitative criteria. The quantitative criteria included (i) cross-correlation of all samples' replicates, (ii) batch VCs, and (iii) percent of features for which the VC across replicate samples increased upon normalization. The qualitative criterion was the spectrum of the normalized data compared with that of the initial data (**Figure 2**).

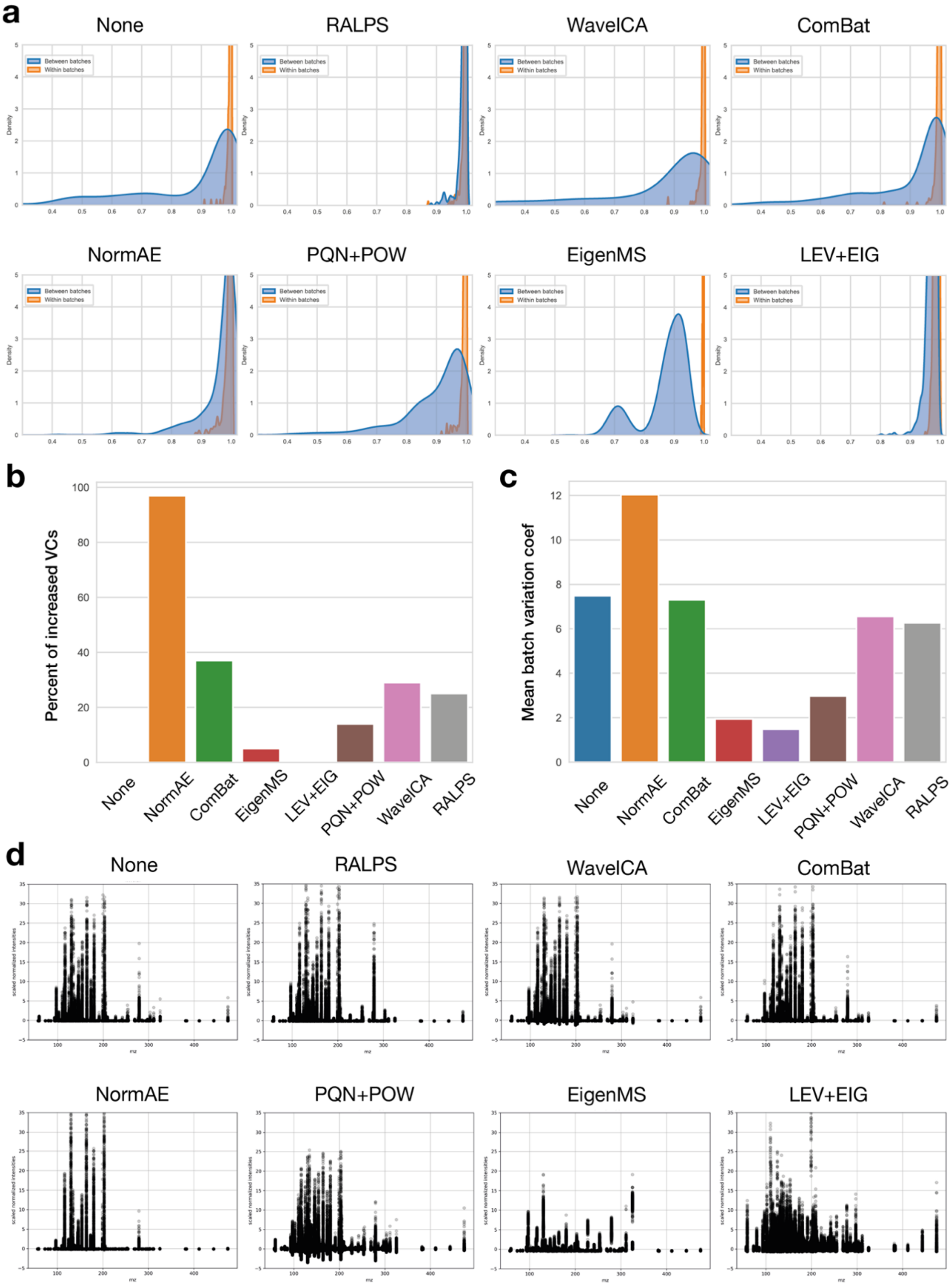


Figure 2. **Comparison of methods for the benchmarking dataset.** (**a**) Distributions of cross-correlation of reference samples between and within batches. (**b**) Full spectra. (**c**) Percent of samples with increased variation coefficients. (**d**) Mean batch variation coefficients.

In summary, RALPS was the only method among the seven tested approaches to excel in the suppression of batch biases while controlling for mean batch variance, replicate VCs, and drastic spectral transformation.

We found that three of seven methods improved the cross-correlation of replicates among batches: RALPS, NormAE, and LEV+EIG (**Figure 2a**). In Supplementary Figure 5, the effect of normalization on cross-correlation is shown in detail for one of the samples that featured prominent bias effects, i.e., P1_FA_0008. This sample was an 8-fold–diluted fatty acid mix that was strongly affected by the common background signal of fatty acids. For this example, RALPS and LEV+EIG achieved the best results, with correlation coefficients > 0.9. In the case of RALPS, a small bias between the first batch and all others remained. A different picture was obtained in terms of improving mean batch VCs. All methods except NormAE resulted in reduced mean batch variance (**Figure 2b**). However, LEV+EIG, EigenMS, and PQN+POW each resulted in a drastic drop that suggests a degenerated solution. In practice, in the attempt to reduce batch effects, these methods also may attenuate biological differences and obscure the information of interest.

The third quantitative evaluation criterion highlights a frequently overlooked caveat of normalization procedures: the fraction of sample whose VC increases by 5% or more upon normalization. The expectation is that normalization removes inter-batch biases without compromising intra-batch reproducibility. However, an apparent improvement in inter-batch reproducibility can also be attained by drastically worsening intra-batch precision, generally confounding all data and biological information. We again found the best results with EigenMS, LEV+EIG, and PQN+POW (**Figure 2c**). NormAE, meanwhile, inflated the VC in > 95% of the cases. RALPS fell in the middle average of the methods and outperformed ComBat and WaveICA. The striking difference between NormAE and RALPS highlights the relevance of the additional term L_v that was introduced in the loss function to control for an increase in variance.

Lastly, we examined the MS spectra obtained by normalization. From an analytical standpoint, the expectation is that normalization will have a minor effect on the overall spectrum. In reality, this is the case only if the ranking of metabolites based on average intensities is roughly maintained. In contrast, a normalization procedure that has a large effect on the value ranges of features will result in a qualitatively different spectrum. Spectra preservation is particularly important if calibrants are included to estimate concentrations. By visual inspection of the resulting spectra (**Figure 2d**), we indeed observed that using methods with the lowest batch VCs (LEV+EIG and EigenMS) completely altered the data. NormAE preserved high-intensity peaks but suppressed low-intensity features, which is in agreement with the previously reported issue of sensitivity to outliers[9]. PQN+POW yielded the opposite pattern and amplified low-abundance ions. RALPS, WaveICA, and ComBat had the most neutral impact on the spectra, preserving both high- and low-intensity features. Of note, WaveICA and ComBat, but not RALPS, can output negative ion intensity values, which can result in complications in downstream data analysis and interpretation.

RALPS corrects biases on multi-batch cancer cell metabolomics data

We further tested the performance of RALPS with data published by Cherkaoui et al.[15]. This dataset includes >1400 untargeted metabolomics measurements for a panel of ca. 180 cancer cell lines, resulting in a matrix with relative abundances for 1817 putative metabolite ions. This dataset is interesting for three reasons. First, it represents real-life, mid-sized untargeted metabolomics studies. Second, its batch effects are associated with sample preparation and not with sample acquisition, as in the benchmarking dataset. This association is present because the limiting step on the study was sample generation and not MS analysis. Because of the tedious procedures necessary to cultivate numerous cell lines in parallel, the entire study was divided into seven batches of samples generated over the span of about a year. Upon preparation, cell pellets were stored at –80 °C, and when the seven sampling batches were complete, all samples were prepared and subjected to sequential MS analysis. The expected batch effects thus are dominated by shifts in cultivation conditions (e.g., media, incubation conditions, handling). The third aspect of interest is that the study did not use a set of reference samples that were included in all batches. Only two cell lines (MDAMB231 and MCF7) were present in multiple batches (5 and 4, respectively) and were used as reference samples in the training phase of RALPS.

All methods seemingly improved cross-correlation of the MDAMB231 sample, except for PQN+POW (**Supplementary Figure 6a**). Additional odd results included a mean batch variance that dropped to almost 0 for LEV+EIG and almost doubled for NormAE

(**Supplementary Figure 6d**). In line with the previous results, spectra were altered by LEV+EIG, EigenMS, PQN+POW, and NormAE. Overall, RALPS and WaveICA were the best methods for normalizing the data from Cherkaoui et al. ComBat was also a good alternative, but the final cross-correlations of MDAMB231 across samples were worse, and more control samples would be needed for more precise conclusions. Among the three best methods, RALPS also appeared to better preserve low-intensity data. ComBat and WaveICA produced negative values that affected ca. 1% of the samples (**Supplementary Table 2**). The normalized negative values exceeded 10^6 counts, resulting in obvious complications for further analysis and interpretation. In contrast, RALPS produced strictly positive values.

Finally, we compared the evaluation times of all approaches (**Supplementary Table 3**). A single run of RALPS took about 3 minutes. A randomized grid search with 50 samples required about 2-3 hours, which was similar to the time required by the slowest algorithms such as NormAE (even using a GPU), EigenMS, or LEV+EIG. However, the total processing time with RALPS can be easily reduced to less than an hour by employing four or more CPUs. We conclude that RALPS, ComBat, and WaveICA showed overall competitive performance on the Cherkaoui et al. data. However, RALPS is the only method that combines top normalization performance with minimum biological information loss and preservation of spectral properties. A closer look at the UMAP embeddings of the normalized data produced by RALPS reveals that virtually all divisions across batches were successfully removed (**Figure 3**).

One important issue is the striking difference between NormAE and RALPS, which are both based on adversarial learning but yielded opposing results. In the Cherkaoui et al. data, NormAE reproducibly creates a collapsed solution driven by the growing classifier loss (**Supplementary Figure 7**). This happens when the batches are already fairly mixed in the initial data and the classifier fails to tell them apart. The classifier loss grows and keeps contributing to the joint loss function of the autoencoder, which ultimately leads to a singular output matrix. In RALPS, this degenerate behavior is prevented by the embedded early stopping.

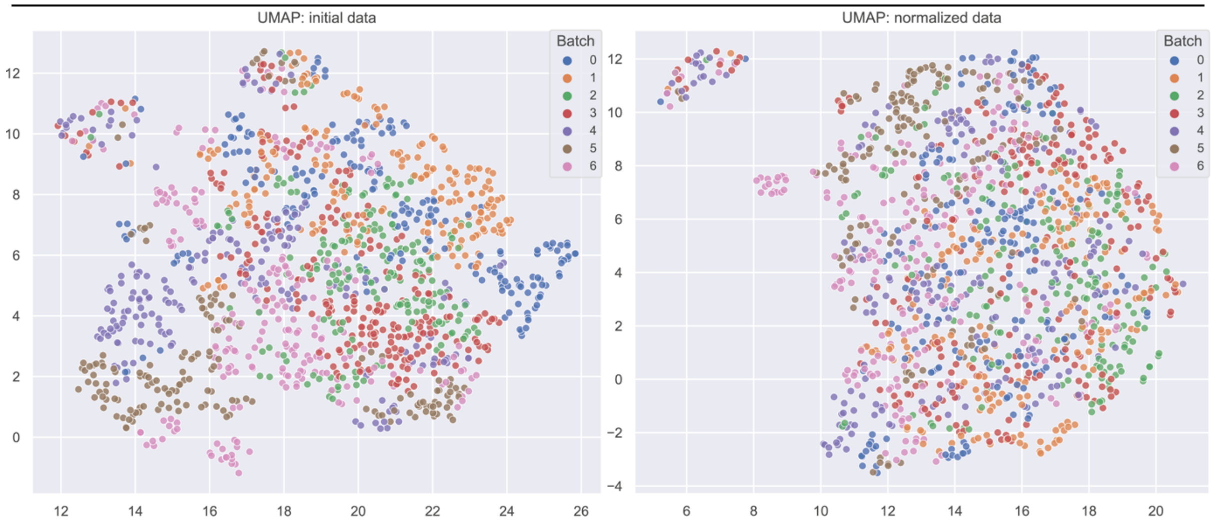


Figure 3. UMAP embeddings for the cancer cell lines dataset by Cherkaoui et al. Initial (left) versus normalized with RALPS (right) data is presented.

Top performance requires three-term loss function

After demonstrating the performance of RALPS in practice, we set out to investigate in more detail how the regularization terms in the composite objective function (i.e., **L** in **Figure 1**) impact the outcomes. For this investigation, we used the benchmarking dataset and two reference sample groups (as in **Supplementary Figure 3**) and tested four scenarios of objective functions with 0 to 3 regularization terms (**Figure 4a**). For example, the case of $\lambda_d = \lambda_g = \lambda_v = 0$ corresponds to an autoencoder that is solely trained to reconstruct the data without any penalty. The last scenario with non-zero λ_d, λ_g, and λ_v reflects the default architecture with all regularization terms. For each scenario, we applied RALPS with default parameters and performed a randomized grid search of size 100, selected the 10 top-performing normalization solutions, and compared their key metrics (Figure 4a). As expected, introducing $\lambda_g > 0$ improved tight clustering of reference samples in the embedded space. Introducing a penalty for the batch classifier with $\lambda_d > 0$ had only subtle effects on tight clustering and replicate cross-correlation. Clearly, top performance on all three criteria shown is achieved only in combination with $\lambda_v > 0$. These results indicate that the three terms in the composite objective function L of RALPS have a synergetic effect on batch normalization.

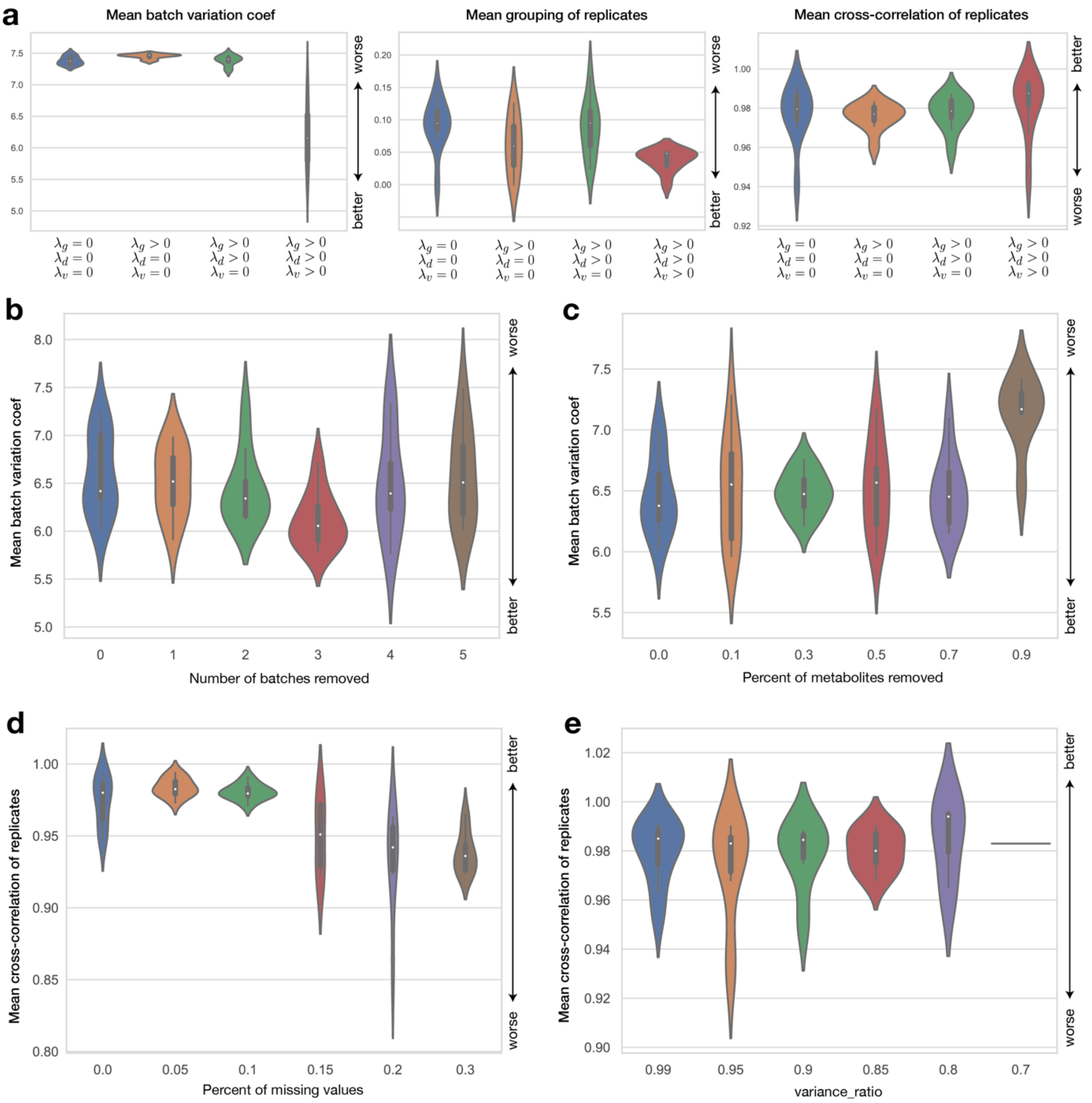


Figure 4. **Ablation experiments.** (**a**) Impact of regularization terms. (**b**) Number of batches removed. (**c**) percent of metabolites removed. (**d**) percent of missing values (**e**) and different values of *variance_ratio* parameter.

RALPS is robust against data ablations

Missing values represent a particular challenge for normalization of untargeted metabolomics data. Values are missing when they are undetectable in a subset of samples, as is common for rare compounds, such as derivatives of drugs or special food additives in human serum, or low concentration compounds near the analytical limits of detection. To verify the robustness of RALPS, we performed multiple ablation experiments. All tests were done on the benchmarking dataset using default parameters and the same two reference sample groups as above. First, we ran RALPS on subsets of the initial dataset. We shrank the dataset batch-wise from seven to two, i.e., down to 29% of the samples. We observed that the mean batch VC was generally constant and comparable to the full data case (**Figure 4b**). We then shrank the dataset by removing random metabolites down to 10% of the initial number. Only in the last point, corresponding to 17 metabolites, did the mean batch variance tangibly worsen (**Figure 4c**).

Next, we tested the robustness to randomly distributed missing values. A user-defined parameter (*min_relevant_intensity*) instructs RALPS on the lowest value to consider. The default is 1000 counts, and missing values are replaced with this minimum value. In our test, we replaced 0%, 5%, 10%, 15%, 20%, and 30% randomly picked values in the data matrix with the default minimum value and then trained RALPS. We observed a decreasing trend in cross-correlation of replicates as the fraction of missing values went up (**Figure 4d**). However, the decrease in mean correlation coefficients was marginal (from 0.98 to 0.94 with 30% missing values), indicating that RALPS is generally robust against missing values.

Finally, we verified how the number of neurons in model architectures affects normalization results. By default, RALPS uses as many neurons as number of principal components necessary to explain at least 90% of variance. We tested multiple values ranging from 99% to 70% variance. We did not observe any clear trends in any of four metrics used to evaluate and compare methods. The mean cross-correlation of replicates stayed above 0.98 in all cases (**Figure 4e**).

Choice of clustering algorithm

One of the key cornerstones of RALPS is preserving similarity of reference samples during the adversarial training loop. To assess this in each epoch, the latent representations of the samples are projected with UMAP and clustered with the HDBSCAN algorithm. As many other clustering approaches exist[16,17], some could be better suited than others for a particular dataset. To illustrate the impact of the clustering method on the benchmarking dataset with two reference sample groups, we tested the performance of RALPS with five alternative clustering methods (**Supplementary Figure 8**). On all of our target metrics, HDBSCAN compared favorably with the other methods. MeanShift[18] delivered slightly better overall results but required almost twice the computational time. Because the clustering is part of the training and repeated at each epoch, the impact on the overall normalization time is substantial. Hence, we selected HDBSCAN as the default method. For the tested dataset, UPGMA[19] and BIRCH[20] were also very competitive because they combined speed with better scores for grouping of the replicates of the reference samples in the embedded space. These results might vary with different datasets, and we recommend that users seeking the best performance test different approaches. Of note, six clustering methods are integrated into RALPS, and any one of them can be reconfigured as default.

Discussion

We introduce RALPS, a novel batch normalization method based on regularized adversarial learning for untargeted metabolomics data. In this work, we demonstrated its performance on two representative datasets with thousands of samples or spectral features. The benchmarking dataset was generated to test the algorithm on MS data produced over several months. In the case of the cancer cell line data by Cherkaoui et al., batches instead were associated with cultivation and sampling of samples over the span of almost a year, whereas MS analysis was done sequentially with all samples. We demonstrated that RALPS outperformed other state-of-the-art methods on several key metrics. RALPS offers additional features such as adaptive network architectures, embedded hyperparameter optimization, automated model selection, and input validation. Together, these features convey flexibility, scalability, usability, and robustness as confirmed by testing with different configurations of reference samples and in ablation experiments.

Historically, the loss function with three terms embedded in RALPS evolved from multiple tests we performed with several datasets, some of which are not described here. The classification term was initially replicated from NormAE and is the component that drives the removal of batch effects. However, a deeper analysis of the

resulting normalized data revealed novel problems that prompted us to introduce additional terms. The observed increase in VCs for supposedly identical samples (e.g., replicates) upon normalization is common to most methods and is a particularly acute problem for NormAE. To our best knowledge, this issue has not previously been acknowledged or addressed. We offer two hypothetical explanations for this gap. The first is the use of mean absolute error for the reconstruction loss, which is sensitive to outliers. The second is that generally increasing the noise, and thus the VC, makes it more difficult for the classifier to separate batches. Hence, there is an apparent beneficial effect on batch discrimination, but detrimental side effects in downstream analyses. Introducing the variation loss L_v mitigated but did not abolish the problem.
A closer investigation of the RALPS results revealed that the increase in VC arose from a small set of samples (1.2%–1.8% of the total for the tested datasets). Filtering these samples by outlier detection (explained below) reversed the increase in VCs for RALPS and EigenMS, but not for NormAE, ComBat, WaveICA, and PQN+POW (**Supplementary Figure 9**), suggesting that the latter methods produced even more outliers. We do not advocate for outlier detection and removal or a particular approach for it, but we recommend that researchers using batch normalization methods carefully evaluate sample-wise VCs and consider correcting them before performing downstream statistical analysis.

The grouping regularization term r_g was the second novel addition to the adversarial training and the key to preserving similarity of supposedly equal samples. Although RALPS relies on clustering of reference samples in the embedded space to assess grouping, alternative approaches could be considered. For example, r_g could be calculated from distances in the latent space. Distance-based metrics would carry several hypothetical advantages and pitfalls. Owing to the fast computation, training would be tangibly shorter than with clustering. We expect that distance-based metrics would perform better with data characterized by subtle batch effects in which clustering fails to separate reference samples. Even in the case of all replicates falling into a single cluster (which happens naturally when most batch effects have been already removed), minimizing distances between all pairs of replicates would still have a regularization effect, whereas clustering would not. Among the potential drawbacks, we would expect an increased sensitivity to single outliers. Alternative paradigms for similarity preservation should be tested in the future.

In terms of experimental design, RALPS requires reference samples across batches, but it is not strictly necessary to have the very same reference samples present in all batches. We illustrated this flexibility in our normalization of the Cherkaoui et al. data, in which two reference sample groups were present in only four and five of the seven batches, respectively. This feature leaves considerable freedom for the experimental design, in particular for the number of replicate groups to include. Based on our tests with the benchmarking dataset and other studies not shown here, using replicates from a pooled study sample in each batch is generally sufficient to correct for typical batch biases arising from untargeted metabolomics measurements performed over different days. If the total number of samples does not become prohibitively large, including a second group of reference samples is beneficial. In this scenario, the recommendation is to spike in a reduced set of compounds of interest and at low concentrations to present a realistic challenge for calculating the grouping term r_g. In contrast, blanks do not constitute a good control group because they can readily be distinguished from all other samples and are ineffective in the training process. The flexibility of RALPS allows it to be tested on existing datasets or applied in experiments that were not specifically designed with its use in mind. In all cases, RALPS is developed to make full use of all available reference samples in correcting multi-batch experiments.

In untargeted metabolomics, batch effects are only one class of possible biases. Two further classes of problems exist: sample-to-sample variability and temporal drifts associated with MS detection. In case such problems emerge within individual batches, we recommend a two-step procedure. First, all single batches should be corrected individually to maximize coherence within each individual batch. Second, RALPS should be applied to harmonize data across batches.

Finally, we see no fundamental problem in using RALPS to normalize any kind of tabular data consisting of samples coming from different batches and with features characterizing the samples. Information about the similarity of samples (for references and beyond) can be encoded via groups in the batch information file. We encourage researchers from other omics fields to challenge RALPS with their own data and drive its further development by reporting experience and issues in the repository.

Methods

Loss functions and optimizers

RALPS uses mean square error loss for the autoencoder and cross-entropy loss for the classifier. Mean square error was preferred over mean absolute error because of faster convergence and better reconstruction quality when trained without the classifier. Consequently, RALPS does not use pretraining epochs; both model networks are trained in turns from the start. Adam optimizer is used for both.

Regularization term preserving similarity of samples

We propose a regularization term r_g to preserve the similarity of reference samples while running an adversarial training loop. This term explicitly takes into account the grouping of reference samples across batches in the embedded space (**Figure 1**). At every training epoch, the representations of the input data are clustered, and the grouping coefficient for each reference sample is computed based on how many replicates of the sample across batches appear to be in the same cluster. The regularization term penalizes the reconstruction loss ($r_g > 0$) when the replicates of the reference samples happen to be in different clusters. Conversely, the regularization cancels out ($r_g = 0$) when reference samples across batches are clustered together. By design, the regularization term r_g is limited from above ($r_g \leq 1$). Before clustering, RALPS uses the UMAP algorithm to generate embeddings of the learned representations. HDBSCAN is the default clustering algorithm parameterized with the number of batches times the number of replicates in the data as the minimum cluster size, and not allowing single clusters.

Adaptive network architecture

Although the backbones of the architectures for the autoencoder and the classifier are fixed, the number of neurons in some of their layers is automatically adjusted to the data. Principal component analysis is used to find the number of components that capture a user-defined percent of variance (selected randomly among 90%, 95%, or 99% by default). The number determines the dimensionality of the autoencoder bottleneck layer and the classifier input layer. In this way, RALPS adapts to the datasets of different sizes and variances.

Randomized grid search of hyperparameters

A few hyperparameters such as learning rates, regularization coefficients, batch size, and minimal variance ratio for principal component analysis are required to run RALPS. If the user does not provide them, they are sampled randomly from the predefined ranges. We also implemented ways for users to define their own ranges of hyperparameters for sampling. Every unique parameter set is assigned its own unique ID that is used to name a directory in the file system, where all the corresponding results will be stored. The user also provides the size of the randomized grid (defaults to 1, a single run). Beyond exploring the hyperparameter space and finding the best normalization solutions, such implementation allows for convenient testing of model robustness. Fixing training hyperparameters and setting the randomized grid size to values >1 facilitates investigations of how randomization during training affects performance of a particular parameter set.

Model selection

Several metrics are tracked during the training process to evaluate model performance and select the best model: (i) grouping coefficient calculated on the clustering results, (ii) statistic reflecting cross-correlation of replicates across batches, (iii) mean batch VC, and (iv) percent of samples with increased VCs. These metrics are calculated at each epoch for the two sample types: the reference samples used for r_g regularization, and for benchmarking samples if specified by the user. Defining benchmarking samples is optional but may be useful to verify that the normalization solution is not overfitted to the reference samples.

Within a single run, the best model (epoch) is supposed to achieve low grouping coefficient of reference samples, high cross-correlation, reduced mean VC of the reference samples, and reduced mean batch VC. The best epoch is selected after the training is complete based on the reference samples and the overall reconstruction loss. First, the epochs with 10% of the lowest grouping coefficients are selected. Among these epochs, the 10% of the highest cross-correlation is selected. Among the remaining epochs, the best one corresponds to the minimal reconstruction loss. To avoid the risk of exploding classifier loss and the over-normalization problem, we implemented an early stopping criterion that interrupts training if the classifier loss starts continuous growing. Such models are marked as stopped early. An exemplary run is illustrated in **Supplementary Figure 10**.

Model selection for the randomized grid search applies to the best epochs selected previously for each parameter set. The best models across all parameter sets are selected using a similar logic:

- The lowest mean batch VC and the highest cross-correlation values.
- The lowest grouping coefficient and the highest cross-correlation values.
- The lowest grouping coefficient and the lowest percent of increased VCs.

The models corresponding to these groups are combined and sorted by the reconstruction loss. The first 10 models are marked as the best in the randomized grid search output file containing all metrics of all evaluated parameter sets. Based on tests with multiple data sets, we recommend using a *grid_size* ≥ 50 to have high chances of finding a (nearly) optimal normalization solution. This heuristic model selection strategy proved to be sufficient to pick the models of the best normalization effects in all applications presented in this study. However, RALPS keeps all model results so that the user can select a different solution, based on a combination of quantitative and qualitative criteria available for each model.

Outlier detection

In the attempt to control for increased VCs, we adopted a version of a boxplot outlier removal approach. For each sample in the normalized data, we removed metabolites with intensities below $Q1 - \alpha \cdot IQR$ or above $Q3 + \alpha \cdot IQR$, where $Q1$ is the 25th intensity percentile, $Q3$ is the 75th percentile, IQR is the interquartile range, and α is a parameter (defaults to 1.5 for a classical boxplot). The script is available in the code repository.

For each application of RALPS described earlier in this paper, we selected α such that only 1% of samples had increased VCs. After filtering the normalized data, we calculated how many metabolites were dropped and then applied exactly the same procedure to the data normalized by other methods to compare their propensity for generating outliers.

Data availability

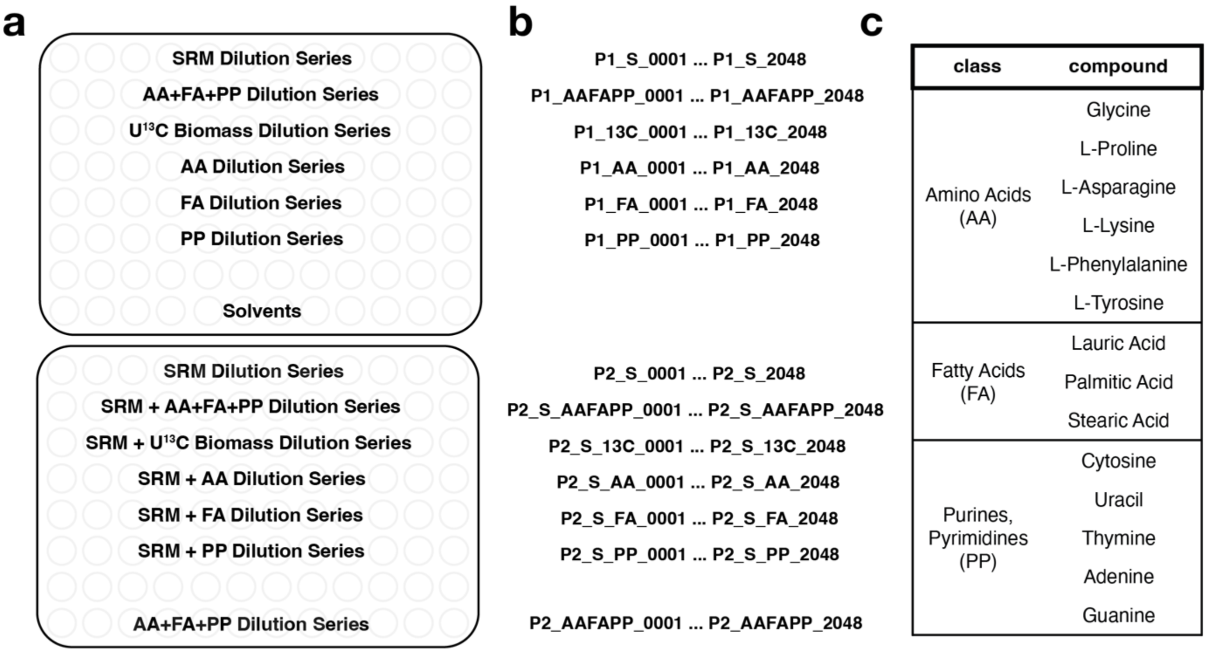
The full multi-batch benchmarking dataset is available at https://doi.org/10.3929/ethz-b-000545373. The filtered data table used in our experiments is provided with the code. The full cancer metabolomics dataset by Cherkaoui et al. is available at https://doi.org/10.3929/ethz-b-000511784.

Code availability

The source code and the description of input parameters, as well as examples of the input files, are available at https://github.com/zamboni-lab/RALPS.

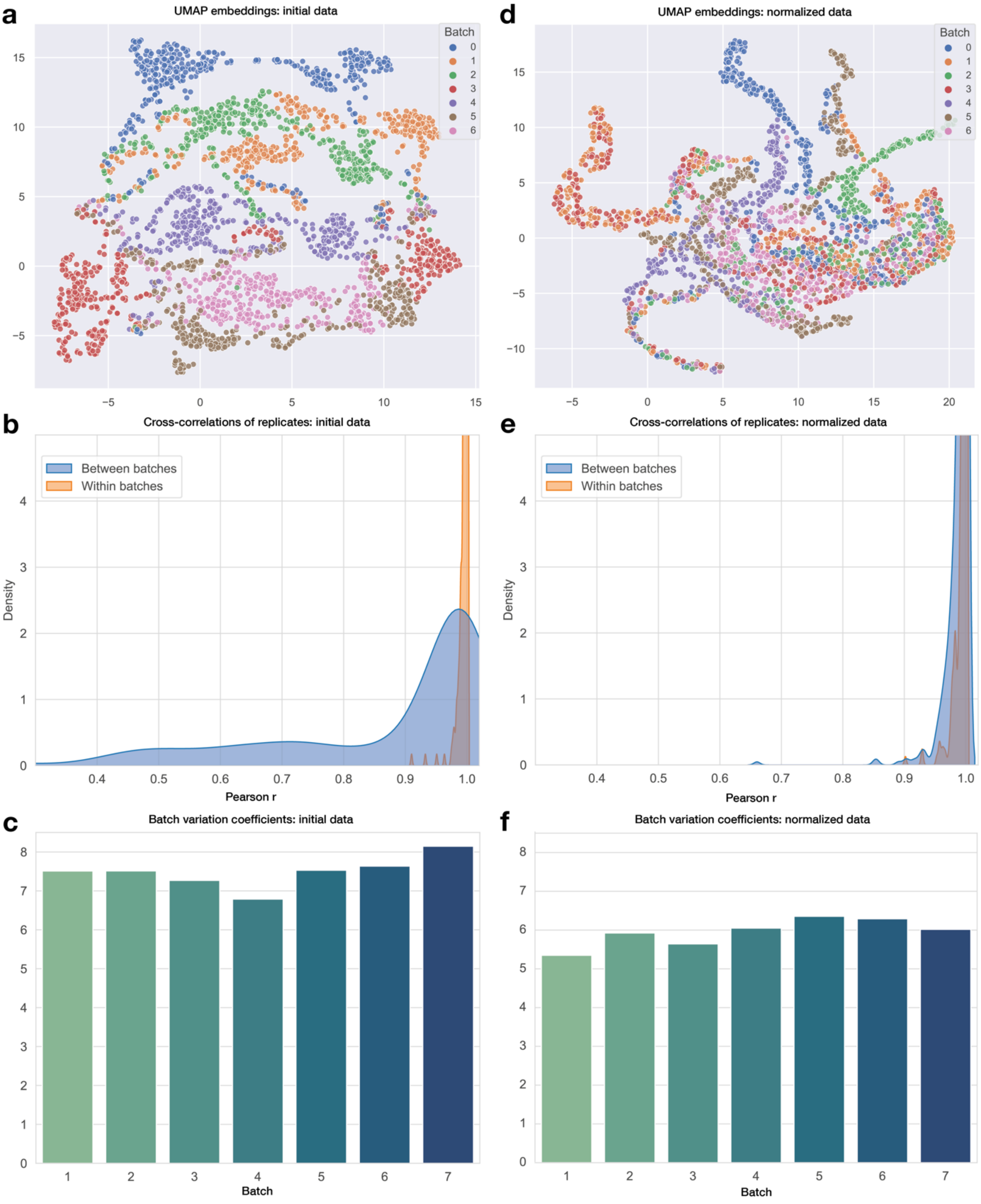
Supplementary material

Supplementary Figure 1. Description of the benchmarking dataset. (**a**) The design of plates 1 (top) and 2 (bottom). (**b**) The labels used for the samples in the benchmarking dataset throughout the text. (**c**) The corresponding compound classes of the spike-ins. Plasma NIST SRM1950 material is abbreviated as SRM. “0001” corresponds to an undiluted sample; further 2x dilutions go up to 2048 for each sample.

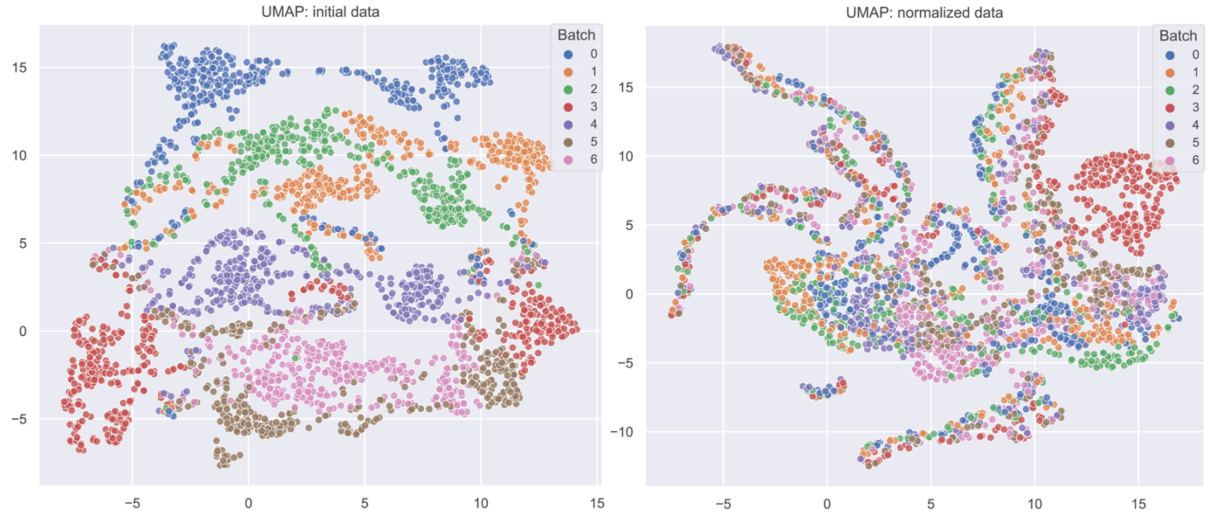

a
SRM Dilution Series
AA+FA+PP Dilution Series
U^{13}C Biomass Dilution Series
AA Dilution Series
FA Dilution Series
PP Dilution Series
Solvents
SRM Dilution Series
SRM + AA+FA+PP Dilution Series
SRM + U^{13}C Biomass Dilution Series
SRM + AA Dilution Series
SRM + FA Dilution Series
SRM + PP Dilution Series
AA+FA+PP Dilution Series
b
P1_S_0001 ... P1_S_2048
P1_AAFAPP_0001 ... P1_AAFAPP_2048
P1_13C_0001 ... P1_13C_2048
P1_AA_0001 ... P1_AA_2048
P1_FA_0001 ... P1_FA_2048
P1_PP_0001 ... P1_PP_2048
P2_S_0001 ... P2_S_2048
P2_S_AAFAPP_0001 ... P2_S_AAFAPP_2048
P2_S_13C_0001 ... P2_S_13C_2048
P2_S_AA_0001 ... P2_S_AA_2048
P2_S_FA_0001 ... P2_S_FA_2048
P2_S_PP_0001 ... P2_S_PP_2048
P2_AAFAPP_0001 ... P2_AAFAPP_2048
c

class	compound
Amino Acids (AA)	Glycine
	L-Proline
	L-Asparagine
	L-Lysine
	L-Phenylalanine
	L-Tyrosine
Fatty Acids (FA)	Lauric Acid
	Palmitic Acid
	Stearic Acid
Purines, Pyrimidines (PP)	Cytosine
	Uracil
	Thymine
	Adenine
	Guanine

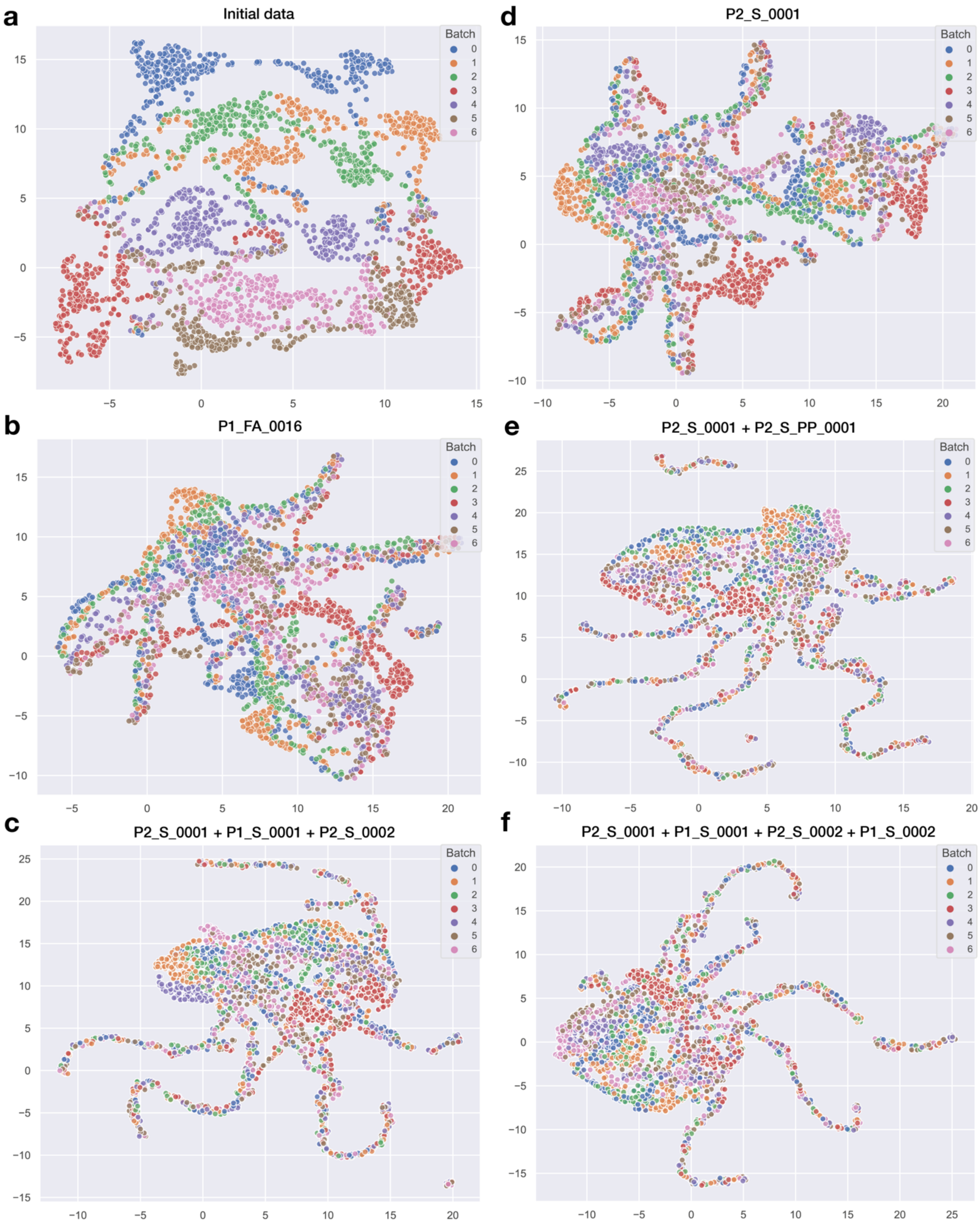
Supplementary Figure 2. Application to the benchmarking dataset using all study samples as references. (**a, d**) UMAP embeddings; (**b, e**) distributions of cross-correlations; (**c, f**) and mean batch variation coefficients. Left panels refer to the initial data, right panels to the normalized data produced by RALPS.



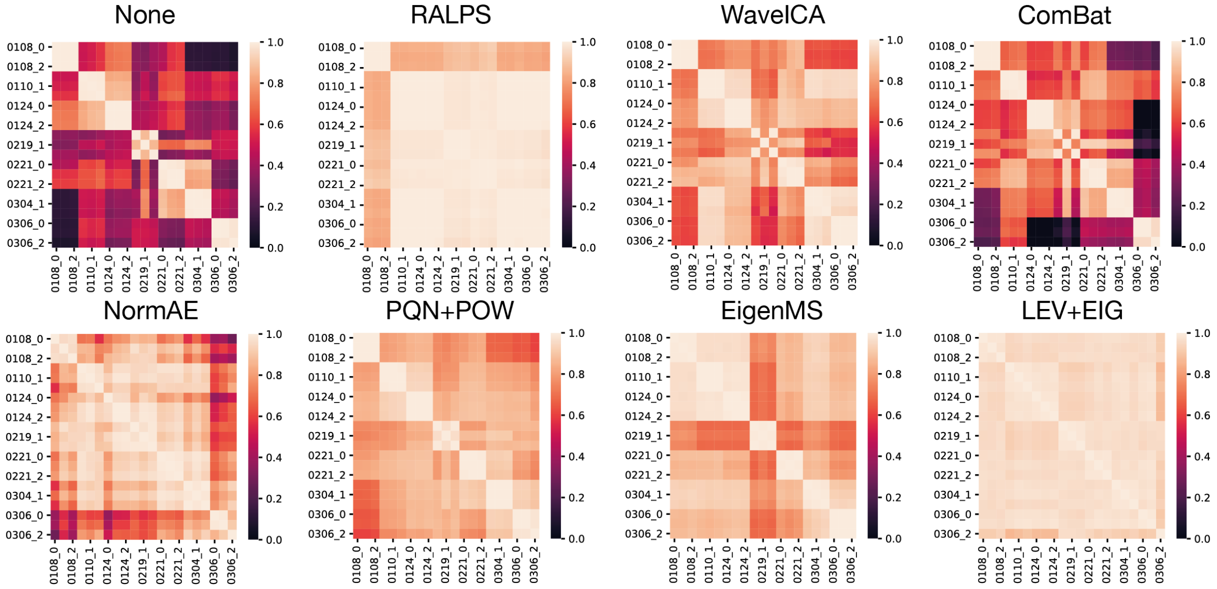
Supplementary Figure 3. UMAP embeddings of the initial (left) and normalized (right) data using P2_S_0001 and P2_S_PP_0001 as reference samples.



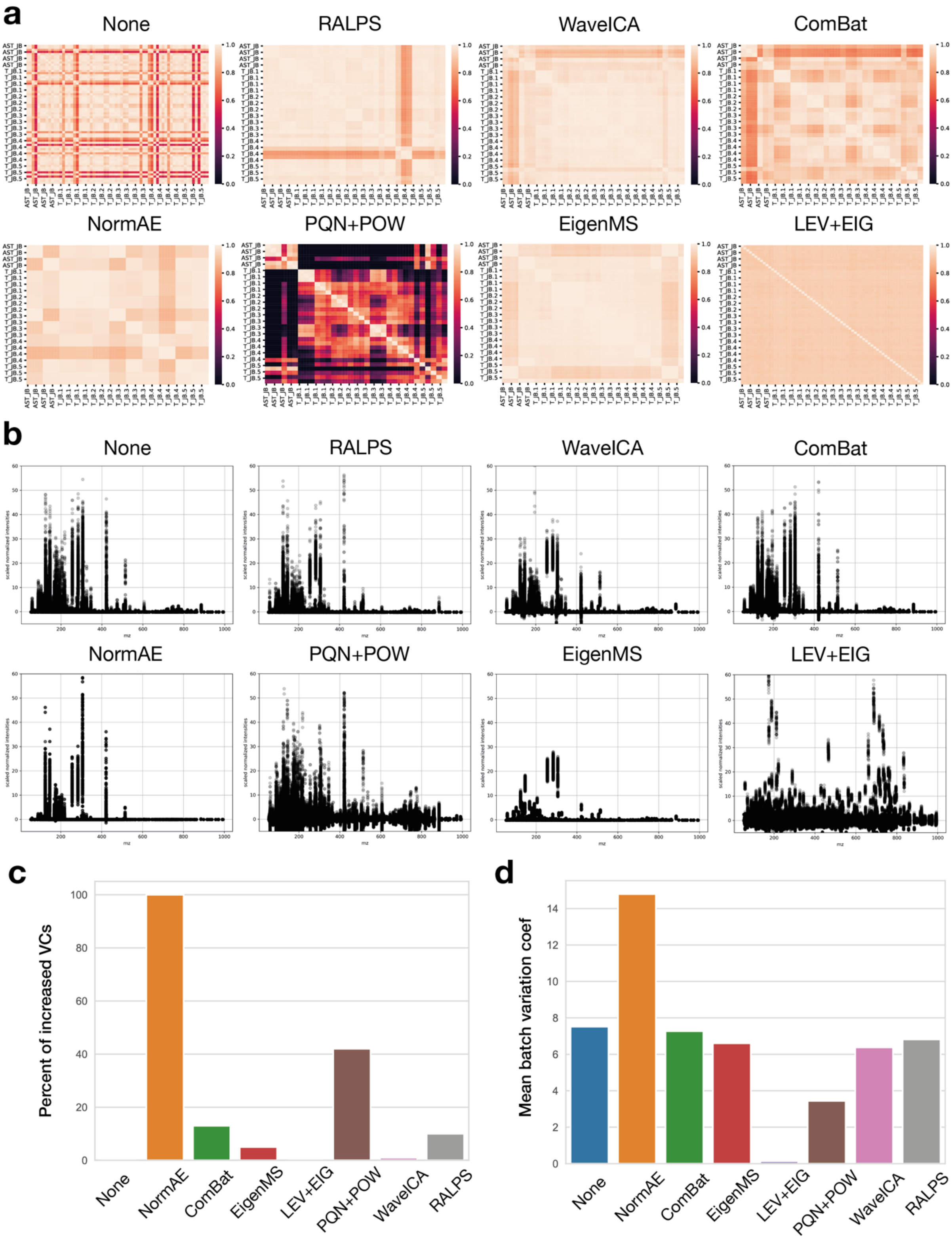
Supplementary Figure 4. (**a**) UMAP embeddings of the initial data; (**b-f**) UMAP embeddings of the data normalized with reference samples groups listed in **Supplementary Table 1**.



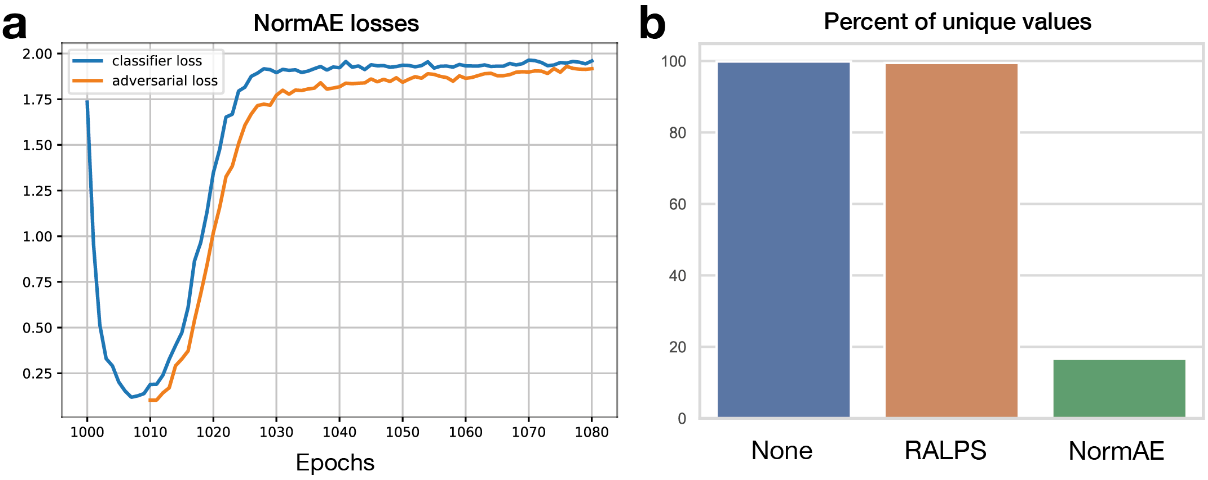
Supplementary Figure 5. Cross-correlation heatmaps for replicate samples of fatty acids sample with 8x dilution (P1_FA_0008).



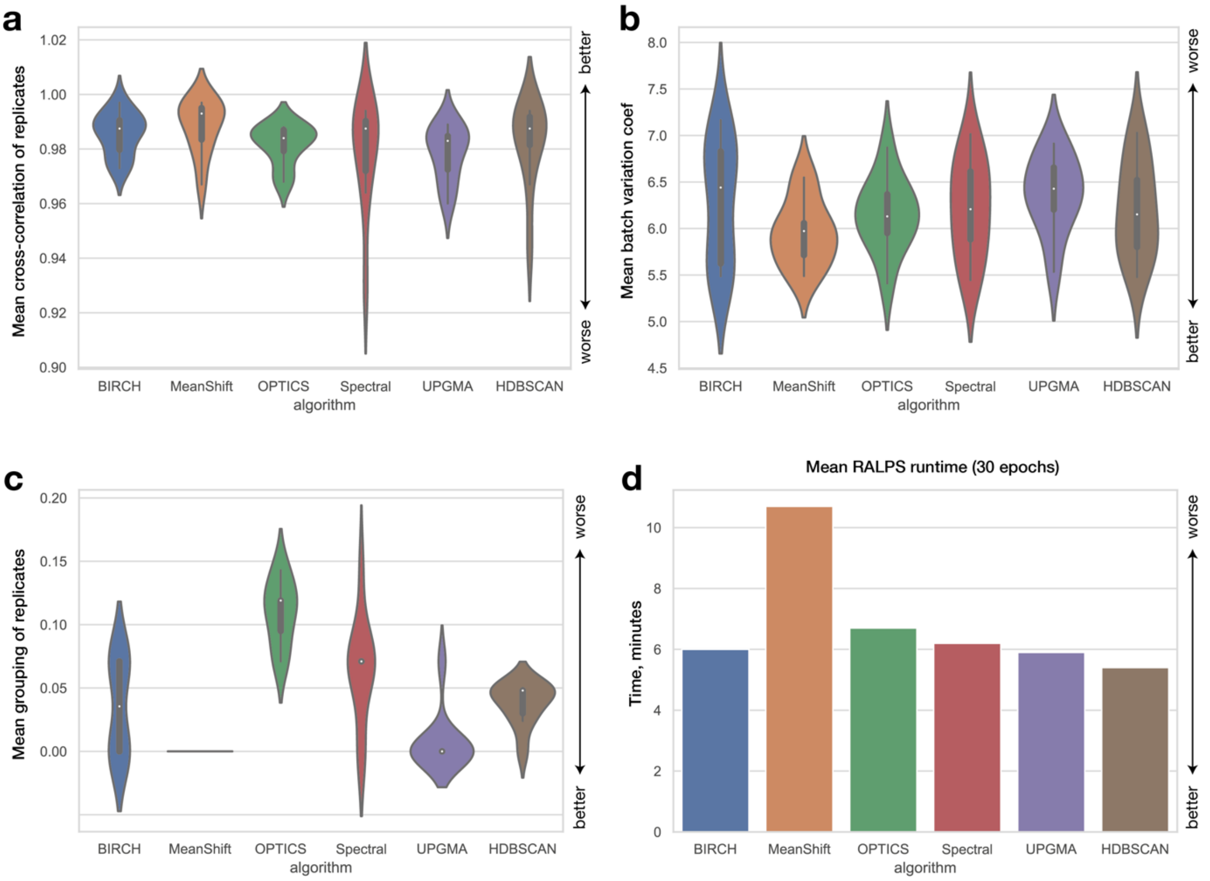
Supplementary Figure 6. Comparison of methods for the cancer cell lines dataset by Cherkaoui et al. (**a**) Cross-correlation heatmaps for MDAMB231 replicates; (**b**) full spectra; (**c**) percent of samples and feature combination with increased variation coefficients; (**d**) and mean batch variation coefficient.



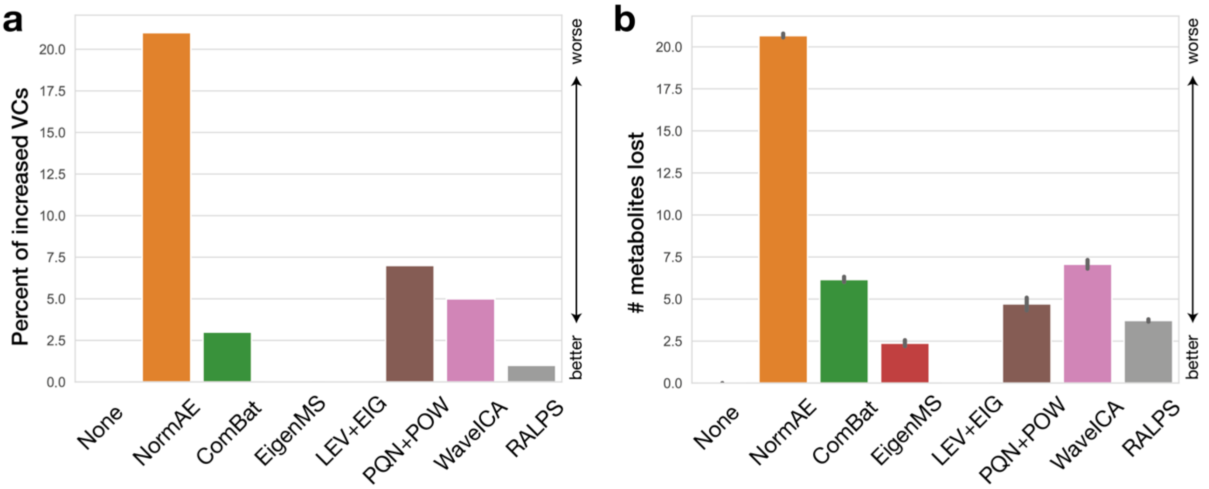
Supplementary Figure 7. Training losses plot generated by NormAE. (**a**) The first 1000 epochs are used for the autoencoder pretraining without the classifier. Then, the adversarial training epochs follow, where the autoencoder and batch classifier are trained in turns. We observe a fast decrease in the classifier loss followed by an “explosion”. Further training leads to the collapsed solution. Percent of unique values in MDAMB231 reference sample before and after normalization. (**b**) We observe that NormAE normalized the data such that it contains less than 20% of unique values. This indicates over-normalization, i.e., most of the biological samples became numerically identical and, therefore, non-comparable among each other.



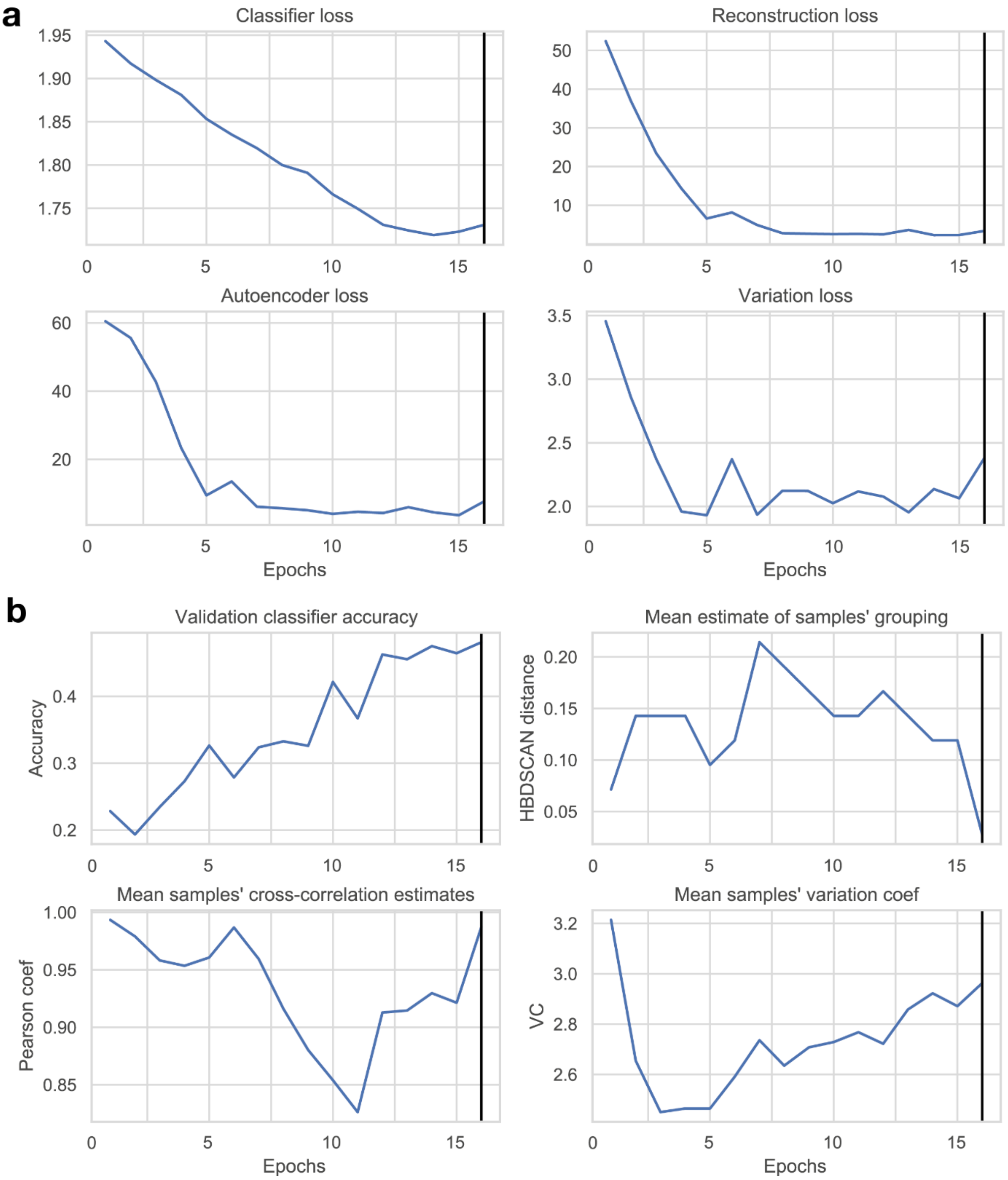
Supplementary Figure 8. Performance of RALPS depending on the underlying clustering algorithm. (**a**) Mean cross-correlation of replicates; (**b**) mean batch variation coefficient; (**c**) mean grouping coefficient for samples' replicates; (**d**) and mean total RALPS runtime.



Supplementary Figure 9. Outlier detection and removal. (**a**) Percent of samples with increased VCs after filtering; (**b**) and number of metabolites removed during filtering. The results were obtained with the same outlier removal procedure applied to the benchmarking dataset after normalization.



Supplementary Figure 10. (a) Losses and (b) metrics for a single run. The vertical line indicates the best epoch (#16) selected automatically by RALPS. The best epoch corresponds to the lowest grouping coefficient (**b**, top-right) and one of the highest correlation coefficients (**b**, bottom-left). Note that this run was stopped early, because the classifier loss started growing after epoch 14 (**a**, top-left).



Supplementary Table 1. Best normalization solutions of different reference samples. Autoencoder reconstruction loss, grouping (*the lower the better*) and Pearson correlation (*the higher the better*) coefficients are presented for samples' replicates. A solution was treated reproducible if at least three other solutions with comparable could be found in an experiment of repetitive training with the same parameter set.

#	Reference samples	reproducible	MSE loss	grouping	correlation
1	P2_S_0001 P2_S_PP_0001	Yes	3.097	0.0	0.969
2	P2_S_0001	No	2.385	0.0	0.969
3	P2_S_0001 P2_S_FA_0001	No	1.632	0.071	0.967
4	P2_S_0001 P2_SF_0001	No	2.028	0.119	0.942
5	P2_S_0001 P2_S_0002 P2_S_0004	Yes	2.241	0.0	0.961
6	P2_S_0001 P1_S_0001 P2_S_0002	Yes	3.684	0.033	0.972
7	P2_S_0001 P2_S_0002 P2_S_0004 P2_S_0008	Yes	2.190	0.094	0.972
8	P2_S_0001 P2_S_0002 P2_S_PP_0001 P2_S_PP_0002	No	1.775	0.113	0.955
9	P2_S_0001 P1_S_0001 P2_S_0002 P1_S_0002	Yes	2.358	0.0	0.970
10	P1_FA_0016	No	17.278	0.095	0.910

Supplementary Table 2. Ion intensity ranges of cancer cell line data by Cherkaoui et al. Only methods that output ion intensities are given.

method	min	max	negative values
None	0	125,663,017	0.0%
ComBat	- 4,354,927	71,056,531	1.0%
EigenMS	402	28,265,811	0.0%
WaveICA	- 15,200,577	121,753,507	1.2%
RALPS	0	95,519,700	0.0%

Supplementary Table 3. Run time for normalization of cancer cell line data by Cherkaoui et al. All methods, except NormAE, were run on 2.2 GHz Intel Core i7 8750H (“Coffee Lake”) under Mac OS Big Sur. * NormAE was run on a Nvidia GeForce RTX 2060 under Windows 10.

	LEV+EIG	**PQN+POW**	**ComBat**	**EigenMS**	**WaveICA**	**NormAE**	**RALPS**
Time (min)	127	6	1	111	5	136*	3

References

(1) Johnson, W. E.; Li, C.; Rabinovic, A. Adjusting Batch Effects in Microarray Expression Data Using Empirical Bayes Methods. *Biostatistics* **2007**, *8* (1), 118–127. https://doi.org/10.1093/biostatistics/kxj037.

(2) Karpievitch, Y. V.; Nikolic, S. B.; Wilson, R.; Sharman, J. E.; Edwards, L. M. Metabolomics Data Normalization with EigenMS. *PLoS One* **2014**, *9* (12), 1–10. https://doi.org/10.1371/journal.pone.0116221.

(3) Dieterle, F.; Ross, A.; Schlotterbeck, G.; Senn, H. Probabilistic Quotient Normalization as Robust Method to Account for Dilution of Complex Biological Mixtures. Application In1H NMR Metabonomics. *Anal. Chem.* **2006**, *78* (13), 4281–4290. https://doi.org/10.1021/ac051632c.

(4) Yang, Q.; Wang, Y.; Zhang, Y.; Li, F.; Xia, W.; Zhou, Y.; Qiu, Y.; Li, H.; Zhu, F. NOREVA: Enhanced Normalization and Evaluation of Time-Course and Multi-Class Metabolomic Data. *Nucleic Acids Res.* **2021**, *48* (1), W436–W448. https://doi.org/10.1093/NAR/GKAA258.

(5) Deng, K.; Zhang, F.; Tan, Q.; Huang, Y.; Song, W.; Rong, Z.; Zhu, Z. J.; Li, Z.; Li, K. WaveICA: A Novel Algorithm to Remove Batch Effects for Large-Scale Untargeted Metabolomics Data Based on Wavelet Analysis. *Anal. Chim. Acta* **2019**, *1061*, 60–69. https://doi.org/10.1016/j.aca.2019.02.010.

(6) Wang, T.; Johnson, T. S.; Shao, W.; Lu, Z.; Helm, B. R.; Zhang, J.; Huang, K. BERMUDA: A Novel Deep Transfer Learning Method for Single-Cell RNA Sequencing Batch Correction Reveals Hidden High-Resolution Cellular Subtypes. *Genome Biol.* **2019**, *20* (1), 1–15. https://doi.org/10.1186/s13059-019-1764-6.

(7) Li, X.; Wang, K.; Lyu, Y.; Pan, H.; Zhang, J.; Stambolian, D.; Susztak, K.; Reilly, M. P.; Hu, G.; Li, M. Deep Learning Enables Accurate Clustering with Batch Effect Removal in Single-Cell RNA-Seq Analysis. *Nat. Commun.* **2020**, *11* (1), 1–14. https://doi.org/10.1038/s41467-020-15851-3.

(8) Lakkis, J.; Wang, D.; Zhang, Y.; Hu, G.; Wang, K.; Pan, H.; Ungar, L.; Reilly, M.; Li, X.; Li, M. A Joint Deep Learning Model Enables Simultaneous Batch Effect Correction, Denoising and Clustering in Single-Cell Transcriptomics. *Genome Res.* **2021**, gr.271874.120. https://doi.org/10.1101/gr.271874.120.

(9) Rong, Z.; Tan, Q.; Cao, L.; Zhang, L.; Deng, K.; Huang, Y.; Zhu, Z. J.; Li, Z.; Li, K. NormAE: Deep Adversarial Learning Model to Remove Batch Effects in Liquid Chromatography Mass Spectrometry-Based Metabolomics Data. *Anal. Chem.* **2020**, *92* (7), 5082–5090. https://doi.org/10.1021/acs.analchem.9b05460.

(10) Wehrens, R.; Hageman, J. A.; van Eeuwijk, F.; Kooke, R.; Flood, P. J.; Wijnker, E.; Keurentjes, J. J. B.; Lommen, A.; van Eekelen, H. D. L. M.; Hall, R. D.; Mumm, R.; de Vos, R. C. H. Improved Batch Correction in Untargeted MS-Based Metabolomics. *Metabolomics* **2016**, *12* (5). https://doi.org/10.1007/s11306-016-1015-8.

(11) Bhojanapalli, S.; Wilber, K.; Veit, A.; Rawat, A. S.; Kim, S.; Menon, A.; Kumar, S. On the Reproducibility of Neural Network Predictions. **2021**, 1–19.

(12) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. **2018**.

(13) Malzer, C.; Baum, M. A Hybrid Approach to Hierarchical Density-Based Cluster Selection. *IEEE Int. Conf. Multisens. Fusion Integr. Intell. Syst.* **2020**, *2020-*

Septe, 223–228. https://doi.org/10.1109/MFI49285.2020.9235263.

(14) Fuhrer, T.; Heer, D.; Begemann, B.; Zamboni, N. High-Throughput, Accurate Mass Metabolome Profiling of Cellular Extracts by Flow Injection-Time-of-Flight Mass Spectrometry. *Anal. Chem.* **2011**, *83* (18), 7074–7080. https://doi.org/10.1021/ac201267k.

(15) Cherkaoui, S.; Durot, S.; Bradley, J.; Critchlow, S.; Dubuis, S.; Masiero, M. M.; Wegmann, R.; Snijder, B.; Othman, A.; Bendtsen, C.; Zamboni, N. Epithelial-Mesenchymal Transition Is the Main Driver of Intrinsic Metabolism in Cancer Cell Lines. *bioRxiv* **2021**, 2021.11.02.466992.

(16) Uw, S. Spectral Clustering. *Encycl. Mach. Learn. Data Min.* **2017**, 1167–1167. https://doi.org/10.1007/978-1-4899-7687-1_100437.

(17) Ankerst, M.; Breunig, M. M.; Kriegel, H.; Sander, J. Ankerst_1999. *ACM SIGMOD Rec.* **1999**, *28* (2), 49–60.

(18) Comaniciu, D.; Meer, P. Mean Shift: A Robust Approach toward Feature Space Analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24* (5), 603–619. https://doi.org/10.1109/34.1000236.

(19) Müllner, D. Modern Hierarchical, Agglomerative Clustering Algorithms. **2011**, No. 1973, 1–29.

(20) Zhang, T.; Ramakrishnan, R.; Livny, M. BIRCH: An Efficient Data Clustering Method for Very Large Databases. *SIGMOD Rec. (ACM Spec. Interes. Gr. Manag. Data)* **1996**, *25* (2), 103–114. https://doi.org/10.1145/235968.233324.

Chapter 4

Prediction of relative metabolite concentrations in human serum analyzed by flow injection time-of-flight mass spectrometry

Andrei Dmitrenko and Nicola Zamboni

Author contributions:

AD and NZ conceived the study and wrote the manuscript. AD formulated machine learning tasks, trained the models and performed all the downstream data analysis.

Abstract

Flow injection analysis-time-of-flight mass spectrometry offers unparalleled throughputs for untargeted metabolomics but comes with fundamental challenges in quantification. First, the lack of chromatographic separation exacerbates the emergence of matrix effects and non-linear responses. In addition, the use of heavy internal standards for quantification doesn't scale beyond a limited number of compounds. Here, we test the potential of quantifying relative metabolome abundances in human serum extracts on the sole basis of data that can be easily collected. This includes dilution series of a pooled study sample and calibration curves for a limited number of compounds, i.e., some amino acids and nucleobases. We formulate calibration tasks as machine learning problems, and test different strategies using own benchmarking reference data. We show that (i) including more dilutions consistently improves performance across tasks; (ii) the model predicting absolute ion counts extrapolates well for spiked-in amino acids but not for purines and pyrimidines; (iii) prediction of relative metabolite abundances has a potential to achieve the best performance in reproducibility across batches. We discuss opportunities to overcome current limitations and deploy machine learning based calibration of FIA-TOF-MS in real applications. To improve extrapolation capability of the current models, we propose to expand the list of metabolites and compound classes, include other biological matrices and complement the dataset with QC features and instrument settings from the system suitability testing platform. These guidelines will be implemented in the design of calibration samples to be included in future studies.

Introduction

The term flow injection analysis – mass spectrometry (FIA-MS) describes the direct injection of a liquid sample to a mass spectrometer. It is equivalent to running a liquid chromatography without a column for separation and under isocratic conditions, that is with constant mobile phase composition and flow. In absence of chromatography, samples elute readily in a single plug. Samples are cleared from the source by the effect of liquid and gas flow, in a process that can take ca. 20-30 sec. This allows injecting thousands of samples per day, with obvious advantages for the analysis of large sample sets.

FIA-MS analysis comes with two important challenges. First, the lack of separation complicates identification of compounds and the analysis of complex samples. This problem is mitigated by MS instruments with high-resolution in the spectral domain, which allow to distinguish mass differences in the range of millidaltons and, hence, resolve chemical formulas with nearly identical molecular weight. One such MS detector is time-of-flight (TOF) MS and has been used in our lab to analyze > 1 Mio samples over the years[1].

The second major challenge of FIA-MS is the acute occurrence of matrix effects. They originate primarily during ionization. It is well established that only a minor fraction of the analytes in the electrospray become completely ionized to gas-phase ions. The vast majority of analytes remains trapped in solvent droplets and is evacuated from the ionization chamber before having a chance of entering the MS. This competitive process provokes an interdependence of ionization across molecules in the sample and non-linear responses. Eventually, matrix effects complicate quantification and comparability across batches. One common way to reduce matrix effects is calibration.

In fact, calibration is an essential part of any quantitative analytical method[2]. In the context of mass spectrometry, it provides a relationship between the real amounts of analytes and the signal measured by the detector. Ideally, the relationship should be linear[3] but it is quite common to observe non-linear calibration curves[4–6]. A few recent works vividly demonstrated the importance of FIA-MS calibration for harmonization, standardization and reproducibility of results in newborn screening. Carling et al. investigated the effect of stable isotope internal standard used for quantitation on inter-laboratory variation[7]. They found that internal calibration reduced the inter-lab variance significantly for most of the analytes. The most recent work of the same author systematically evaluated the effect of multiple calibration approaches used with 5 different instruments in 7 labs[6]. They observed consistent decrease in percentage relative standard deviation between replicate injections and instruments, which points to improved reproducibility and inter-lab comparability of FIA-MS measurements.

MIIS calibration method[8] proposed to predict relative concentrations of compounds was an attempt to quantify ion suppression to gain deeper understanding of matrix effects. The idea to use a single compound as an ion suppression indicator enabled accurate prediction of dilution factors in urinary metabolomics data measured with FIA-TOF-MS. The same system was used to develop a dilute and shoot acquisition method for metabolic phenotyping[9] with quantification accuracy and precision comparable to well-established LC-MS/MS, while producing an order of magnitude higher throughput. The authors leveraged reduction of matrix effects in diluted samples[10] to accurately quantify amino acids in microbial cultivation supernatants.

Overall, low day-to-day reproducibility of ionization made it quite common in FIA-MS and LC-MS experiments to generate calibration curves each time. This is feasible for targeted studies with up to a few hundred compounds, but not for untargeted experiments that try to quantify hundreds to possibly thousands of compounds. Only a limited number of compounds is available as pure compounds, and there are practical limitations on how many can be mixed without inducing novel matrix effects or on how many calibrant can be run in parallel to study samples. It prompts for alternative and innovative approaches that allow to guess concentrations of metabolites in complex matrices from a limited set of calibrants.

Here, we follow up on the efforts to improve calibration of FIA-ESI-MS methods and thereby come closer to high-throughput quantitative analysis of untargeted metabolomics. We aim at predicting relative and absolute amounts of amino acids and nucleobases in a complex biological matrix, such as human serum. We use an untargeted metabolomics dataset of FIA-TOF-MS comprising NIST SRM1950 with multiple spike-ins and dilution series and evaluate machine learning approaches to predict dilution factors, absolute ion counts and relative abundances of the spiked-in metabolites. Further, we assess the reproducibility of predictions among individual batches and compare calibration with the batch correction method developed in **Chapter 3** (RALPS).

Problem formulation

We set out to investigate the performance of calibration for untargeted FIA-TOF-MS data. In particular, we wondered to which extent it is possible to quantify metabolites in a complex sample that is largely affected by matrix effects. In principle, an untargeted metabolomics experiment can include pooled study samples, dilutions thereof (linear or serial), and a limited of calibration curves obtained by spiking chemical standards in a pooled study sample (i.e., on top of the unknown, endogenous concentration) or in water. The number of calibration curves is limited by the availability of chemicals, but also by the simple fact that it is not possible to combine too many compounds in a single spike because it introduces additional matrix effects which differ from those present in a spiked study sample. Hence, the fundamental problem is to investigate whether it is possible extrapolate calibrations from a small number of calibrants or across compounds of the same class. If the latter holds true, few standards would be needed to guess the (relative) abundance of a larger set of compounds. To test that, we formulated three coherent tasks compatible with the available data (**Figure 1**). We approached each task with machine learning by assembling a feature matrix X and training models to predict a target vector Y.

The first test case attempts to learn a function $f(X,\theta)$ that maps measured ion counts and minimal meta-information of a metabolite to the corresponding dilution factor of the dilution series, i.e., some form of a calibration curve (**Figure 1a**). The function f of the feature matrix X and parameters of the machine learning model θ is, therefore, a regression model trained to minimize the mean squared error (MSE) and/or maximize coefficient of determination (R^2) between true and predicted dilution factors. Both statistics are well-established measures of performance of linear regressions. The model was trained in three different setups to evaluate its generalizing power: (i) using 80% of randomized data, (ii) using only a fraction of spike-ins, (iii) using only a fraction of dilutions as a training set. The model performance was then evaluated on the test set comprising the remaining data (never used for fitting the model).

The second test task was the prediction of absolute ion counts of spiked-in metabolites in the serum standard reference material (NIST SRM1950) based on the minimal meta-information, such as dilution factor, batch id, m/z and metabolite identity (**Figure 1b**). To find the best performing setup, we supplemented the feature matrix X with ion counts in SRM1950 and water spike-ins as additional feature columns. However, even with all available features in X, the task remains challenging due to strong matrix effects for which the calibration obtained in water doesn't transfer directly to serum extracts (**Figure 2**). It is easy to see that ion counts corresponding to the same dilution factors of SRM1950 spike-in and SRM1950 baseline samples do not correlate consistently. It is, therefore, impossible to establish a linear relationship between the two. Including more information, as we propose here, might help finding a more complex relationship though. The model for this task was trained and evaluated similarly as the previous one, except the MSE was calculated in the log-scale to account for vast differences in absolute ion counts among metabolites.

Finally, the third test case aims at predicting relative metabolite abundances (**Figure 1c**). We define relative abundance as ratio of detected ion counts between a diluted and an undiluted sample. For a perfectly quantitative method, we could expect the ratio between ion counts of a 4-fold diluted sample and an undiluted sample to be exactly 0.25. Due to ion suppression effects, this is far from being the case for FIA-TOF-MS data. Nonetheless, we sought to predict relative abundances for selected metabolites leveraging available metabolite-specific meta-information. Notably, this problem formulation allows complete exclusion of SRM1950 spike-in data from the training set. In the ideal case, relative abundances of SRM1950 spike-ins could be predicted just based on the relative abundances of the same metabolites in SRM1950 baseline samples. In reality, we also supplemented X with fractions of SRM1950 spike-in dilutions data as additional training samples to achieve the best performance. The model was evaluated on the held-out test set using MSE and R^2.

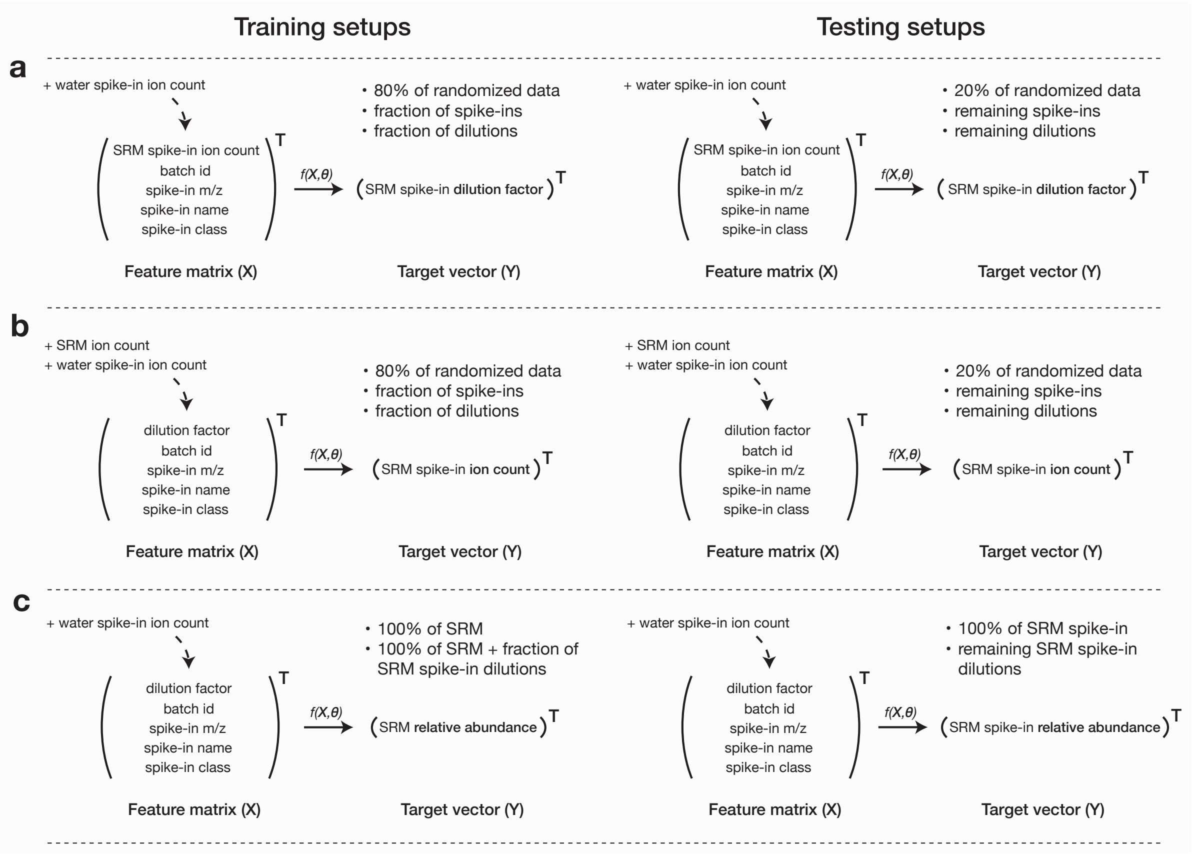


Figure 1. Test cases for predicting dilution factors (**a**), absolute ion counts (**b**) and relative abundances (**c**) of metabolites spiked into serum extracts (NIST SRM1950, abbreviated to SRM).

Data

To address the aforementioned machine learning tasks, we used the benchmarking dataset described in detail in **Chapter 3**. The samples were based on either the extracts from serum (NIST standard reference material SRM1950) or water. Each batch of samples included dilution series (from undiluted to 64-fold diluted), which allow estimating the slope of each detectable compound. In addition, each batch included several classes of spike-ins. We found 5 amino acids (AA) and 4 nucleobases (PP) producing robust ion peaks appearing in every batch and dilution. Visualizing the distribution of ion counts as a function of dilution factor for each metabolite, we observed linear relationships for amino acids and exponential relationships for most of the nucleobases in water (**Figure 2a**). Interestingly, both classes of analytes produced exponential calibration curves in human serum (**Figure 2a, b**), which is a consequence of matrix effects and ion suppression. Even more so are the calibration curves of the SRM1950 baseline sample (without spike-ins), where dilution factor correlates positively with detector response for several metabolites. Additionally, we noted the tendency for increased number of outliers in SRM1950 spike-ins compared to water.

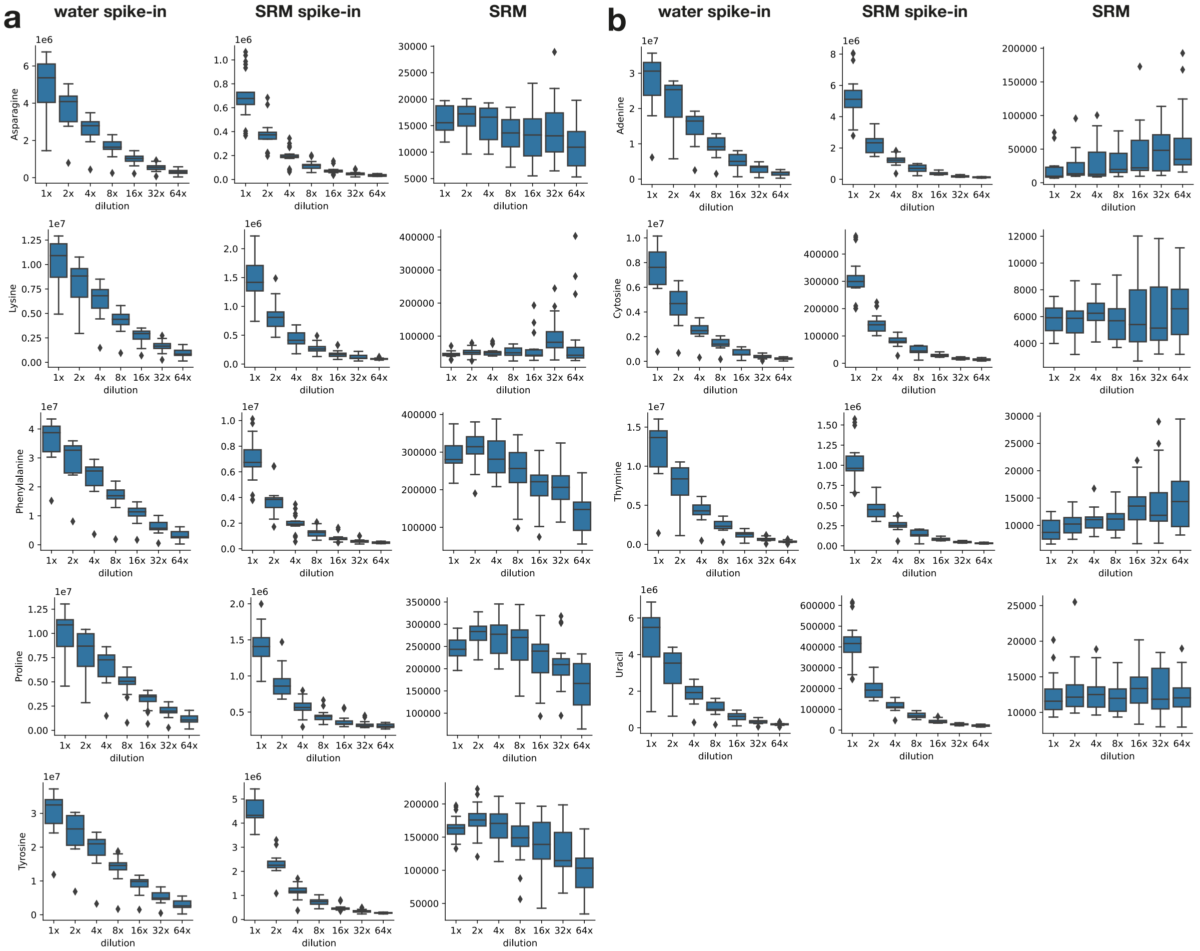


Figure 2. Calibration curves for amino acids (**a**) and nucleobases (**b**) spiked into water, spiked into NIST SRM1950, and in unspiked NIST SRM1950 extracts.

Results

Prediction of dilution factors

Given the strength of ion suppression effects (**Figure 2**), predicting dilution factors for serum spike-ins solely based on their dilutions factors in water or using just serum dilution series is unrealistic. We confirmed that by training several machine learning models and consistently getting R^2 scores near zero. The reasons for that are, likely, order of magnitude differences in ion counts between the same dilution factors, as well as linear versus exponential calibration curves for a half of metabolites. Therefore, we used serum spike-in ion counts in the feature matrix for training (**Figure 1a)**. Additionally, we wondered whether calibration curves for water spike-ins can improve prediction of dilutions factors of the same spike-ins in serum.

To test that, we assembled the dataset containing the following features for each metabolic replicate: ion counts of water and serum spike-ins, m/z, batch id, as well as 11 one-hot encoded features of compound names (9) and classes (2). The full feature matrix had 1323 samples, as for 9 compounds measured in 3 replicates for 7 dilution factors in each of 7 batches. This matrix was used to train an SVR model to predict dilutions in the $\log_2$-scale. **Figure 3a** shows predicted versus true dilutions for the test set. We observed highly accurate predictions with $R^2 = 0.94$ and a few outliers for low dilution factors only. Repeating the exercise with water spike-in ion counts excluded gives $R^2 = 0.813$ only and mean squared error tripled (**Figure 3b**), although the model hyperparameters and random seeds for training and splitting the data were kept identical. The scatter of points grew visibly for all concentrations and the linearity was impaired in high dilution factors. These results demonstrate a significant improvement in predicting relative concentrations of serum spike-ins by adding water spike-in samples. We speculate that the improvement is caused by lower prediction errors for metabolites which have exponential calibration curves in both sample types.

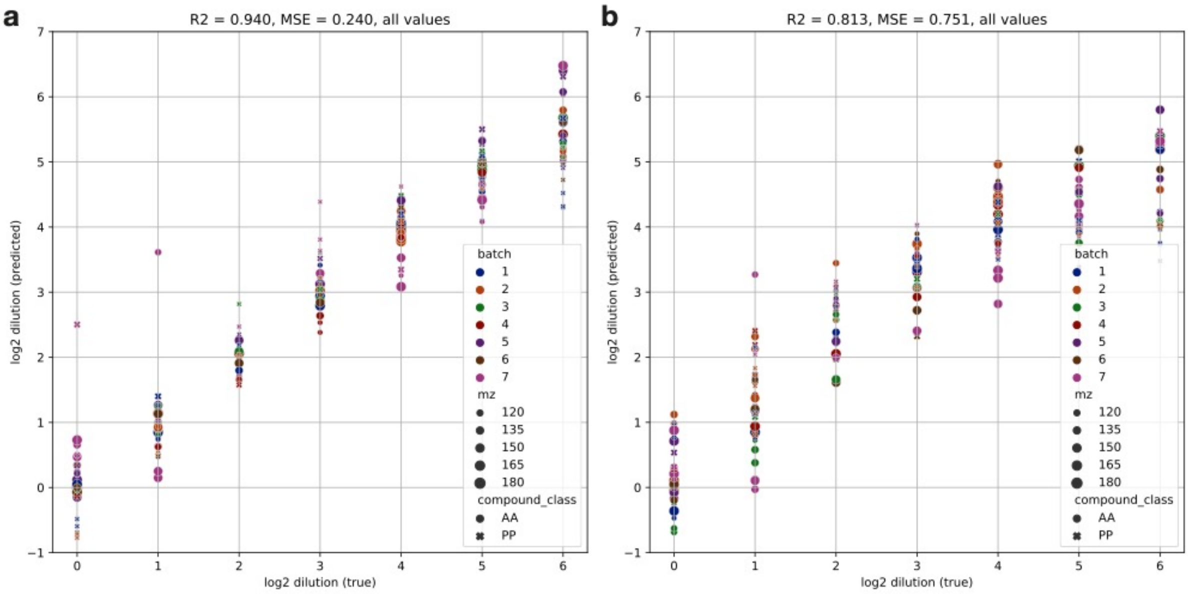


Figure 3. Prediction of relative concentrations with full data (**a**) and excluding spike-ins in water (**b**).

Training with missing dilutions and spike-ins

Next, we asked how many dilutions are necessary for training to accurately predict the remaining ones. In other words, we wanted to find out the minimum number of relative concentrations for spike-ins that have to be measured in water and serum to recover the full calibration curve. We employed the same training setup, now applied to subsets of dilutions. We evaluated the model for each combination of dilutions present in the training set to predict other dilutions appearing in the test set only.

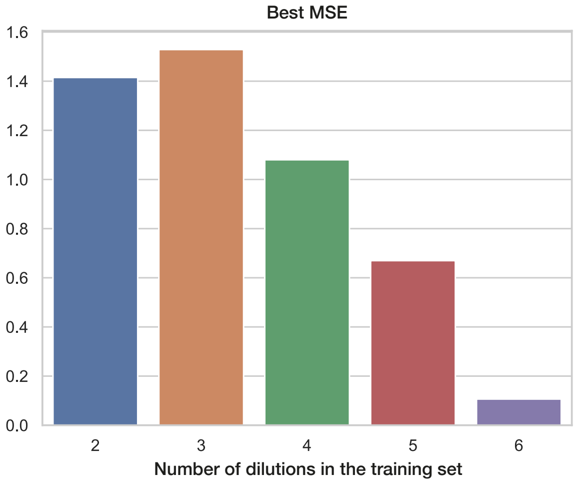


Figure 4. Lowest mean squared error achieved for subsets of dilutions.

We found that with increasing number of dilutions included in the training set the mean squared error of predictions drops, which was expected (**Figure 4**). However, the coefficient of determination R^2 hardly reached 0.7 in all cases, suggesting that the information in the training set was insufficient to achieve the performance of the full data case.

Finally, we experimented with metabolites missing in the training data to see if the model can generalize within compound classes. This would allow to spike only a subset of compounds in a set of calibrants and extrapolate calibration on similar compounds. Therefore, we attempted to train the model on subsets of metabolites and further test it on the remaining compounds of the same class. Like in the missing dilutions case, we observed improvements for both metrics as the number of training metabolites increased (**Figure 5**). Top performing R^2 scores were around 0.8 for cases of one metabolite missing in the class, which is comparable to results of full data without ion counts in water. However, the corresponding mean squared errors stayed relatively high, indicating large number of predicted outliers. It means that with current setup is not yet possible to efficiently extrapolate calibration based on structural similarity of compounds. In the following, we discuss ideas on how to achieve that in the future.

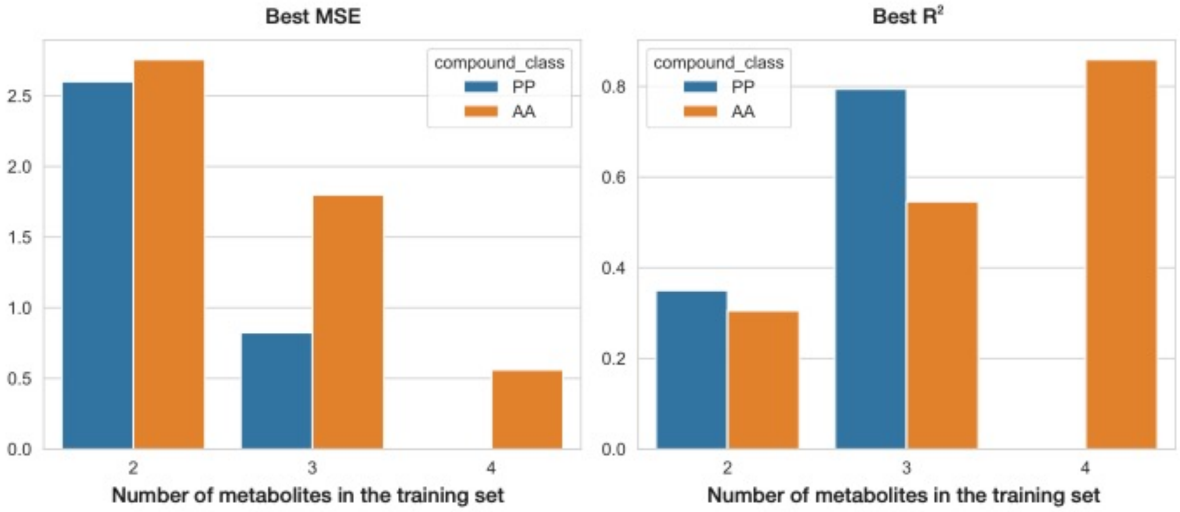


Figure 5. Best regression scores achieved for subsets of metabolites within compound classes.

Prediction of absolute ion counts

We attempted to predict absolute ion counts in serum spike-in samples using the dilution series data of water spike-in and serum samples (**Figure 1b)**. First, we used all samples in the training set (covering all dilutions and metabolites) and experimented with feature columns to achieve top performance. We trained the model with ion counts of water spike-ins only (**Figure 6a**), with ion counts of SRM1950 only (**Figure 6b**) and with both of these feature columns included (**Figure 6c**). We were able to obtain good performance measures in all cases with R^2 = 0.832 in the worst case. Not unexpectedly, combination of both features produced the best result with the lowest MSE score and R^2 effectively reaching 0.9. Interestingly, training with unspiked SRM1950 samples produced higher R^2 compared to using water spike-in samples, suggesting that biological matrix as a factor for prediction absolute ion counts is more important than the spiked-in metabolite.

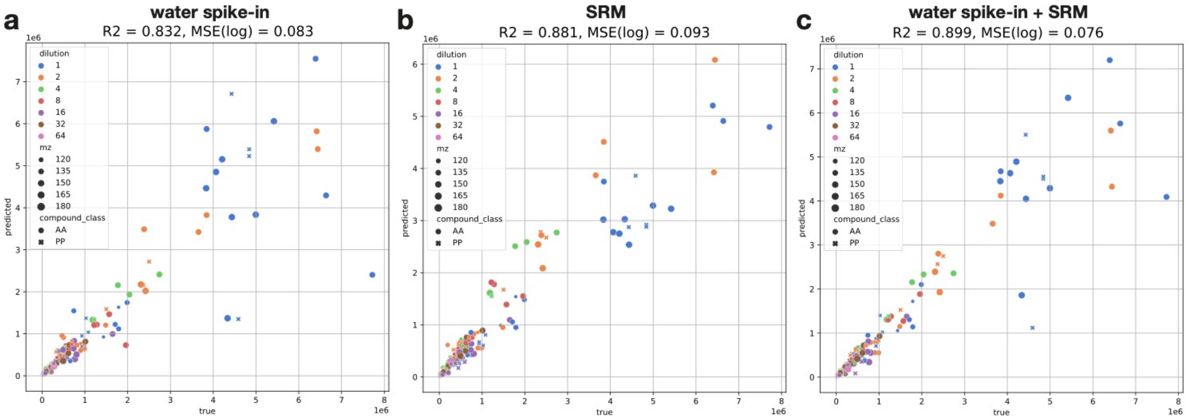


Figure 6. Prediction of absolute ion counts with water spike-in samples only (**a**), SRM1950 samples only (**b**) and both sample types (**c**) included in the training set.

Training with missing dilutions and spike-ins

Next, we repeated the exercise of decreasing numbers of spike-ins and dilutions in the training set to assess to which extent the model can extrapolate for the missing data. Overall, we observed the same trend as in the task of predicting dilution factors. For instance, we found that including 4 dilutions out of 7 (i.e., an undiluted, 4x, 8x and 64x diluted samples) delivers R^2 = 0.876, which is well comparable to the full data results. However, the prediction error grows to 0.237, which is roughly three times higher than in the full data case. Ultimately, including three dilutions only (an undiluted, 8x and 64x diluted samples) represents the most practical example for applications where this level of error is still acceptable. Prediction of absolute ion counts for unseen 2x, 4x, 16x and 32x diluted samples produced MSE of 0.24. Here we conclude that the model interpolates reasonably well for the missing dilutions, and the performance can be further improved by increasing the samples size.

Decreasing the number of amino acids in the training set worked surprisingly well in this problem formulation. We were able to achieve R^2 = 0.823 using Proline and Tyrosine data to predict ion counts of the other three amino acids. However, the lowest MSE of 0.082 could only be reached in case of single amino acids missing in the training data, i.e., predicting ion counts of Tyrosine. In the same setup for nucleobases, R^2 reached as high as 0.685 predicting Cytosine ion counts and the lowest MSE of 0.249 predicting Uracil ion counts. Therefore, our model extrapolates better for amino acids, which was expected simply because more data was available for this compound class (thus, more samples and additional informative feature columns).

Evaluation of reproducibility across batches

Finally, we sought to compare the best calibration model with the batch correction method developed in **Chapter 3** in terms of reproducibility of quantification. For that, we calculated variation coefficients (VCs) for predicted ion counts in each data batch. The variation coefficient is a measure of noise in the sample. Therefore, better reproducibility is associated with lower VCs that are constant across batches, since each batch contains identical set of samples.

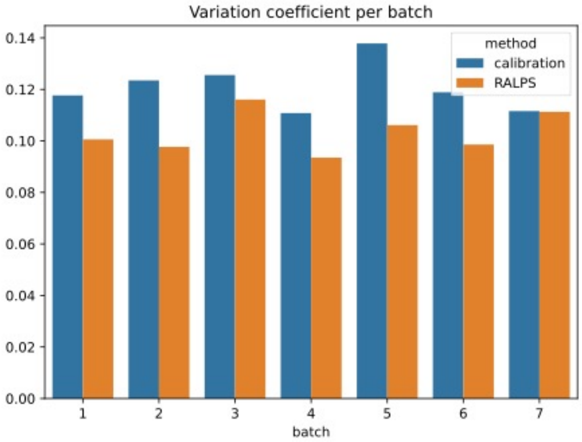


Figure 7. Variation coefficient of ion counts per batch for calibration and batch correction.

We compared the calibration model of the full data case with normalization results produced by RALPS (**Figure 7**). We observed that VCs across batches were relatively stable for both methods, being slightly more consistent for RALPS. In terms of actual values, RALPS delivered lower VCs for most batches. Compared to initial data, median batch VC went down from 0.123 to 0.119 for calibration and to 0.101 for RALPS. These results highlight better reproducibility achieved with RALPS in the analysis of absolute ion counts.

Prediction of relative abundances

As mentioned earlier in the problem formulation, predicting relative abundances is a substantially different task compared to the previous two. By design, it ignores the serum spike-ins data completely while training the model. This was not possible for predicting ion counts since absolute values differed drastically between the three sample types (**Figure 2**). However, we hypothesized that relative abundances might be conserved between the samples. In this case, it would be possible to train the model to predict relative abundances in unspiked serum samples and further reuse it for prediction of relative abundances in serum spike-in samples.

We implemented this idea and trained the model running grid search over sets of hyperparameters similarly to all the previous tasks. However, rather poor performance was achieved on the test set of serum data with $R^2 = 0.501$ only (**Figure 8a**). Testing the model on the unseen data of serum spike-ins generated no linear trend between true and predicted values and a large mean squared error (**Figure 8b**). Retrospectively, we can explain this with inconsistent calibration curves for serum samples: some of the metabolites were detected in larger amounts (ion counts) in samples of larger dilution factors (**Figure 2**), which is an artifact of semi-quantitative analysis likely caused by ion suppression.

To better understand the potential of predicting relative abundances, we repeated the exercise using serum spike-in data only. We split the data randomly into train (80%) and test (20%) sets and trained the model with a grid search as before. The best performing hyperparameter set produced rather high $R^2 = 0.852$ and an exceptionally low MSE (**Figure 9a**). A few predicted outliers decreased the coefficient of determination, but the MSE of 0.015 clearly indicates that, overall, predictions were highly accurate.

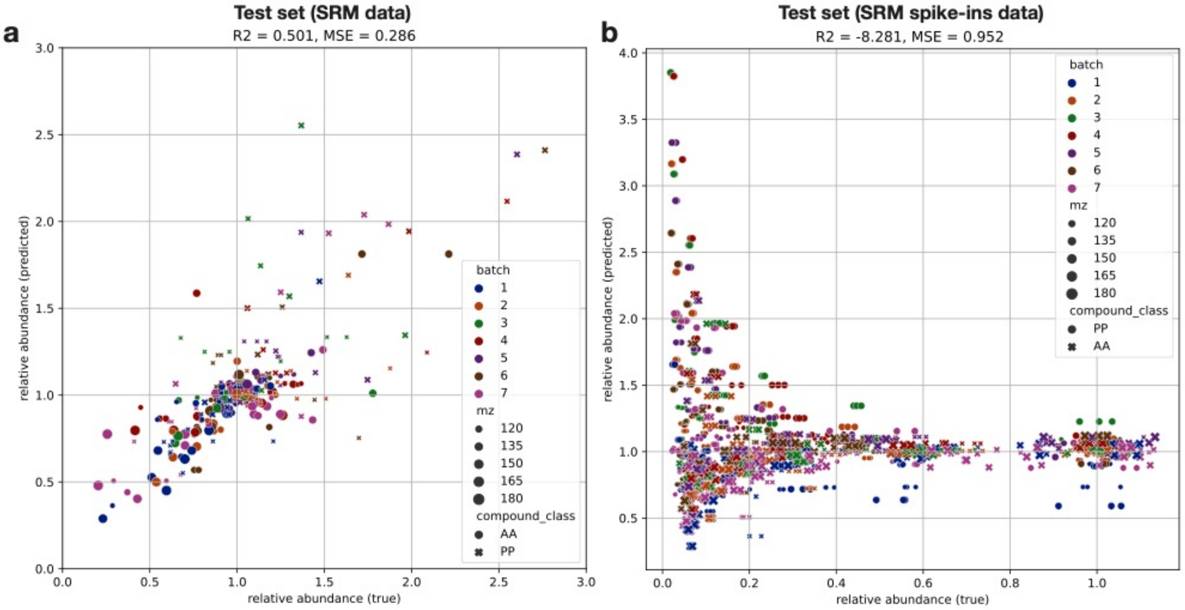


Figure 8. Prediction of relative abundances for serum (**a**) and serum spike-ins (**b**) data with model trained on serum data only.

We further evaluated reproducibility of relative abundances across batches and compared the model with RALPS. In this task, we observed better consistency of variation coefficients for calibration (**Figure 9b**). These results suggest that transitioning from absolute ion counts to relative abundances leads to improved reproducibility of analysis obtained with calibration curves. For practical applications, however, a calibration model pretrained on sufficient data is required, which might be a limiting factor.

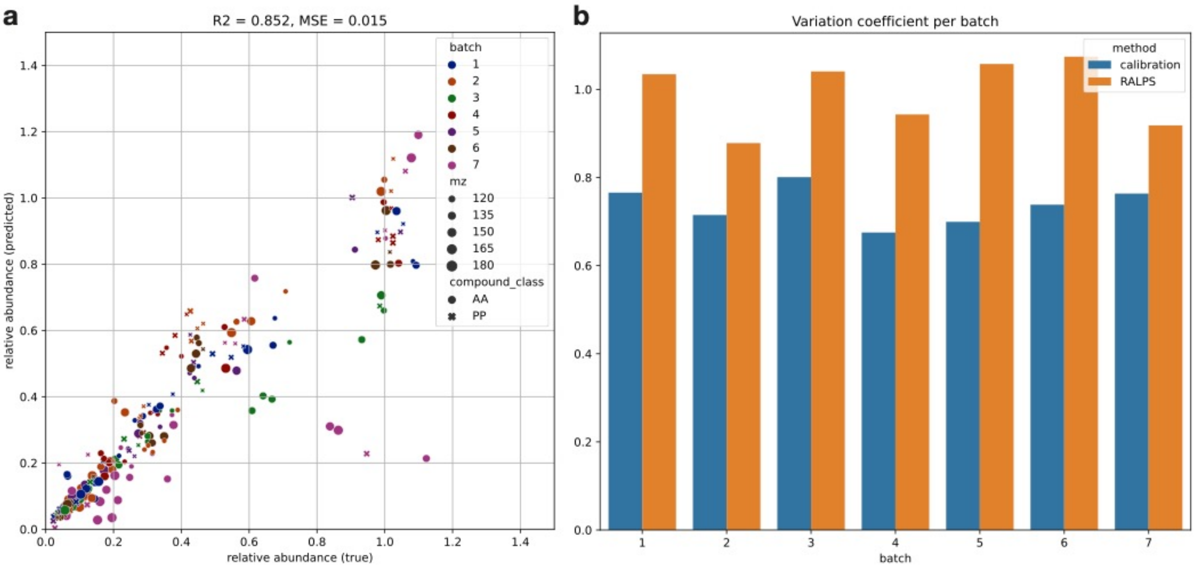


Figure 9. Prediction of relative abundances for serum spike-ins (**a**), and variation coefficient of relative abundances per batch for calibration and batch correction (**b**).

Discussion

Analyzing the benchmarking dataset of untargeted metabolomics acquired with FIA-TOF-MS, we observed substantial matrix effects reflected in non-linearities of calibration curves. Not unexpectedly, some metabolites deviate from linear relationship between relative concentrations and the currents recorded by the MS detector. Previous works[8–10] demonstrated how dilution series can mitigate matrix effects and improve calibration. Our empirical results are in agreement with this notion, as we were able to achieve better performance of the machine learning setups by including more dilutions in the training data. This was seen both for the prediction of dilution factors and the prediction of absolute ion counts. Additionally, we showed that calibration for spiked-in metabolites in human serum extracts can be improved significantly by including measurements of the same spike-ins in water. Although somewhat surprising in nature, this result can be explained by consistent decay of the instrument response with increasing dilution factors in water, a trend which is largely conserved among tested sample types and reinforced with data scaling as part of the training pipeline. We speculate that adding more matrices to the training set could further boost performance and improve quantification of metabolites generally, across matrices.

Given the benchmarking dataset, we attempted to predict relative concentrations and absolute ion counts for partially missing metabolites, which mimics the actual practical application to some extent. However, the amount of available data was limited and likely not sufficient to extrapolate calibration for unseen but structurally similar metabolites. Although we were able to successfully predict ion counts using only 40% of amino acids in the training set, rather poor performance was achieved for purines and pyrimidines. Extrapolation capacity of the model predicting dilution factors was rather modest as well, as we obtained R^2 scores around 0.8 only in cases of single metabolites missing in each compound class.

Expanding the number of metabolites within compound classes and growing the sample size accordingly can definitely improve prediction of calibration curves. Introducing biophysical and/or structural information might be beneficial as well.

However, measuring enough of diverse compounds to cover a reasonably large chemical space is an ambiguous and tedious task. Instead, the set of QC features described in **Chapter 2** as a snapshot of the system state could bring a lot of relevant information. The instrument tunes and readouts could complement the feature space linking detector response directly to the state of mass filters, lenses, mirrors, etc. Certainly, with large number of features in the training set, it becomes essential to select the most informative ones. A variety of approaches exist[11], including model-specific[12,13], sequential[14], recursive[15], statistical[16] feature selection, etc.

Therefore, we envision a follow-up study involving an expanded list of metabolites (and, ideally, compound classes) measured in water, blood, plasma, urine, bacterial extracts in dilutions series for up to a factor 64x, where each analytical triplicate is accompanied with the QC sample and the corresponding instrument settings. Based on our observations, a machine learning model achieving high performance on such a dataset should be capable of quantifying unseen but structurally similar compounds across biological matrices provided the actual QC features and instrument settings are available during inference. In addition, quantification of relative abundances must be revisited to investigate reproducibility of calibration across biological matrices and demonstrate a real application example.

Methods

Data

The dataset of multi-batch untargeted metabolomics data used in this study is available at https://doi.org/10.3929/ethz-b-000545373. From each batch, we retrieved the samples of the following prefixes: “P1_AA” for samples with spiked-in amino acids in water; “P1_PP” for samples with spiked-in purines and pyrimidines in water; “P2_SRM” for baseline SRM1950 samples; “P2_SAA” for samples with spiked-in amino acids in SRM1950; “P2_SPP” for samples with spiked-in purines and pyrimidines in SRM1950. Dilution factors for up to 64 were used to train models. Folds of dilution are also encoded in sample names (e.g., “0016” corresponds to 16-fold dilution).

Machine learning

Several machine learning models were tested for predicting relative concentrations and absolute ion counts in the regression settings, including LASSO, Ridge, ElasticNet and SVR implementations from scikit-learn[17]. Several data scaling approaches were integrated into the training pipelines, including Standard, MinMax, MaxAbs scalers. Grid search with model-specific hyperparameters was implemented for each combination of model and scaler. Coefficient of determination R^2 was optimized in stratified k-fold cross validation to achieve the best performance. Mean squared error between true and predicted values was evaluated along with R^2 in all setups. We used 5-fold splits for the full data settings and 3-fold for partially missing data.

Standard scaler in combination with SVR consistently outperformed other methods and was picked for final training and evaluation. Grid search over three hyperparameters was implemented: C, epsilon and kernel. Linear, sigmoid and radial basis function (RBF) kernels were tested to account for non-linearities in the data. In the results section, the performance on the test set is reported. Test cases for each task are defined on **Figure 1**. Random seeds were fixed for all data splits and machine learning models, where possible.

Data and code availability

The code snippets for each of the three tasks are available at https://gitlab.ethz.ch/andreidm/calibration. To ensure reproducibility of results, the initial dataset and the one normalized with RALPS are also available in the same repository as csv files.

References

(1) Fuhrer, T.; Heer, D.; Begemann, B.; Zamboni, N. High-Throughput, Accurate Mass Metabolome Profiling of Cellular Extracts by Flow Injection-Time-of-Flight Mass Spectrometry. *Anal. Chem.* **2011**, *83* (18), 7074–7080. https://doi.org/10.1021/ac201267k.

(2) Moosavi, Seyed Mojtaba; Ghassabian, S. Linearity of Calibration Curves for Analytical Methods: A Review of Criteria for Assessment of Method Reliability. In *Calibration and Validation pf Analytical Methods*; 2018. https://doi.org/10.5772/intechopen.72932.

(3) Martin, J.; Gracia, A. R.; Asuero, A. G. Fitting Nonlinear Calibration Curves: No Models Perfect. *J. Anal. Sci. Methods Instrum.* **2017**, *07* (01), 1–17. https://doi.org/10.4236/jasmi.2017.71001.

(4) Asnin, L. D. Peak Measurement and Calibration in Chromatographic Analysis. *TrAC - Trends Anal. Chem.* **2016**, *81*, 51–62. https://doi.org/10.1016/j.trac.2016.01.006.

(5) Henderickx, H. J. W.; Duchateau, A. L. L.; Raemakers-Franken, P. C. Chiral Liquid Chromatography-Mass Spectrometry for High-Throughput Screening of Enzymatic Racemase Activity. *J. Chromatogr. A* **2003**, *1020* (1), 69–74. https://doi.org/10.1016/S0021-9673(03)00427-8.

(6) Carling, R. S.; Whyte, E.; John, C.; Garstone, R.; Goddard, P.; Greenfield, T.; Hogg, S. L.; Le Masurier, C.; Cowen, S.; Moat, S. J.; Hopley, C. Improving Harmonization and Standardization of Expanded Newborn Screening Results by Optimization of the Legacy Flow Injection Analysis Tandem Mass Spectrometry Methods and Application of a Standardized Calibration Approach. *Clin. Chem.* **2022**, *1083*, 1075–1083. https://doi.org/10.1093/clinchem/hvac070.

(7) Carling, R. S.; John, C.; Goddard, P.; Griffith, C.; Cowen, S.; Hopley, C.; Moat, S. J. Evaluation of a Common Internal Standard Material to Reduce Inter-Laboratory Variation and Ensure the Quality, Safety and Efficacy of Expanded Newborn Screening Results When Using Flow Injection Analysis Tandem Mass Spectrometry with Internal Calibration. *Int. J. Neonatal Screen.* **2020**, *6* (4). https://doi.org/10.3390/ijns6040092.

(8) Chen, G. yuan; Liao, H. wei; Tseng, Y. J.; Tsai, I. lin; Kuo, C. hua. A Matrix-Induced Ion Suppression Method to Normalize Concentration in Urinary Metabolomics Studies Using Flow Injection Analysis Electrospray Ionization Mass Spectrometry. *Anal. Chim. Acta* **2014**, *864* (33), 21–29. https://doi.org/10.1016/j.aca.2015.01.022.

(9) Reiter, A.; Herbst, L.; Wiechert, W.; Oldiges, M. Need for Speed: Evaluation of Dilute and Shoot-Mass Spectrometry for Accelerated Metabolic Phenotyping in Bioprocess Development. *Anal. Bioanal. Chem.* **2021**, *413* (12), 3253–3268. https://doi.org/10.1007/s00216-021-03261-3.

(10) Stahnke, H.; Kittlaus, S.; Kempe, G.; Alder, L. Reduction of Matrix Effects in Liquid Chromatography-Electrospray Ionization-Mass Spectrometry by Dilution of the Sample Extracts: How Much Dilution Is Needed? *Anal. Chem.* **2012**, *84* (3), 1474–1482. https://doi.org/10.1021/ac202661j.

(11) Azhar, M. A.; Thomas, P. A. Comparative Review of Feature Selection and Classification Modeling. *2019 6th IEEE Int. Conf. Adv. Comput. Commun. Control. ICAC3 2019* **2019**. https://doi.org/10.1109/ICAC347590.2019.9036816.

(12) Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B* **1996**, *58* (1), 267–288.
(13) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32.
(14) Ferri, F. J.; Pudil, P.; Hatef, M.; Kittler, J. Comparative Study of Techniques for Large-Scale Feature Selection. *Mach. Intell. Pattern Recognit.* **1994**, *16* (C), 403–413. https://doi.org/10.1016/B978-0-444-81892-8.50040-7.
(15) Isabelle, G.; Jason, W.; Stephen, B.; Vladimir, V. Gene Selection for Cancer Classification Using Support Vector Machines. *Mach. Learn.* **2002**, *46*, 389–422.
(16) Paninski, L. Estimation of Entropy and Mutual Information. *Neural Comput.* **2003**, *15* (6), 1191–1253. https://doi.org/10.1162/089976603321780272.
(17) Pedregosa, F. et al. Scikit-Learn: Machine Learning in Python Fabian. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830. https://doi.org/10.1289/EHP4713.

Chapter 5

Comparing representations of biological data learned with different AI paradigms, augmenting and cropping strategies

Andrei Dmitrenko, Mauro M. Masiero and Nicola Zamboni

This manuscript is accepted for publication in *Proceedings of Machine Learning Research* as a full paper presented at MIDL 2022 conference (poster).

Author contributions:

MM collected the imaging dataset. AD conceived the study, developed and trained the deep learning models, formulated the downstream tasks to evaluate their performance and wrote the paper describing the comparative analysis. NZ supervised the study.

Abstract

Recent advances in computer vision and robotics enabled automated large-scale biological image analysis. Various machine learning approaches have been successfully applied to phenotypic profiling. However, it remains unclear how they compare in terms of biological feature extraction. In this study, we propose a simple CNN architecture and implement 4 different representation learning approaches. We train 16 deep learning setups on the 770k cancer cell images dataset under identical conditions, using different augmenting and cropping strategies. We compare the learned representations by evaluating multiple metrics for each of three downstream tasks: i) distance-based similarity analysis of known drugs, ii) classification of drugs versus controls, iii) clustering within cell lines. We also com- pare training times and memory usage. Among all tested setups, multi-crops and random augmentations generally improved performance across tasks, as expected. Strikingly, self-supervised (implicit contrastive learning) models showed competitive performance being up to 11 times faster to train. Self-supervised regularized learning required the most of memory and computation to deliver arguably the most informative features. We observe that no single combination of augmenting and cropping strategies consistently results in top performance across tasks and recommend prospective research directions. Keywords: Representation learning, self-supervised learning, regularized learning, com- parison, memory constraints, cancer research, microscopy imaging.

Introduction

With recent advances in robotics and deep learning methods, automated large-scale biological image analysis has become possible. Different microscopy technologies allow to collect imaging data of samples under various treatment conditions. Then, images are processed to extract meaningful biological features and compare samples across cohorts. As opposed to carefully engineered features used in the past, deep learning approaches are widespread and automatically distil relevant information directly from the data[1].

A lot of approaches, following different paradigms of machine learning, have been successfully applied to image-based phenotypic profiling: from fully supervised approaches[2,3] to generative adversarial learning[4–6] and self-supervision[7,8]. However, it remains unclear how these approaches align with each other in terms of biological feature extraction. The direct comparison is close to impossible, as many aspects differ between the studies: imaging technologies, datasets, learning approaches and model architectures, implementations and hardware. We discuss related works in more depth in a section below.

In the emergent field of self-supervised learning, a key role of random data augmentations and multiple image views has recently been shown[9]. Their synergetic impact on learning image representations has not yet been rigorously studied. In this paper, we compare different deep learning setups in their ability to learn representations of drug-treated cancer cells. We propose a simple CNN architecture and implement several approaches to learn representations: the weakly-supervised learning (WSL), the implicit contrastive learning (ICL) and classical self-supervised learning without (SSL) and with regularization (SSR). We train four models on the same dataset of 770k images of cancer cells growing in 2D cultures in a drug screening campaign. We use four settings for each model: with and without random augmentations, with single and multi-crops. The other training conditions are kept identical. We compare the learned representations in three downstream analysis tasks, discuss their performance and provide the comparison summary table.

Therefore, our main contributions are:

- implementations of 16 deep learning setups, including state-of-the-art methods trainable within limited resources (the source code and the trained models are available at https://github.com/dmitrav/morpho-learner),
- a systematic comparison of learned representations.

Related work

Weak supervision has been a popular choice to learn medical image representations and has proven its efficiency[10,11]. When analyzing samples corresponding to different treatments, patients, or any experimental conditions, those are often used as weak labels. In our case, there are 693 conditions with different combinations of drugs and cell lines. However, the effects of those combinations are largely unknown, so we restrict ourselves into using two labels only: drug vs control (supposedly, effect vs no effect).

A recent approach to understand morphological features of cancer cells by Longden et al. follows an unsupervised perspective[12]. The authors apply a deep autoencoder to learn 27 continuous morphological features. However, their model does not work with raw images. It uses 624 extracted numerical features as input, and applies a series of linear layers to reconstruct them. Here, we use a convolutional autoencoder instead, to learn more features directly from the data.

Several approaches for learning representations of cell images are based on generative adversarial networks[13,14]. Such models often have two components: the generator and the discriminator network, trained simultaneously in a competitive manner. In this work, we implement a similar idea in the form of regularization: we use a deep convolutional autoencoder as generator, and a weakly-supervised classifier as discriminator. Both networks share the same stack of layers, responsible for learning representations, while optimizing different loss functions. In this setting, the computational time and memory usage remain comparable to the aforementioned approaches.

Finally, self-supervision has recently emerged in bioinformatics to address problems like cell segmentation, annotation and clustering[15,16]. Most recently, a self-supervised contrastive learning framework has been proposed by Ciortan & Defrance to learn representations of scRNA-seq data[17]. The authors follow SimCLR[18] in the implementation of contrastive loss and show that their approach compares favorably with state-of-the-art (SOTA) methods in a downstream clustering task. Here, we train a self-supervised CNN backbone, following BYOL[19]. Unlike SimCLR, this approach does not need negative pairs, yet it was shown to have a superior performance.

In spite of the great interest in deep-learning-based approaches to learning representations of biological data, there have been very few attempts to fairly compare those. A comparison of AI-based methods to predict cell function has come out lately[20]. However, it was primarily focused on collating traditional machine learning versus deep learning. Brief general comparisons of recent AI approaches can be found in reviews and surveys[1,21,22], but they lack details and cannot inform decision making. Recently, a thorough comparison of data-efficient image classification models has been published by Brigato et al. The authors evaluated 10 models on 6 different datasets[23]. Eventually, this analysis focuses on classification tasks only. Although we see papers illustrating how a single study can benefit from multiple AI paradigms[24], it remains unclear which approach is preferable in a particular representation learning task. In this study, we attempt to address this question for analysis of images of cancer cell lines growing in tissue cultures.

Data

The initial dataset comprises 1.1M high-resolution grey-scale images of drug-treated cancer cell populations growing adherently *in vitro*. It captures 693 unique combinations of 21 cell lines and 31 drugs at 5 different drug concentrations, multiple time points and biological replicates. Details are given in **Supplementary Text 1**.

We carefully subset the initial data to obtain a balanced dataset of two labels: strong drug effect (i.e., the highest drug concentration, the latest time point) and control (no drugs, any time point). We end up with about 770k image crops of size 64 × 64. It is important to note that some drugs did not provoke any effect on resistant cell lines, so the corresponding images of drugs and controls look similar. Some other drugs showed growth arrest only, which resulted in drug-treated images being similar to early time point controls, where the cells have not grown yet. By balancing the dataset to contain such cases (**Figure 1**), we expected the models to learn specific morphological differences, instead of superficial features like cell location in the crop, cell population density, amount of grey, etc.

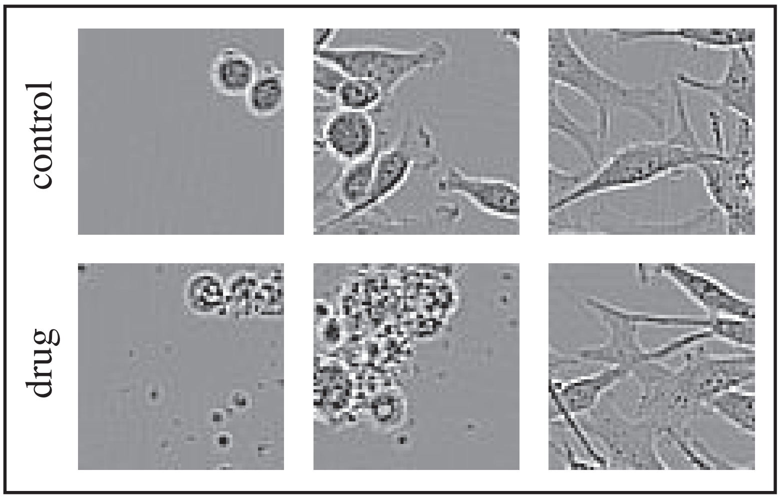


Figure 1. Examples of control and drug images (M14 cell line). On the left, an early time point of the control (cells have not grown yet) is shown against a strong drug effect (fragmented or dead cells). On the right, the end time points for the control and an ineffective drug are depicted. In the middle, an example of intermediate growth of a control sample versus another cytotoxic drug is given.

Methods

Model architectures

To learn image representations, we implemented the following models:

a. deep classifier with two output labels only (WSL),
b. convolutional autoencoder with classic encoder-decoder architecture (SSL),
c. models **a** and **b** in the joint encoder-classifier-decoder architecture (SSR),
d. CNN backbone, trained with BYOL (ICL).

As seen on **Figure 2**, the four architectures contain the same CNN backbone, which is used to produce image representations for the downstream analysis. It was important to use the same stack of layers to ensure fair comparison of methods. However, image representations are learned solving substantially different tasks.

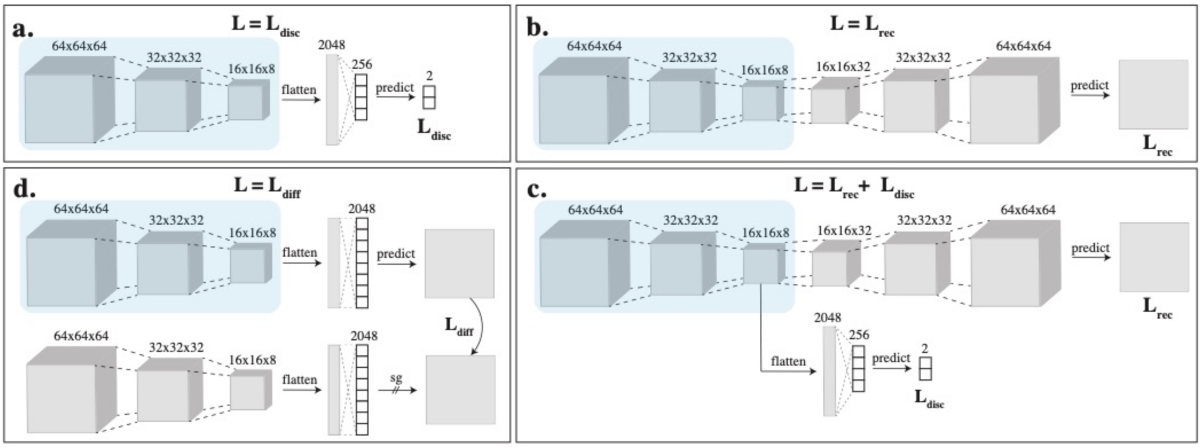


Figure 2. Graphical overview of models. **a.** A weakly-supervised deep classifier with a categorical cross-entropy discrimination loss. **b.** A convolutional autoencoder with a binary cross-entropy reconstruction loss as in Creswell et al.[25] **c.** A regularized convolutional autoencoder: models a and b, sharing the CNN backbone, trained simultaneously. **d.** A self-supervised CNN backbone with a mean squared error difference loss.

For the SSR model, we adopted a particular implementation where a classifier and an autoencoder are trained in turns, optimizing different loss functions (**Figure 2c**). Our idea was to encourage the autoencoder to learn representations that would bear differences between drug and control images, while still delivering high quality image reconstructions. In this formulation, the classifier acts as a regularizer. Although similar models have been utilized in chemo- and bioinformatics tasks[26,27], to our knowledge this architecture has not been tested previously in the analysis of biological images.

Training setups

Each model was trained under 4 conditions of presence and absence of random image augmentations and multi-crops, giving rise to 16 training setups in total. Since the dataset is naturally grayscale, we only applied random resized crops, horizontal flips and Gaussian blurs to augment. Note that data augmentations are intrinsic to the self-supervised approach. Therefore, we tested single and double augmenting (while preprocessing and/or while training) for the ICL model. In the one-crop setting, we used single 64 × 64 images. For multi-crop, we added 4 random resized crops applied to 64 × 64 images: 2 of about half-size, and 2 more of about quarter-size (5 crops in total).

We implemented the 16 described setups and trained them using Nvidia GeForce RTX 2060 with 6 GB only. We chose the CNN backbone architecture, batch size and other common hyperparameters by running grid search and finding the best average performance across models, achievable within reasonable training time and hardware memory constraints.

For the ICL model, we additionally optimized BYOL parameters: projection size, projections hidden size and moving average decay. We trained the model 100 times, sampling parameters from predefined ranges. We found that equal number of neurons for hidden and projection layers consistently delivered the lowest MSE loss.

We trained all models for 50 epochs, using Adam optimizer with a constant learning rate of 0.0001. A batch size of 256 was used. We defined the same early stopping criterion, which checks a simple divergence condition on the loss function. We used the same data splits with 10% for the validation set to test classification accuracy and reconstruction quality.

Validation and evaluation

We validated the models by monitoring loss functions, classification accuracy and image reconstruction quality for training and validation sets (**Supplementary Figure 1**). Evaluation metrics for the downstream tasks are described below.

Distance metrics

First, we compared the learned representations in their ability to capture similarity of known drugs. Let S_1 and S_2 be the sets of images of two drugs, known to have similar effects, and C be the set of control images. We quantify similarity between S_1 and S_2 as follows:

- $D(S_1, S_2) = \underset{u \in S_1, v \in S_2}{\mathrm{median}}(\|u - v\|)$,
  i.e., the median Euclidean distance between any two images *(u, v)* of two sets.
- $d(S_1, S_2) = \frac{\widehat{D} - D(S_1, S_2)}{\widehat{D}}$, where $\widehat{D} = \frac{1}{2}[D(S_1, C) + D(S_2, C)]$,
  i.e., the normalized difference between drug-to-control and drug-to-drug distances.

The motivation for using Euclidean distance is given on **Supplementary Figure 2**.

Classification metrics

Next, we performed binary classification. We used a pretrained stack of layers of each model to generate latent representations and then trained a two-layer classifier to differentiate between drugs and controls. We used the same data splits and trained for 25 epochs with SGD optimizer and batch size of 1024. We ran grid search over learning rate, momentum and weight decay to achieve the best validation accuracy. To comprehensively evaluate the performance, we calculated several metrics individually for each cell line: accuracy, precision, recall and area under ROC.

Clustering metrics

Finally, we performed clustering to quantify how similar drug effects group in the latent space. For each cell line, we obtained image representations, reduced their dimensionality with UMAP[28], and clustered their embeddings with HDBSCAN[29]. We evaluated several metrics on the partitions: number of identified clusters and percent of noise points, Silhouette score and Davies-Bouldin similarity.

For each cell line, we ran grid search over two parameters: i) n neighbors, responsible for constraining the size of local neighborhood in UMAP, and ii) min cluster size, representing the smallest grouping size in HDBSCAN. We adopted the following procedure to find the best partitions: 1) select Silhouette scores above median, 2) select Davies-Bouldin scores below median, 3) select the lowest percent of noise, 4) pick a parameter set of max number of clusters. This logic was motivated by zero correlation between Silhouette and Davies-Bouldin measures, and by the objective to find as many "clean" clusters as possible.

Results

Distance-based drug similarity analysis

Pemetrexed (PTX) and Methotrexate (MTX) are two drugs that have similar chemical structures and both inhibit folate-related enzymes. Over the years, they have been successfully applied to cure many types of cancer[30]. We applied distance-based analysis to evaluate how close PTX and MTX are to each other in terms of learned features, and how distant they both are from controls (images of cells under no treatment).

We picked all images related to PTX and MTX drugs from the validation set. Then, we randomly picked the same number of control images (DMSO). We calculated *D*(MTX, PTX), *D*(MTX, DMSO), *D*(PTX, DMSO) on image representations, which resulted in around 3600 distances for each cell line and pair on average. Based on a-priori knowledge of efficiency and similarity of the drugs, we expected MTX-PTX distances to be consistently lower than of MTX-DMSO and PTX-DMSO. Analysis of M14 cell line shows it was not the case for all models (**Supplementary Figure 3**).

We repeated the same analysis for each of 21 cell lines. We found that with the exception of the WSL model, all produced lower average MTX-PTX distances, compared to MTX-DMSO and PTX-DMSO. This suggests that the space of learned features of the WSL model is likely to contain more trivial information about the drug effects, rather than features of altered morphology. Interestingly, the median normalized difference *d* turned out to be the largest for the ICL model (**Table 2**).

Classification of drugs versus controls

All models showed comparable classification performance, crossing 0.6 accuracy bottom line and reaching 0.7 in many cases. However, it is only the WSL model that achieved 0.8 accuracy for some cell lines (**Supplementary Figure 4**) and delivered consistently higher performance in all setups. This was expected due to identical problem formulation during representation learning. Notably, the other three models have shown rival performance on this task. That implies that all models have a potential in detecting drug effects in time-series imaging data (e.g., to predict drug onset times for different concentrations).

Table 2 contains four classification metrics for each training setup, evaluated on the entire dataset. Median performance for 21 cell lines is reported. The SSR model with single crops and augmentations showed the highest overall accuracy (0.76 ± 0.07) and ROAUC (0.76 ± 0.06), though the WSL model was the most robust across settings. The WSL and ICL models improved performance with multi-crops.

Clustering analysis within cell lines

Supplementary Figure 5 shows a particular example of clustering analysis of HCT cell line obtained with ICL model. **Figure 3A** presents numbers of identified clusters across all models and settings. Varying the clustering parameters resulted in relatively large confidence intervals. However, even the lower bounds exceeded n=2 clusters, which would correspond to the trivial case of differentiating between drugs and controls (effect vs no effect), in the majority of cases. That indicates that the learned representations allow studying the data in more depth (e.g., finding similarities in concentration-dependent morphological drug effects).

Although mean numbers of clusters look similar, the quality of partitions differed substantially across cell lines, as follows from the Silhouette score barplots (**Figure 3B**). The WSL model produced the poorest scores for the three picked cell lines. Close-to-zero and even negative values suggest that the clusters were mainly overlapping. In such cases, obtained partitions are far less trustworthy and any follow-up analysis on them is controversial. The top performance was shown by the SSL and the SSR models. Statistics across all cell lines are given in **Table 2**.

Training times and memory usage

All models were trained using Nvidia GeForce RTX 2060 with 6 GB memory. With batch size of 256, steady memory consumption was around 4.3 and 4.7 GB for single and multi- crops, respectively. Batch size of 512 resulted in cuda-out-of-memory error in all setups.

Table 1. Training time (hours).

	SSL	**ICL**	**WSL**	**SSR**
one_crop	7	1.5	2.5	9
multi_crop	35	4	11	45

Unlike memory usage, training times differed largely for model architectures and cropping strategies (**Table 1**). The ICL model was the only one to meet the early stopping criterion, which resulted in remarkably small training times. The one crop training stopped after 16/50 epochs, whereas multi crop made 7/50 epochs only. The other models were trained for all 50 epochs.

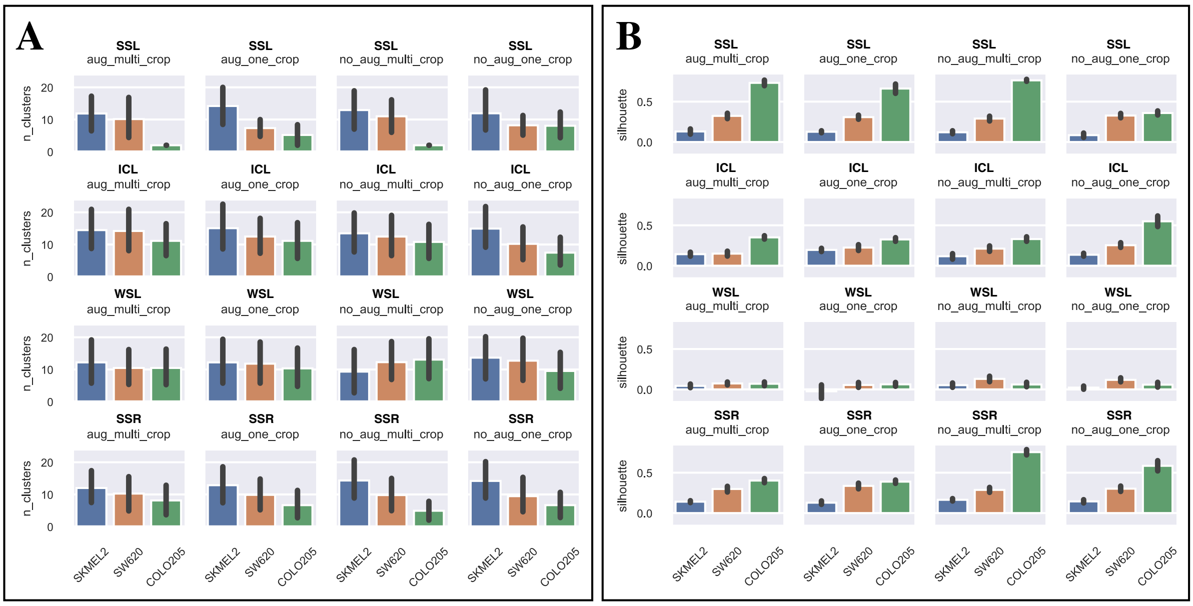


Figure 3. Clustering analysis for three picked cell lines: SKMEL2, SW620, COLO205. Mean numbers of identified clusters (**A**) and mean Silhouette scores (**B**) are shown with confidence intervals. Model architectures are (from top to down): SSL, ICL, WSL and SSR. Different setups are (from left to right): augmentations + multi-crops, augmentations + single crops, no augmentations + multi-crops, no augmentations + single crops.

Discussion

In this study, we used the distance-based analysis to validate and compare models. We took images of two drugs (PTX and MTX), known to be structurally and functionally similar, evaluated and compared their distances to control images in the space of learned features. However, this analysis stays limited to the choice of drugs. Although PTX and MTX made the best example for this dataset to use *a-priori* knowledge in validation and comparison of learned features, the results cannot be generalized for any pair of drugs.

A common practice to evaluate learned representations is to apply them to different tasks and datasets. Often, linear evaluation and transfer learning scenarios are tested. However, this is the case when representations are learned from multi-class general purpose datasets (e.g., ImageNet). On the contrary, biological imaging datasets are specific. It has been reported that even SOTA models trained on ImageNet drop their performance significantly on such datasets[19]. In this study, we had a large imbalanced unlabeled dataset of 1.1M cell images under 693 different conditions over time. We sampled from it in the way to formulate a balanced binary classification problem, which in turn drastically limited further transfer learning applications.

To the date, no consensual measure to evaluate clustering results has been proposed[31]. A number of metrics, such as Adjusted Rand Index, Silhouette score, Normalized Mutual Information, etc., are typically used together to compare results. Most metrics, however, require the ground truth labelling, which were not available in this study. Besides, the clustering itself can be approached in many different ways, using the classical or the newly developed deep-learning based algorithms[17]. In this study, we only intended to fairly compare clustering results, obtained under identical conditions (same algorithm, grid search parameters, evaluation metrics, etc.)

In this study, we have demonstrated a number of ways to analyze large biological datasets with different representation learning paradigms. Similar approaches can be applied to address actual problems in healthcare and biotech industry (e.g., deriving drug onset times, characterizing concentration-dependent pharmacodynamics, exploring opportunities for combination therapy, etc.) In this context, it is important for the scientific community to see that SOTA methods (such as BYOL) can be successfully trained on large datasets within reasonable time using limited resources.

Conclusion

We applied different AI paradigms to learn representations of images of drug treated cancer cell lines. We implemented and trained 16 deep learning setups under identical conditions to ensure fair comparison of learned representations. We evaluated them on 3 tasks using multiple metrics to quantify performance. We made the following observations:

- Multi-crops and augmentations generally improve performance in downstream tasks, as expected. Of 40 rows in the comparison **Table 2**, only 6 show superior performance with no augmentations and single crops (bold values in the rightmost column only).
- The CNN backbone trained with BYOL (ICL) showed competitive performance and was the fastest to train. Strikingly, we managed to train it on the 770k dataset using a moderate GPU within 1.5 and 4 hours only (for single and multi-crops). Additionally, double augmenting resulted in improved performance on 2 of 3 downstream tasks.
- Overall, the regularized autoencoder (SSR) produced the most informative features. It delivered the best accuracy and ROAUC in the classification task and the best quality of partitions in the clustering task. However, it required more time to train.
- No single combination of model (architecture) and setting (augmenting and cropping strategy) consistently outperformed the others. Within each model, the top performance on downstream tasks was often shown by different settings.

Our results suggest a combination of contrastive learning and domain-specific regularization as the most promising way to efficiently learn semantically meaningful representations. To achieve top performance in a particular application, we recommend to extensively evaluate the strength of regularization, as well as augmenting and cropping strategies.

Table 2. Summary of comparison. Median drug similarity distances and drug-vs-control classification metrics are given with median absolute deviations. Mean clustering analysis metrics are given with standard deviations. All metrics satisfy the higher the better. Top performance for each model and task is highlighted in bold.

	SSL			
	aug		**no_aug**	
	multi_crop	**one_crop**	**multi_crop**	**one_crop**
***d*(MTX, PTX)**	0.17 ± 0.00	0.17 ± 0.00	0.16 ± 0.00	**0.20 ± 0.00**
D^{-1}(MTX, PTX)	**0.11 ± 0.02**	0.08 ± 0.01	**0.11 ± 0.02**	0.09 ± 0.01
Accuracy	0.72 ± 0.06	0.70 ± 0.06	0.72 ± 0.05	**0.75 ± 0.07**
Precision	0.80 ± 0.06	0.75 ± 0.05	0.75 ± 0.04	**0.86 ± 0.06**
Recall	0.66 ± 0.11	0.69 ± 0.10	**0.72 ± 0.11**	0.65 ± 0.12
ROAUC	0.72 ± 0.06	0.69 ± 0.06	0.70 ± 0.05	**0.75 ± 0.06**
# clusters	**4 ± 2**	**4 ± 2**	**4 ± 2**	3 ± 1
Not noise, %	93 ± 6	93 ± 5	**94 ± 5**	**94 ± 5**
Silhouette	0.32 ± 0.14	0.34 ± 0.17	**0.35 ± 0.16**	0.32 ± 0.08
(Davies-Bouldin)$^{-1}$	0.92 ± 0.79	**0.99 ± 0.83**	0.94 ± 0.89	0.80 ± 0.27
	ICL			
***d*(MTX, PTX)**	**0.27 ± 0.00**	0.24 ± 0.00	0.25 ± 0.00	0.20 ± 0.00
D^{-1}(MTX, PTX)	0.22 ± 0.02	**0.69 ± 0.06**	0.26 ± 0.02	0.64 ± 0.09
Accuracy	**0.62 ± 0.05**	0.60 ± 0.04	0.61 ± 0.04	0.61 ± 0.05
Precision	**0.69 ± 0.05**	0.63 ± 0.04	**0.69 ± 0.05**	**0.69 ± 0.05**
Recall	0.54 ± 0.04	**0.63 ± 0.10**	0.56 ± 0.12	0.55 ± 0.13
ROAUC	**0.62 ± 0.04**	0.59 ± 0.03	0.61 ± 0.04	0.61 ± 0.05
# clusters	**5 ± 3**	4 ± 3	4 ± 2	3 ± 1
Not noise, %	93 ± 4	94 ± 4	**95 ± 5**	**95 ± 4**
Silhouette	0.29 ± 0.09	0.32 ± 0.06	**0.34 ± 0.09**	**0.34 ± 0.12**
(Davies-Bouldin)$^{-1}$	0.74 ± 0.15	0.75 ± 0.14	0.86 ± 0.47	**0.92 ± 0.63**
	WSL			
***d*(MTX, PTX)**	-0.15 ± 0.00	**0.03 ± 0.00**	-0.18 ± 0.00	0.01 ± 0.00
D^{-1}(MTX, PTX)	0.14 ± 0.03	**1.47 ± 0.26**	0.10 ± 0.02	1.20 ± 0.19
Accuracy	0.73 ± 0.05	0.73 ± 0.05	**0.75 ± 0.05**	0.73 ± 0.05
Precision	0.73 ± 0.05	**0.75 ± 0.04**	**0.75 ± 0.04**	**0.75 ± 0.04**
Recall	**0.77 ± 0.12**	0.75 ± 0.11	**0.77 ± 0.11**	**0.77 ± 0.10**
ROAUC	0.72 ± 0.05	0.73 ± 0.05	**0.74 ± 0.05**	0.73 ± 0.05
# clusters	5 ± 4	3 ± 1	4 ± 1	**6 ± 5**
Not noise, %	90 ± 5	**91 ± 7**	89 ± 8	87 ± 7
Silhouette	0.13 ± 0.08	**0.14 ± 0.09**	0.13 ± 0.08	0.12 ± 0.07
(Davies-Bouldin)$^{-1}$	0.55 ± 0.17	0.54 ± 0.14	**0.56 ± 0.17**	0.53 ± 0.10
	SSR			
***d*(MTX, PTX)**	0.17 ± 0.00	**0.19 ± 0.00**	0.15 ± 0.00	0.18 ± 0.00
D^{-1}(MTX, PTX)	**0.12 ± 0.02**	0.08 ± 0.01	0.09 ± 0.02	0.08 ± 0.01
Accuracy	0.73 ± 0.07	**0.76 ± 0.07**	0.70 ± 0.05	0.72 ± 0.06
Precision	0.79 ± 0.05	**0.83 ± 0.05**	0.75 ± 0.04	0.80 ± 0.06
Recall	**0.70 ± 0.11**	0.68 ± 0.11	0.66 ± 0.11	0.66 ± 0.09
ROAUC	0.73 ± 0.06	**0.76 ± 0.06**	0.69 ± 0.05	0.72 ± 0.06
# clusters	**4 ± 1**	**4 ± 2**	3 ± 1	**4 ± 2**
Not noise, %	93 ± 5	93 ± 4	**94 ± 5**	**94 ± 5**
Silhouette	0.32 ± 0.06	0.30 ± 0.09	**0.35 ± 0.15**	0.33 ± 0.12
(Davies-Bouldin)$^{-1}$	0.76 ± 0.20	0.77 ± 0.26	**0.99 ± 0.92**	0.94 ± 0.68
Total bold	11 (20%)	14 (25.5%)	**16 (29%)**	14 (25.5%)

Supplementary material

Supplementary Text 1. Description of the dataset.

To cover a wide range of phenotypic effects in experimental and FDA-approved anticancer drugs, we selected drugs that displayed at least 3 cell lines as resistant and 3 cell lines sensitive in the NCI-60 cancer cell line panel (**Supplementary Table 1**), with a threshold in the $\log_{10}$(GI50) of 1% between the sensitive and resistant groups. The list comprised 31 experimental and FDA-approved anticancer drugs, covering several modes of action of clinical and research interest (**Supplementary Table 2**). The cancer cell lines were grown in RPMI-1640 GlutaMax medium (ThermoFischer) with supplementation of 1% of Penicylin-Streptomycin (Gibco), and 5% of dialyzed fetal bovine serum (Sigma-Aldrich) at 37◦C in an atmosphere of 5% CO2. The seeding density to achieve a confluence of 70% was determined in Nunc 96 well plates (ThermoFischer), and that seeding density was used for experiments with a factor of four correction for the reduction in area between the 96 and 384 well plates, where cells were seeded in 45 uL of medium. Cells were incubated and imaged every two hours in the Incucyte S3 (Sartorious) 10x phase contrast mode from for up to 48 hours before drug addition, in order to achieve optimal cell adherence and starting experimental conditions. To reduce evaporation effects, the plates were sealed with Breathe-Easy sealing membrane (Diversified Biotech).

To allow a broad coverage of effects on time, we collected the time information about when the drugs were treated for each cell line, and corrected the analysis based on the drug treatment. Drugs were resuspended in the appropriate solvent (DMSO or water), and the same amount of DMSO (check amount) was added across all wells, including controls. The randomized 384 drug source plates were generated with Echo Liquid Handling System (Integra-Biosciences), and then transferred in 5uL of medium to Nunc 384 well plates (ThermoFischer) with the AssistPlus liquid handler (Integra Biosciences).

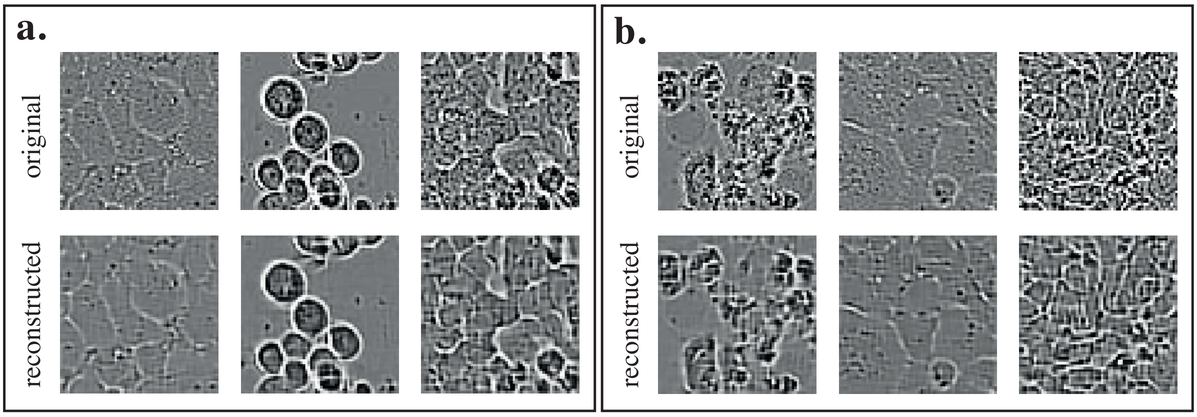
Supplementary Table 1. Cell lines and inoculation densities for 96 well plate format.

Cell line	Panel	Inoculation density
EKVX	Non-Small Cell Lung	11000
HOP-62	Non-Small Cell Lung	9000
COLO 205	Colon	15000
HCT-15	Colon	12000
HT29	Colon	12000
SW-620	Colon	24000
SF-539	CNS	10000
LOX IMVI	Melanoma	8500
MALME-3M	Melanoma	8500
M14	Melanoma	5000
SKMEL-2	Melanoma	10000
UACC-257	Melanoma	20000
IGR-OV1	Ovarian	10000
OVCAR-4	Ovarian	10000
OVCAR-5	Ovarian	15000
A498	Renal	3200
ACHN	Renal	8200
MDA-MB-231 / ATCC	Breast	20000
HS 578T	Breast	13000
BT-549	Breast	10000
T-47D	Breast	15000

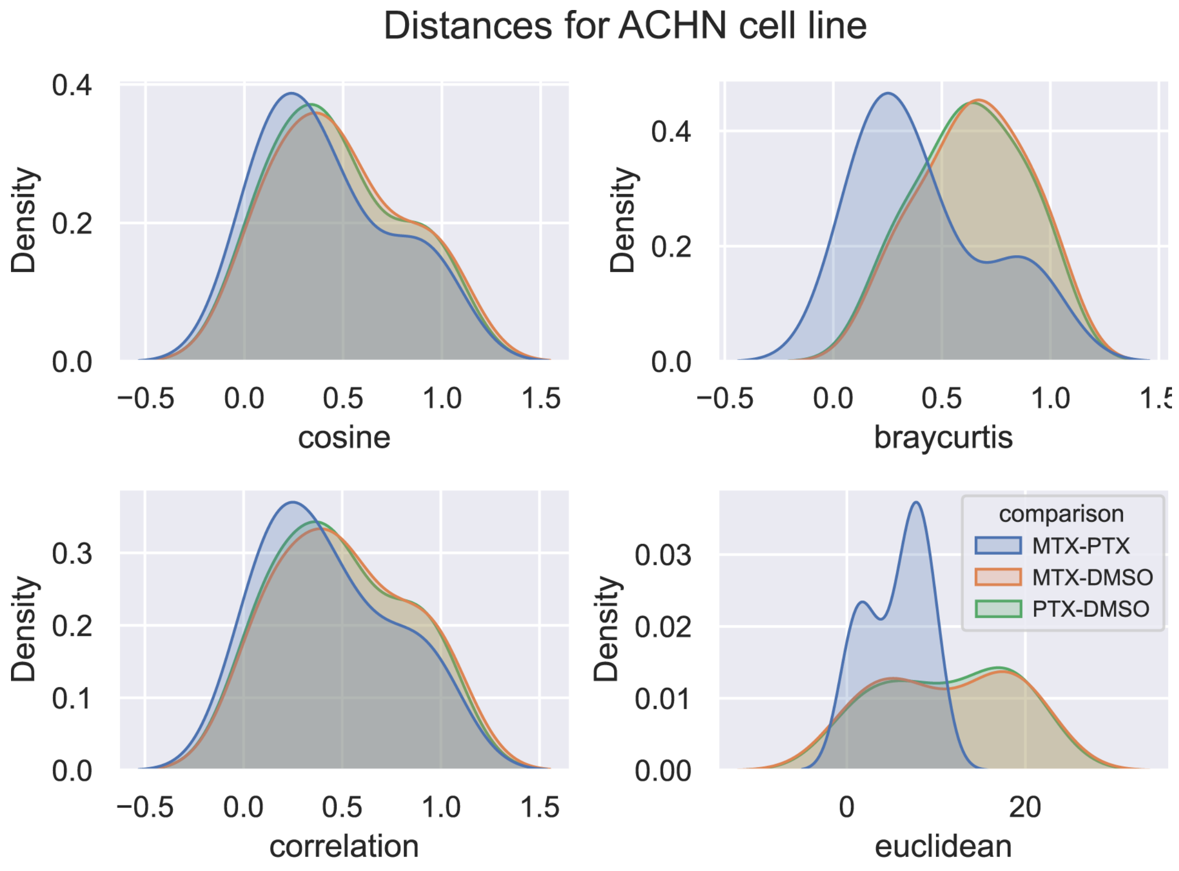
Supplementary Table 2. Drugs, solvents, CAS registry numbers and maximum concentrations. The other four concentrations for each drug were 10x serial dilutions of the maximum concentration.

Drug	Fluid	CAS	Concentration
Erlotinib	DMSO	183321-74-6	10 uM
Irinotecan	DMSO	100286-90-6	10 uM
Clofarabine	DMSO	123318-82-1	10 uM
Fluorouracil	DMSO	51-21-8	10 uM
Pemetrexed	Water	150399-23-8	10 uM
Docetaxel	DMSO	148408-66-6	1 uM
Everolimus	DMSO	159351-69-6	1 uM
Chlormethine	DMSO	55-86-7	10 uM
BPTES	DMSO	314045-39-1	10 uM
Oligomycin A	DMSO	579-13-5	1 uM
UK-5099	DMSO	NA	10 uM
Panzem (2-ME2)	DMSO	362-07-2	10 uM
MEDICA16	DMSO	87272-20-6	10 uM
Gemcitabine	Water	122111-03-9	1 uM
17-AAG	DMSO	75747-14-7	10 uM
Lenvatinib	DMSO	417716-92-8	10 uM
Topotecan	DMSO	119413-54-6	1 uM
Cladribine	DMSO	4291-63-8	10 uM
Mercaptopurine	DMSO	6112-76-1	10 uM
Decitabine	DMSO	2353-33-5	10 uM
Methotrexate	DMSO	59-05-2	1 uM
Paclitaxel	DMSO	33069-62-4	1 uM
Rapamycin	DMSO	53123-88-9	0.1 uM
Oxaliplatin	DMSO	61825-94-3	10 uM
Omacetaxine	DMSO	26833-87-4	1 uM
Metformin	Water	1115-70-4	10 uM
YC-1	DMSO	170632-47-0	10 uM
Etoximir	DMSO	828934-41-4	10 uM
Oxfenicine	DMSO	32462-30-9	2.5 uM
Trametinib	DMSO	871700-17-3	1 uM
Asparaginase	Water	9015-68-3	0.00066 units / uL

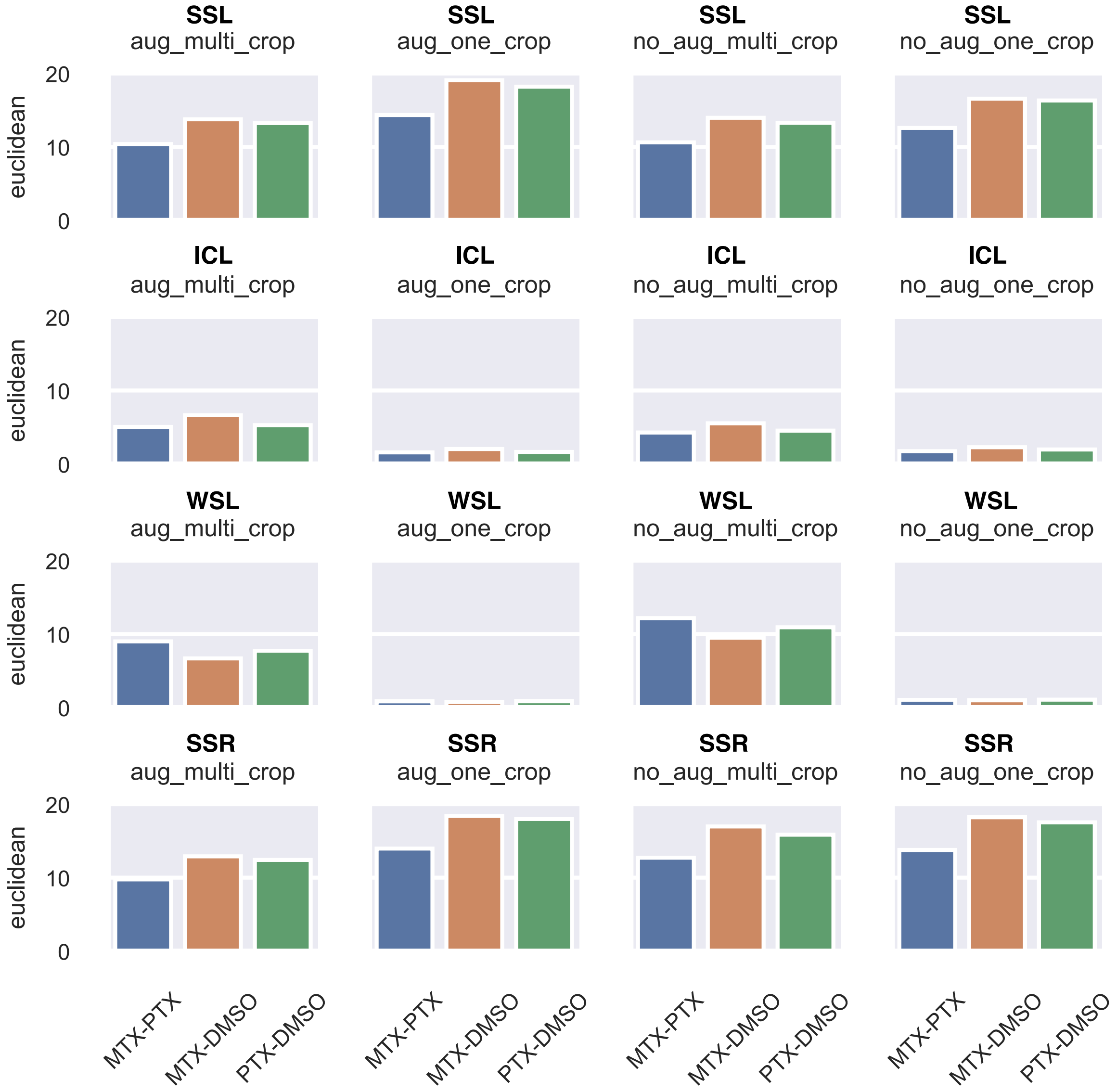
Supplementary Figure 1. **Quality of image reconstructions**. Random examples of reconstructed and original images for the unsupervised (**a**) and the regularized (**b**) models. Regularization did not harm the quality of reconstructions. The learning capacity of the CNN backbone was sufficient to capture normal and altered morphology of the cells.



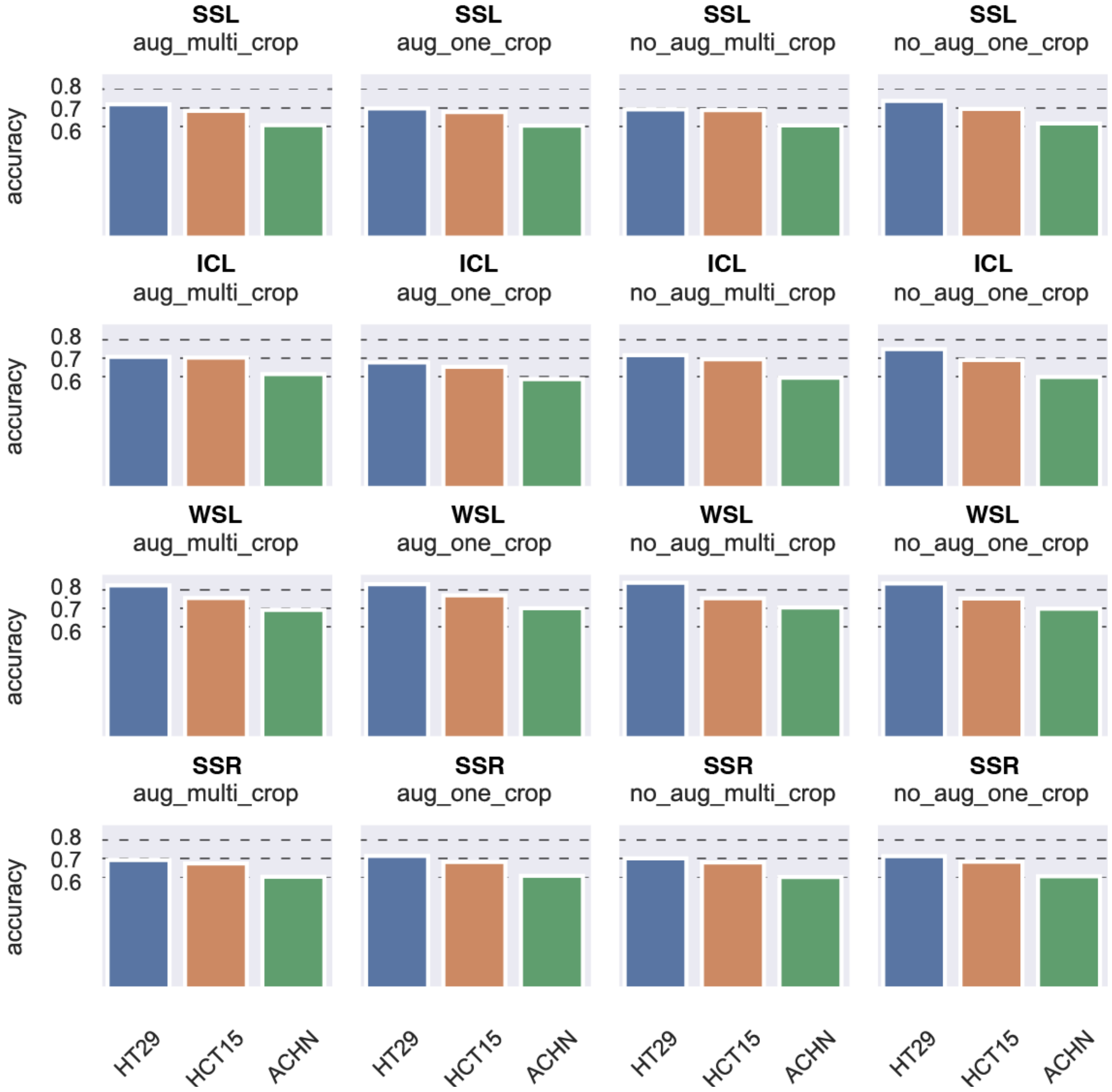
Supplementary Figure 2. Motivation for Euclidean distance in the similarity analysis. We tested several distances to investigate how close the two drugs (PTX and MTX) were to each other and both distant from control (DMSO) in the space of learned features. We found a number of cases, where cosine and correlation distances could not differentiate between drugs and controls, i.e., *D*(PTX, MTX) ≈ *D*(PTX, DMSO) ≈ *D*(MTX, DMSO). Whereas Bray-Curtis and Euclidean distances both resulted in *D*(PTX, MTX) < *D*(PTX, DMSO) and *D*(PTX, MTX) < *D*(MTX, DMSO). The figure explains it very clearly: although distributions of cosine and correlation distances are both slightly shifted towards zero for MTX-PTX comparison (blue), these effects are much stronger for Bray-Curtis and Euclidean distances. From this we concluded that Euclidean distance was the most informative for the drug similarity analysis of PTX and MTX.



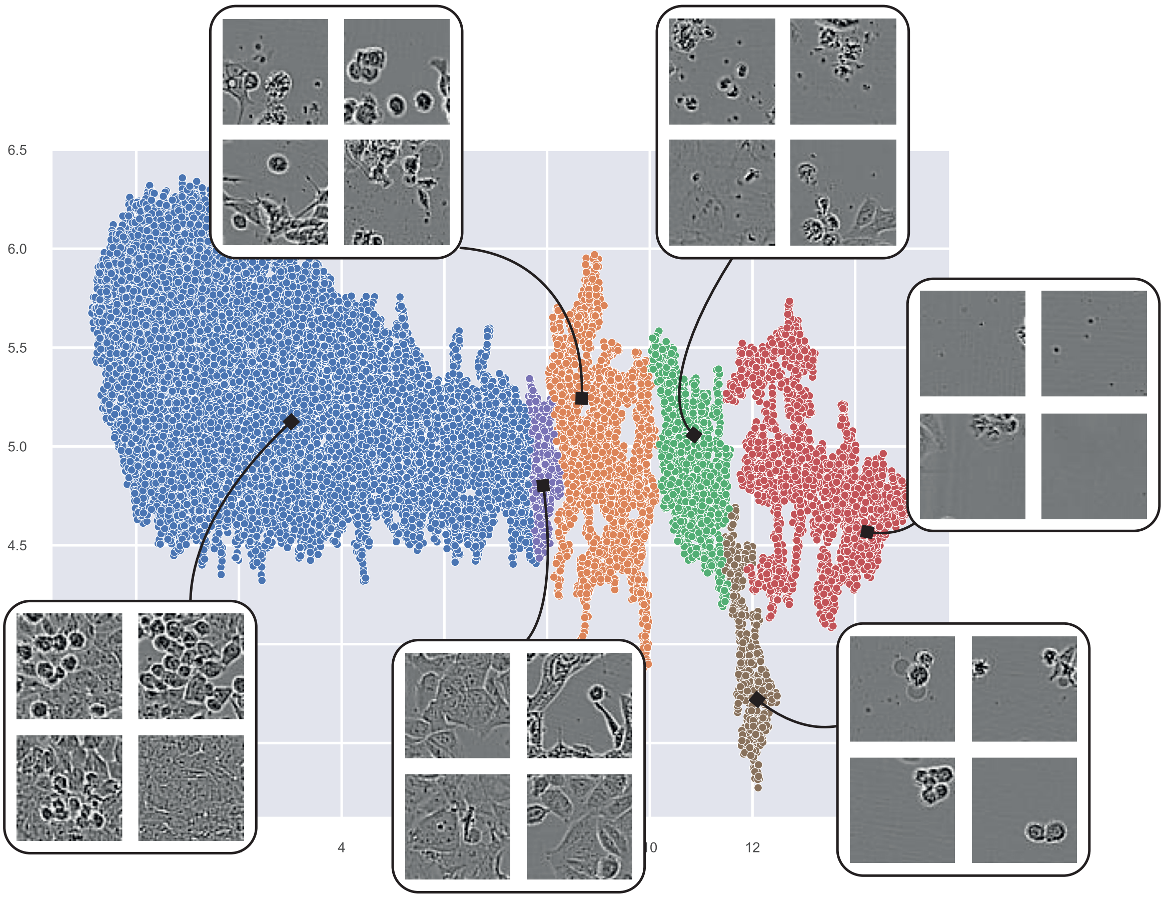
Supplementary Figure 3. Distance-based similarity analysis for M14. Analyzing distances for M14 cell line, we observed that in the latent space the two drugs (PTX and MTX) were closer to each other than either of them to controls (DMSO) for two models only: the unsupervised (row 1) and the regularized (row 4) ones. The distances for the ICL and WSL models (rows 2 and 3) were rather on the same level. Strikingly, the one-crop setup for both of them (columns 2 and 4) resulted in distances close to zero, which implies that information in the learned representations was insufficient to characterize drug effects. Multi-crop setting, in turns, caused large increase in distances, which suggests information gain. Nonetheless, it was not enough to capture dissimilarity between drugs and controls in this case.



Supplementary Figure 4. Binary classification for HT29, HCT15 and ACHN. Binary classification accuracy (drug vs control) for three picked cell lines: HT29, HCT15, ACHN. Weakly-supervised architecture was the only one to reach 0.8 accuracy for HT29 and cross 0.7 accuracy in all settings for HCT15 and ACHN.



Supplementary Figure 5. Clustering of HCT cell line representations. Clustering example of 20480 images (HCT15 cell line) with random cluster representatives. Each point is a 2D UMAP embedding of the learned image representations (self-supervised model). Clusters found by HDBSCAN are highlighted in colors. The left cluster (blue) contains drugs of no effect on HCT15. The right cluster (red) contains the drugs of the strongest effect.



References

(1) Moen, E.; Bannon, D.; Kudo, T.; Graf, W.; Covert, M.; Van Valen, D. Deep Learning for Cellular Image Analysis. *Nat. Methods* **2019**, *16* (12), 1233–1246. https://doi.org/10.1038/s41592-019-0403-1.

(2) Godinez, W. J.; Hossain, I.; Lazic, S. E.; Davies, J. W.; Zhang, X. A Multi-Scale Convolutional Neural Network for Phenotyping High-Content Cellular Images. *Bioinformatics* **2017**, *33* (13), 2010–2019. https://doi.org/10.1093/bioinformatics/btx069.

(3) Kraus, O. Z.; Grys, B. T.; Ba, J.; Chong, Y.; Frey, B. J.; Boone, C.; Andrews, B. J. Automated Analysis of High-content Microscopy Data with Deep Learning. *Mol. Syst. Biol.* **2017**, *13* (4), 924. https://doi.org/10.15252/msb.20177551.

(4) Hu, B.; Tang, Y.; Chang, E. I. C.; Fan, Y.; Lai, M.; Xu, Y. Unsupervised Learning for Cell-Level Visual Representation in Histopathology Images with Generative Adversarial Networks. *IEEE J. Biomed. Heal. Informatics* **2019**, *23* (3), 1316–1328. https://doi.org/10.1109/JBHI.2018.2852639.

(5) Goldsborough, P.; Pawlowski, N.; Caicedo, J. C.; Singh, S.; Carpenter, A. E. CytoGAN: Generative Modeling of Cell Images. *bioRxiv* **2017**, No. Nips, 227645.

(6) Radford, A.; Metz, L.; Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *4th Int. Conf. Learn. Represent. ICLR 2016 - Conf. Track Proc.* **2016**, 1–16.

(7) Robitaille, M. C. . et al. A Self-Supervised Machine Learning Approach for Objective Live Cell Segmentation and Analysis Michael. *bioRxiv* **2021**.

(8) Li, X.; Wang, K.; Lyu, Y.; Pan, H.; Zhang, J.; Stambolian, D.; Susztak, K.; Reilly, M. P.; Hu, G.; Li, M. Deep Learning Enables Accurate Clustering with Batch Effect Removal in Single-Cell RNA-Seq Analysis. *Nat. Commun.* **2020**, *11* (1), 1–14. https://doi.org/10.1038/s41467-020-15851-3.

(9) Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; Joulin, A. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. *Adv. Neural Inf. Process. Syst.* **2020**, *2020-Decem* (NeurIPS), 1–13.

(10) Caicedo, J. C.; Weakly Supervised Learning of Single-Cell Feature Embeddings. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit.* **2018**, 9309–9318. https://doi.org/10.1109/CVPR.2018.00970.Weakly.

(11) Lu, M. Y.; Williamson, D. F. K.; Chen, T. Y.; Chen, R. J.; Barbieri, M.; Mahmood, F. Data-Efficient and Weakly Supervised Computational Pathology on Whole-Slide Images. *Nat. Biomed. Eng.* **2021**, *5* (6), 555–570. https://doi.org/10.1038/s41551-020-00682-w.

(12) Longden, J.; Robin, X.; Engel, M.; Ferkinghoff-Borg, J.; Kjær, I.; Horak, I. D.; Pedersen, M. W.; Linding, R. Deep Neural Networks Identify Signaling Mechanisms of ErbB-Family Drug Resistance from a Continuous Cell Morphology Space. *Cell Rep.* **2021**, *34* (3), 108657. https://doi.org/10.1016/j.celrep.2020.108657.

(13) Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein GAN. **2017**.

(14) Gulrajani, I. et al. Improved Training OfWasserstein GANs Ishaan. *Proc. 31st Int. Conf. Neural Inf. Process. Syst.* **2017**.

(15) Lu, A. X.; Kraus, O. Z.; Cooper, S.; Moses, A. M. Learning Unsupervised Feature Representations for Single Cell Microscopy Images with Paired Cell Inpainting. *PLoS Comput. Biol.* **2019**, *15* (9), 1–27. https://doi.org/10.1371/journal.pcbi.1007348.

(16) Santos-Pata, D.; Amil, A. F.; Raikov, I. G.; Rennó-Costa, C.; Mura, A.; Soltesz, I.; Verschure, P. F. M. J. Entorhinal Mismatch: A Model of Self-Supervised Learning in the Hippocampus. *iScience* **2021**, *24* (4). https://doi.org/10.1016/j.isci.2021.102364.

(17) Ciortan, M.; Defrance, M. Contrastive Self-Supervised Clustering of ScRNA-Seq Data. *BMC Bioinformatics* **2021**, *22* (1), 1–27. https://doi.org/10.1186/s12859-021-04210-8.

(18) Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. E. A Simple Framework for Contrastive Learning of Visual Representations. BT - Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event. **2020**, No. Figure 1, 1597–1607.

(19) Grill, J. B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P. H.; Buchatskaya, E.; Doersch, C.; Pires, B. A.; Guo, Z. D.; Azar, M. G.; Piot, B.; Kavukcuoglu, K.; Munos, R.; Valko, M. Bootstrap Your Own Latent a New Approach to Self-Supervised Learning. *Adv. Neural Inf. Process. Syst.* **2020**, *2020-Decem* (NeurIPS), 1–14.

(20) Padi, S.; Manescu, P.; Schaub, N.; Hotaling, N.; Simon, C.; Bharti, K.; Bajcsy, P. Comparison of Artificial Intelligence Based Approaches to Cell Function Prediction. *Informatics Med. Unlocked* **2020**, *18* (October 2019), 100270. https://doi.org/10.1016/j.imu.2019.100270.

(21) Chandrasekaran, S. N.; Ceulemans, H.; Boyd, J. D.; Carpenter, A. E. Image-Based Profiling for Drug Discovery: Due for a Machine-Learning Upgrade? *Nat. Rev. Drug Discov.* **2021**, *20* (2), 145–159. https://doi.org/10.1038/s41573-020-00117-w.

(22) Nguyen, G.; Dlugolinsky, S.; Bobák, M.; Tran, V.; López García, Á.; Heredia, I.; Malík, P.; Hluchý, L. Machine Learning and Deep Learning Frameworks and Libraries for Large-Scale Data Mining: A Survey. *Artif. Intell. Rev.* **2019**, *52* (1), 77–124. https://doi.org/10.1007/s10462-018-09679-z.

(23) Brigato, L.; Barz, B.; Iocchi, L.; Denzler, J. Tune It or Don't Use It: Benchmarking Data-Efficient Image Classification. *Proc. IEEE Int. Conf. Comput. Vis.* **2021**, *2021-Octob*, 1071–1080. https://doi.org/10.1109/ICCVW54120.2021.00125.

(24) Chen, L.; Zhai, Y.; He, Q.; Wang, W.; Deng, M. Integrating Deep Supervised, Self-Supervised and Unsupervised Learning for Single-Cell Rna-Seq Clustering and Annotation. *Genes (Basel).* **2020**, *11* (7), 1–20. https://doi.org/10.3390/genes11070792.

(25) Creswell, A.; Arulkumaran, K.; Bharath, A. A. On Denoising Autoencoders Trained to Minimise Binary Cross-Entropy. **2017**, No. 1, 2–9.

(26) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.;

Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4* (2), 268–276. https://doi.org/10.1021/acscentsci.7b00572.

(27) Rong, Z.; Tan, Q.; Cao, L.; Zhang, L.; Deng, K.; Huang, Y.; Zhu, Z. J.; Li, Z.; Li, K. NormAE: Deep Adversarial Learning Model to Remove Batch Effects in Liquid Chromatography Mass Spectrometry-Based Metabolomics Data. *Anal. Chem.* **2020**, *92* (7), 5082–5090. https://doi.org/10.1021/acs.analchem.9b05460.

(28) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. **2018**.

(29) McInnes, L.; Healy, J.; Astels, S. Hdbscan: Hierarchical Density Based Clustering. *J. Open Source Softw.* **2017**, *2* (11), 205. https://doi.org/10.21105/joss.00205.

(30) Ruszkowski, M.; Sekula, B.; Ruszkowska, A.; Contestabile, R.; Nogues, I.; Angelaccio, S.; Szczepaniak, A.; Dauter, Z. Structural Basis of Methotrexate and Pemetrexed Action on Serine Hydroxymethyltransferases Revealed Using Plant Models. *Sci. Rep.* **2019**, *9* (1), 1–14. https://doi.org/10.1038/s41598-019-56043-4.

(31) Palacio-Niño, J.-O.; Berzal, F. Evaluation Metrics for Unsupervised Learning Algorithms. **2019**.

Chapter 6

Self-supervised learning for analysis of temporal and morphological drug effects in cancer cell imaging data

Andrei Dmitrenko, Mauro M. Masiero and Nicola Zamboni

This manuscript is accepted for publication in *Proceedings of Machine Learning Research* as a full paper presented at MIDL 2022 conference (oral).

Author contributions:

AD and MM conceived the study. MM collected the imaging dataset. AD developed, trained and validated the deep learning models. AD and MM developed the workflow to analyze temporal drug effects. AD developed the workflow to analyze morphological drug effects. AD and MM wrote the paper. NZ supervised the study.

Abstract

In this work, we propose two novel methodologies to study temporal and morphological phenotypic effects caused by different experimental conditions using imaging data. As a proof of concept, we apply them to analyze drug effects in 2D cancer cell cultures. We train a convolutional autoencoder on 1M images dataset with random augmentations and multi-crops to use as feature extractor. We systematically compare it to the pretrained state-of-the-art models. We further use the feature extractor in two ways. First, we apply distance-based analysis and dynamic time warping to cluster temporal patterns of 31 drugs. We identify clusters allowing annotation of drugs as having cytotoxic, cytostatic, mixed or no effect. Second, we implement an adversarial/regularized learning setup to improve classification of 31 drugs and visualize image regions that contribute to the improvement. We increase top-3 classification accuracy by 8% on average and mine examples of morphological feature importance maps. We provide the feature extractor and the weights to foster transfer learning applications in biology. We also discuss utility of other pretrained models and applicability of our methods to other types of biomedical data.

Introduction

Deep learning has been extensively applied to the analysis of biological images[1–4]. Learning cellular features from imaging data in an automated way, instead of designing them manually with expert knowledge, resulted in a remarkable progress across many tasks, such as classification and segmentation, object tracking and others[5].

Among many studies based on deep representation learning, Yang et al. investigated cell trajectories in the feature space along the time axis[6]. Lu et al. exploited distance measures in the feature space to quantify similarity of cells[7]. However, no study applied distance-based analysis of temporal drug effects using learned representations. In this study, we develop a workflow to analyze effects of anti-cancer drugs with time.

Many efforts have gone into improving interpretability of deep learning for biomedical applications[8,9]. Several methods have been used to study cellular phenotypes using variational autoencoders[10] (VAEs) and generative adversarial networks[11] (GANs). Here, we propose another way to gain insights into morphological features of cells driving drug classification. As a proof of concept, we apply it to improve classification of anti-cancer drugs and visualize image regions contributing to that improvement.

Therefore, our main contributions are:

- We train a convolutional autoencoder (ConvAE) on 1M cancer cell images using random augmentations and multi-crops. We provide the source code and the model for future transfer learning applications at https://github.com/dmitrav/pheno-ml.
- We propose a workflow to study temporal drug effects using learned representations of images with distance-based clustering analysis.
- We propose an adversarial/regularized learning setup to improve multiclass classification of drugs and visualize morphological features driving classifier decisions.

Related work

State-of-the-art (SOTA) general purpose pretrained models (e.g., ResNet-50 trained with SwAV[12] or DINO[13]) are often used for transfer learning applications[14]. However, their performance may drop significantly on specific datasets[15] such as ours. Models trained on biological data are available, but they are usually trained on smaller datasets. Services and tools exist to assist on biological image analysis (such as CellProfiler[16] or DeepImageJ[17]). However, they are not designed to handle high-throughput and often do not provide direct access to extracted feature vectors. In this work, we train ConvAE on 1M image dataset comprising 21 cell lines and 31 cancer drugs on 5 concentrations. We use random augmentations and multi-crops, prove the representations contain meaningful biological information and provide the trained model with minimal API to extract features.

A number of approaches to improve interpretability of deep learning are based on autoencoders[18,19]. Often, they are used to localize and visualize pathologies or lesions[20]. Perhaps, the closest approach to ours is the one by Chen et al. The authors train a VAE on healthy subjects and then use it to detect outliers with MAP-based restoration[21]. That is, the lesions are detected as noise in the process of image restoration. The detected regions are then visualized by calculating the difference between input and restored images. In this work, we use ConvAE as a feature extractor in a regularized learning setup to train the lens model[22], conceptually introduced by Sajjadi et al. The resulting morphological feature importance maps are then obtained by calculating the difference between the reconstructed and the lensed images.

Data

We used advanced robotics, assay miniaturization and high-throughput imaging to acquire a dataset comprising 1M phase-contrast images, covering 21 cancer cell lines exposed to 31 experimental and FDA-approved clinical cancer drugs at 5 logs of concentration, where every condition was imaged every 2 hours for up to 6 days (**Figure 1**). Detailed description of the data is given in **Supplementary Text 1**.

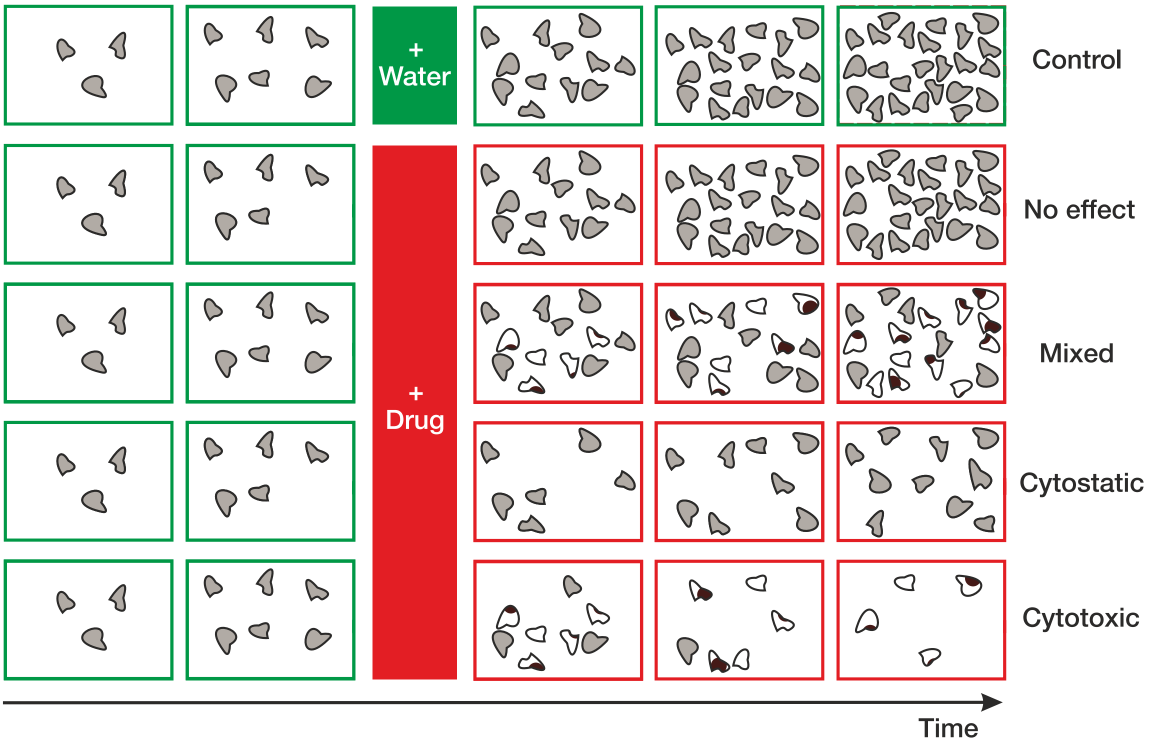


Figure 1. Schematic of the dataset and the expected drug effects over time (on the right).

Methods

Learning representations

We adopted a convolutional autoencoder (ConvAE) to learn image representations. We experimented with architectures to achieve good reconstruction quality and reasonable training time, as we used a single Nvidia GeForce RTX 2060 with 6GB only (see **Supplementary Text 2** for details). We ended up with an architecture of 3 convolutional layers for encoder and decoder parts, having a relatively large receptive field (maps of 32 × 32 pixels in the bottleneck layer). The total number of parameters stayed rather low (190k), which allowed faster training and feature extraction, as well as lower memory consumption.

Augmenting and cropping strategies

Since we had naturally grayscale images, we only applied random resized crops (RRCs), augmented with random Gaussian blur and horizontal flip. However, we tested a number of cropping strategies. The initial 256 × 256 images were randomly cropped and resized to 128 × 128, but the scale of RRCs varied. We tested combinations of full images and square crops of about half size and about quarter size (e.g., the following 3-crop strategy: 1 full image, 1 square crop of random size between 128-256 pixels, 1 square crop of random size between 64-128 pixels). We tested 12 cropping strategies, always having a full image and up to 4 additional RRCs of different sizes.

Evaluation and comparison to the pretrained models

We compared image representations obtained with ConvAE and general-purpose SOTA models pretrained on ImageNet: i) supervised ResNet-50, ii) self-supervised ResNet-50 (SwAV), iii) self-supervised ResNet-50 (DINO), iv) self-supervised ViT-B/8 (DINO). We evaluated performance of each model on 3 downstream tasks using multiple metrics.

Similarity of biological replicates

First, we analyzed similarity of biological replicates in the latent space. For that, we picked the images of drugs at maximum concentrations and latest time points, where the strongest effect must be observable if present. We did that for each cell line and calculated distances between every pair of images of the same drug. We used the following distances to estimate similarity: Euclidean, cosine, correlation and Bray-Curtis. Since biological replicates are expected to display the same effects, we expected the distances to be lower for those methods that capture the similarity well.

Clustering of drug effects

Next, we performed clustering of images within each cell line. We retrieved latent representations, reduced dimensions with UMAP[23] and ran HDBSCAN[24] clustering over multiple parameter sets. Since the true labels of drug effects were not available in this study, we evaluated the quality of partitions with the following metrics: percent of noise points, Silhouette score, Davies-Bouldin measure, Calinski-Harabasz index.

We picked the best clustering performance over parameters sets and averaged them across cell lines.

Classification of drugs vs controls

Finally, we formulated a classification problem to differentiate between drugs and controls. We assigned label 1 to the images of maximum drug concentrations and label 0 to the control (no drug) images. We trained two-layer classifiers and calculated a few standard metrics: accuracy, recall, precision, specificity. The resulting setting is only weakly supervised, since some drugs did not in fact provoke any effect.

Analysis of temporal drug effects

To characterize temporal drug effects, we calculated distances between drug and control image representations at every time point and clustered trajectories of distances over time. More specifically, we aligned images of drugs and controls along the time axis first. Then, we retrieved their latent representations, averaged features across biological replicates and calculated distance to control for each drug at every time point. Finally, we normalized distances for each experimental condition, applied dynamic time warping (DTW) and k-means to cluster temporal patterns (**Figure 2**). In this setting, rapidly growing distance (fast divergence from control) is expected for immediate strong drug effect. And vice versa, low distance to control along the entire timeline is expected for no observable effect. We tested several distance metrics. For *k*-means, we incremented *k* by 1 to find the minimum number of clusters covering the expected biological patterns.

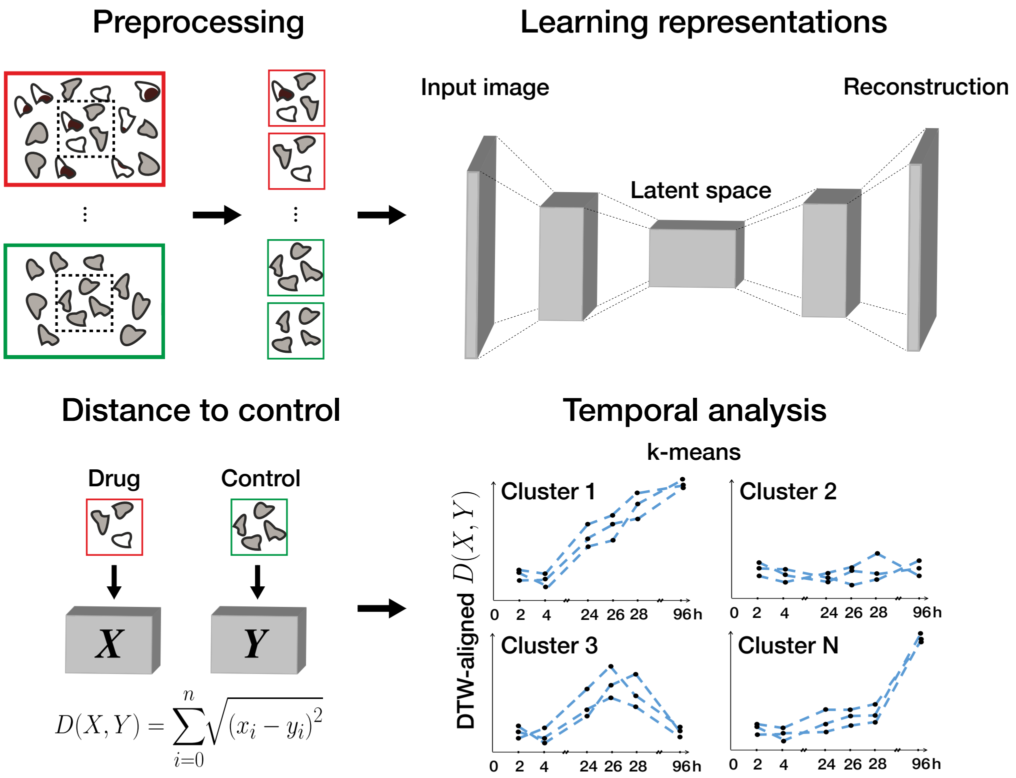


Figure 2. An overview of temporal analysis.

Analysis of morphological drug effects

Following the idea of shortcut removal[25], we leveraged an adversarial learning setup to improve multiclass classification of drugs using the best pretrained model as feature extractor (**Figure 3**). The lens was trained on the images of the highest drug concentrations and the latest time points using the following loss function: $L = L_{rec} - \alpha \cdot L_{disc}$, where L_{rec} is the image reconstruction loss, L_{disc} is the drug discrimination loss, and α is an adversary coefficient. We used the same ConvAE architecture for the lens and ran grid search for $\alpha \in [-60,60]$, evaluating classification accuracy on the lensed images. Negative values of α correspond to the regularized learning.

In cases of improved classification accuracy, we visualized regions on the images perturbed by the lens. We did that by plotting the absolute difference between the lensed and the reconstructed images. The resulting regions serve as morphological feature importance maps, as they highlight regions of altered cell morphology important for drug classification.

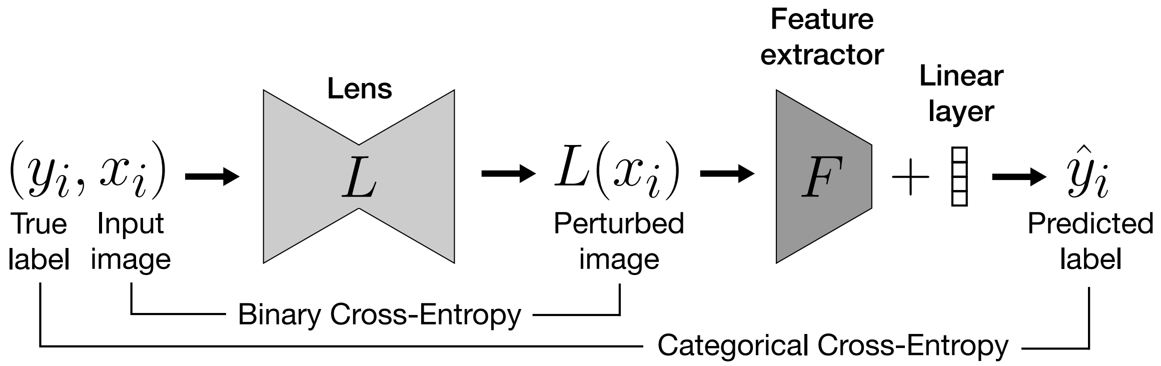


Figure 3. A schematic of the lens setup.

Results

Multi-crops improve performance on downstream tasks

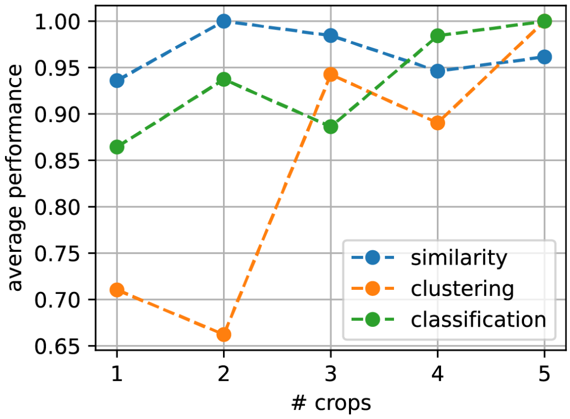


Figure 4. Lowest mean squared error achieved for subsets of dilutions.

We experimented with scales of RRCs and averaged their performance for each n-crop strategy, where n ∈ {2, 3, 4, 5}. We used multiple metrics corresponding to a particular task to average. We further normalized performance on each task, so that the top performance equals to 1. As expected, we observed that increasing number of multi-crops improves the performance across tasks on average (**Figure 4**).

However, different scales of RRCs sometimes led to sporadic drops in performance on particular tasks. The best scores across tasks were achieved by the following 5-crop strategy: 1 full size image, 1 square crop of random size between 128-256 pixels, 3 square crops of random size between 64-128 pixels. That strategy was used for training ConvAE on the entire dataset and further evaluations.

Comparison of pretrained state-of-the-arts

We compared several pretrained models with the ConvAE on three downstream tasks as described earlier. Median metrics are reported in **Table 1**.

Table 1. Comparison of pretrained models. All metrics are the higher the better.

	Similarity		**Clustering**		**Classification**	
	(Euclidean)$^{-1}$	(Cosine)$^{-1}$	Silhouette	(Davies)$^{-1}$	Accuracy	F1
ResNet-50	0.10 ± 0.01	1.21 ± 0.03	0.25 ± 0.10	0.46 ± 0.18	0.59 ± 0.02	0.59 ± 0.05
ResNet-50 (SwAV)	**2.75 ± 0.63**	**5.37 ± 2.39**	**0.47 ± 0.13**	**0.85 ± 0.71**	0.77 ± 0.01	0.78 ± 0.01
ResNet-50 (DINO)	1.07 ± 0.02	1.12 ± 0.02	0.34 ± 0.11	0.52 ± 0.27	0.72 ± 0.00	0.73 ± 0.00
ViT-B/8 (DINO)	2.18 ± 0.42	4.57 ± 1.91	0.44 ± 0.12	0.70 ± 0.52	0.81 ± 0.00	0.82 ± 0.00
ConvAE (trained)	2.26 ± 0.68	1.53 ± 0.27	0.30 ± 0.11	0.39 ± 0.24	**0.85 ± 0.05**	**0.85 ± 0.05**

Surprisingly, ResNet-50 pretrained on ImageNet with SwAV algorithm showed the best performance on similarity and clustering tasks. That indicates high level of consistency of the learned representations, obtained with SwAV. On the other hand, the best classification accuracy (drug vs control) and F1 score were shown by our model, followed by ViT-B/8 pretrained with DINO. Therefore, features extracted by the pretrained models lacked some domain-specific information to better differentiate between drug and control images. Notably, a small model such as ours (ConvAE) can show rival performance with pretrained state-of-the-arts when trained on large enough dataset.

Proof-of-concept: studying temporal drug effects

First, we analyzed representations learned by ConvAE and made sure they exhibit expected spatial and temporal separation patterns (see **Supplementary Figure 1**). Then, we calculated Euclidean distance to control for each experimental condition at each time point. We further scaled, DTW-aligned and clustered the distances as described earlier.

We found three clear patterns: i) no response, where the distance between drug-treated condition and control either stays constant or decays, so that both conditions become indistinguishable; ii) temporary response (cytostatic effect), where an initial divergence from control is observed, but ultimately reduced towards the end time points; iii) constant response (cytotoxic effect), where the distance grows throughout the experiment (**Figure 5A**). Red lines are mean cluster representatives. Full clustering is shown on **Supplementary Figure 2**.

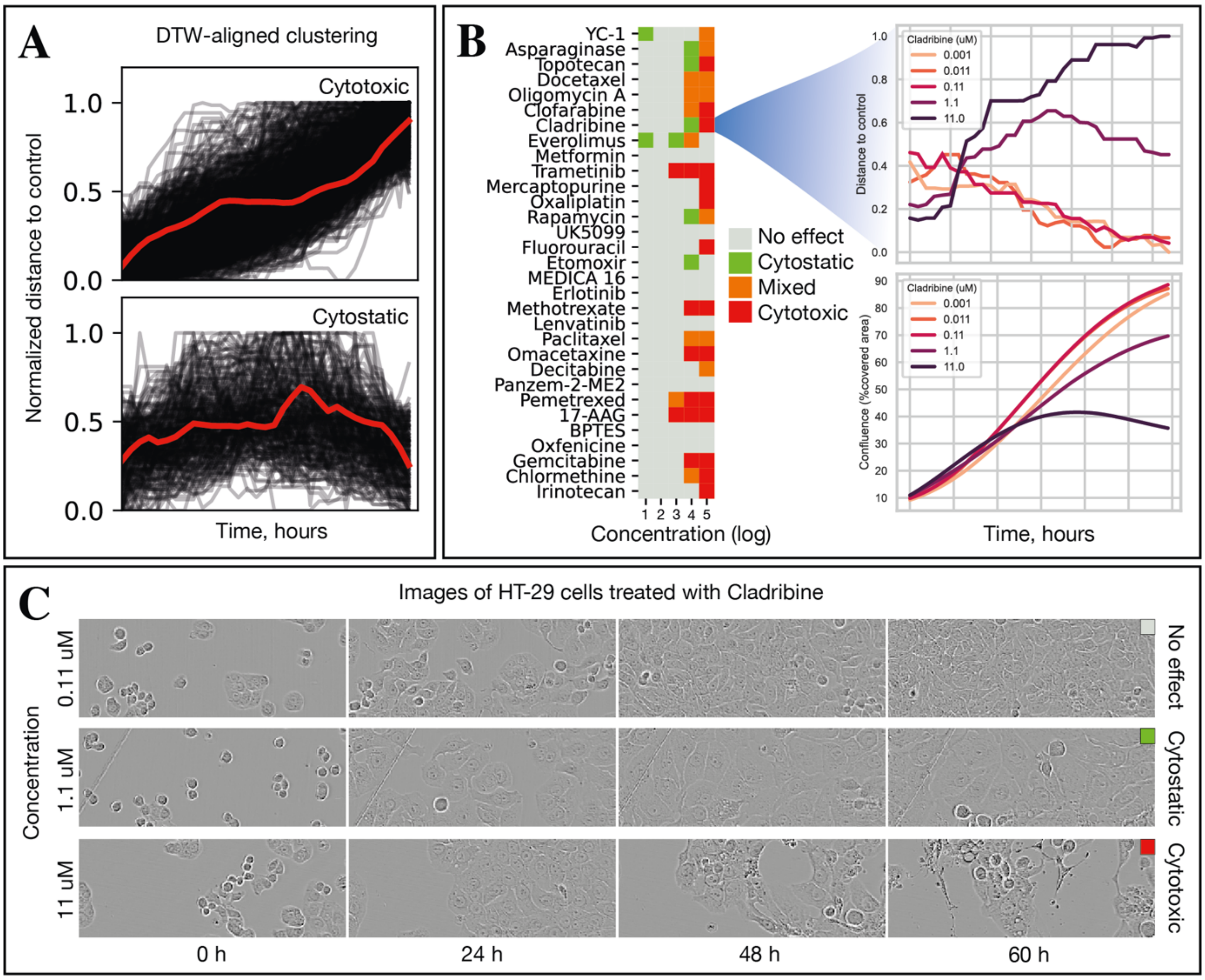


Figure 5. Analysis of temporal drug effects.

Analyzing these patterns, we were able to annotate concentration-dependent effects for all drugs in the dataset (**Figure 5B**). Interestingly, we observed that some drugs (e.g., Cladribine) switched between cytostatic and cytotoxic modes of action, as the concentration was increased. Notably, this was impossible to detect analyzing classical growth curves, as the confluence grew for all concentrations, but the highest (**Figure 5B**, bottom-right). Conversely, distance-based analysis of learned representations allowed picking up another distinct response pattern (**Figure 5B**, top-right).

To validate such patterns, we visually inspected the corresponding images with time, as shown on **Figure 5C** for HT-29 cell line and three concentrations of Cladribine. The first row shows no effect, these images look identical to controls. In the last row, irregular cell morphology features (such as granules and bubbles) associated with cytotoxic effect can be seen. For 1.1 µM concentration in the middle, we indeed observed temporary proliferation arrest, accompanied with increased cell sizes. It proves that our method can distinguish between different response patterns and thus, can be used for studying temporal drug effects.

Proof-of-concept: exploring morphological drug effects

We used the trained ConvAE as feature extractor to train the lens. We observed negligible lens effects with $|\alpha| \leq 1$. Increasing $|\alpha|$ up to 60 we were able to obtain consistent improvement of top-3 classification metrics (**Figure 6**). With $\alpha = -60$, we were able to improve classification accuracy by 8%, which is significant. We looked into examples of improved classification and plotted differences between the reconstructed and the lensed images.

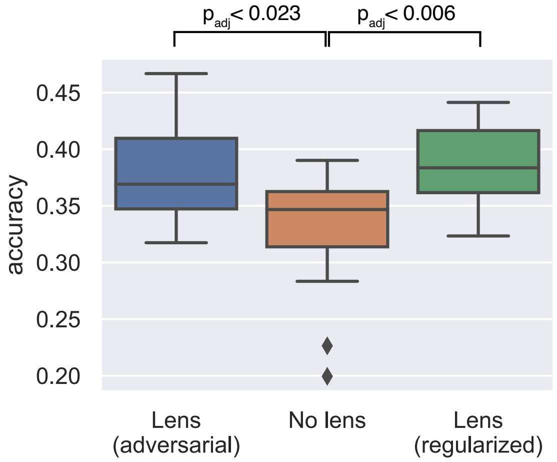


Figure 6. Multi-class classification accuracy with no ($\alpha = 0$), adversarial ($\alpha = 60$) and regularizing ($\alpha = -60$) lens.

By design, such differences highlight regions on the image, that caused changes in the classification results.

We considered three cases of improved classification with lens as the most useful to study morphological drug effects: i) when the classifier initially confused a drug with a control, and the lens resolved the issue; ii) when the classifier initially confused a drug with another drug, and the lens resolved the issue; iii) when the classifier initially had low probability of correct class, and the lens dramatically increased that probability. **Figure 7** gives examples of the first two cases.

An image of Topotecan (drug) was incorrectly classified as DMSO (control) without the lens, likely due to high confluence (cell population density) on the crop (**Figure 7A**). The output probability for DMSO was quite low though. After the lens was applied, the classifier got it right with very high confidence. Feature importance map highlights the regions of altered cell morphology that improved classification. **Figure 7B** shows an image with confluence below 50% corresponding to BPTES (drug), misclassified as Chlormethine (another drug). However, the lens identified regions of altered morphology that led to correct classification with high confidence. We provide more examples for the increased probability case on **Supplementary Figure 3** and **Supplementary Figure 4**.

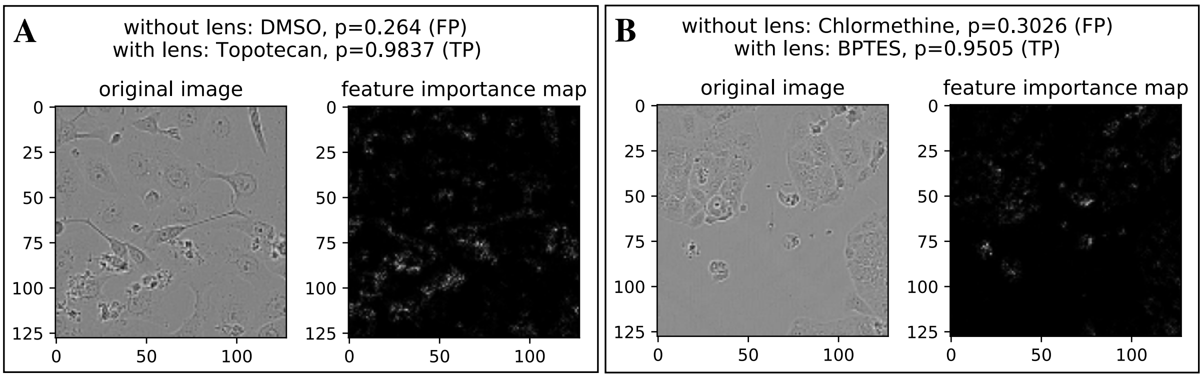

A
without lens: DMSO, p=0.264 (FP)
with lens: Topotecan, p=0.9837 (TP)
original image
feature importance map
0
25
50
75
100
125
0
50
100
B
without lens: Chlormethine, p=0.3026 (FP)
with lens: BPTES, p=0.9505 (TP)
original image
feature importance map

Figure 7. Classification results and morphological feature importance maps.

Discussion

Both applications were developed using ConvAE pretrained on a large dataset of drug-treated cancer cell lines. However, we were able to obtain temporal patterns similar to those on **Figure 5A** for individual drugs using ResNet-50 pretrained with SwAV (**Supplementary Figure 5**). Thus, based on our empirical findings (**Table 1**), we speculate that ResNet-50 pretrained with SwAV could be used instead of ConvAE to study temporal and morphological drug effects with little information loss at no additional training cost. Although ViT-B/8 pretrained with DINO produced the closest to ConvAE classification results, using visual transformers may still be prohibitive, because of their size. Using a Nvidia GeForce RTX 2060, we estimated the forward pass of 1M images with ViT-B/8 to take around 200 hours, with ResNet-50 – 32 hours, with ConvAE – 40 minutes.

Although we demonstrated the utility of our methods on a single biological dataset only, related works discussed earlier in this paper show comparable approaches applied to many types of biomedical imaging data. Therefore, we hope our results will contribute broadly to further development of deep learning methods for fundamental and clinical research.

Conclusion

In this work, we proposed two workflows to study phenotypic changes of experimental conditions using pretrained models. As a proof of concept, we applied them to study temporal and morphological drug effects on cancer cell lines. Besides, we trained a CNN model on a 1M images dataset comprising 21 cancer cell lines and 31 drugs at 5 concentrations. We validated the learned representations and provided the model to enable transfer learning applications. Overall, our findings suggest that pretrained models can be used for efficient and interpretable deep learning applications in biological and biomedical image analysis.

Supplementary material

Supplementary Text 1. Description of the dataset.

To cover a wide range of phenotypic effects in experimental and FDA-approved anticancer drugs, we selected drugs that displayed at least 3 cell lines as resistant and 3 cell lines sensitive in the NCI-60 cancer cell line panel (**Supplementary Table 1**), with a threshold in the $\log_{10}$(GI50) of 1% between the sensitive and resistant groups. The list comprised 31 experimental and FDA-approved anticancer drugs, covering several modes of action of clinical and research interest (**Supplementary Table 2**). The cancer cell lines were grown in RPMI-1640 GlutaMax medium (ThermoFischer) with supplementation of 1% of Penicylin-Streptomycin (Gibco), and 5% of dialyzed fetal bovine serum (Sigma-Aldrich) at 37°C in an atmosphere of 5% CO2. The seeding density to achieve a confluence of 70% was determined in Nunc 96 well plates (ThermoFischer), and that seeding density was used for experiments with a factor of four correction for the reduction in area between the 96 and 384 well plates, where cells were seeded in 45 uL of medium. Cells were incubated and imaged every two hours in the Incucyte S3 (Sartorious) 10x phase contrast mode from for up to 48 hours before drug addition, in order to achieve optimal cell adherence and starting experimental conditions. To reduce evaporation effects, the plates were sealed with Breathe-Easy sealing membrane (Diversified Biotech).

To allow a broad coverage of effects on time, we collected the time information about when the drugs were treated for each cell line, and corrected the analysis based on the drug treatment. Drugs were resuspended in the appropriate solvent (DMSO or water), and the same amount of DMSO (check amount) was added across all wells, including controls. The randomized 384 drug source plates were generated with Echo Liquid Handling System (Integra-Biosciences), and then transferred in 5uL of medium to Nunc 384 well plates (ThermoFischer) with the AssistPlus liquid handler (Integra Biosciences).

Supplementary Text 2. Selection of the ConvAE architecture.

To be able to demonstrate novel applications described in the results section, we needed a compact model trainable within limited resources (Nvidia GeForce RTX 2060 with 6 GB only). We tested 10 other CNN architectures and compared them by reconstruction quality. The architectures had the same number of layers as the final ConvAE, but differed in number of neurons and pooling strategies to keep the same dimensionality of the bottleneck layer. The choice of an architecture was constrained by the need to fit the data and the model into the GPU memory (especially important for the lens framework). Therefore, an integration test for each architecture was performed for the lens training setup. The final architecture of ConvAE and the weights are now available on GitHub.

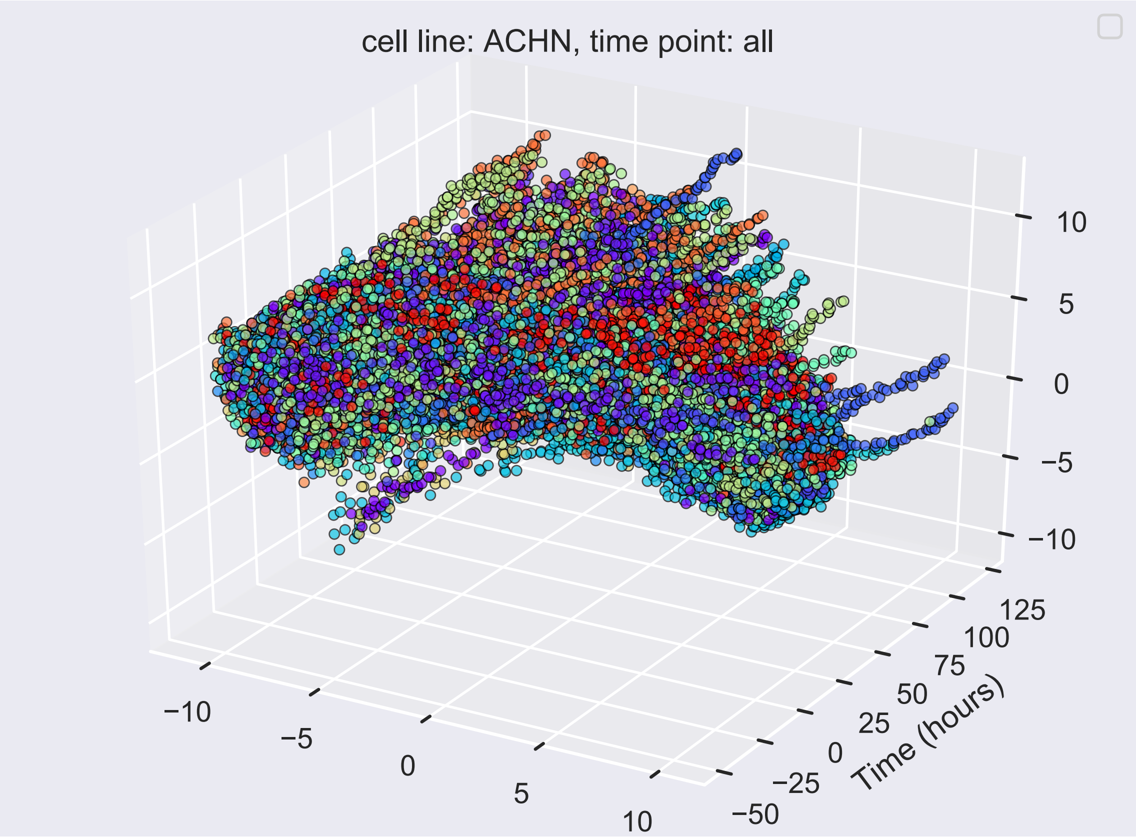
Supplementary Table 1. Cell lines and inoculation densities for 96 well plate format.

Cell line	Panel	Inoculation density
EKVX	Non-Small Cell Lung	11000
HOP-62	Non-Small Cell Lung	9000
COLO 205	Colon	15000
HCT-15	Colon	12000
HT29	Colon	12000
SW-620	Colon	24000
SF-539	CNS	10000
LOX IMVI	Melanoma	8500
MALME-3M	Melanoma	8500
M14	Melanoma	5000
SKMEL-2	Melanoma	10000
UACC-257	Melanoma	20000
IGR-OV1	Ovarian	10000
OVCAR-4	Ovarian	10000
OVCAR-5	Ovarian	15000
A498	Renal	3200
ACHN	Renal	8200
MDA-MB-231 / ATCC	Breast	20000
HS 578T	Breast	13000
BT-549	Breast	10000
T-47D	Breast	15000

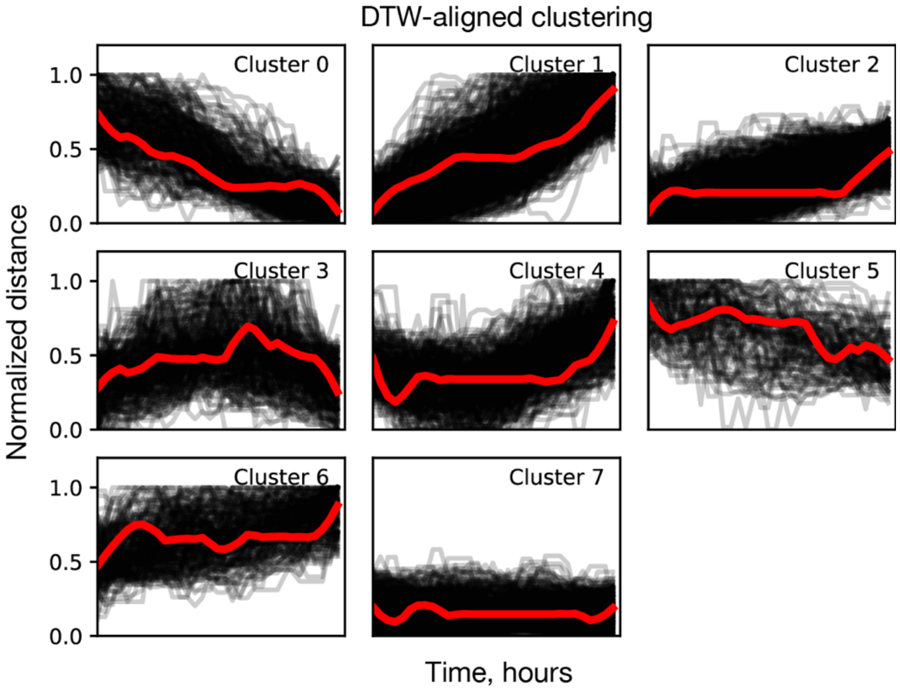
Supplementary Table 2. Drugs, solvents, CAS registry numbers and maximum concentrations. The other four concentrations for each drug were 10-fold serial dilutions of the maximum concentration.

Drug	Fluid	CAS	Concentration
Erlotinib	DMSO	183321-74-6	10 uM
Irinotecan	DMSO	100286-90-6	10 uM
Clofarabine	DMSO	123318-82-1	10 uM
Fluorouracil	DMSO	51-21-8	10 uM
Pemetrexed	Water	150399-23-8	10 uM
Docetaxel	DMSO	148408-66-6	1 uM
Everolimus	DMSO	159351-69-6	1 uM
Chlormethine	DMSO	55-86-7	10 uM
BPTES	DMSO	314045-39-1	10 uM
Oligomycin A	DMSO	579-13-5	1 uM
UK-5099	DMSO	NA	10 uM
Panzem (2-ME2)	DMSO	362-07-2	10 uM
MEDICA16	DMSO	87272-20-6	10 uM
Gemcitabine	Water	122111-03-9	1 uM
17-AAG	DMSO	75747-14-7	10 uM
Lenvatinib	DMSO	417716-92-8	10 uM
Topotecan	DMSO	119413-54-6	1 uM
Cladribine	DMSO	4291-63-8	10 uM
Mercaptopurine	DMSO	6112-76-1	10 uM
Decitabine	DMSO	2353-33-5	10 uM
Methotrexate	DMSO	59-05-2	1 uM
Paclitaxel	DMSO	33069-62-4	1 uM
Rapamycin	DMSO	53123-88-9	0.1 uM
Oxaliplatin	DMSO	61825-94-3	10 uM
Omacetaxine	DMSO	26833-87-4	1 uM
Metformin	Water	1115-70-4	10 uM
YC-1	DMSO	170632-47-0	10 uM
Etoximir	DMSO	828934-41-4	10 uM
Oxfenicine	DMSO	32462-30-9	2.5 uM
Trametinib	DMSO	871700-17-3	1 uM
Asparaginase	Water	9015-68-3	0.00066 units / uL

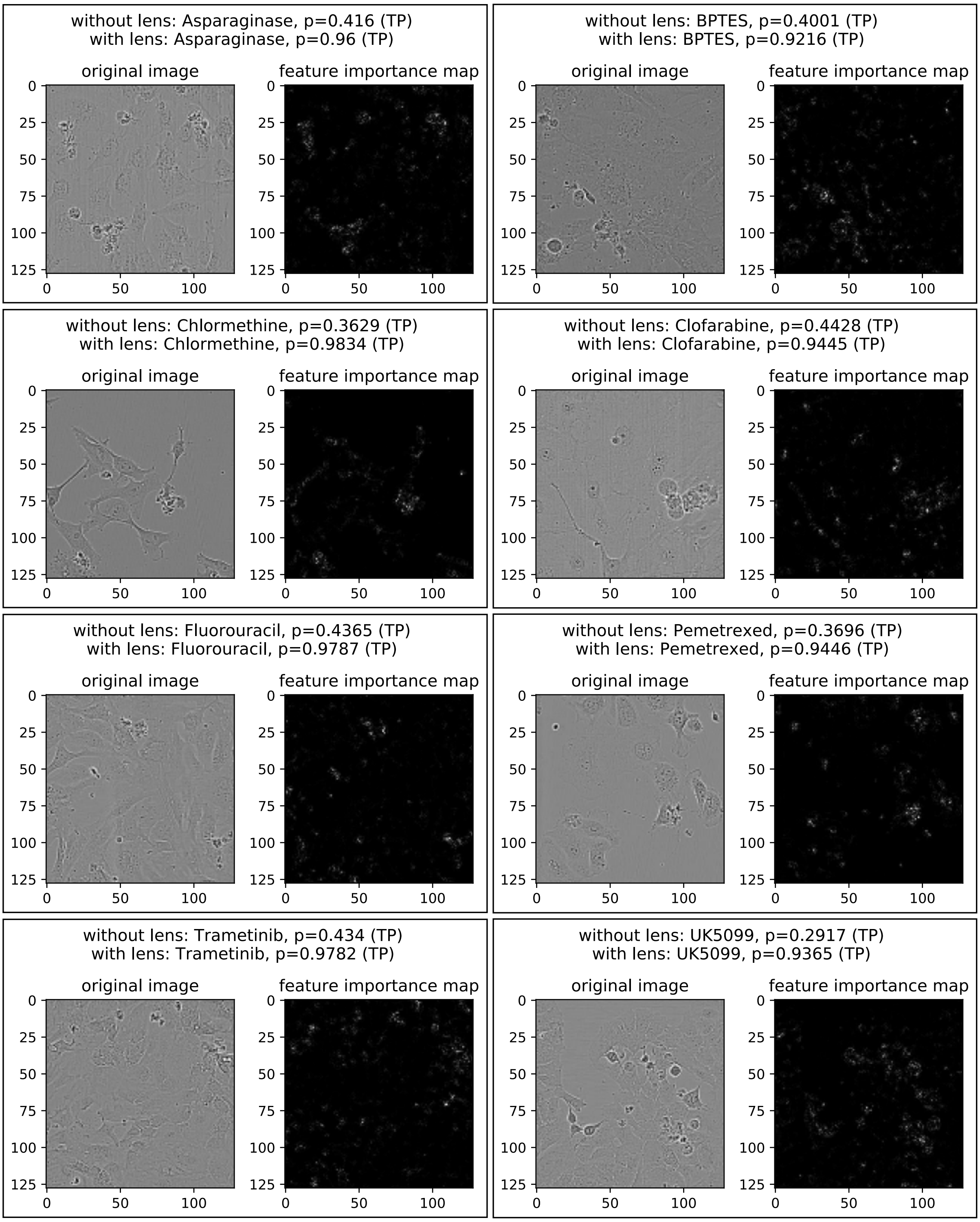
Supplementary Figure 1. Spatial and temporal separation of UMAP embeddings. We plotted UMAP embeddings of image representations. Taking the latest time points only, we observed gradual transition from drug clusters of no effect to clusters of strong cytotoxic drug effects. Taking UMAP embeddings of all time points, we saw spatial and temporal separation of images even more clearly. Single drug tracks can be seen in colors, and some of them diverged dramatically from the initial locus (points of time < 0), demonstrating a variety of drug effects, distinct in nature and intensity.



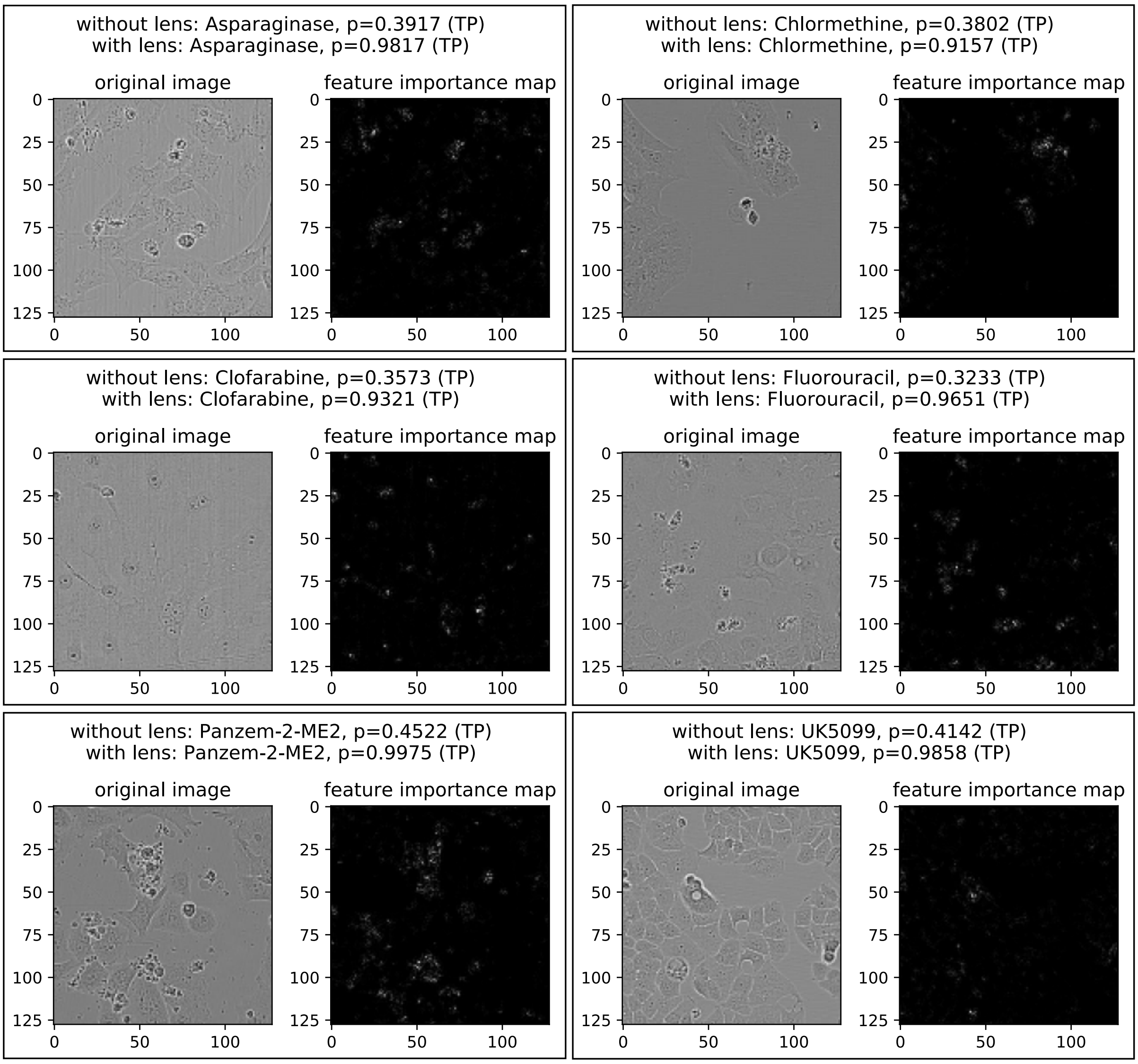
Supplementary Figure 2. **Clusters of temporal drug effects**. In total, we found 8 clusters, characterizing expected types of response, intensity and speed of divergence from controls. We examined random cluster representatives to validate the analysis and interpret identified patterns. We found that cluster 1 is associated with a strong cytotoxic effect, as we observed a lot of cell deaths on the corresponding images. Cluster 3 represents cytostatic effect, as the images showed temporary cell growth arrest. Clusters 2 and 6 are related to mixed effects, as the images displayed both patterns. We labeled clusters 0, 4, 5, 7 as showing no effect. Images of cluster 7 stayed indistinguishable to controls at all time points. Clusters 4 and 5 had only 30% of images showing weak cytotoxic or mixed effect. Cluster 0 is an expected artifact of the distance-based analysis: when the cell population density is low (images of early time points), the distance may be large due to varying localization of cells on the crop. With growing cell population, the distance gradually drops unless there is a drug effect.



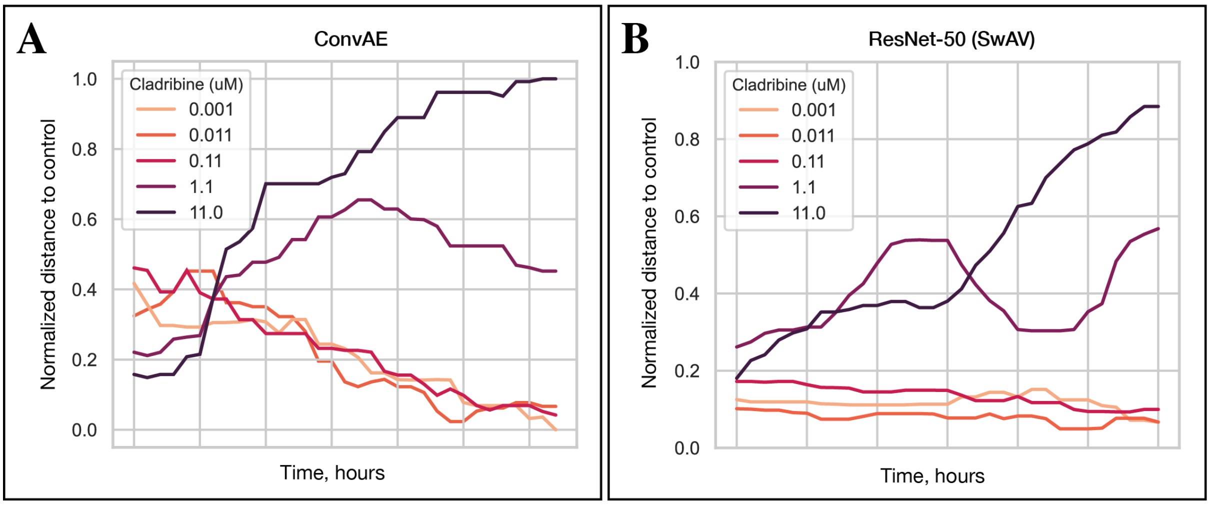
Supplementary Figure 3. Morphological feature importance maps for increased classification probability.


without lens: Asparaginase, p=0.416 (TP)
with lens: Asparaginase, p=0.96 (TP)
without lens: BPTES, p=0.4001 (TP)
with lens: BPTES, p=0.9216 (TP)
without lens: Chlormethine, p=0.3629 (TP)
with lens: Chlormethine, p=0.9834 (TP)
without lens: Clofarabine, p=0.4428 (TP)
with lens: Clofarabine, p=0.9445 (TP)
without lens: Fluorouracil, p=0.4365 (TP)
with lens: Fluorouracil, p=0.9787 (TP)
without lens: Pemetrexed, p=0.3696 (TP)
with lens: Pemetrexed, p=0.9446 (TP)
without lens: Trametinib, p=0.434 (TP)
with lens: Trametinib, p=0.9782 (TP)
without lens: UK5099, p=0.2917 (TP)
with lens: UK5099, p=0.9365 (TP)
original image
feature importance map
0
25
50
75
100
125
0
50
100

Supplementary Figure 4. Morphological feature importance maps for increased classification probability.


without lens: Asparaginase, p=0.3917 (TP)
with lens: Asparaginase, p=0.9817 (TP)
original image
feature importance map
without lens: Chlormethine, p=0.3802 (TP)
with lens: Chlormethine, p=0.9157 (TP)
original image
feature importance map
without lens: Clofarabine, p=0.3573 (TP)
with lens: Clofarabine, p=0.9321 (TP)
original image
feature importance map
without lens: Fluorouracil, p=0.3233 (TP)
with lens: Fluorouracil, p=0.9651 (TP)
original image
feature importance map
without lens: Panzem-2-ME2, p=0.4522 (TP)
with lens: Panzem-2-ME2, p=0.9975 (TP)
original image
feature importance map
without lens: UK5099, p=0.4142 (TP)
with lens: UK5099, p=0.9858 (TP)
original image
feature importance map
0
25
50
75
100
125
0
50
100

Supplementary Figure 5. Similar temporal patterns of Cladribine obtained with different models. We repeated distance-based analysis for the Cladribine case emphasized on **Figure 5B**. This time, we used representations obtained with ResNet-50 pretrained with SwAV on ImageNet. We observed many similarities between temporal patterns previously identified with ConvAE (**A**) and the ones of ResNet-50 + SwAV (**B**). Three lowest concentrations showed decrease of distance to control over time for both models. The highest concentration, in turns, caused constant growth of distances. For the 1.1 µM concentration, both models produced an initial increase of the distance, followed by the phase of decay. However, ResNet-50 + SwAV additionally presents an increase of the distance in the latest timepoints. This artifact is likely caused by the ImageNet biases. We, therefore, recommend to use general-purpose pretrained models to analyze more specific datasets with caution.



References

(1) Adam, G.; Rampášek, L.; Safikhani, Z.; Smirnov, P.; Haibe-Kains, B.; Goldenberg, A. Machine Learning Approaches to Drug Response Prediction: Challenges and Recent Progress. *npj Precis. Oncol.* **2020**, *4* (1), 1–10. https://doi.org/10.1038/s41698-020-0122-1.

(2) Kan, A. Machine Learning Applications in Cell Image Analysis. *Immunol. Cell Biol.* **2017**, *95* (6), 525–530. https://doi.org/10.1038/icb.2017.16.

(3) Meijering, E. A Bird's-Eye View of Deep Learning in Bioimage Analysis. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 2312–2325. https://doi.org/10.1016/j.csbj.2020.08.003.

(4) Suganyadevi, S.; Seethalakshmi, V.; Balasamy, K. A Review on Deep Learning in Medical Image Analysis. *Int. J. Multimed. Inf. Retr.* **2022**, *11* (1), 19–38. https://doi.org/10.1007/s13735-021-00218-1.

(5) Moen, E.; Bannon, D.; Kudo, T.; Graf, W.; Covert, M.; Van Valen, D. Deep Learning for Cellular Image Analysis. *Nat. Methods* **2019**, *16* (12), 1233–1246. https://doi.org/10.1038/s41592-019-0403-1.

(6) Yang, K. D.; Damodaran, K.; Venkatachalapathy, S.; Soylemezoglu, A. C.; Shivashankar, G. V.; Uhler, C. Predicting Cell Lineages Using Autoencoders and Optimal Transport. *PLoS Comput. Biol.* **2020**, *16* (4), 1–20. https://doi.org/10.1371/journal.pcbi.1007828.

(7) Lu, A. X.; Kraus, O. Z.; Cooper, S.; Moses, A. M. Learning Unsupervised Feature Representations for Single Cell Microscopy Images with Paired Cell Inpainting. *PLoS Comput. Biol.* **2019**, *15* (9), 1–27. https://doi.org/10.1371/journal.pcbi.1007348.

(8) Huff, D. T. . et al. Interpretation and Visualization Techniques for Deep Learning Models in Medical Imaging Daniel. *Phys Med Biol.* **2021**, *66* (4).

(9) Singh, A.; Sengupta, S.; Lakshminarayanan, V. Explainable Deep Learning Models in Medical Image Analysis. *J. Imaging* **2020**, *6* (6), 1–19. https://doi.org/10.3390/JIMAGING6060052.

(10) Lafarge, M. W.; Juan Caicedo, T. C.; Anne Carpenter, B. E.; Josien Pluim, B. P.; Shantanu Singh, T.; Veta, M. Capturing Single-Cell Phenotypic Variation via Unsupervised Representation Learning. *Proc. Mach. Learn. Res.* **2019**, *102*, 315–325.

(11) Goldsborough, P.; Pawlowski, N.; Caicedo, J. C.; Singh, S.; Carpenter, A. E. CytoGAN: Generative Modeling of Cell Images. *bioRxiv* **2017**, No. Nips, 227645.

(12) Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; Joulin, A. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. *Adv. Neural Inf. Process. Syst.* **2020**, *2020-Decem* (NeurIPS), 1–13.

(13) Caron, M.; Touvron, H.; Misra, I.; Jegou, H.; Mairal, J.; Bojanowski, P.; Joulin, A. Emerging Properties in Self-Supervised Vision Transformers. *Proc. ICCV* 9650–9660.

(14) Chandrasekaran, S. N.; Ceulemans, H.; Boyd, J. D.; Carpenter, A. E. Image-Based Profiling for Drug Discovery: Due for a Machine-Learning Upgrade? *Nat. Rev. Drug Discov.* **2021**, *20* (2), 145–159. https://doi.org/10.1038/s41573-020-00117-w.

(15) Grill, J. B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P. H.; Buchatskaya, E.; Doersch, C.; Pires, B. A.; Guo, Z. D.; Azar, M. G.; Piot, B.; Kavukcuoglu, K.;

Munos, R.; Valko, M. Bootstrap Your Own Latent a New Approach to Self-Supervised Learning. *Adv. Neural Inf. Process. Syst.* **2020**, *2020-Decem* (NeurIPS), 1–14.

(16) Carpenter, A. E.; Jones, T. R.; Lamprecht, M. R.; Clarke, C.; Kang, I. H.; Friman, O.; Guertin, D. A.; Chang, J. H.; Lindquist, R. A.; Moffat, J.; Golland, P.; Sabatini, D. M. CellProfiler: Image Analysis Software for Identifying and Quantifying Cell Phenotypes. *Genome Biol.* **2006**, *7* (10). https://doi.org/10.1186/gb-2006-7-10-r100.

(17) Gómez-de-Mariscal, E.; García-López-de-Haro, C.; Ouyang, W.; Donati, L.; Lundberg, E.; Unser, M.; Muñoz-Barrutia, A.; Sage, D. DeepImageJ: A User-Friendly Environment to Run Deep Learning Models in ImageJ. *Nat. Methods* **2021**, *18* (10), 1192–1195. https://doi.org/10.1038/s41592-021-01262-9.

(18) Biffi, C.; Cerrolaza, J. J.; Tarroni, G.; Bai, W.; De Marvao, A.; Oktay, O.; Ledig, C.; Le Folgoc, L.; Kamnitsas, K.; Doumou, G.; Duan, J.; Prasad, S. K.; Cook, S. A.; O'Regan, D. P.; Rueckert, D. Explainable Anatomical Shape Analysis through Deep Hierarchical Generative Models. *IEEE Trans. Med. Imaging* **2020**, *39* (6), 2088–2099. https://doi.org/10.1109/TMI.2020.2964499.

(19) Hou, L. et al. Sparse Autoencoder for Unsupervised Nucleus Detection and Representation in Histopathology Images. *Pattern Recognit.* **2019**, *86*, 188–200. https://doi.org/10.1016/j.patcog.2018.09.007.

(20) Uzunova, H.; Ehrhardt, J.; Jacob, F.; Frydrychowicz, A.; Handels, H. Multi-Scale GANs for Memory-Effcient Generation of High Resolution Medical Images. *arXiv* **2019**, 19.

(21) Chen, X.; You, S.; Tezcan, K. C.; Konukoglu, E. Unsupervised Lesion Detection via Image Restoration with a Normative Prior. *Med. Image Anal.* **2020**, *64*. https://doi.org/10.1016/j.media.2020.101713.

(22) Sajjadi, M. S. M.; Parascandolo, G.; Mehrjou, A.; Schölkopf, B. Tempered Adversarial Networks. *6th Int. Conf. Learn. Represent. ICLR 2018 - Work. Track Proc.* **2018**.

(23) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. **2018**.

(24) McInnes, L.; Healy, J.; Astels, S. Hdbscan: Hierarchical Density Based Clustering. *J. Open Source Softw.* **2017**, *2* (11), 205. https://doi.org/10.21105/joss.00205.

(25) Minderer, M.; Bachem, O.; Houlsby, N.; Tschannen, M. Automatic Shortcut Removal for Self-Supervised Representation Learning. *37th Int. Conf. Mach. Learn. ICML 2020* **2020**, *PartF168147-9*, 6883–6893.

Chapter 7

Concluding remarks

Reproducibility of metabolomics

Reproducibility of experiment is a cornerstone of the scientific method[1]. Failure to reproduce measurements, computations, or results of a previous study is perceived as a lack of rigor and undermines the validity of study and its claims. Omics technologies are not immune to these challenges[2]. The reproducibility problems in metabolomics are rooted in the nature of the process that underly ionization and detection of ions. Sources of variation include day-to-day variability of the instrument response, differences in sample preparation, matrix effects, contamination and accumulating dirt in the system during long acquisition sequences. Together, these factors affect the measurements non-linearly and thereby hamper long-term reproducibility of results. Emerging clinical applications of metabolomics make it extremely important to overcome the reproducibility issue[3–6].

In **Chapters 2-4** of this thesis, we addressed it from multiple standpoints. We proposed means of quality assurance and quality control for data acquisition, investigated calibration strategies and developed a batch correction method to improve comparability of FIA-TOF-MS measurements across acquisition batches. More specifically:

1. We presented the concept and the actual implementation of a system suitability testing platform for monitoring the status of a high-resolution QTOF mass spectrometer. The setup consists of a QC mixture, an acquisition method, software to extract a detailed ensemble of quantitative features describing spectral properties, and a simple R Shiny front-end for real-time visualization. We operated the testing platform in a pilot lasting for 21 months and including 153 individual measurements of the QC mixture. We demonstrated instrument monitoring by a set of quality indicators and the implementation of routines for trend and outlier detection. The platform, therefore, is capable of in-depth evaluation of the instrument readiness to measure biological samples.
2. We introduced RALPS, a novel batch normalization method based on regularized adversarial learning for untargeted metabolomics data. We demonstrated its performance on two representative datasets with thousands of samples or spectral features. RALPS consistently decreased batch variation

coefficients while keeping the biological information intact, which enables direct comparison of biological replicates across distant-in-time batches. The benchmarking dataset was generated to test the algorithm on MS data produced over several months. In the case of the cancer cell line data by Cherkaoui et al., batches instead were associated with cultivation and sampling of samples over the span of almost a year, whereas MS analysis was done sequentially with all samples. We demonstrated that RALPS outperformed state-of-the-art methods on several key metrics. RALPS offers additional features such as adaptive network architectures, embedded hyperparameter optimization, automated model selection, and input validation. Together, these features convey flexibility, scalability, usability, and robustness as confirmed by testing with different configurations of reference samples and in ablation experiments.

3. We investigated the performance of calibration using the benchmarking dataset of FIA-ESI-MS. We wondered to which extent quantification of selected metabolites spiked into a complex biological matrix, such as human serum, is possible and formulated several tasks to test that. We approached each task with machine learning, trained the models and evaluated their extrapolation capability, e.g., to predict ion counts of metabolites never used for training. We observed that (i) including dilution series can, indeed, improve calibration, (ii) machine learning models can extrapolate prediction of absolute ion counts of amino acids well, and (iii) prediction of relative abundances has a potential to beat the most advanced batch correction methods in terms of reproducibility when enough data is available for training.

The proposed methods and tools present individual advances in reproducible flow injection analysis with time-of-flight mass spectrometry that are complementary to each other. It is clear that they must synergize to solve the reproducibility issue for untargeted metabolomics. Systematic monitoring of the instrument performance should become routine to avoid sudden drops in the quality indicators of particular importance for the experiment. But also, it must be accounted for during calibration or batch correction with statistical relationships established between quality indicators and instrument settings. Follow-up studies must explore and exploit this potential by implementing system suitability testing and including enough of QC and reference samples in the study design to enable further developments in calibration and batch correction.

The presented SST platform setup itself offers ample room for further improvements. In particular, the long-term stability of the QC mixture should be verified. The feature extraction is generic and easily transferrable to TOF instruments from other vendors. However, adaptation to high-resolution instruments would require additional features capturing artifacts of Fourier Transform spectra: harmonic peaks, coalescence, etc. Further, we illustrated how collection of instrument setpoints and readbacks allows to derive the causal relationships between instrument settings and instrument performance measured with the QC mixture. Beyond assistance in scheduling instrument maintenance and diagnostic applications, we envisage that with more data available it could be possible to recommend instrument settings to attain a particular value of a quality indicator in real-time.

The central idea behind RALPS is the grouping regularization term r_g responsible for preserving similarity of supposedly equal samples. Although RALPS relies on clustering of reference samples in the embedded space to assess grouping, alternative approaches could be considered. For example, r_g could be calculated from distances in the latent space. Distance-based metrics would carry several hypothetical advantages and pitfalls. Owing to the fast computation, training would be tangibly faster. We expect that distance-based metrics would perform better with data characterized by subtle batch effects in which clustering fails to separate reference samples. Even in the case of all replicates falling into a single cluster (which happens naturally when most batch effects have been already removed), minimizing distances between all pairs of replicates would still have a regularization effect, whereas clustering would not. Among the potential drawbacks, we would expect an increased sensitivity to single outliers. Alternative paradigms for similarity preservation should be tested in the future.

Another aspect we have not been able to test in batch correction is inclusion of the SST data. The experimental design for the benchmarking dataset assumed very few QC samples, which was enough to ensure normal operating state of the instrument but insufficient QC data to use as additional reference samples in the training process of RALPS. Therefore, this test could not be implemented, though it could bring another layer of data harmonization in terms of consistency of the QC features across batches integrated into the model selection logic. On the other hand, stability of the QC features could be optimized directly with an additional regularization term in the joint loss function controlling their variation. These ideas need rigorous evaluation and comparison to the existing implementation, but present particular opportunities to achieve synergy between the proposed SST platform and RALPS.

The amount of data we used for calibration of FIA-TOF-MS was also limited and likely not sufficient to extrapolate for unseen but structurally similar metabolites. Although we were able to successfully predict ion counts using only 40% of amino acids in the training set, rather poor performance was achieved in other setups. We envision a follow-up study involving an expanded list of metabolites (and, ideally, compound classes) measured in water, blood, plasma, urine, bacterial extracts in dilutions series for up to a factor of 64, where each analytical triplicate is accompanied with the QC sample and the corresponding instrument settings to provide auxiliary information linking detector response directly to the state of mass filters, lenses, mirrors, etc. Additionally, quantification of relative abundances must be revisited to investigate reproducibility of calibration across biological matrices and demonstrate a real application example.

Multi-task representation learning for applications in metabolomics

In the last decade, the field of artificial intelligence has been growing rapidly. It penetrated and significantly advanced many fields of scientific research. In untargeted metabolomics, deep learning frameworks have been proposed to improve peak integration[7], predict retention times[8], remove batch effects[9] and generate structures *de novo* using mass spectra[10]. Altogether, such developments enhance the capabilities of analytical methods and provide deeper insights into the data.

Latest batch correction approaches heavily exploited multi-task representation learning[9,11–13], wherein two or more deep neural networks are trained simultaneously to solve multiple coherent tasks expressed as individual terms in the joint loss function. This approach is designed to learn such data representations that reflect the most important properties of the data. In **Chapter 3**, we developed RALPS with an idea to retain similarity of biological samples while alleviating batch-related biases. The central challenge there was to find the suitable mathematical problem formulation and set up the corresponding multi-task representation learning. In **Chapters 5-6** of this thesis, we demonstrated superiority of multi-task representation learning for another data modality (i.e., microscopy images of cell cultures) and used them to develop novel AI applications to cancer research. More specifically:

1. We trained 16 deep learning setups on 1M cancer cell images to fairly compare representation learning approaches. We evaluated the learned representations on 3 independent tasks using multiple metrics to quantify performance. We made several key observations: (i) multi-crops and random augmentations generally enrich representations with relevant information, which results in improved performance in downstream tasks; (ii) some implicit contrastive learning setups can be trained strikingly fast, but need further refinements to achieve top performance across tasks; (iii) the regularized autoencoder model, which represents a multi-task representation learning approach, produced the most informative features with the best accuracy and ROAUC in the classification task and the best quality of partitions in the clustering task.

2. We proposed two workflows to study phenotypic changes of experimental conditions using pretrained deep learning models. As a proof of concept, we applied them to study temporal and morphological drug effects on cancer cell lines using the same dataset of 1M images comprising 21 cancer cell lines and 31 drugs at 5 concentrations. Analyzing distances between retrieved representations of drug and control images, we were able to identify temporal patterns impossible to detect with conventional approaches (i.e., analyzing growth curves). Designing yet another multi-task representation learning approach (the lens setup), we were able to highlight morphological cell features of importance to predict the correct drug labels in a multi-class classification setting. The latter presents an explainable AI application frequently desired in the clinics.

We conclude that multi-task representation learning is a powerful technique that reaches state-of-the-art performance across data modalities. For metabolomics, it opens countless opportunities for follow-up research integrating mass spectrometry data with other types of omics or imaging data. For instance, metabolic phenotype could be used to assign labels in the lens setup presented in **Chapter 6**. That is, mass spectrometry metabolomics data corresponding to the cancer cells exposed to drug treatments could be used for clustering and deriving groups of metabolic phenotypes, serving as class labels for microscopy images. In this case, identical formulation of the lens setup would provide insights into associations between exterior morphological features and the underlying chemical composition of the cells. Alternative formulations are possible exploiting representations of individual data modalities to investigate their interdependencies or aggregate information for more sophisticated applications.

References

(1) Staddon, J. *Scientific Method: How Science Works, Fails to Work, and Pretends to Work*; 2017. https://doi.org/10.4324/9781315100708.

(2) Tarazona, S.; Balzano-Nogueira, L.; Gómez-Cabrero, D.; Schmidt, A.; Imhof, A.; Hankemeier, T.; Tegnér, J.; Westerhuis, J. A.; Conesa, A. Harmonization of Quality Metrics and Power Calculation in Multi-Omic Studies. *Nat. Commun.* **2020**, *11* (1), 1–13. https://doi.org/10.1038/s41467-020-16937-8.

(3) Wishart, D. S. Emerging Applications of Metabolomics in Drug Discovery and Precision Medicine. *Nat. Rev. Drug Discov.* **2016**, *15* (7), 473–484. https://doi.org/10.1038/nrd.2016.32.

(4) Koen, N.; Du Preez, I.; Loots, D. T. *Metabolomics and Personalized Medicine*, 1st ed.; Elsevier Inc., 2016; Vol. 102. https://doi.org/10.1016/bs.apcsb.2015.09.003.

(5) Jacob, M.; Lopata, A. L.; Dasouki, M.; Abdel Rahman, A. M. Metabolomics toward Personalized Medicine. *Mass Spectrom. Rev.* **2019**, *38* (3), 221–238. https://doi.org/10.1002/mas.21548.

(6) Trivedi, D. K.; Goodacre, R. *The Role of Metabolomics in Personalized Medicine*; Elsevier Inc., 2020. https://doi.org/10.1016/b978-0-12-812784-1.00011-6.

(7) Liebal, U. W.; Phan, A. N. T.; Sudhakar, M.; Raman, K.; Blank, L. M. Machine Learning Applications for Mass Spectrometry-Based Metabolomics. *Metabolites* **2020**, *10* (6), 1–23. https://doi.org/10.3390/metabo10060243.

(8) Giese, S. H.; Sinn, L. R.; Wegner, F.; Rappsilber, J. Retention Time Prediction Using Neural Networks Increases Identifications in Crosslinking Mass Spectrometry. *Nat. Commun.* **2021**, *12* (1), 1–11. https://doi.org/10.1038/s41467-021-23441-0.

(9) Niu, J.; Xu, W.; Wei, D.; Qian, K.; Wang, Q. Deep Learning Framework for Integrating Multibatch Calibration, Classification, and Pathway Activities. *Anal. Chem.* **2022**. https://doi.org/10.1021/acs.analchem.2c00601.

(10) Stravs, M. A.; Dührkop, K.; Böcker, S.; Zamboni, N. MSNovelist: De Novo Structure Generation from Mass Spectra. *Nat. Methods* **2022**, *19* (July). https://doi.org/10.1038/s41592-022-01486-3.

(11) Rong, Z.; Tan, Q.; Cao, L.; Zhang, L.; Deng, K.; Huang, Y.; Zhu, Z. J.; Li, Z.; Li, K. NormAE: Deep Adversarial Learning Model to Remove Batch Effects in Liquid Chromatography Mass Spectrometry-Based Metabolomics Data. *Anal. Chem.* **2020**, *92* (7), 5082–5090. https://doi.org/10.1021/acs.analchem.9b05460.

(12) Lakkis, J.; Wang, D.; Zhang, Y.; Hu, G.; Wang, K.; Pan, H.; Ungar, L.; Reilly, M.; Li, X.; Li, M. A Joint Deep Learning Model Enables Simultaneous Batch Effect Correction, Denoising and Clustering in Single-Cell Transcriptomics. *Genome Res.* **2021**, gr.271874.120. https://doi.org/10.1101/gr.271874.120.

(13) Li, X.; Wang, K.; Lyu, Y.; Pan, H.; Zhang, J.; Stambolian, D.; Susztak, K.; Reilly, M. P.; Hu, G.; Li, M. Deep Learning Enables Accurate Clustering with Batch Effect Removal in Single-Cell RNA-Seq Analysis. *Nat. Commun.* **2020**, *11* (1), 1–14. https://doi.org/10.1038/s41467-020-15851-3.

Acknowledgments

This thesis reflects four years of scientific work in the Zamboni lab of the Institute of Molecular Systems Biology. I'm grateful to many good people that helped, supported and accompanied me on the way. Foremost, I would like to thank **Nicola** for giving me the chance to join the lab. I appreciated a lot the freedom I had to pursue my ideas and the opportunity to chat about life and science almost any time. I think I learned a lot from you, and I'm very thankful for that.

I thank my PhD committee members for open scientific discussions with approval and criticism of my work: **Uwe Sauer**, **Bernd Wollscheid** and **Juho Rousu**. Thank you for sharing your experience and knowledge with me.

I would like to thank **Michelle Reid**, who collaborated with me on three main projects of this thesis on the experimental side. And big thanks to **Mauro Masiero** for approaching me once with an amazing imaging dataset and tons of ideas that resulted in two already accepted publications. It was really fun working with you two!

I would like to thank all past and present Sauer/Zamboni/Zampieri lab members, who made my PhD time great despite the crazy times in the world. Thanks to **Ohad** and **Mattia** for help in teaching. Thanks to **Michiel** and **Alaa** for sharing the office for the most (enjoyable) time. Thanks to **Alexis** and **Sarah** for being cool (French-speaking) neighbors. Thanks to **Tomek** and **Philipp** for profound philosophical discussions and taking courses together. Thanks to **Jane** for being my buddy in leaning German and always being an expert about things. Thanks to **Stephan** for sharing my passion to martial arts and not kicking me too hard. Thanks to **Karin M**. for exclusive coffee breaks and reading my fiction stories in German. Thanks to **Karin O**. for cute pictures of birds (gift me one, please). Thanks to **Daniela** for being a companion in my first skiing experience. Thanks to **Toby**, **Duncan, Laurentz** and **Chris** for organizing many events I enjoyed so much. Thanks to **Maren** for a nice PhD thesis template!

Last but not least, I thank all my friends outside the lab and my family. Especially my lovely wife **Ruzanna** — for supporting me throughout the years of the journey.

Curriculum vitae

Andrei Dmitrenko, M.Sc.

Contacts

Phone: +41 76 449 3967
E-mail: dmitrav@inbox.ru
GitHub: github.com/dmitrav
LinkedIn: linkedin.com/in/andrei-dmitrenko

Address

Albisriederstrasse 352
8047 Zurich
Switzerland

Education

2018-2022 (PhD)	ETH Zürich / Swiss Federal Institute of Technology (Switzerland), Institute of Molecular Systems Biology
2015-2017 (Master)	Peter the Great St. Petersburg Polytechnic University (Russia), Institute Of Applied Mathematics and Mechanics, specialty: "Bioinformatics"
2011-2015 (Bachelor)	St. Petersburg State University (Russia), Faculty of Applied Mathematics and Control Processes, specialty: "Applied Mathematics, Physics and Computer Science"

Experience

2018-2022 at ETH Zürich, IMSB, Zamboni lab (Switzerland):

- Developed a new method for multi-batch mass-spectrometry data normalization, based ono regularized adversarial learning. This work is submitted for a publication.
- Designed and implemented a system suitability testing platform for Agilent 6550 iFunnel Q-TOF mass-spectrometer. This work is submitted for a publication.

- Developed AI applications to study temporal and morphological drug effects using cancer cell imaging data. This work is accepted for publication in *PMLR*.
- Systematically compared different AI paradigms in learning representations of biological data. This work is accepted for publication in *PMLR*.

2016-2018 at BIOCAD, Department of Computational Biology (Russia):

- Performed modelling of pharmacokinetics (PK) and pharmacodynamics (PD) of therapeutics reaching clinical trials.
- Developed a deep learning model to classify 4 breast cancer stages using microscopy images of immunohistochemistry experiments.
- Developed and maintained a statistical tool to automate immunogenicity assessment of antibodies.
- Developed approaches to increase precision of Gibbs energy calculation of protein complexes using OPLS-type force field.
- Headed up a systems biology group of four people, trained the team to deliver analysis of PK and PD within limited timelines.
- Established interactions among the group and the other units, managed and supervised PK/PD modelling of 10 products.
- Managed development of an internal computational tool for PK/PD modelling.

2015-2016 at Peter the Great St. Petersburg Polytechnic University,
Mathematical Biology lab (Russia):

- Developed a stochastic model of regulation of the *gap* genes in Drosophila. This work was presented on local conferences.

Awards

- "Project of the year" prize (2017).
- "The best student research project" scholarship (2016).

Languages

- Russian: C2 (native)
- English: C1 (advanced)
- Deutsch: B2-C1 (advanced)
- Français: A1-A2 (beginner)